2021

Algorithmic Analysis And Statistical Inference Of Sparse Models In High Dimension

Zhiqi Bu
University of Pennsylvania

Algorithmic Analysis And Statistical Inference Of Sparse Models In High Dimension

Abstract

The era of machine learning features large datasets that have high dimension of features. This leads to the emergence of various algorithms to learn efficiently from such high-dimensional datasets, as well as the need to analyze these algorithms from both the prediction and the statistical inference viewpoint. To be more specific, an ideal model is expected to predict accurately on the unseen new data, and to provide valid inference so as to harness the uncertainty in the model. Unfortunately, the high dimension of features poses a great challenge on the analysis of many prevalent models, rendering them either inapplicable or difficult to study.

This thesis leverages the approximate message passing (AMP) algorithm, the optimization theory, and the Sorted L-One Penalized Estimation (SLOPE) to study several important problems of the sparse models.

The first chapter introduces various ℓ_1 penalties including but not limited to the SLOPE, a relatively new convex optimization procedure via the sorted ℓ_1 penalty, in the general machine learning models. We then focus on the linear models and demonstrate some basic properties of SLOPE, especially its advantages over the Lasso. Next, we cover the AMP algorithm in terms of convergence behavior and asymptotic statistical characterization.

The second chapter extends the AMP algorithms from Lasso to SLOPE and provides an asymptotically tight characterization of the SLOPE solution. Note that SLOPE is a relatively new convex optimization procedure for high-dimensional linear regression via the sorted ℓ_1 penalty: the larger the rank of the fitted coefficient, the larger the penalty. This non-separable penalty renders many existing techniques invalid or inconclusive in analyzing the SLOPE solution. We develop an asymptotically exact characterization of the SLOPE solution under Gaussian random designs through solving the SLOPE problem using approximate message passing (AMP). This algorithmic approach allows us to approximate the SLOPE solution via the much more amenable AMP iterates. Explicitly, we characterize the asymptotic dynamics of the AMP iterates relying on a recently developed state evolution analysis for non-separable penalties, thereby overcoming the difficulty caused by the sorted ℓ_1 penalty. Moreover, we prove that the AMP iterates converge to the SLOPE solution in an asymptotic sense, and numerical simulations show that the convergence is surprisingly fast. Our proof rests on a novel technique that specifically leverages the SLOPE problem. In contrast to prior literature, our work not only yields an asymptotically sharp analysis but also offers an algorithmic, flexible, and constructive approach to understanding the SLOPE problem.

The third chapter builds on top of the asymptotic characterization of SLOPE to study the trade-off between true positive proportion (TPP) and false discovery proportion (FDP) or, equivalently, between measures of type I error and power. Assuming a regime of linear sparsity and working under Gaussian random designs, we obtain an upper bound on the optimal trade-off for SLOPE, showing its capability of breaking the Donoho--Tanner power limit. To put it into perspective, this limit is the highest possible power that the Lasso, which is perhaps the most popular ℓ_1-based method, can achieve even with arbitrarily strong effect sizes. Next, we derive a tight lower bound that delineates the fundamental limit of sorted ℓ_1 regularization in optimally trading theFDP off for the TPP. Finally, we show that on any problem instance, SLOPE with a certain regularization sequence outperforms the Lasso, in the sense of having a smaller FDP, larger TPP, and smaller ℓ_2 estimation risk simultaneously. Our proofs are based on a novel technique that reduces a calculus of variations problem to a class of infinite-dimensional convex optimization problems and a very recent result from approximate message passing theory.

The fourth chapter works on the practical application of SLOPE by efficiently designing the SLOPE penalty sequence in the finite dimension, by restricting the number of unique values in the SLOPE penalty to be small. SLOPE's magnitude-dependent regularization requires an input of penalty sequence $\blam$, instead of a scalar penalty as in the Lasso case, thus making the design extremely expensive in computation. We propose two efficient algorithms to design the possibly high-dimensional SLOPE penalty, in order to minimize the mean squared error. For Gaussian data matrices, we propose a first-order Projected Gradient Descent (PGD) under the Approximate Message Passing regime. For general data matrices, we present a zeroth-order Coordinate Descent (CD) to design a sub-class of SLOPE, referred to as the k-level SLOPE. Our CD allows a useful trade-off between accuracy and computation speed. We demonstrate the performance of SLOPE with our designs via extensive experiments on synthetic data and real-world datasets.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Applied Mathematics

First Advisor
Weijie J. Su

Second Advisor
Qi Long

Keywords
high-dimensional statistics, linear models, machine learning, optimization algorithm, sparse models, statistical inference

Subject Categories
Applied Mathematics | Computer Sciences | Statistics and Probability

ALGORITHMIC ANALYSIS AND STATISTICAL INFERENCE OF SPARSE MODELS IN HIGH DIMENSION

Zhiqi Bu

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

____

Weijie Su
Assistant Professor in Wharton Statistics and Data Science

Co-Supervisor of Dissertation

____

Qi Long
Professor of Biostatistics in Biostatistics and Epidemiology

Graduate Group Chairperson

____

Robin Pemantle, Merriam Term Professor of Mathematics

Dissertation Committee:
Edward I. George, Universal Furniture Professor Emeritus of Statistics and Data Science
Ian J. Barnett, Assistant Professor of Biostatistics
Weijie Su, Assistant Professor in Wharton Statistics and Data Science
Qi Long, Professor of Biostatistics in Biostatistics and Epidemiology

Acknowledgments

I truly believe that the success of a Ph.D. student is attributed to many factors beyond one's own endeavor. Therefore, I owe thanks to so many people that are impossible to be fully listed here, but I will give my best try.

First and foremost, I would like to thank my family. Without their support, I would not have the opportunity to come to the United States and join the Ph.D. program at University of Pennsylvania. Their encouragement and understanding help me go through the hardest days and think bigger at all times.

I would like to thank my advisors, Weijie Su and Qi Long, for the constant support on my academic research and beyond. Especially, I sincerely appreciate the time and ideas contributed to my research, without which the completion of this dissertation would not be possible. I especially want to thank them for their rigor and vision in their research areas, as well as the critical insight and the abundant experience that they passed onto me. As an old saying goes: 'well begun is half done', my advisors' guidance is key to the fruitful results of my research. Also, I would like to thank them for the great support on my industry internship, which brought new angles to my research and directly formulated my career plan.

I am grateful to Edward I. George and Ian J. Barnett for kindly serving at my thesis defense committee. I have enjoyed our discussion during my oral exam and thank you for the commitment especially during the COVID period.

I would also like to thank all my collaborators. I want to thank Jason M. Klusowski and Cynthia Rush for contributing heavily to the main content of this thesis. I owe special thanks to Janardhan Kulkarni, Sergey Yekhanin and Gopi Sivakanth at Microsoft Research for hosting my first industry internship on differentially private algorithms. My thanks extend to Rich Caruana, Yin-Tat Lee, Ping Li, Zha Sheng, and Yu-xiang Wang during other internships and great discussion.

I would like to express my gratitude to many friends, including Jinshuo Dong, Yiliang Zhang, Kan Chen, Shiyun Xu, Hua Wang, Zongyu Dai, Matteo Sordello, Qiyiwen Zhang, Xiang Zhao, Harsha Nori, Xinyu Liao, and many others, for the academic and non-academic discussions. My thanks also go to University of Pennsylvania, particularly my program chairs Charles Epstein and Robin Pemantle, the program coordinator Ammarah Aftab, the academic writing staffs, and the providers of my scholarship and grants. I am grateful to the entire community for a great eco-system, including the editors and especially all anonymous reviewers who are extremely helpful to improve my research.

Finally, thanks to the 2017-2021 myself, who kept up the good work and never gave up on his style under the pressure.

ABSTRACT

ALGORITHMIC ANALYSIS AND STATISTICAL INFERENCE OF SPARSE MODELS IN HIGH DIMENSION

Zhiqi Bu

Weijie Su

Qi Long

The era of machine learning features large datasets that have high dimension of features. This leads to the emergence of various algorithms to learn efficiently from such high-dimensional datasets, as well as the need to analyze these algorithms from both the prediction and the statistical inference viewpoint. To be more specific, an ideal model is expected to predict accurately on the unseen new data, and to provide valid inference so as to harness the uncertainty in the model. Unfortunately, the high dimension of features poses a great challenge on the analysis of many prevalent models, rendering them either inapplicable or difficult to study.

This thesis leverages the approximate message passing (AMP) algorithm, the optimization theory, and the Sorted L-One Penalized Estimation (SLOPE) to study several important problems of the sparse models.

The first chapter introduces various ℓ_1 penalties including but not limited to the SLOPE, a relatively new convex optimization procedure via the sorted ℓ_1 penalty, in the general machine learning models. We then focus on the linear models and

demonstrate some basic properties of SLOPE, especially its advantages over the Lasso. Next, we cover the AMP algorithm in terms of convergence behavior and asymptotic statistical characterization.

The second chapter extends the AMP algorithms from Lasso to SLOPE and provides an asymptotically tight characterization of the SLOPE solution. Note that SLOPE is a relatively new convex optimization procedure for high-dimensional linear regression via the sorted ℓ_1 penalty: the larger the rank of the fitted coefficient, the larger the penalty. This non-separable penalty renders many existing techniques invalid or inconclusive in analyzing the SLOPE solution. We develop an asymptotically exact characterization of the SLOPE solution under Gaussian random designs through solving the SLOPE problem using approximate message passing (AMP). This algorithmic approach allows us to approximate the SLOPE solution via the much more amenable AMP iterates. Explicitly, we characterize the asymptotic dynamics of the AMP iterates relying on a recently developed state evolution analysis for non-separable penalties, thereby overcoming the difficulty caused by the sorted ℓ_1 penalty. Moreover, we prove that the AMP iterates converge to the SLOPE solution in an asymptotic sense, and numerical simulations show that the convergence is surprisingly fast. Our proof rests on a novel technique that specifically leverages the SLOPE problem. In contrast to prior literature, our work not only yields an asymptotically sharp analysis but also offers an algorithmic, flexible, and constructive approach to understanding the SLOPE problem.

The third chapter builds on top of the asymptotic characterization of SLOPE to study the trade-off between true positive proportion (TPP) and false discovery proportion (FDP) or, equivalently, between measures of type I error and power. Assuming a regime of linear sparsity and working under Gaussian random designs, we obtain an upper bound on the optimal trade-off for SLOPE, showing its capability of breaking the Donoho–Tanner power limit. To put it into perspective, this limit is the highest possible power that the Lasso, which is perhaps the most popular ℓ_1-based method, can achieve even with arbitrarily strong effect sizes. Next, we derive a tight lower bound that delineates the fundamental limit of sorted ℓ_1 regularization in optimally trading the FDP off for the TPP. Finally, we show that on any problem instance, SLOPE with a certain regularization sequence outperforms the Lasso, in the sense of having a smaller FDP, larger TPP, and smaller ℓ_2 estimation risk simultaneously. Our proofs are based on a novel technique that reduces a calculus of variations problem to a class of infinite-dimensional convex optimization problems and a very recent result from approximate message passing theory.

The fourth chapter works on the practical application of SLOPE by efficiently designing the SLOPE penalty sequence in the finite dimension, by restricting the number of unique values in the SLOPE penalty to be small. SLOPE's magnitude-dependent regularization requires an input of penalty sequence $\boldsymbol{\lambda}$, instead of a scalar penalty as in the Lasso case, thus making the design extremely expensive in computation. We propose two efficient algorithms to design the possibly high-

dimensional SLOPE penalty, in order to minimize the mean squared error. For Gaussian data matrices, we propose a first-order Projected Gradient Descent (PGD) under the Approximate Message Passing regime. For general data matrices, we present a zeroth-order Coordinate Descent (CD) to design a sub-class of SLOPE, referred to as the *k*-level SLOPE. Our CD allows a useful trade-off between accuracy and computation speed. We demonstrate the performance of SLOPE with our designs via extensive experiments on synthetic data and real-world datasets.

Contents

Chapter 1

Introduction

In this era of big data, machine learning methods are often applied on datasets with high dimensions of features, at the scale of hundreds to millions. From a practical perspective, large models are difficult to work with: it is time-consuming to train and tune hyperparameters for large models; the memory cost to store and deploy can be infeasible. As a consequence, sparse models that possess comparable performance in comparison to the dense models are desired. This thesis focuses on applying and analyzing ℓ_1-based regularizations which are applied to modern machine learning to obtain sparse models, starting from the classical Lasso (standing for least absolute shrinkage and selection operator) problem that traces back to 1990s. The Lasso problem applies an ℓ_1 regularization on a linear model:

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

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a known measurement matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown signal or prior, and $\boldsymbol{w} \in \mathbb{R}^n$ is the measurement noise.

The Lasso is a convex minimization problem with penalty scalar $\lambda \in \mathbb{R}$,

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{b}|_i.$$

Researches have shown that Lasso enjoys many desirable properties like exact support recovery and asymptotic consistency, under working assumptions. Especially in high dimensions, when the ordinary least squares (OLS) fail to work due to the infinite number of solutions, the Lasso can have a unique solution. In fact, the Lasso penalty can be applied to a much wider class than linear models, which covers support vector machine and neural networks. Additionally, variants of Lasso emerge to achieve better prediction and inference performance. Some prevalent examples are elastic net [ZH05a], adaptive Lasso [Zou06a], group Lasso [YL06a], and many others.

Recently in 2015, sorted L-One penalty estimation (SLOPE) [Bog+15a] is proposed to control the false discovery rate in the case of independent predictors and shown to achieve minimax estimation without any knowledge of the sparsity degree of $\boldsymbol{\beta}$. SLOPE is also a convex minimization problem, with a penalty vector $\boldsymbol{\lambda} \in \mathbb{R}^p$,

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \lambda_i |\boldsymbol{b}|_{(i)}.$$

Here $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and $|\boldsymbol{b}|_{(1)} \geq \cdots \geq |\boldsymbol{b}|_{(p)}$ are the sorted absolute values of the fitted coefficients. The idea is to penalize the larger coefficient more heavily, similar to how tax works.

One challenge to analyzing the Lasso and SLOPE is that these problems do not have closed-form solutions in general. In the past few years, the approximate message passing (AMP) algorithm for the Lasso is proposed with an asymptotic characterization theory. AMP is a class of computationally efficient and easy-to-implement algorithms for a broad range of statistical estimation problems, including compressed sensing and the LASSO. Under certain assumptions where the AMP algorithm is guaranteed to work, the AMP theory establishes the equivalence in distribution between the Lasso solution $\hat{\boldsymbol{\beta}}$ and a proximal operator:

$$\hat{\beta}_j = \eta(B + \tau Z; \alpha\tau),$$

where B is the distribution from which elements of $\boldsymbol{\beta}$ are i.i.d. drawn, Z is an independent standard Gaussian, and (α, τ) are two scalars derived from the AMP algorithm. The η function is the proximal operator which is known as the soft-thresholding function in the case of the Lasso. With this tight characterization, statistical quantities such as true positive proportion (TPP), false discovery proportion (FDP), and estimiation risk can be computed and analyzed.

To develop the corresponding characterization for SLOPE, we have to overcome many technical difficulties mainly caused by the non-separability of the sorted ℓ_1 norm penalty in SLOPE. To be a bit more specific, the non-separability makes the proximal operator of SLOPE to be much less analytical than that of the Lasso and requires us to guarantee the convergence of the AMP algorithm, as well as the statistical characterization of SLOPE, in a more complicated way. Another challenge lies in the practical application of SLOPE: the length p SLOPE penalty vector is computationally expensive to design than a single scalar penalty in the Lasso.

In summary, this thesis tackles the above challenges and makes the following contributions:

1. We propose the SLOPE AMP, an optimization algorithm that provably solves the SLOPE problem (Chapter 2).

2. We present the SLOPE AMP theory that characterizes the SLOPE solution asymptotically tightly (Chapter 2).

3. We use the characterization to analyze TPP, FDP, and estimation risk of SLOPE. Consequently, we can describe the TPP-FDP trade-off and show that SLOPE overcomes the Donoho-Tanner power limit, in fact achieving full power (Chapter 3).

4. We show the outperformance of SLOPE over the Lasso in the fixed prior scenario, by using 2-level SLOPE (Chapter 3).

5. We show SLOPE can achieve significantly smaller estimation risk than the Lasso asymptotically. We further illustrate a practically efficient method to design the k-level SLOPE that empirically achieves low estimation risk in the finite dimension. (Chapter 4)

While the previous introduction mainly focuses on linear models, which are preferred for their simplicity and interpretability in many areas such as healthrelated data analysis and economic forecasts, we believe SLOPE and its variants (e.g. k-level SLOPE and group SLOPE) can work compatibly with other machine learning models.

Chapter 2

Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing

This chapter is based on "Zhiqi Bu, Jason M. Klusowski, Cynthia Rush, and Weijie J. Su. "Algorithmic analysis and statistical estimation of SLOPE via approximate message passing." IEEE Transactions on Information Theory 67, no. 1 (2020): 506-537.".

2.1 Introduction

Consider observing linear measurements $\boldsymbol{y} \in \mathbb{R}^n$ that are modeled by the equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tag{2.1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a known measurement matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown signal, and $\boldsymbol{w} \in \mathbb{R}^n$ is the measurement noise. Among numerous methods that seek to recover the signal $\boldsymbol{\beta}$ from the observed data, especially in the setting where $\boldsymbol{\beta}$ is sparse and p is larger than n, SLOPE has recently emerged as a useful procedure that allows for estimation and model selection [Bog+15a]. This method reconstructs the signal by solving the minimization problem

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \lambda_i |\boldsymbol{b}|_{(i)}, \tag{2.1.2}$$

where $\|\cdot\|$ denotes the ℓ_2 norm, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ (with at least one strict inequality) is a sequence of thresholds, and $|\boldsymbol{b}|_{(1)} \geq \cdots \geq |\boldsymbol{b}|_{(p)}$ are the order statistics of the fitted coefficients in absolute value. The regularizer $\sum \lambda_i |\boldsymbol{b}|_{(i)}$ is a *sorted* ℓ_1*-norm* (denoted as $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ henceforth), which is *non-separable* due to the sorting operation involved in its calculation. Notably, SLOPE has two attractive features that are not simultaneously present in other methods for linear regression including the LASSO [Tib96a] and knockoffs [BC+15]. Explicitly, on the estimation side, SLOPE achieves minimax estimation properties under certain random designs *without* requiring any knowledge of the sparsity degree of $\boldsymbol{\beta}$ [SC16; BLT18]. On the testing side, SLOPE controls the false discovery rate in the case of independent predictors [Bog+15a; Brz+18]. For completeness, we remark that [BR08; ZF14; FN16] proposed similar non-separable regularizers to encourage grouping of correlated predictors.

This work is concerned with the algorithmic aspects of SLOPE through the lens of *approximate message passing* (AMP) [BM11a; DMM09a; Krz+12; Ran11].

AMP is a class of computationally efficient and easy-to-implement algorithms for a broad range of statistical estimation problems, including compressed sensing and the LASSO [BM11c]. When applied to SLOPE, AMP takes the following form: at initial iteration $t = 0$, assign $\boldsymbol{\beta}^0 = \mathbf{0}, \boldsymbol{z}^0 = \boldsymbol{y}$, and for $t \geq 0$,

$$\boldsymbol{\beta}^{t+1} = \operatorname{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t), \tag{2.1.3a}$$

$$\boldsymbol{z}^{t+1} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{t+1} + \frac{\boldsymbol{z}^t}{n}\Big[\nabla \operatorname{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t)\Big]. \tag{2.1.3b}$$

The non-increasing sequence $\boldsymbol{\theta}_t$ is proportional to $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)$ and will be given explicitly in Section 2.2. Here, $\operatorname{prox}_{J_{\boldsymbol{\theta}}}$ is the proximal operator of the sorted ℓ_1 norm, that is,

$$\operatorname{prox}_{J_{\boldsymbol{\theta}}}(\boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{b}\|^2 + J_{\boldsymbol{\theta}}(\boldsymbol{b}), \tag{2.1.4}$$

and $\nabla \operatorname{prox}_{J_\theta}$ denotes the divergence of the proximal operator (see an equivalent, but more explicit form, of this algorithm in Section 2.2 and further discussion of SLOPE and the prox operator in Section 2.5.1). Compared to the proximal gradient descent (ISTA) [Cha+98; DDDM04; PB14], AMP has an extra correction term in its residual step that adjusts the iteration in a non-trivial way and seeks to provide improved convergence performance [DMM09a].

The *empirical* performance of AMP in solving SLOPE under i.i.d. Gaussian matrix $\boldsymbol{X}$ is illustrated in Figure 2.1 and Table 2.1, which suggest the superiority of AMP over ISTA and FISTA [BT09]—perhaps the two most popular proximal gradient descent methods—in terms of speed of convergence in this setting. However, the vast AMP literature thus far remains silent on whether AMP *provably* solves SLOPE and,

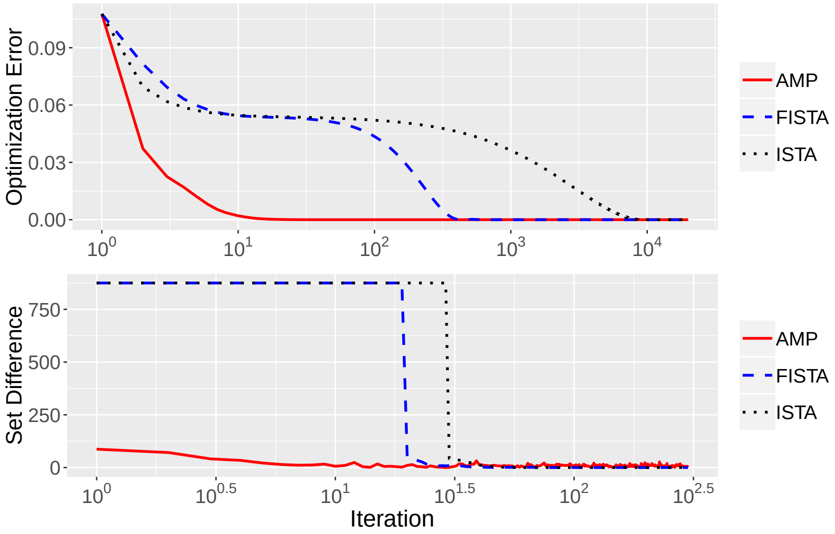


Figure 2.1: Optimization errors, $||\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}||^2/p$, and (symmetric) set difference of $\text{supp}(\boldsymbol{\beta}^t)$ and $\text{supp}(\widehat{\boldsymbol{\beta}})$.

		Optimization errors				
	Set Diff	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
ISTA	60	4048	7326	8569	9007	9161
FISTA	47	275	374	412	593	604
AMP	30	6	13	22	32	40

Table 2.1: First iteration t for which there is zero set difference or optimization error $||\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}||^2/p$ falls below a threshold.

Figure 2.1 and Table 2.1 Details: Design X is 500×1000 with i.i.d. $\mathcal{N}(0, 1/500)$ entries. True signal $\boldsymbol{\beta}$ is i.i.d. Gaussian-Bernoulli: $\mathcal{N}(0, 1)$ with probability 0.1 and 0 otherwise. Noise variance $\sigma_w^2 = 0$. A careful calibration between the thresholds $\boldsymbol{\theta}_t$ in AMP and $\boldsymbol{\lambda}$ is SLOPE is used (details in Sec. 2.2).

if so, whether one can leverage AMP to get insights into the statistical properties of SLOPE. This vacuum in the literature is due to the *non-separability* of the SLOPE regularizer, making it a major challenge to apply AMP to SLOPE directly. In stark contrast, AMP theory has been rigorously applied to the LASSO [BM11c], showing both good empirical performance and nice theoretical properties of solving the LASSO using AMP. Moreover, AMP in this setting allows for asymptotically exact statistical characterization of its output, which converges to the LASSO solution, thereby providing a powerful tool in fine-grained analyses of the LASSO [BEM13; SBC17; MMB+18a; RV18].

Main contributions. In this work, we prove that the AMP algorithm (2.1.3) solves the SLOPE problem in an asymptotically *exact* sense under independent Gaussian random designs. Our proof uses the recently extended AMP theory for non-separable denoisers [BMN20] and applies this tool to derive the state evolution that describes the asymptotically exact behaviors of the AMP iterates $\boldsymbol{\beta}^t$ in (2.1.3). The next step, which is the core of our proof, is to relate the AMP estimates to the SLOPE solution. This presents several challenges that *cannot* be resolved only within the AMP framework. In particular, unlike the LASSO, the number of nonzeros in the SLOPE solution can exceed the number of observations. This fact imposes substantially more difficulties on showing that the distance between the SLOPE solution and the AMP iterates goes to zero than in the LASSO case due to the possible *non-strong convexity* of the SLOPE problem, even restricted to the solution support. To overcome these challenges, we develop novel techniques that are tailored to the characteristics of the SLOPE solution. For example, our proof relies on the crucial property of SLOPE that the *unique* nonzero components of its solution never outnumber the observation units.

As a byproduct, our analysis gives rise to an *exact* asymptotic characterization of the SLOPE solution under independent Gaussian random designs through leveraging the statistical aspect of the AMP theory. In more detail, the probability distribution of the SLOPE solution is completely specified by a few parameters that are the solution to a certain fixed-point equation in an asymptotic sense. This provides a powerful tool for fine-grained statistical analysis of SLOPE as it was for the LASSO problem. We note that a recent paper [HL19a]—which takes an entirely different path—gives an asymptotic characterization of the SLOPE solution that matches our asymptotic analysis deduced from our AMP theory for SLOPE. However, our AMP-based approach is more algorithmic in nature and offers a more concrete connection between the finite-sample behaviors of the SLOPE problem and its asymptotic distribution via the computationally efficient AMP algorithm.

Paper outline. In Section 2.2 we develop an AMP algorithm for finding the SLOPE estimator in (2.1.2). Specifically, it is through the threshold values $\boldsymbol{\theta}_t$ in the AMP algorithm in (2.1.3) that one can ensure the AMP estimates converge to the SLOPE estimator with parameter $\boldsymbol{\lambda}$, so in Section 2.2 we provide details for how one should calibrate the thresholds of the AMP iterations in (2.1.3) in order for the algorithm to solve SLOPE cost in (2.1.2). Then in Section 2.3, we state theoretical guarantees showing that the AMP algorithm solves the SLOPE optimization asymptotically and we leverage theoretical guarantees for the AMP algorithm to exactly characterize the mean square error (more generally, any pseudo-Lipschitz error) of the SLOPE estimator in the large system limit. This is done by applying recent theoretical results for AMP algorithms that use a non-separable non-linearity [BMN20], like the one in (2.1.3). Finally, Sections 2.4-2.7 prove rigorously the theoretical results stated in Section 2.3 and we end with a discussion in Section 2.8.

2.2 Algorithmic Development

To begin with, we state assumptions under which our theoretical results will hold and give some preliminary ideas about SLOPE that will be useful in the development of the AMP algorithm.

Assumptions. Concerning the linear model (2.1.1) and parameter vector in (2.1.2), we assume:

(A1) The measurement matrix $\boldsymbol{X}$ has independent and identically-distributed (i.i.d.) Gaussian entries that have mean 0 and variance $1/n$.

(A2) The signal $\boldsymbol{\beta}$ has elements that are i.i.d. B, with $\mathbb{E}(B^2 \max\{0, \log B\}) < \infty$.

(A3) The noise $\boldsymbol{w}$ is elementwise i.i.d. W, with $\sigma_w^2 := \mathbb{E}(W^2) < \infty$.

(A4) The vector $\boldsymbol{\lambda}(p) = (\lambda_1, \ldots, \lambda_p)$ is elementwise i.i.d. Λ, with $\mathbb{E}(\Lambda^2) < \infty$ and $\min\{\boldsymbol{\lambda}(p)\} > 0$.

(A5) The ratio n/p approaches a constant $\delta \in (0, \infty)$ in the large system limit, as $n, p \to \infty$.

Remark: **(A4)** can be relaxed as $\lambda_1, \ldots, \lambda_p$ having an empirical distribution that converges weakly to probability measure Λ on $\mathbb{R}$ with $\mathbb{E}(\Lambda^2) < \infty$ and $\|\boldsymbol{\lambda}(p)\|^2/p \to \mathbb{E}(\Lambda^2)$ and $\min\{\boldsymbol{\lambda}(p)\} > 0$. A similar relaxation can be made for the distributional assumptions **(A2)** and **(A3)**.

SLOPE preliminaries. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, the divergence of the proximal operator, $\nabla \operatorname{prox}_f(\boldsymbol{v})$, is given by the following:

$$\nabla \operatorname{prox}_f(\boldsymbol{v}) := \sum_{i=1}^{p} \frac{\partial}{\partial v_i} [\operatorname{prox}_f(\boldsymbol{v})]_i = \left(\frac{\partial}{\partial v_1}, \frac{\partial}{\partial v_2}, \ldots, \frac{\partial}{\partial v_p} \right) \cdot \operatorname{prox}_f(\boldsymbol{v}), \tag{2.2.1}$$

where [SC16, proof of Fact 3.4],

$$\frac{\partial [\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i}{\partial v_j} = \begin{cases} \frac{\operatorname{sign}([\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i) \cdot \operatorname{sign}([\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j)}{\#\{1 \leq k \leq p : |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_k| = |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j|\}}, & \text{if } |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j| = |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i|, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2.2}$$

Hence the divergence takes the simplified form

$$\nabla \operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}) = \| \operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}) \|_0^*, \tag{2.2.3}$$

where $\| \cdot \|_0^*$ counts the unique non-zero magnitudes in a vector, e.g. $\|(0, 1, -2, 0, 2)\|_0^*$ $= 2$. This explicit form of divergence not only waives the need to use approximation in calculation but also speed up the recursion, since it only depends on the proximal operator as a whole instead of on $\boldsymbol{\theta}_{t-1}, \boldsymbol{X}, \boldsymbol{z}^{t-1}, \boldsymbol{\beta}^{t-1}$. Therefore, we have

Lemma 2.2.1. *In AMP,* (2.1.3b) *is equivalent to* $\boldsymbol{z}^{t+1} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{t+1} + \frac{\boldsymbol{z}^t}{\delta p}\|\boldsymbol{\beta}^{t+1}\|_0^*$.

Other details and background on SLOPE and the prox operator are found in Section 2.5.1. Now we discuss the details of an AMP algorithm that can be used for finding the SLOPE estimator in (2.1.2).

2.2.1 AMP Background

An attractive feature of AMP is that its statistical properties can be exactly characterized at each iteration t, at least asymptotically, via a one-dimensional recursion known as state evolution [BM11a; BMN20; RV18; JM13]. Specifically, it can be shown that the pseudo-data, meaning the input $\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t$ for the estimate of the unknown signal in (2.1.3a), is asymptotically equal in distribution to the true signal plus independent, Gaussian noise, i.e. $\boldsymbol{\beta} + \tau_t \boldsymbol{Z}$, where the noise variance τ_t is defined by the state evolution. For this reason, the function used to update the estimate in (2.1.3a), in our case, the proximal operator, $\text{prox}_{J_{\boldsymbol{\theta}_t}}(\cdot)$, is usually referred to as a 'denoiser' in the AMP literature.

This statistical characterization of the pseudo-data was first rigorously shown to be true in the case of 'separable' denoisers by Bayati and Montanari [BM11a], and an analysis of the rate of this convergence was given in [RV18]. A 'separable' denoiser is one that applies the same (possibly non-linear) function to each element of its input. Recent work [BMN20] proves that the pseudo-data has distribution $\boldsymbol{\beta} + \tau_t \boldsymbol{Z}$ asymptotically, even when the 'denoisers' used in the AMP algorithm are non-separable, like the SLOPE prox operator in (2.1.3a).

As mentioned previously, the dynamics of the AMP iterations are tracked by a recursive sequence referred to as the state evolution, defined as follows. For $\boldsymbol{B}$ elementwise i.i.d. B independent of $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$, let $\tau_0^2 = \sigma_w^2 + \mathbb{E}[B^2]/\delta$ and for $t \geq 0$,

$$\tau_{t+1}^2 = \sigma_w^2 + \lim_p \frac{1}{\delta p} \mathbb{E}\|\text{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{B} + \tau_t \boldsymbol{Z}) - \boldsymbol{B}\|^2. \tag{2.2.4}$$

Below we make rigorous the way that the recursion in (2.2.4) relates to the AMP iteration (2.1.3).

We note that throughout, we let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian density with mean μ and variance σ^2 and we use $\mathbb{I}_p$ to indicate a $p \times p$ identity matrix.

2.2.2 Analysis of the AMP State Evolution

As the state evolution (2.2.4) predicts the performance of the AMP algorithm (2.1.3) (the pseudo-data, $\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t$, is asymptotically equal in distribution $\boldsymbol{\beta} + \tau_t \boldsymbol{Z}$), it is of interest to study the large t asymptotics of (2.2.4). Moreover, recall that through the sequence of thresholds $\boldsymbol{\theta}_t$, one can relate the AMP algorithm to the SLOPE estimator in (2.1.2) for a specific $\boldsymbol{\lambda}$, and the explicit form of this calibration, given in Section 2.2.3, is motivated by such asymptotic analysis of the state evolution.

It turns out that a finite-size approximation, which we denote $\tau_t^2(p)$, will be easier to analyze than (2.2.4). The definition of $\tau_{t+1}^2(p)$ is stated explicitly in (2.2.5) below. Throughout the work, we will define thresholds $\boldsymbol{\theta}_t := \boldsymbol{\alpha}\tau_t(p)$ for every iteration t where the vector $\boldsymbol{\alpha}$ is fixed via a calibration made explicit in Section 2.2.3. We can interpret this to mean that within the AMP algorithm, $\boldsymbol{\alpha}$ plays the role of the regularizer parameter, $\boldsymbol{\lambda}$. Now we define $\tau_{t+1}^2(p)$, for large p, as a finite-sample approximation to (2.2.4), namely

$$\tau_{t+1}^2(p) = \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau_t(p)}}(\boldsymbol{\beta} + \tau_t(p)\boldsymbol{Z}) - \boldsymbol{\beta}\|^2, \tag{2.2.5}$$

where the difference between (2.2.5) and the state evolution (2.2.4) is via the large system limit in p. When we refer to the recursion in (2.2.5), we will always specify the p dependence explicitly as $\tau_t(p)$. An analysis of the limiting properties (in t) of (2.2.5) is given in Theorem 1 below, after which it is then argued that because interchanging limits and differentiation is justified, the large t analysis of (2.2.5) holds for (2.2.4) as well. Before presenting Theorem 1, however, we give the following result which motivates why the AMP iteration should relate at all to the SLOPE estimator.

Lemma 2.2.2. *Any stationary point $\widehat{\boldsymbol{\beta}}$ (with corresponding $\widehat{\boldsymbol{z}}$) in the AMP algorithm* (2.1.3a)-(2.1.3b) *with $\boldsymbol{\theta}_* = \boldsymbol{\alpha}\tau_*$ is a minimizer of the SLOPE cost function in* (2.1.2) *with*

$$\boldsymbol{\lambda} = \boldsymbol{\theta}_*\Big(1 - \frac{1}{\delta p}\left(\nabla \text{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}})\right)\Big) = \boldsymbol{\theta}_*\Big(1 - \frac{1}{n}\left\|\text{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}})\right\|_0^*\Big).$$

Proof of Lemma 2.2.2. Denote, $w := (\nabla \text{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}}))/(\delta p)$. Now, by stationarity,

$$\widehat{\boldsymbol{\beta}} = \operatorname{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}}), \qquad \text{and} \qquad \widehat{\boldsymbol{z}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \frac{\widehat{\boldsymbol{z}}}{\delta p}(\nabla \operatorname{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}})). \tag{2.2.6}$$

From (2.2.6), notice that $\widehat{\boldsymbol{z}} = \frac{\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{1-w}$. By Fact 2.5.2, $\boldsymbol{X}^\top \widehat{\boldsymbol{z}} \in \partial J_{\boldsymbol{\theta}_*}(\widehat{\boldsymbol{\beta}})$, where $\partial J_{\boldsymbol{\theta}_*}(\widehat{\boldsymbol{\beta}})$ is the subgradient of $J_{\boldsymbol{\theta}_*}(\cdot)$ at $\widehat{\boldsymbol{\beta}}$ (a precise definition of a subgradient is given in Section 2.5.1). Then, $\boldsymbol{X}^\top \widehat{\boldsymbol{z}} = \frac{\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})}{1-w} \in J_{\boldsymbol{\theta}_*}(\widehat{\boldsymbol{\beta}})$, and therefore $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in J_{\boldsymbol{\theta}_*(1-w)}(\widehat{\boldsymbol{\beta}})$ which is *exactly* the stationary condition of SLOPE with regularization parameter $\boldsymbol{\lambda} = (1-w)\boldsymbol{\theta}_*$, as desired. □

Now we present Theorem 1, which provides results about the t asymptotics of the recursion in (2.2.5) and its proof is given in Appendix 2.9.1. First, some notation must be introduced: let $\boldsymbol{A}_{\min}(\delta)$ be the set of solutions to

$$\begin{aligned} &\delta = f(\boldsymbol{\alpha}), \\ \text{where} \quad &f(\boldsymbol{\alpha}) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{\Big(1 - |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_i| \sum_{j \in I_i} \alpha_j\Big) \Big/ [\boldsymbol{D}(\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z}))]_i\Big\}. \end{aligned} \tag{2.2.7}$$

Here $\odot$ represents elementwise multiplication of vectors and for vector $\boldsymbol{v} \in \mathbb{R}^p$, $\boldsymbol{D}$ is defined elementwise as $[\boldsymbol{D}(\boldsymbol{v})]_i = \#\{j : |v_j| = |v_i|\}$ if $v_i \neq 0$ and ∞ otherwise. Let $I_i = \{j : 1 \leq j \leq p \text{ and } |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_j| = |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_i|\}$. The expectation in (2.2.7) is taken with respect to $\boldsymbol{Z}$, a p-length vector of i.i.d. standard Gaussians. Finally, for $\boldsymbol{u} \in \mathbb{R}^m$, the notation $\langle \boldsymbol{u} \rangle := \sum_{i=1}^m u_i/m$ and we say $\boldsymbol{u}$ is strictly larger than $\boldsymbol{v} \in \mathbb{R}^m$ if $u_i > v_i$ for all elements $i \in \{1, 2, \ldots, m\}$. For the simple case of $p = 2$, we illustrate an example of the set $\boldsymbol{A}_{\min}(\delta)$ in Figure 2.2.

Theorem 1. *For any $\boldsymbol{\alpha}$ strictly larger than at least one element in the set $\boldsymbol{A}_{\min}(\delta)$, the recursion in* (2.2.5) *has a unique fixed point that we denote as $\tau_*^2(p)$. Then $\tau_t(p) \to \tau_*(p)$ monotonically for any initial condition. Define a function $\mathsf{F} : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ as*

$$\mathsf{F}\Big(\tau^2(p), \boldsymbol{\alpha}\tau(p)\Big) := \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau(p)}}(\boldsymbol{B} + \tau(p)\boldsymbol{Z}) - \boldsymbol{B}\|^2, \tag{2.2.8}$$

where $\boldsymbol{B}$ is elementwise i.i.d. B independent of $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$, so that $\tau_{t+1}^2(p) = \mathsf{F}(\tau_t^2(p), \boldsymbol{\alpha}\tau_t(p))$. Then $|\frac{\partial \mathsf{F}}{\partial \tau^2(p)}(\tau^2(p), \boldsymbol{\alpha}\tau(p))| < 1$ at $\tau(p) = \tau_*(p)$. Moreover, for $f(\boldsymbol{\alpha})$ defined in (2.2.7)*, we show that $f(\boldsymbol{\alpha}) = \delta \lim_{\tau(p)\to\infty} d\mathsf{F}/d\tau^2(p)$.*

Beyond providing the large t asymptotics of the state evolution sequence, notice that Theorem 1 gives necessary conditions on the calibration vector $\boldsymbol{\alpha}$ under which the recursion in (2.2.5), and equivalently, the calibration detailed in Section 2.2.3 below are well-defined.

Recall that it is actually the state evolution in (2.2.4) (and not that in (2.2.5)) that predicts the performance of the AMP algorithm, and therefore we would really like a version of Theorem 1 studying the large system limit in p. We argue that because interchanging differentiation and the limit, the proof of Theorem 1 analyzing (2.2.5), can easily be used to give an analogous result for (2.2.4). In particular analyzing (2.2.4) via the strategy given in the proof of Theorem 1 requires that we study the partial derivative of $\lim_p \mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z}) - \boldsymbol{B}\|^2/(\delta p)$, with respect to τ^2. Indeed, to directly make use our proof for the finite-p case given in Theorem 1, it is enough that

$$\frac{\partial}{\partial \tau^2} \lim_p \mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau \boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p) = \lim_p \frac{\partial}{\partial \tau^2}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau \boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p). \tag{2.2.9}$$

Note that we already have an argument (based on dominated convergence for fixed p, see (2.9.1) and Lemma 2.9.1) showing that

$$\frac{\partial}{\partial \tau^2}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau \boldsymbol{Z})-\boldsymbol{B}\|^2 = \mathbb{E}\Big\{\frac{\partial}{\partial \tau^2}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau \boldsymbol{Z})-\boldsymbol{B}\|^2\Big\}.$$

The next lemma gives us a roadmap for how to proceed (c.f., [Rud+64, Theorem 7.17]) to justify the interchange in (2.2.9).

Lemma 2.2.3. *Suppose $\{g_m\}$ is a sequence of functions that converge pointwise to g on a compact domain D and whose derivatives $\{g_m'\}$ converge uniformly to a function h on D. Then $h = g'$ on D.*

Therefore, taking $\{g_p\} = \{\mathsf{F}(\tau^2(p), \boldsymbol{\alpha}\tau(p))\}$, it suffices to show that if

$$\frac{\partial \mathsf{F}}{\partial \tau^2(p)}(\tau^2(p), \boldsymbol{\alpha}\tau(p)) = \frac{\partial}{\partial \tau^2(p)}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau(p)}}(\boldsymbol{B}+\tau(p) \boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p),$$

then the sequence $\{\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)\}_p$ converges uniformly as $p \to \infty$. The main tool for proving such a result is given in the following lemma.

Lemma 2.2.4. *Suppose $\{g_m\}$ is a sequence of L-Lipschitz functions (where L is independent of m) that converge pointwise to a function g on a compact domain D. Then, the convergence is also uniform on D.*

Using this lemma, the essential idea is to show that there exists a constant $L > 0$, independent of p, such that for all p and all τ_1, τ_2 in a bounded set $D = \{\tau : 0 < r \leq |\tau| \leq R\}$,

$$\Big|\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau_1^2, \boldsymbol{\alpha}\tau_1) - \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau_2^2, \boldsymbol{\alpha}\tau_2)\Big| \leq L|\tau_1 - \tau_2|.$$

This follows by the mean value theorem and (2.9.14), with

$$L = \sup_{p,\tau \in D} |\frac{\partial}{\partial \tau^2}\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)| < +\infty$$

.

Remark 2.2.5. The boundedness of $\{\tau_t(p)\}$ is guaranteed by Proposition 2.2.6. In particular, since $\boldsymbol{\alpha}$ satisfies the assumption of Theorem 1, Proposition 2.2.6 guarantees $\boldsymbol{\lambda}$ is bounded and, consequently, so is τ (see the calibration in (2.2.10) below).

2.2.3 Threshold Calibration

Motivated by Lemma 2.2.2 and the result of Theorem 1, we define a calibration from the regularization parameter $\boldsymbol{\lambda}$, to the corresponding threshold $\boldsymbol{\alpha}$ used to define the AMP algorithm. In practice, we will be given finite-length $\boldsymbol{\lambda}$ and then we want to design the AMP iteration to solve the corresponding SLOPE cost. We do this by choosing $\boldsymbol{\alpha}$ as the vector that solves $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\alpha})$ where

$$\boldsymbol{\lambda}(\boldsymbol{\alpha}) := \boldsymbol{\alpha}\tau_*(p)\Big(1 - \frac{1}{n}\mathbb{E}\| \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*(p)}}(\boldsymbol{B} + \tau_*(p)\boldsymbol{Z})\|_0^*\Big), \tag{2.2.10}$$

where $\boldsymbol{B}$ is elementwise i.i.d. B independent of $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$ and $\tau_*(p)$ is the limiting value defined in Theorem 1. We note the fact that the calibration in (2.2.10) sets $\boldsymbol{\alpha}$ as a vector *in the same direction* as $\boldsymbol{\lambda}$, but that is scaled by a constant value (for each p), where the scaling constant value is $\tau_*(p)(1 - \mathbb{E}\| \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*(p)}}(\boldsymbol{B} + \tau_*(p)\boldsymbol{Z})\|_0^*/n)$.

In Proposition 2.2.6 we show that the calibration (2.2.10) and its inverse $\boldsymbol{\lambda} \mapsto \boldsymbol{\alpha}(\boldsymbol{\lambda})$ are well-defined and in Algorithm 1 we show that determining the calibration is straightforward in practice.

Proposition 2.2.6. *The function* $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ *defined in* (2.2.10) *is continuous on* $\{\boldsymbol{\alpha} : f(\boldsymbol{\alpha}) < \delta\}$ *for* $f(\cdot)$ *defined in* (2.2.7) *with* $\boldsymbol{\lambda}(\boldsymbol{A}_{\min}) = -\infty$ *and* $\lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = \infty$ *(where the limit is taken elementwise). Therefore the function* $\boldsymbol{\lambda} \mapsto \boldsymbol{\alpha}(\boldsymbol{\lambda})$ *satisfying* (2.2.10) *exists. As* $p \to \infty$*, the function* $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ *becomes invertible (given* $\boldsymbol{\lambda}$*,* $\boldsymbol{\alpha}$ *satisfying* (2.2.10) *exists uniquely). Furthermore, the inverse function is continuous non-decreasing for any* $\boldsymbol{\lambda} > \mathbf{0}$.

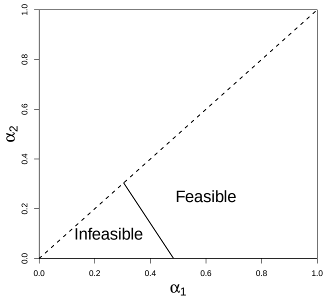


Figure 2.2: $\boldsymbol{A}_{\min}$ (black curve) when $p = 2$ and $\delta = 0.6$.

In [BM11c, Proposition 1.4 and Corollary 1.7] this is proven rigorously for the analogous LASSO calibration and in Appendix 2.9.1 we show how to adapt this proof to SLOPE case. This proposition motivates Algorithm 1 which uses a bisection method to find the unique

$\boldsymbol{\alpha}$ for each $\boldsymbol{\lambda}$. It suffices to find two guesses of $\boldsymbol{\alpha}$ parallel to $\boldsymbol{\lambda}$ that, when mapped via (2.2.10), sandwich the true $\boldsymbol{\lambda}$.

Algorithm 1 Calibration from $\boldsymbol{\lambda} \to \boldsymbol{\alpha}$

```
1. Initialize $\alpha_1 = \alpha_{\min}$ such that $\alpha_{\min}\boldsymbol{\ell} \in \boldsymbol{A}_{\min}$, where $\boldsymbol{\ell} := \boldsymbol{\lambda}/\lambda_1$; Initialize $\alpha_2 = 2\alpha_1$
while $L(\alpha_2) < 0$ where $L : \mathbb{R} \to \mathbb{R}; \alpha \mapsto \text{sign}(\boldsymbol{\lambda}(\alpha\boldsymbol{\ell}) - \boldsymbol{\lambda})$ do
    2. Set $\alpha_1 = \alpha_2, \alpha_2 = 2\alpha_2$
3. return BISECTION $(L(\alpha), \alpha_1, \alpha_2)$
```

Remark: $\text{sign}(\boldsymbol{\lambda}(\cdot) - \boldsymbol{\lambda}) \in \mathbb{R}$ is well-defined since $\boldsymbol{\lambda}(\cdot) \parallel \boldsymbol{\lambda}$ implies all entries share the same sign. The function "**BISECTION**(L, a, b)" finds the root of L in $[a, b]$ via the bisection method.

The calibration in (2.2.10) is exact when $p \to \infty$, so we study the mapping between $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ in this limit. Recall from **(A4)**, that the sequence of vectors $\{\boldsymbol{\lambda}(p)\}_{p\geq 0}$ are drawn i.i.d. from distribution Λ. It follows that the sequence $\{\boldsymbol{\alpha}(p)\}_{p\geq 0}$ defined for each p by the finite-sample calibration (2.2.10) are i.i.d. from a distribution A, where A satisfies $\mathbb{E}(A^2) < \infty$, and is defined via

$$\Lambda = A\tau_*\Big(1 - \lim_{p} \frac{1}{\delta p}\mathbb{E}|| \operatorname{prox}_{J_{\boldsymbol{A}(p)\tau_*}}(\boldsymbol{B} + \tau_*\boldsymbol{Z})||_0^*\Big), \tag{2.2.11}$$

We note, moreover, that the calibrations presented in this section are well-defined:

Fact 2.2.7. *The limits in* (2.2.4) *and* (2.2.11) *exist.*

This fact is proven in Appendix 2.9.3. One idea used in the proof of Fact 2.2.7 is that the prox operator is *asymptotically* separable, a result shown by [HL19a, Proposition 1]. Specifically, for sequences of input, $\{\boldsymbol{v}(p)\}$, and thresholds, $\{\boldsymbol{\lambda}(p)\}$, having empirical distributions that weakly converge to distributions V and Λ, respectively, then there exists a limiting scalar function $h(\cdot) := h(\boldsymbol{v}(p); V, \Lambda)$ (determined by V and Λ) of the proximal operator $\text{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}(p))$. Further details are given in Lemma 2.3.3 in Section 2.3. Using $h(\cdot) := h(\cdot; B + \tau_* Z, A\tau_*)$, this argument implies that (2.2.4) can be represented as

$$\tau_*^2 := \sigma_w^2 + \frac{1}{\delta}\mathbb{E}(h(B + \tau_* Z) - B)^2,$$

and if we denote m as the Lebesgue measure, then the limit in (2.2.11) can be represented as

$$\mathbb{P}\left(B + \tau_* Z \in \left\{x \,\middle|\, h(x) \neq 0 \quad \text{and} \quad m\{z \mid |h(z)| = |h(x)|\} = 0\right\}\right). \qquad (2.2.12)$$

In other words, the limit in (2.2.11) is the Lebesgue measure of the domain of the quantile function of h for which the quantile of h assumes unique values (i.e., is not flat).

2.3 Asymptotic Characterization of SLOPE

2.3.1 AMP Recovers the SLOPE Estimate

Here we show that the AMP algorithm converges in ℓ_2 to the SLOPE estimator, implying that the AMP iterates can be used as a surrogate for the global optimum of the SLOPE cost function. The schema of the proof is similar to [BM11c, Lemma 3.1], however, major differences lie in the fact that the proximal operator used in the AMP updates (2.1.3a)-(2.1.3b) is non-separable. We sketch the proof here, and a forthcoming article will be devoted to giving a complete and detailed argument.

Theorem 2. *Under assumptions **(A1)** - **(A5)**, for the output of the AMP algorithm in* (2.1.3a) *and the SLOPE estimate* (2.1.2)*,*

$$\operatorname*{plim}_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2 = c_t, \quad \textit{where} \quad \lim_{t\to\infty} c_t = 0. \tag{2.3.1}$$

The proof of Theorem 2 can be found in Section 2.4. At a high level, the proof requires dealing carefully with the fact that the SLOPE cost function, $\mathcal{C}(\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{b})$, given in (2.1.2) is *not* necessarily strongly convex, meaning that we could encounter the undesirable situation where $\mathcal{C}(\widehat{\boldsymbol{\beta}})$ is close to $\mathcal{C}(\boldsymbol{\beta})$ but $\widehat{\boldsymbol{\beta}}$ is not close to $\boldsymbol{\beta}$, meaning the statistical recovery of $\boldsymbol{\beta}$ would be poor.

In the LASSO case, one works around this challenge by showing that the (LASSO) cost function does have nice properties when considering just the elements of the non-zero support of $\boldsymbol{\beta}^t$ at any (large) iteration t. In the LASSO case, the non-zero support of $\boldsymbol{\beta}$ has size no larger than $n < p$.

In the SLOPE problem, however, it is possible that the support set has size exceeding n, and therefore the LASSO analysis is not immediately applicable. Our proof develops novel techniques that are tailored to the characteristics of the SLOPE solution. Specifically, when considering the SLOPE problem, one can show nice properties (similar to those in the LASSO case) by considering a support-like set, that being the *unique* non-zeros in the estimate $\boldsymbol{\beta}^t$ at any (large) iteration t. In other words, if we define an equivalence relation $x \sim y$ when $|x| = |y|$, then entries of AMP estimate at any iteration t are partitioned into equivalence classes. Then we observe from (2.2.10), and the non-negativity of $\boldsymbol{\lambda}$, that the number of equivalence classes is no larger than n. We see an analogy between SLOPE's equivalence class (or 'maximal atom' as described in Appendix 2.5.1) and LASSO's support set. This approach allows us to deal with the lack of a strongly convex cost.

Theorem 2 ensures that the AMP algorithm solves the SLOPE problem in an asymptotic sense. To better appreciate the convergence guarantee, it calls for elaboration on (2.3.1). First, it implies that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2/p$ converges in probability to a constant, say c_t. Next, (2.3.1) says $c_t \to 0$ as $t \to \infty$.

2.3.2 Exact Asymptotic Characterization of the SLOPE Estimate

A consequence of Theorem 2.4.1, is that the SLOPE estimator $\widehat{\boldsymbol{\beta}}$ inherits performance guarantees provided by the AMP state evolution, in the sense of Theorem 3 below. Theorem 3 provides as asymptotic characterization of pseudo-Lipschitz loss between $\widehat{\boldsymbol{\beta}}$ and the truth $\boldsymbol{\beta}$.

Definition 2.3.1. Uniformly pseudo-Lipschitz functions *[BMN20]: For $k \in \mathbb{N}_{>0}$, a function $\phi : \mathbb{R}^d \to \mathbb{R}$ is* pseudo-Lipschitz of order k *if there exists a constant*

L, such that for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{b})\| \leq L\Big(1 + (\|\boldsymbol{a}\|/\sqrt{d})^{k-1} + (\|\boldsymbol{b}\|/\sqrt{d})^{k-1}\Big)\Big(\|\boldsymbol{a} - \boldsymbol{b}\|/\sqrt{d}\Big). \tag{2.3.2}$$

A sequence (in p) of pseudo-Lipschitz functions $\{\phi_p\}_{p \in \mathbb{N}_{>0}}$ is uniformly pseudo-Lipschitz *of order k if, denoting by L_p the pseudo-Lipschitz constant of ϕ_p, $L_p < \infty$ for each p and $\limsup_{p\to\infty} L_p < \infty$.*

Theorem 3. *Under assumptions **(A1)** - **(A5)**, for any uniformly pseudo-Lipschitz sequence of functions $\psi_p : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ and for $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$,*

$$\operatorname*{plim}_p \psi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_t \operatorname*{plim}_p \mathbb{E}_{\boldsymbol{Z}}[\psi_p(\text{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}), \boldsymbol{\beta})],$$

where τ_t is defined in (2.2.4) *and the expectation is taken with respect to $\boldsymbol{Z}$.*

Theorem 3 tells us that under uniformly pseudo-Lipschitz loss, in the large system limit, distributionally the SLOPE optimizer acts as a 'denoised' version of the truth corrupted by additive Gaussian noise where the denoising function is given by the proximal operator, i.e. within uniformly pseudo-Lipschitz loss $\widehat{\beta}$ can be replaced with $\text{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})$ for large p, t.

The proof of Theorem 3 can be found in Section 2.4. We show that Theorem 3 follows from Theorem 2 and recent AMP theory dealing with the state evolution analysis in the case of non-separable denoisers [BMN20], which can be used to demonstrate that the state evolution given in (2.2.4) characterizes the performance of the SLOPE AMP (2.1.3) via pseudo-Lipschitz loss functions.

We note that [HL19a, Theorem 1] follows by Theorem 3 and their separability result [HL19a, Proposition 1]. To see this, we use the following lemma that is a simple application of the Law of Large Numbers.

Lemma 2.3.2. *For any function $f : \mathbb{R}^p \to \mathbb{R}$ that is asymptotically separable, in the sense that there exists some function $\widetilde{f} : \mathbb{R} \to \mathbb{R}$, such that*

$$\left| f(\boldsymbol{\beta}) - \frac{1}{p}\sum_{i=1}^{n} \widetilde{f}(\beta_i) \right| \to 0, \qquad as \qquad p \to \infty,$$

where $\widetilde{f}(B)$ is Lebesgue integrable then $\operatorname{plim}_p \Big(f(\boldsymbol{\beta}) - \mathbb{E}_{\boldsymbol{B}}[\widetilde{f}(\boldsymbol{B})] \Big) = 0$, *where $\boldsymbol{B} \sim$ i.i.d. B.*

Now to show the result [HL19a, Theorem 1], consider a special case of Theorem 3 where $\psi_p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\sum \psi(x_i, y_i)$ for function $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is pseudo-Lipschitz of order $k = 2$. It is easy to show that $\psi_p(\cdot, \cdot)$ is uniformly pseudo-Lipschitz of order $k = 2$. The result of Theorem 3 then says that

$$\operatorname*{plim}_p \frac{1}{p}\sum_{i=1}^{p} \psi(\widehat{\beta}_i, \beta_i) = \lim_t \operatorname*{plim}_p \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}_{\boldsymbol{Z}}[\psi([\operatorname{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})]_i, \beta_i)].$$

Then [HL19a, Theorem 1] follows by [HL19a, Proposition 1], restated below in Lemma 2.3.3, the Law of Large Numbers, and Theorem 1. Now we restate in Lemma 2.3.3, the result given in [HL19a, Proposition 1], which says that $\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ becomes asymptotically separable as $p \to \infty$, for convenience.

Lemma 2.3.3 (Proposition 1, [HL19a])**.** *For an input sequence $\{\boldsymbol{v}(p)\}$, and a sequence of thresholds $\{\boldsymbol{\lambda}(p)\}$, both having empirical distributions that weakly converge* *to distributions V and Λ, respectively, then there exists a limiting scalar function h (determined by V and Λ) such that as $p \to \infty$,*

$$\|\text{prox}_{J_{\boldsymbol{\lambda}(p)}}(\boldsymbol{v}(p)) - h(\boldsymbol{v}(p); V, \Lambda)\|^2/p \to 0, \tag{2.3.3}$$

where h applies $h(\cdot; V, \Lambda)$ coordinate-wise to $\boldsymbol{v}(p)$ (hence it is separable) and h is Lipschitz(1).

Then [HL19a, Theorem 1] follows from Theorem 3 by using the asymptotic separability of the prox operator. Namely, the result of Lemma 2.3.3 (using that $\boldsymbol{\alpha}(p)\tau_t$ has an empirical distribution that converges weakly to $A\tau_t$ for A defined by (2.2.11)), along with Cauchy-Schwarz and the fact that ψ is pseudo-Lipschitz, allow us to apply a dominated convergence argument (see Lemma 2.9.3), from which it follows for some limiting scalar function h^t as specified by Lemma 2.3.3,

$$\frac{1}{p}\Big|\sum_{i=1}^{p}\mathbb{E}_{\boldsymbol{Z}}[\psi([\text{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta}+\tau_t\boldsymbol{Z})]_i, \beta_i)] - \sum_{i=1}^{p}\mathbb{E}_{\boldsymbol{Z}}[\psi([h^t(\boldsymbol{\beta}+\tau_t\boldsymbol{Z})]_i, \beta_i)]\Big| \to 0.$$

Then the above allows us to apply Lemma 2.3.2 and the Law of Large Numbers to show

$$\begin{aligned}\operatorname*{plim}_p \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}_{\boldsymbol{Z}}[\psi([\text{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta}+\tau_t\boldsymbol{Z})]_i, \beta_i)] &= \lim_p \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{B}}[\psi(h^t([\boldsymbol{B}+\tau_t\boldsymbol{Z}]_i), B_i)]\\ &= \mathbb{E}_{Z,B}[\psi(h^t(B+\tau_t Z), B)],\end{aligned}$$

Finally we note that the result of [HL19a, Theorem 1] follows since

$$\lim_t \mathbb{E}_{Z,B}[\psi(h^t(B+\tau_t Z), B)] = \mathbb{E}_{Z,B}[\psi(h^*(B+\tau_* Z), B)].$$

We highlight that our Theorem 3 allows the consideration of a non-asymptotic case in t. While Theorem 1 motivates an algorithmic way to find a value $\tau_t(p)$ which approximates $\tau_*(p)$ well, Theorem 3 guarantees the accuracy of such approximation for use in practice. One particular use of Theorem 3 is to design the optimal sequence $\boldsymbol{\lambda}$ that achieves the minimum τ_* and equivalently minimum error [HL19a], though a concrete algorithm for doing so is still under investigation.

Finally we show how we use Theorem 3 to study the asymptotic mean-square error between the SLOPE estimator and the truth [CM19].

Corollary 2.3.4. *Under assumptions **(A1)**– **(A5)**,* $\operatorname{plim}_p \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p = \delta(\tau_*^2 - \sigma_w^2)$.

Proof. Applying Theorem 3 to the pseudo-Lipschitz loss function $\psi^1 : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, defined as $\psi^1(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^2/p$, we find $\operatorname{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \lim_t \operatorname{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}) - \boldsymbol{\beta}\|^2]$. The desired result follows since $\lim_t \operatorname{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}) - \boldsymbol{\beta}\|^2] = \delta(\tau_*^2 - \sigma_w^2)$. To see this, note that $\lim_t \delta(\tau_{t+1}^2 - \sigma_w^2) = \delta(\tau_*^2 - \sigma_w^2)$ and

$$\begin{aligned}&\operatorname*{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}) - \boldsymbol{\beta}\|^2] \\ &= \lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{B}}[\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{B} + \tau_t \boldsymbol{Z}) - \boldsymbol{B}\|^2] = \delta(\tau_{t+1}^2 - \sigma_w^2),\end{aligned}$$

for $\boldsymbol{B}$ elementwise i.i.d. B independent of $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$. A rigorous argument for the above requires showing that the assumptions of Lemma 2.3.2 are satisfied and follows similarly to that used to prove property **(P2)** stated in Section 2.4 and proved in Appendix 2.9.2. □

2.4 Proof for Asymptotic Characterization of the SLOPE Estimate

In this section we prove Theorem 3. To do this, we use a result guaranteeing that the state evolution given in (2.2.4) characterizes the performance of the SLOPE AMP algorithm (2.1.3b), given in Lemma 2.4.1 below. Specifically, Lemma 2.4.1 relates the state evolution (2.2.4) to the output of the AMP iteration (2.1.3b) for pseudo-Lipschitz loss functions. This result follows from [BMN20, Theorem 14], which is a general result relating state evolutions to AMP algorithm with non-separable denoisers. In order to apply [BMN20, Theorem 14], we need to demonstrate that our denoiser, i.e. the proximal operator $\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ defined in (2.1.4), satisfies two additional properties labeled **(P1)** and **(P2)** below.

Define a sequence of denoisers $\{\eta_p^t\}_{p\in\mathbb{N}_{>0}}$ where $\eta_p^t : \mathbb{R}^p \to \mathbb{R}^p$ to be those that apply the proximal operator $\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ defined in (2.1.4), i.e. for a vector $\boldsymbol{v} \in \mathbb{R}^p$, define

$$\eta_p^t(\boldsymbol{v}) := \text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{v}). \tag{2.4.1}$$

(P1) For each t, denoisers $\eta_p^t(\cdot)$ defined in (2.4.1) are uniformly Lipschitz (i.e. uniformly pseudo-Lipschitz of order $k = 1$) per Definition 2.3.1.

(P2) For any s, t with $(\boldsymbol{Z}, \boldsymbol{Z}')$ a pair of length-p vectors, where for $i \in \{1, 2, \ldots, p\}$, the pair (Z_i, Z_i') i.i.d. $\sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ any 2×2 covariance matrix, the following limits exist and are finite.

$$\operatorname*{plim}_{p\to\infty} \frac{1}{p}\|\boldsymbol{\beta}\|, \quad \operatorname*{plim}_{p\to\infty} \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})],$$
$$\text{and } \operatorname*{plim}_{p\to\infty} \frac{1}{p}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}[\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})].$$

We will show that properties **(P1)** and **(P2)** are satisfied for our problem in Appendix 2.9.2.

Lemma 2.4.1. *[BMN20, Theorem 14] Under assumptions **(A1)** - **(A4)**, given that **(P1)** and **(P2)** are satisfied, for the AMP algorithm in* (2.1.3b) *and for any uniformly pseudo-Lipschitz sequence of functions $\phi_n : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $\psi_p : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, let $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_n)$ and $\boldsymbol{Z}' \sim \mathcal{N}(0, \mathbb{I}_p)$, then*

$$\operatorname*{plim}_{n} \Big(\phi_n(\boldsymbol{z}^t, \boldsymbol{w}) - \mathbb{E}_{\boldsymbol{Z}}[\phi_n(\boldsymbol{w} + \sqrt{\tau_t^2 - \sigma_w^2}\boldsymbol{Z}, \boldsymbol{w})]\Big) = 0,$$
$$\operatorname*{plim}_{p} \Big(\psi_p(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{z}^t, \boldsymbol{\beta}) - \mathbb{E}_{\boldsymbol{Z}'}[\psi_p(\boldsymbol{\beta} + \tau_t\boldsymbol{Z}', \boldsymbol{\beta})]\Big) = 0,$$

where τ_t is defined in (2.2.4)*.*

We now show that Theorem 3 follows from Lemma 2.4.1 and Theorem 2.

Proof of Theorem 3. First, for any fixed n and t, the following bound uses that ψ_n is uniformly pseudo-Lipschitz of order k and the Triangle Inequality,

$$\begin{aligned}
&\Big|\psi_p(\boldsymbol{\beta}^t, \boldsymbol{\beta}) - \psi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\Big| \\
&\le L\Big(1 + \Big(\frac{\|(\boldsymbol{\beta}^t, \boldsymbol{\beta})\|}{\sqrt{2p}}\Big)^{k-1} + \Big(\frac{\|(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\|}{\sqrt{2p}}\Big)^{k-1}\Big)\frac{1}{\sqrt{2p}}\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\| \\
&\le L\Big(1 + \Big(\frac{\|\boldsymbol{\beta}^t\|}{\sqrt{2p}}\Big)^{k-1} + \Big(\frac{\|\widehat{\boldsymbol{\beta}}\|}{\sqrt{2p}}\Big)^{k-1} + \Big(\frac{\|\boldsymbol{\beta}\|}{\sqrt{2p}}\Big)^{k-1}\Big)\frac{1}{\sqrt{2p}}\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\|.
\end{aligned}$$

Now we take limits on either side of the above, first with respect to p and then with respect to t. We note that the term $\frac{1}{\sqrt{n}}\|\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}\|$ vanishes by Theorem 2. Then as long as

$$\lim_t \operatorname*{plim}_p \left(\|\boldsymbol{\beta}^t\|/\sqrt{p}\right)^{k-1}, \quad \operatorname*{plim}_p \left(\|\widehat{\boldsymbol{\beta}}\|/\sqrt{p}\right)^{k-1}, \quad \text{and} \qquad \operatorname*{plim}_p \left(\|\boldsymbol{\beta}\|/\sqrt{p}\right)^{k-1}, \tag{2.4.2}$$

are all finite, we have $\operatorname{plim}_p \psi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_t \operatorname{plim}_p \psi_p(\boldsymbol{\beta}^t, \boldsymbol{\beta})$. But by Theorem 2.4.1 we also know that

$$\lim_t \operatorname*{plim}_p \psi_p(\boldsymbol{\beta}^t, \boldsymbol{\beta}) = \lim_t \operatorname*{plim}_p \mathbb{E}[\psi_p(\eta^t(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}), \boldsymbol{\beta})],$$

giving the desired result.

Finally we convince ourself that the limits in (2.4.2) are finite. Since k finite, that the third term in (2.4.2) is finite follows by property **(P2)**. Bounds for the first and second term are demonstrated in Lemma 2.7.1 found in Appendix 2.6.

□

2.5 Proof AMP Finds the SLOPE Solutions

In this section we aim to prove Theorem 2. Define the SLOPE cost function as follows,

$$\mathcal{C}(\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{b}), \tag{2.5.1}$$

where $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ is the sorted ℓ_1-norm. The proof of Theorem 2 relies on a technical lemma, Lemma 2.5.5, stated in Section 2.5.2 below, that deals carefully with the fact that the SLOPE cost function given in (2.5.1) is *not* necessarily strongly convex.

In the LASSO case, one works around this challenge by showing that the (LASSO) cost function does have nice properties when considering just the elements of the non-zero support of $\boldsymbol{\beta}^t$ at any (large) iteration t, using that the non-zero support of $\boldsymbol{\beta}$ has size no larger than $n < p$.

In the SLOPE problem, however, it is possible that the support set has size exceeding n, and therefore the LASSO analysis is not immediately applicable. Our proof develops novel techniques that are tailored to the characteristics of the SLOPE solution. Specifically, when considering the SLOPE problem, one can show nice properties (similar to those in the LASSO case) by considering a support-like set, that being the *unique* non-zeros in the estimate $\boldsymbol{\beta}^t$ at any (large) iteration t.

In other words, our strategy is to define an equivalence relation $x \sim y$ when $|x| = |y|$ and partition the entries of the AMP estimate at any iteration t into equivalence classes. This allows us to observe, using (2.2.10) and the non-negativity of $\boldsymbol{\lambda}$, that the number of equivalence classes is no larger than n. (Recall that $\|\cdot\|_0^*$ counts the unique non-zero magnitudes in a vector.) We see an analogy between SLOPE's equivalence class (or 'maximal atom' as described in Section 2.5.1) and LASSO's support set. This approach, taken in Lemma 2.5.5 below, allows us to deal with the fact that we are not guaranteed to have a strongly convex cost. Then Lemma 2.5.5 is used to prove Theorem 3.

Before we state Lemma 2.5.5, we include some useful preliminary information on

SLOPE that will be needed for the upcoming work. In particular, we introduce in more details the idea of equivalence classes of elements having the same magnitude, a mapping of vector ranking denoted as $\hat{\Pi}$, and a polytope-related mapping whose image is the set of subgradients denoted as $\mathcal{P}$. These definitions are all given in more detail in Section 2.5.1.

2.5.1 Preliminaries on SLOPE

In general, we refer to the function $\mathcal{C}(\cdot)$ stated in (2.5.1) as the SLOPE cost function and the SLOPE estimator $\hat{\boldsymbol{\beta}}$ is the one that minimizes the SLOPE cost. We note that the SLOPE cost function $\mathcal{C}(\cdot)$ depends on both $\boldsymbol{y}$ and $\boldsymbol{\lambda}$, so technically a notation like $\mathcal{C}_{(\boldsymbol{y},\boldsymbol{\lambda})}(\cdot)$ would be more rigorous, however, we don't think that dropping the explicit dependence on $(\boldsymbol{y},\boldsymbol{\lambda})$ will cause any confusion.

For a convex function $f : \mathbb{R}^p \to \mathbb{R}$, we denote the subgradient of f at a point $\boldsymbol{x} \in \mathbb{R}^p$ as $\partial f(\boldsymbol{x})$. We will be interested, particularly, in the subgradient of the SLOPE cost $\partial\mathcal{C}(\boldsymbol{b})$ which forces us to study the subgradient of the SLOPE norm $\partial J_{\boldsymbol{\lambda}}(\boldsymbol{b})$. In particular,

Fact 2.5.1. $\partial\mathcal{C}(\boldsymbol{b}) = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) + \partial J_{\boldsymbol{\lambda}}(\boldsymbol{b})$.

We will now describe explicitly the relevant subgradient, $\partial J_{\boldsymbol{\lambda}} \subset \mathbb{R}^p$. We note that the proximal operator given in (2.1.4) is linked to the subgradient of the SLOPE norm in the following way.

Fact 2.5.2. *If* $\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}_1) = \boldsymbol{v}_2$, *then* $\boldsymbol{v}_1 - \boldsymbol{v}_2 \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{v}_2)$.

Define a function $\Pi_{\boldsymbol{x}} : \mathbb{R}^p \to \mathbb{R}^p$ to be a mapping (not necessarily unique) that sorts its input by magnitude in descending order according to absolute values of entries in $\boldsymbol{x}$. For example, if $\boldsymbol{x} = (5, 2, -3, -5)$, then there are two possible such mappings $\Pi_{\boldsymbol{x}}(\boldsymbol{b}) = (|b_1|, |b_4|, |b_3|, |b_2|)$ or $\Pi_{\boldsymbol{x}}(\boldsymbol{b}) = (|b_4|, |b_1|, |b_3|, |b_2|)$. Using this notation, we can rewrite the SLOPE norm as $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \boldsymbol{\lambda} \cdot \Pi_{\boldsymbol{b}}(\boldsymbol{b})$. Since such mapping may not be unique, the inverse may not exist and we therefore define a pseudo-inverse mapping, $\hat{\Pi}_{\boldsymbol{x}}^{-1}$, that is based on the function $\hat{\Pi}_{\boldsymbol{x}} : \mathbb{R}^p \to \{\text{maximal atoms}\}$. In words, $\hat{\Pi}_{\boldsymbol{x}}$ finds the maximal atoms of ranking of the absolute values of $\boldsymbol{x}$. Then $\hat{\Pi}_{\boldsymbol{x}}$ corresponds to the mapping

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ \{1,2\} & 4 & 3 & \{1,2\} \end{pmatrix}$$

with $\hat{\Pi}_{\boldsymbol{x}}(\boldsymbol{x}) = (\{5,-5\}, \{5,-5\}, -3, 2)$ and $\hat{\Pi}_{\boldsymbol{x}}^{-1}(\boldsymbol{\lambda}) = (\{\lambda_1, \lambda_2\}, \lambda_4, \lambda_3, \{\lambda_1, \lambda_2\})$. Then it is not hard to see that there exists $\hat{\boldsymbol{\lambda}} \in \hat{\Pi}_{x}^{-1}(\boldsymbol{\lambda})$ such that $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \boldsymbol{\lambda} \cdot \Pi_{\boldsymbol{b}}(\boldsymbol{b}) = \hat{\boldsymbol{\lambda}} \cdot |\boldsymbol{b}|$. In words, this says there are two equivalent ways to consider the calculation of $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ when $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. First $\boldsymbol{\lambda} \cdot \Pi_{\boldsymbol{b}}(\boldsymbol{b})$ computes the inner product between $\boldsymbol{\lambda}$ and the *sorted* magnitudes of $\boldsymbol{b}$, and in the second case, $\hat{\boldsymbol{\lambda}}^\top |\boldsymbol{b}|$ computes the inner product between the magnitudes of $\boldsymbol{b}$ (unsorted), with a rearrangement of the $\boldsymbol{\lambda}$ vector (based on $\boldsymbol{b}$) that pairs the values in $\boldsymbol{\lambda}$ with the values of $|\boldsymbol{b}|$ by magnitude.

Now we define an equivalence relation $x \sim y$ if $|x| = |y|$. Then $\hat{\Pi}_{\boldsymbol{x}}$ partitions elements in $\boldsymbol{x}$ into different equivalence classes I. The motivation of using equivalence classes roots from AMP. In calibrating the AMP to the SLOPE problem, we need to calculate $\nabla\,\mathrm{prox}$, which equals the number of non-zero equivalence classes. For example, $\frac{\partial\,\mathrm{prox}}{\partial \boldsymbol{v}}|_{\boldsymbol{v}=(1,0,-1,3)} = (\frac{1}{2}, 0, \frac{1}{2}, 1)$ has a sum of 2.

Now we note that the subgradient of the SLOPE norm can be represented using the idea of the equivalence classes. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we use the notation $\boldsymbol{v}_I$ to be the elements of the vector $\boldsymbol{v}$ belonging to equivalence class I. Then,

Fact 2.5.3.

$$\partial J_{\boldsymbol{\lambda}}(\boldsymbol{s}) = \left\{ \boldsymbol{v} \in \mathbb{R}^p : \text{ for each equivalent class } I, \begin{cases} \textit{if } \boldsymbol{s}_I \neq 0 \implies \boldsymbol{v}_I \in \mathcal{P}([\,\hat{\Pi}_{\boldsymbol{s}}^{-1}(\boldsymbol{\lambda})]_I\,)\ \mathrm{sign}(\boldsymbol{s}_I); \\ \\ \textit{if } \boldsymbol{s}_I = 0 \implies |\boldsymbol{v}_I| \in \mathcal{P}_0([\,\hat{\Pi}_{\boldsymbol{s}}^{-1}(\boldsymbol{\lambda})\,]_I) \end{cases} \right\}.$$

In the above, $\mathcal{P}, \mathcal{P}_0$ are polytope-related mappings,

$$\mathcal{P}(\boldsymbol{u}) := \{\boldsymbol{y} : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{u} \text{ for some doubly stochastic matrix } \boldsymbol{A}\}$$

$$\mathcal{P}_0(\boldsymbol{u}) := \{\boldsymbol{y} : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{u} \text{ for some doubly sub-stochastic matrix } \boldsymbol{A}\}$$

By definition, the doubly stochastic matrix, a.k.a. a Birkhoff polytope, is a square matrix of non-negative real numbers, whose row and column sums equal 1. For example,

$$\boldsymbol{A} = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/6 & 1/3 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix} \tag{2.5.2}$$

is a doubly stochastic matrix. Similarly, a doubly sub-stochastic matrix is defined as a square matrix of non-negative real numbers, whose row and column sums are at most 1. Note that if all entries of $\boldsymbol{\lambda}$ take the same value, the subgradient in Fact 2.5.3 gives the usual subgradient of the ℓ_1 norm.

Using the subgradient definition in Fact 2.5.3, consider $\mathcal{P}((\lambda_1, \lambda_2, \lambda_3))$, relating to a non-zero equivalence class having three entries. Then $\boldsymbol{A}$ in (2.5.2) is one possible matrix considered in defining the set $\mathcal{P}((\lambda_1, \lambda_2, \lambda_3))$ and it has the following interpretation. The rows of $\boldsymbol{A}$ determine how the subgradient $\boldsymbol{v}_I$ values are calculated by averaging the corresponding threshold values $\boldsymbol{\lambda}$, for example, the first entry of $\boldsymbol{v}_I$ is a weighted average with 1/3 its weight in λ_1 and 2/3 in λ_2; the second entry of $\boldsymbol{v}_I$ is a weighted average with 1/6 its weight in λ_1, 1/3 in λ_2, and 1/2 in λ_2, etc. You can think of this as determining the threshold each input value $\boldsymbol{s}_I$ receives, as some weighted combination of all the possible threshold values $\boldsymbol{\lambda}$ corresponding to this equivalence class. Similarly, the columns of the doubly-stochastic matrix considered in the mapping $\mathcal{P}$ define how the thresholds $\boldsymbol{\lambda}$ are spread out amongst each element of the subgradient, for example, 1/3 of λ_1's value goes to the first element of $\boldsymbol{v}_I$, 1/6 to the second value, and 1/2 to the third value, etc.

To see why $\partial J_{\boldsymbol{\lambda}}(\boldsymbol{s})$ takes the form given in Fact 2.5.3, let's consider again the $\mathcal{P}$ used in the case that $\boldsymbol{s}_I \neq 0$. Recall the $\boldsymbol{s}_I$ looks at only the indices of $\boldsymbol{s}$ appearing in the equivalence class I, so all elements of $\boldsymbol{s}_I$ have the same absolute value. This means that there are many ways to share the corresponding $\boldsymbol{\lambda}$ threshold values among them. We can think of this as an assignment problem: assign jobs (thresholds $\boldsymbol{\lambda}$) to workers (s_i) where as assignment according to a doubly stochastic matrix is a natural one (all workers take on the same load, and all jobs must be completed). On the other hand, $\mathcal{P}_0$ does not require that the sharing of the threshold values $\boldsymbol{\lambda}$ amongst the entries of $\boldsymbol{s}_I$ be strict: row and/or column sums can be smaller than one. This difference is rooted in the subgradient of ℓ_1 norm: i.e. $\partial|x| = \text{sign}(x)$ when $x \neq 0$ and $\partial|x| \in [-1, 1]$ when $x = 0$.

For a rigorous proof of Fact 2.5.3, we refer the reader to [RW09, Exercise 8.31], but we give a quick sketch here in the case of $\boldsymbol{s}_I \neq 0$. The proof uses that $\mathcal{P}(\boldsymbol{u})$ is a permutohedron, meaning a convex hull with vertices corresponding to permuted entries of $\boldsymbol{u}$. Notice that we can rewrite $J_{\boldsymbol{\lambda}}(\boldsymbol{s})$ as a finite max function $J_{\boldsymbol{\lambda}}(\boldsymbol{s})$: $\max\{\boldsymbol{\lambda}^\top f_1(\boldsymbol{s}), ..., \boldsymbol{\lambda}^\top f_m(\boldsymbol{s})\}$, where $\{f_i(\boldsymbol{s})\}_{1\leq i\leq m}$ is the collection of all possible permutations for the entries of $|\boldsymbol{s}|$. Notice that the permutation that sorts the magnitudes will be chosen by the maximum function. For such a function (see [RW09, Exercise 8.31]) the subgradient takes the form of a convex hull of the partial derivatives of the maximizing elements:

$$\partial J_{\boldsymbol{\lambda}}(\boldsymbol{s}) \in \text{conv}\{\nabla_{\boldsymbol{s}}(\boldsymbol{\lambda}^\top f_i(\boldsymbol{s})) : i \in A(\boldsymbol{s})\} \equiv \text{conv}\{f_i^{-1}(\boldsymbol{\lambda}) : i \in A(\boldsymbol{s})\}, \tag{2.5.3}$$

where $A(\boldsymbol{s}) = \{i \in \{1, 2, \ldots, m\} : \boldsymbol{\lambda}^\top f_i(\boldsymbol{s}) = J_{\boldsymbol{\lambda}}(\boldsymbol{s})\}$ and in our case, the partial derivatives correspond to permutations of the thresholds. Now, without loss of generality, let's consider an input that has only one non-zero equivalence class, i.e. $\boldsymbol{s} = (s, s, ..., s) \in \mathbb{R}^d$. Then clearly there are $m = d!$ possible permutations. Therefore,

$$\partial J_{\boldsymbol{\lambda}}(\boldsymbol{s}) \in \text{conv}\{f_i^{-1}(\boldsymbol{\lambda}) : i \in \{1, 2, ..., d!\}\} \equiv \text{conv}\{f_i(\boldsymbol{\lambda}) : i \in \{1, 2, ..., d!\}\}.$$

In other words, the partial derivative lies in the set that is the convex combination of all possible permutations of the threshold $\boldsymbol{\lambda}$. By definition, this is a permutohedron. So, in our case, the subgradient is a convex hull whose vertices are the permutated thresholds, i.e. an image of Birkhoff polytope under the thresholds, which can be characterized by doubly stochastic matrices.

2.5.2 Main Technical Lemma

Now we state and prove the main technical lemma that will be used to prove Theorem 2. Before we state Lemma 2.5.5, let us introduce a very important definition:

Definition 2.5.4. *Given a vector $\boldsymbol{v} \in \mathbb{R}^p$, a set $I \subset \{1, \dots, p\}$ is said to be a maximal atom of indices of $\boldsymbol{v}$ if $|v_i| = |v_j|$ for all $i, j \in I$ and $|v_i| \neq |v_k|$ for $i \in I$ and all $k \notin I$. With this definition in place, we define the star support of the vector $\boldsymbol{v}$ as*

$$\operatorname{supp}^\star(\boldsymbol{v}) := \{I : I \subset \{1, \dots, p\} \textit{ is a maximal atom of indices of } \boldsymbol{v} \textit{ and } \boldsymbol{v}_I \neq 0\}.$$

For example, if $\boldsymbol{v} = (1, 1, -1, 0, 2, -1)$, then $\operatorname{supp}^\star(\boldsymbol{v}) = \{\{1, 2, 3, 6\}, \{5\}\}$. Now we state and prove Lemma 2.5.5.

Lemma 2.5.5. *For constants $c_1, ..., c_5 > 0$, if the following conditions are satisfied,*

(1) $\frac{1}{\sqrt{p}} \|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}\| \leq c_1$,

(2) There exists a subgradient $sg(\mathcal{C}, \boldsymbol{\beta}^t) \in \partial \mathcal{C}(\boldsymbol{\beta}^t)$ such that $\frac{1}{\sqrt{p}} \|sg(\mathcal{C}, \boldsymbol{\beta}^t)\| \leq \epsilon$,

(3) *Let* $\boldsymbol{\nu}^t := \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) + sg(\mathcal{C}, \boldsymbol{\beta}^t) \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ *(where* $sg(\mathcal{C}, \boldsymbol{\beta}^t)$ *is the subgradient from Condition (2)). Denote* $s_t(c_2) := \{I \subset [p] : |\boldsymbol{\nu}^t_I| \succeq [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I(1 - c_2)\}$ *and* $S_t(c_2) := \{i \in I : I \in s(c_2)\}$*, where the equivalence classes,* I*, for both sets are defined via the AMP estimation* $\boldsymbol{\beta}^t$*, and for a vector* $\boldsymbol{x} \in \mathbb{R}^d$ *and a set* $\boldsymbol{A} \subset \mathbb{R}^d$*, the notation* $\boldsymbol{x} \succeq \boldsymbol{A}$ *means there exists some* $\boldsymbol{y} \in \boldsymbol{A}$ *such that* $\boldsymbol{x} \geq \boldsymbol{y}$ *elementwise. Then for* s' *being* any *set of maximal atoms in* $[p]$ *with* $|s'| \leq c_3 p$ *and* $S' := \{i \in I : I \in s'\}$*, we have* $\sigma_{min}(\boldsymbol{X}_{S_t(c_2) \cup S'}) \geq c_4$*.*

(4) *The minimum non-zero and maximum singular value of* $\boldsymbol{X}$*, denoted as* $\hat{\sigma}^2_{min}(\boldsymbol{X})$ *and* $\sigma^2_{max}(\boldsymbol{X})$*, are bounded: i.e.* $\hat{\sigma}^2_{min}(\boldsymbol{X}) \geq \frac{1}{c_5}$ *and* $\sigma^2_{max}(\boldsymbol{X}) \leq c_5$*.*

(5) *Define* $\mathcal{C}_{\boldsymbol{x}}(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^p \hat{\lambda}_i |b_i|$ *for some* $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{x}}(\boldsymbol{\lambda}))$*. Then* $\mathcal{C}(\boldsymbol{\beta}^t) \geq \mathcal{C}_{\boldsymbol{\beta}^t}(\hat{\boldsymbol{\beta}})$*.*

then for some function $f(\epsilon) := f(\epsilon, c_1, c_2, c_3, c_4, c_5)$ *such that* $f(\epsilon) \to 0$ *as* $\epsilon \to 0$*,*

$$\frac{1}{\sqrt{p}}\|\boldsymbol{\beta}^t - \hat{\boldsymbol{\beta}}\| < f(\epsilon).$$

We wrap up this section by proving Lemma 2.5.5. Once we have proved Lemma 2.5.5, we will be able to prove Theorem 2. The major piece of work in proving Theorem 2 is in showing that the five assumptions of Lemma 2.5.5 are satisfied. Then the result of Theorem 2 is immediate. We show the five assumptions are met in Sections 2.7.1 - 2.7.5. Now we prove the Lemma.

Proof of Lemma 2.5.5. Throughout the proof, we denote $\xi_1, \xi_2, \dots$ as functions of the constants $c_1, \dots, c_5 > 0$ and of ϵ such that $\xi_i(\epsilon) \to 0$ as $\epsilon \to 0$ (we omit the dependence of ξ_i on ϵ). We will think of t as a fixed iteration and we denote the residual we are interested in studying as $\boldsymbol{r} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t$.

The proof strategy is to show that $\frac{1}{p}\|\boldsymbol{X}\boldsymbol{r}\|^2 \leq \xi(\epsilon)$ from which a similar result for $\frac{1}{p}\|\boldsymbol{r}\|^2$ follows when we have control of the singular values of $\boldsymbol{X}$ as we do with Condition (4). Structurally, the proof is similar to that in the LASSO case (cf. [BM11c, Lemma 3.1]), with the main difference coming through Condition (3), where we need to use star support instead of the support when bounding the minimum singular value of a selection of columns of $\boldsymbol{X}$.

For a fixed iteration t, let $S = \{i \in [p] : i \in I \text{ and } I \in \text{supp}^*(\boldsymbol{\beta}^t)\}$, i.e. S is the collection of (unique) indices belonging to the star support of the AMP estimate at iteration t. Then for a vector $\boldsymbol{v} \in \mathbb{R}^p$ we denote $\boldsymbol{v}_S$ to mean the vector indexed only over the indices in the set S and we let $\bar{S}$ denote the complement of S. In what follows, we drop the t-dependence on $\boldsymbol{\nu}^t$, writing $\boldsymbol{\nu} = \boldsymbol{\nu}^t$ and for p-length vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, define $\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \frac{1}{p}\sum_i u_i v_i$.

First,

$$
\begin{aligned}
0 &\overset{(a)}{\geq} \frac{1}{p}(\mathcal{C}_{\boldsymbol{\beta}^t}(\widehat{\boldsymbol{\beta}}) - \mathcal{C}(\boldsymbol{\beta}^t)) \overset{(b)}{=} \frac{1}{2p}(\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t\|^2) + \langle \hat{\boldsymbol{\lambda}}, |\widehat{\boldsymbol{\beta}}| - |\boldsymbol{\beta}^t| \rangle \\
&\overset{(c)}{=} \langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}^t_S + \boldsymbol{r}_S| - |\boldsymbol{\beta}^t_S| \rangle + \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle + \frac{1}{2p}(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t - \boldsymbol{X}\boldsymbol{r}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t\|^2) \\
&\overset{(d)}{=} \Big[\langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}^t_S + \boldsymbol{r}_S| - |\boldsymbol{\beta}^t_S| \rangle - \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S \rangle\Big] + \Big[\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle\Big] + \langle \boldsymbol{\nu}, \boldsymbol{r} \rangle \\
&- \langle \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t, \boldsymbol{X}\boldsymbol{r} \rangle + \frac{\|\boldsymbol{X}\boldsymbol{r}\|^2}{2p} \\
&\overset{(e)}{=} \Big[\langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}^t_S + \boldsymbol{r}_S| - |\boldsymbol{\beta}^t_S| \rangle - \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S \rangle\Big] + \Big[\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle\Big] \\
&+ \langle sg(\mathcal{C}, \boldsymbol{\beta}^t), \boldsymbol{r} \rangle + \frac{\|\boldsymbol{X}\boldsymbol{r}\|^2}{2p}.
\end{aligned}
$$

In the above, step (a) follows immediately from Condition (5) and step (b) holds *for any* $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$ by the definition of $\mathcal{C}_{\boldsymbol{\beta}^t}(\widehat{\boldsymbol{\beta}})$, noticing that $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t) = \hat{\boldsymbol{\lambda}}^\top |\boldsymbol{\beta}^t|$ in the SLOPE cost (2.5.1) *since* $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$. Below we will select a specific $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$ based on the definition of $\boldsymbol{\nu}$. Step (c) follows by replacing $\widehat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}^t + \boldsymbol{r}$ and noticing that $\boldsymbol{\beta}^t_{\bar{S}} = \boldsymbol{0}$. Step (d) follows since $\langle \boldsymbol{\nu}, \boldsymbol{r} \rangle = \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S \rangle + \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle$ and step (e) from the definition of $\boldsymbol{\nu}$.

Using Conditions (1) and (2), we get by Cauchy-Schwarz

$$
\begin{aligned}
&\Big[\langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}^t_S + \boldsymbol{r}_S| - |\boldsymbol{\beta}^t_S| \rangle - \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S \rangle\Big] + \Big[\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle\Big] + \frac{\|\boldsymbol{X}\boldsymbol{r}\|^2}{2p} \\
&\leq \frac{\|sg(\mathcal{C}, \boldsymbol{\beta}^t)\| \|\boldsymbol{r}\|}{p} \leq c_1 \epsilon. \qquad (2.5.4)
\end{aligned}
$$

We now show all three terms on the left side of (2.5.4) are non-negative. The idea is then: if all three terms are non-negative and their sum tends to 0 as $\epsilon \to 0$, it must be true that each term tends to 0 too. The third term in (2.5.4), $\frac{1}{2p}\|\boldsymbol{X}\boldsymbol{r}\|^2$, is trivially non-negative, so we focus on the first two.

To show that the other terms are non-negative, we consider choosing a specific vector $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))$ such that on the support, $\hat{\boldsymbol{\lambda}}_S = |\boldsymbol{\nu}_S|$, and off the support $\hat{\boldsymbol{\lambda}}_{\bar{S}} \geq |\boldsymbol{\nu}_{\bar{S}}|$, meaning $\hat{\boldsymbol{\lambda}}_I$ is parallel to $|\boldsymbol{\nu}_I|$ for each equivalence class I of $\boldsymbol{\beta}^t$. That such a $\hat{\boldsymbol{\lambda}}$ exists in the set $\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))$ follows since $\boldsymbol{\nu}$ is a valid subgradient of $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ (see Fact 2.5.3).

Using this $\hat{\boldsymbol{\lambda}}$, notice that the sets defined in Condition (3) are equivalent to the following: $s_t(c_2) := \{I \subset [p] : |\boldsymbol{\nu}_I| \geq (1 - c_2)\hat{\boldsymbol{\lambda}}_I\}$ and $S_t(c_2) := \{i : |\nu_i| \geq (1 - c_2)\hat{\lambda}_i\}$, where both use equivalence classes, I, defined for $\boldsymbol{\beta}^t$. To see that this is the case, note that if I is a non-zero equivalence class, by Fact 2.5.3, since $|\boldsymbol{\nu}_I| \in [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I$, we know that $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I(1 - c_2)$ and similarly, since $\hat{\boldsymbol{\lambda}}_S = |\boldsymbol{\nu}_S|$ we know that $|\boldsymbol{\nu}_I| \geq (1 - c_2)\hat{\boldsymbol{\lambda}}_I$, so I clearly belongs to $s_t(c_2)$ for both definitions. If I is the zero equivalence class, if $|\boldsymbol{\nu}_I| \geq (1-c_2)\hat{\boldsymbol{\lambda}}_I$ then obviously $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I(1-c_2)$ since $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))$. In the other direction, if the non-zero equivalence class I is such that $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I(1 - c_2)$ then there exists a vector $\tilde{\boldsymbol{\nu}}_I \in [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I$ such that $|\boldsymbol{\nu}_I| \geq \tilde{\boldsymbol{\nu}}_I(1 - c_2)$ elementwise. However since $\tilde{\boldsymbol{\nu}}_I \in [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I$, this implies that $|\boldsymbol{\nu}_I| \geq (1 - c_2)\hat{\boldsymbol{\lambda}}_I$ is also true since $\hat{\boldsymbol{\lambda}}_I \in [\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))]_I$ in the same direction as $|\boldsymbol{\nu}_I|$.

To visualize the choice of $\hat{\boldsymbol{\lambda}}$, we consider an example where $\boldsymbol{\nu}_I = (-1, 2)$ for equivalence class $I = \{1, 2\}$ with $\boldsymbol{\lambda}_I = (4, 1)$ in Figure 2.3. In the figure, the blue shaded region indicates possible subgradient values for zero elements and the black line are possible subgradients for zero elements. In this example, the equivalence class is that for zero elements, so we notice that $\boldsymbol{\nu}_I$ lies in the blue region. Then $\boldsymbol{\lambda}_I$ is in the same direction as $|\boldsymbol{\nu}_I|$ but lies on the black line (since $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\beta^t}(\boldsymbol{\lambda}))$).

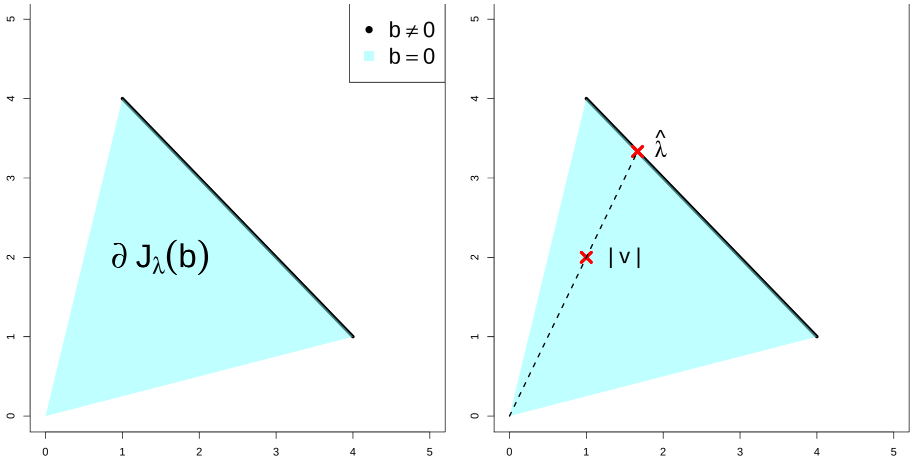


Figure 2.3: The blue area contained by the black line segment is the set of subgradients; Red crosses are examples of $\boldsymbol{\nu}_I$ and $\hat{\boldsymbol{\lambda}}_I$ correspondingly when $\boldsymbol{b}_I = \mathbf{0}$.

Now we would like to show that the first term in (2.5.4) is non-negative. Specifically, our choice of $\hat{\boldsymbol{\lambda}}$ gives $\nu_i = \operatorname{sign}(\beta_i^t)\hat{\lambda}_i$, for each $i \in S$, and then it suffices, in order to prove the non-negativity of $\langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}_S^t + \boldsymbol{r}_S| - |\boldsymbol{\beta}_S^t|\rangle - \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S\rangle$, to show

$$\begin{aligned}0 &\leq (|\beta_i^t + r_i| - |\beta_i^t|) - \operatorname{sign}(\beta_i^t) r_i \\ &= (\beta_i^t + r_i)\operatorname{sign}(\beta_i^t + r_i) - \beta_i^t \operatorname{sign}(\beta_i^t) - r_i \operatorname{sign}(\beta_i^t) \\ &= (\beta_i^t + r_i)\Big[\operatorname{sign}(\beta_i^t + r_i) - \operatorname{sign}(\beta_i^t)\Big],\end{aligned}$$

which follows since each $(\beta_i^t + r_i)\,[\operatorname{sign}(\beta_i^t + r_i) - \operatorname{sign}(\beta_i^t)]$ is either equal to 0 (when $\operatorname{sign}(\beta_i^t) = \operatorname{sign}(\beta_i^t + r_i)$) or equal to $2|\beta_i^t + r_i|$ otherwise.

Finally, the second term in (2.5.4) is also non-negative. It suffices to show for each $i \in \bar{S}$, we have $0 \leq \hat{\lambda}_i|r_i| - \nu_i r_i$, or equivalently $0 \leq \hat{\lambda}_i - \nu_i \operatorname{sign}(r_i) = \hat{\lambda}_i(1 - \operatorname{sign}(\beta_i^t)\operatorname{sign}(r_i))$ which is clearly true. Since all three terms in (2.5.4) are non-negative and their sum tends to 0 as $\epsilon \to 0$, it must be true that each term tends to 0,

$$\begin{aligned} \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle &\leq \xi_1(\epsilon), &(2.5.5)\\ \|\boldsymbol{X}\boldsymbol{r}\|^2 &\leq p\xi_1(\epsilon). &(2.5.6)\end{aligned}$$

We now make use of these inequalities to construct the bound for $\frac{1}{p}\|\boldsymbol{r}\|^2$.

Decompose $\boldsymbol{r}$ as $\boldsymbol{r} = \boldsymbol{r}^{\perp} + \boldsymbol{r}^{\|}$, with $\boldsymbol{r}^{\|} \in \ker(\boldsymbol{X})$ and $\boldsymbol{r}^{\perp} \in \ker^{\perp}(\boldsymbol{X})$ so that $\boldsymbol{X}\boldsymbol{r} = \boldsymbol{X}\boldsymbol{r}^{\perp}$. We will now use (2.5.5) and (2.5.6) to obtain bounds for $\|\boldsymbol{r}^{\perp}\|^2$ and $\|\boldsymbol{r}^{\|}\|^2$. First notice that by (2.5.6) and Condition (4) we have $\frac{1}{c_5}\|\boldsymbol{r}^{\perp}\|^2 \leq \hat{\sigma}^2_{min}(\boldsymbol{X})\|\boldsymbol{r}^{\perp}\|^2 \leq \|\boldsymbol{X}\boldsymbol{r}^{\perp}\|^2 = \|\boldsymbol{X}\boldsymbol{r}\|^2 \leq p\xi_1(\epsilon)$.

In the case $\ker(\boldsymbol{X}) = \{0\}$, the proof is concluded. Otherwise, we prove a similar bound for $\|\boldsymbol{r}^{\|}\|^2$. To bound $\|\boldsymbol{r}^{\|}\|^2$, we use the fact that that this can be done if there exists sets $Q \in [p]$ and $\bar{Q} \in [p]/Q$ such that we can bound $\|\boldsymbol{r}^{\|}_{\bar{Q}}\|^2$ and show a high probability lower bound for $\sigma^2_{min}(\boldsymbol{X}_Q)$.

In (2.5.5), decompose $\boldsymbol{r}_{\bar{S}} = \boldsymbol{r}^{\perp}_{\bar{S}} + \boldsymbol{r}^{\|}_{\bar{S}}$ and observe that by Cauchy Schwarz inequality and the bound just obtained,

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\perp}_{\bar{S}}|\rangle \leq \frac{1}{p}\|\hat{\boldsymbol{\lambda}}_{\bar{S}}\|\|\boldsymbol{r}^{\perp}_{\bar{S}}\| \leq \frac{1}{p}\|\hat{\boldsymbol{\lambda}}\|\|\boldsymbol{r}^{\perp}\| \leq \frac{1}{\sqrt{p}}\|\hat{\boldsymbol{\lambda}}\|\sqrt{c_5\xi_1(\epsilon)}. \tag{2.5.7}$$

Then we use the fact that

$$
\begin{aligned}
&\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\|}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}^{\|}_{\bar{S}}\rangle = \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}} - \boldsymbol{r}^{\perp}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} - \boldsymbol{r}^{\perp}_{\bar{S}}\rangle \\
&\leq \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}|\rangle + \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\perp}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle + \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}^{\perp}_{\bar{S}}\rangle \\
&= \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}|\rangle + \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\perp}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle + \langle \hat{\boldsymbol{\lambda}}_{\bar{S}} \operatorname{sign}(\boldsymbol{\beta}^{t}_{\bar{S}}), \boldsymbol{r}^{\perp}_{\bar{S}}\rangle \\
&\leq \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\perp}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}\rangle + 2\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\perp}_{\bar{S}}|\rangle,
\end{aligned}
$$

to get from (2.5.5) and (2.5.7) that

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\|}_{\bar{S}}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}^{\|}_{\bar{S}}\rangle \leq \xi_2(\epsilon). \tag{2.5.8}$$

Next we would like to show

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}, |\boldsymbol{r}^{\|}_{\bar{S}(c_2)}|\rangle - \langle \boldsymbol{\nu}_{\bar{S}(c_2)}, \boldsymbol{r}^{\|}_{\bar{S}(c_2)}\rangle (1 - c_2)^{-1} \geq 0. \tag{2.5.9}$$

Note that it suffices again to prove this elementwise for each $i \in \bar{S}(c_2)$. Specifically, note that $(1 - c_2)^{-1}|\nu_i| < \hat{\lambda}_i$ for each $i \in \bar{S}(c_2)$ by the set's definition and therefore $\hat{\lambda}_i |r^{\|}_i| - \nu_i r^{\|}_i (1 - c_2)^{-1} \geq |\nu_i||r^{\|}_i|(1 - c_2)^{-1} - \nu_i r^{\|}_i (1 - c_2)^{-1} \geq 0$. Therefore,

$$
\begin{aligned}
&\langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}, |\boldsymbol{r}^{\|}_{\bar{S}(c_2)}|\rangle \\
&\overset{(a)}{\leq} \frac{1}{c_2}\langle \boldsymbol{\lambda}_{\bar{S}(c_2)}, |\boldsymbol{r}^{\|}_{\bar{S}(c_2)}|\rangle - \frac{1}{c_2}\langle \boldsymbol{\nu}_{\bar{S}(c_2)}, \boldsymbol{r}^{\|}_{\bar{S}(c_2)}\rangle \\
&= \frac{1}{c_2}\langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)} - \boldsymbol{\nu}_{\bar{S}(c_2)} \operatorname{sign}(\boldsymbol{r}^{\|}_{\bar{S}(c_2)}), |\boldsymbol{r}^{\|}_{\bar{S}(c_2)}|\rangle \\
&\overset{(b)}{\leq} \frac{1}{c_2}\langle \hat{\boldsymbol{\lambda}}_{\bar{S}} - \boldsymbol{\nu}_{\bar{S}} \operatorname{sign}(\boldsymbol{r}^{\|}_{\bar{S}}), |\boldsymbol{r}^{\|}_{\bar{S}}|\rangle = \frac{1}{c_2}\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}^{\|}_{\bar{S}}|\rangle - \frac{1}{c_2}\langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}^{\|}_{\bar{S}}\rangle \overset{(c)}{\leq} c_2^{-1}\xi_2(\epsilon).
\end{aligned}
\tag{2.5.10}
$$

In particular, step (a) follows by (2.5.9), step (b) since $S \subseteq S_t(c_2)$ implies $\bar{S}_t(c_2) \subseteq \bar{S}$ along with the fact that $\hat{\boldsymbol{\lambda}}_{\bar{S}} - \boldsymbol{\nu}_{\bar{S}} \operatorname{sign}(\boldsymbol{r}^{\|}_{\bar{S}}) \geq 0$ elementwise (for each $i \in \bar{S}$, we have

$\hat{\lambda}_i - \nu_i \operatorname{sign}(r_i^{\|}) > 0$ by $\hat{\lambda}_i \geq |\nu_i|$). Finally step (c) holds by (2.5.8). We now use the bound in (2.5.10) to bound components of $\boldsymbol{r}^{\|}$.

In order to bound $\|\boldsymbol{r}^{\|}\|^2$, we would like to exploit a relationship between the ℓ_1 and ℓ_2 norms. To do this, we consider an ordering of the elements of the vector $\boldsymbol{r}^{\|}$ by magnitude. Recall that $\bar{S}_t(c_2) \subseteq \bar{S}$ and we first assume $|\bar{S}_t(c_2)| \geq pc_3/2$. Now we partition $\bar{S}_t(c_2) = \cup_{\ell=1}^{K} S_\ell$, where $(pc_3/2) \leq |S_\ell| \leq pc_3$, and such that for each $i \in S_\ell$ and $j \in S_{\ell+1}$, it follows that $|r_i^{\|}| \geq |r_j^{\|}|$. Finally, define $\bar{S}_+ := \cup_{\ell=2}^{K} S_\ell \subseteq \bar{S}_t(c_2)$, i.e. the set union of all the partitions except the first one corresponding to the indices containing the largest elements in $\boldsymbol{r}^{\|}$. Now we note for any $i \in S_\ell$, we have $|\boldsymbol{r}_i^{\|}| \leq \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|/|S_{\ell-1}|$, that is, in terms of absolute value, for any i in group ℓ, it should be smaller than the average of all the elements in the previous group $\ell - 1$.

Then,

$$
\begin{aligned}
\|\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2 &\overset{(a)}{=} \sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_\ell}^{\|}\|^2 \overset{(b)}{\leq} \sum_{\ell=2}^{K} |S_\ell| \frac{\|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1^2}{|S_{\ell-1}|^2} \\
&\overset{(c)}{\leq} \frac{4}{pc_3} \sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1^2 \leq \frac{4}{pc_3} \Big[\sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1 \Big]^2 \\
&\overset{(d)}{\leq} \frac{4}{pc_3} \|\boldsymbol{r}_{\bar{S}(c_2)}^{\|}\|_1^2 \overset{(e)}{\leq} \frac{4\xi_2(\epsilon)^2 p}{c_2^2 c_3 (\min \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)})^2} =: p\xi_3(\epsilon).
\end{aligned}
\tag{2.5.11}
$$

In the above, step (a) follows from the definition of $\bar{S}_+$, step (b) from the fact that for $i \in S_\ell$, we have $|\boldsymbol{r}_i^{\|}| \leq \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|/|S_{\ell-1}|$, step (c) since $(pc_3/2) \leq |S_\ell| \leq pc_3$, and step (d) since $\sum_{\ell=2}^{K} S_\ell \subset \sum_{\ell=1}^{K} S_\ell = \bar{S}_t(c_2)$. Finally step (e) follows using that $\frac{1}{p} \min\{\hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}\} \|\boldsymbol{r}_{\bar{S}(c_2)}^{\|}\|_1 \leq \langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}, |\boldsymbol{r}_{\bar{S}(c_2)}^{\|}| \rangle$.

Now, recalling $S_+ = S_t(c_2) \cup S_1$ and $|S_1| \leq pc_3$, by Condition (3), $\sigma_{min}(\boldsymbol{X}_{S_+}) \geq c_4$

and therefore,

$$c_4^2\|\boldsymbol{r}^{\|}_{S_+}\|^2 \leq \sigma^2_{min}(\boldsymbol{X}_{S_+})\|\boldsymbol{r}^{\|}_{S_+}\|^2 \leq \|\boldsymbol{X}_{S_+}\boldsymbol{r}^{\|}_{S_+}\|^2 \overset{(a)}{=} \|\boldsymbol{X}_{\bar{S}_+}\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2 \overset{(b)}{\leq} 2c_5\|\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2. \quad (2.5.12)$$

In the above, in step (a) we use that $\boldsymbol{0} = \boldsymbol{X}\boldsymbol{r}^{\|} = \boldsymbol{X}_{S_+}\boldsymbol{r}^{\|}_{S_+} + \boldsymbol{X}_{\bar{S}_+}\boldsymbol{r}^{\|}_{\bar{S}_+}$. In step (b) we use Condition (4) and the fact that $\|\boldsymbol{X}_{\bar{S}_+}\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2 \leq \sigma^2_{\max}(\boldsymbol{X})\|\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2$. Therefore, to conclude the proof, it is sufficient to prove a bound for $\|\boldsymbol{r}^{\|}_{S_+}\|^2$.

Decomposing $\|\boldsymbol{r}^{\|}\|^2 = \|\boldsymbol{r}^{\|}_{S_+}\|^2 + \|\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2$, we find from (2.5.11) and (2.5.11) the desired bound:

$$\|\boldsymbol{r}^{\|}\|^2 \leq \|\boldsymbol{r}^{\|}_{S_+}\|^2 + \|\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2 \leq \Big(\frac{2c_5}{c_4^2} + 1\Big)\|\boldsymbol{r}^{\|}_{\bar{S}_+}\|^2 \leq \Big(\frac{2c_5}{c_4^2} + 1\Big)p\xi_3(\epsilon).$$

This finishes the proof when $|\bar{S}_t(c_2)| \geq pc_3/2$. When $|\bar{S}_t(c_2)| < pc_3/2$, we can take $\bar{S}_+ = \emptyset$ and $S_+ = [p]$. Hence, the result holds as a special case of the above inequality. □

2.6 Expansion of the AMP State Evolution Ideas

In this section, we develop ideas and notation specifically for the SLOPE AMP algorithm given in (2.1.3). Most are adapted from the work in [BMN20] that studies general non-separable AMP algorithms. These results relate to the performance analysis of the AMP algorithm and will be useful in proving Lemma 2.5.5. Throughout this section, we use the $\{\eta_p^t\}_{p\in\mathbb{N}_{>0}}$ notation introduced in Section 2.4 and defined in (2.4.1). Namely, we consider a sequence of denoisers $\eta_p^t : \mathbb{R}^p \to \mathbb{R}^p$ to be those that apply the proximal operator $\text{prox}_{J_{\alpha\tau_t}}(\cdot)$ defined in (2.1.4), i.e. $\eta_p^t(\boldsymbol{v}) := \text{prox}_{J_{\alpha\tau_t}}(\boldsymbol{v})$ for a vector $\boldsymbol{v} \in \mathbb{R}^p$.

Given $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, define sequences of column vectors $\boldsymbol{h}^{t+1} \in \mathbb{R}^p$ and $\boldsymbol{m}^t \in \mathbb{R}^n$ for $t \geq 0$. At each iteration t, the sequence $\boldsymbol{h}^{t+1}$ measures the difference between the truth $\boldsymbol{\beta}$ and the pseudo-data $\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t$, that is the input to the denoiser, and the sequence $\boldsymbol{m}^t$ measures the difference between the noise $\boldsymbol{w}$ and the AMP residual $\boldsymbol{z}^t$. Namely, define $\boldsymbol{m}^t, \boldsymbol{h}^{t+1}$: for $t \geq 0$,

$$\boldsymbol{h}^{t+1} = \boldsymbol{\beta} - (\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t) \quad \text{and} \quad \boldsymbol{m}^t = \boldsymbol{w} - \boldsymbol{z}^t. \tag{2.6.1}$$

We next introduce a generalization to the state evolution given in (2.2.4), that will be useful in studying the limiting properties of functions of the AMP estimates $\boldsymbol{\beta}^s$ and $\boldsymbol{\beta}^t$ at different iterations s and t. To do this, we will recursively define covariances $\{\Sigma_{s,t}\}_{s,t\geq 0}$: for $\boldsymbol{B}$ elementwise i.i.d. $\sim B$, set $\Sigma_{0,0} = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[B^2]$ and

$$\Sigma_{0,t+1} = \sigma_w^2 + \lim_p \frac{1}{\delta p}\mathbb{E}\{-\boldsymbol{B}^\top[\eta_p^t(\boldsymbol{B} + \tau_t \boldsymbol{Z}_t) - \boldsymbol{B}]\}, \tag{2.6.2}$$

for $\boldsymbol{Z}_t \sim \mathcal{N}(0, \mathbb{I})$ independent of $\boldsymbol{B}$. Then for each $t \geq 0$, given $(\Sigma_{s,r})_{0\leq s,r\leq t}$, define

$$\Sigma_{s+1,t+1} = \sigma_w^2 + \lim_p \frac{1}{\delta p}\mathbb{E}\Big\{[\eta_p^s(\boldsymbol{B} + \tau_s \boldsymbol{Z}_s) - \boldsymbol{B}]^\top[\eta_p^t(\boldsymbol{B} + \tau_t \boldsymbol{Z}_t) - \boldsymbol{B}]\Big\}, \tag{2.6.3}$$

where $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_r$ are length$-p$ jointly Gaussian vectors, independent of $\boldsymbol{B} \sim B$ i.i.d. elementwise, with $\mathbb{E}[\boldsymbol{Z}_s] = \mathbb{E}[\boldsymbol{Z}_r] = \boldsymbol{0}$, $\mathbb{E}\{([\boldsymbol{Z}_s]_i)^2\} = \mathbb{E}\{([\boldsymbol{Z}_r]_i)^2\} = 1$ for any element $i \in [p]$, and $\mathbb{E}\{[\boldsymbol{Z}_s]_i[\boldsymbol{Z}_r]_j\} = \frac{\Sigma_{s,r}}{\tau_r\tau_s}\mathbb{I}\{i = j\}$. Note that $\Sigma_{t,t} = \tau_t^2$ defined in (2.2.4).

Using the above covariances, we have the following result that characterizes the asymptotic empirical distributions of the difference vectors defined in (2.9.41) and generalizes Lemma (2.4.1). This result follows by [BMN20, Theorem 1].

Lemma 2.6.1. *[BMN20, Theorem 1] Assuming that $\Sigma_{0,0}, \ldots, \Sigma_{t+1,t+1} > \sigma_w^2$, then for any deterministic sequence $\phi_p : (\mathbb{R}^p \times \mathbb{R}^n)^t \times \mathbb{R}^p \to \mathbb{R}$ of uniformly pseudo-Lipschitz functions of order k,*

$$\begin{aligned}&\operatorname*{plim}_p \Big(\phi_p(\boldsymbol{\beta}, \boldsymbol{m}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{m}^t, \boldsymbol{h}^{t+1}) \\ &- \mathbb{E}[\phi_p(\boldsymbol{\beta}, \sqrt{\tau_0^2 - \sigma_w^2}\boldsymbol{Z}_0', \tau_0\boldsymbol{Z}_0, \ldots, \sqrt{\tau_t^2 - \sigma_w^2}\boldsymbol{Z}_t', \tau_t\boldsymbol{Z}_t)]\Big) = 0,\end{aligned}$$

for $(\boldsymbol{Z}_0, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t)$ defined in (2.6.3) *in dependent of $(\boldsymbol{Z}_0', \boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_t')$ and the expectation is taken with respect to the collection $(\boldsymbol{Z}_0, \boldsymbol{Z}_0', \boldsymbol{Z}_1, \boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_t', \boldsymbol{Z}_t)$. We note that $\boldsymbol{Z}_s'$ and $\boldsymbol{Z}_r'$ are length$-n$ jointly Gaussian vectors, with $\mathbb{E}[\boldsymbol{Z}_s'] = \mathbb{E}[\boldsymbol{Z}_r'] = \mathbf{0}$, $\mathbb{E}\{([\boldsymbol{Z}_s']_i)^2\} = \mathbb{E}\{([\boldsymbol{Z}_r']_i)^2\} = 1$ for any element $i \in [n]$, and $\mathbb{E}\{[\boldsymbol{Z}_s']_i[\boldsymbol{Z}_r']_j\} = (\Sigma_{s,r} - \sigma_w^2)((\tau_r^2 - \sigma_w^2)(\tau_s^2 - \sigma_w^2))^{-1/2}\mathbb{I}\{i = j\}$.*

We use Lemma 2.6.1 to explicitly state asymptotic characterizations of AMP quantities that will be useful in our analysis.

Lemma 2.6.2. *Under the condition of Theorem 3, for $\boldsymbol{z}^t$ and $\boldsymbol{\beta}^{t+1}$ defined in* (2.1.3) *and the generalized state evolution sequence defined in* (2.6.3),

$$\operatorname*{plim}_n \Big(\frac{1}{n}\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|^2 - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2)\Big) = 0, \tag{2.6.4}$$

$$\operatorname*{plim}_p \Big(\frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2)\Big) = 0. \tag{2.6.5}$$

Proof. The major tools in proving (2.6.4)-(2.6.5) are first recognizing that we can write the differences $\boldsymbol{z}^t - \boldsymbol{z}^{t-1}$ and $\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t$ as a function of the values $(\boldsymbol{\beta}, \boldsymbol{m}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{m}^t, \boldsymbol{h}^{t+1})$ defined in (2.9.41) and finally making an appeal to the Law of Large Numbers. We prove (2.6.5) and (2.6.4) follows similarly.

By (2.1.3a), $\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t = \eta_p^t(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{z}^t) - \eta_p^{t-1}(\boldsymbol{\beta}^{t-1} + \boldsymbol{X}^\top \boldsymbol{z}^{t-1}) = \eta_p^t(\boldsymbol{\beta} - \boldsymbol{h}^{t+1}) - \eta_p^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t)$. Therefore, we will appeal to Lemma 2.6.1 for the uniformly pseudo-Lipschitz function

$$\begin{aligned}
&\phi_p(\boldsymbol{\beta}, \boldsymbol{m}^0, \boldsymbol{h}^1, \ldots, \boldsymbol{m}^t, \boldsymbol{h}^{t+1}) \\
&= \frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 = \frac{1}{\delta p}\|\eta_p^t(\boldsymbol{\beta} - \boldsymbol{h}^{t+1}) - \eta_p^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t)\|^2.
\end{aligned}$$

We note that it easy to show that the above function is uniformly pseudo-Lipschitz, though we don't do this here. Then by Lemma 2.6.1,

$$\operatorname*{plim}_p \left(\frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 - \frac{1}{\delta p}\mathbb{E}\|\eta_p^t(\boldsymbol{\beta} - \tau_t \boldsymbol{Z}_t) - \eta_p^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|^2\right) = 0. \quad (2.6.6)$$

Now to prove result (2.6.4), we note that by Lemma 2.3.2,

$$\begin{aligned}
&\operatorname*{plim}_{\delta p} \frac{1}{p}\mathbb{E}\|\eta_p^t(\boldsymbol{\beta} - \tau_t \boldsymbol{Z}_t) - \eta_p^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|^2 \\
&= \lim_p \frac{1}{\delta p}\mathbb{E}\|\eta_p^t(\boldsymbol{B} - \tau_t \boldsymbol{Z}_t) - \eta_p^{t-1}(\boldsymbol{B} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|^2,
\end{aligned}$$

where $\boldsymbol{B} \sim B$ i.i.d. elementwise independent of $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-1}$. The argument for showing that the assumptions of Lemma 2.3.2 are met follows like that used in Appendix 2.9.2 in the proof of Proposition **(P2)** introduced in Section 2.4. Then, $\lim_p \frac{1}{\delta p}\mathbb{E}\|\eta_p^t(\boldsymbol{B} - \tau_t \boldsymbol{Z}_t) - \eta_p^{t-1}(\boldsymbol{B} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|^2 = \Sigma_{t,t} - 2\Sigma_{t,t-1} + \Sigma_{t-1,t-1}$.

□

We finally state a lemma that characterizes the asymptotic value of the normalized ℓ_2 norm of the residuals in AMP algorithm (2.1.3b) following from Lemma 2.4.1.

Lemma 2.6.3. *For $\boldsymbol{z}^t$ defined in* (2.1.3b) *and τ_t^2 given in* (2.2.4)*,*

$$\operatorname*{plim}_n \left(\|\boldsymbol{z}^t\|^2/n - \tau_t^2\right) = 0. \tag{2.6.7}$$

Proof. This follows from Lemma 2.4.1, using the uniformly pseudo-Lipschitz (of order 2) sequence of functions $\phi_n(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{n}\|\boldsymbol{a}\|^2$ to get, $\text{plim}_n \|\boldsymbol{z}^t\|^2/n = \text{plim}_n \mathbb{E}_{\boldsymbol{Z}}[\|\boldsymbol{w} + \sqrt{\tau_t^2 - \sigma_w^2}\boldsymbol{Z}\|^2]/n$ for $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I})$. Then the final result follows by noticing that $\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{w} + \sqrt{\tau_t^2 - \sigma_w^2}\boldsymbol{Z}\|^2 = \|\boldsymbol{w}\|^2 + (\tau_t^2 - \sigma_w^2)\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{Z}\|^2 = \|\boldsymbol{w}\|^2 + n(\tau_t^2 - \sigma_w^2)$, and therefore, using that $\text{plim}_n \|\boldsymbol{w}\|^2/n = \sigma_w^2$ by the Law of Large Numbers,

$$\operatorname*{plim}_n \frac{1}{n}\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{w} + \sqrt{\tau_t^2 - \sigma_w^2}\boldsymbol{Z}\|^2 = (\tau_t^2 - \sigma_w^2) + \operatorname*{plim}_n \frac{1}{n}\|\boldsymbol{w}\|^2 = \tau_t^2.$$

□

2.7 Verification of Main Technical Lemma Conditions

We now verify that the Lemma 2.5.5 conditions 1-5 are met for the SLOPE cost function and the associated AMP algorithm. We note that conditions 1, 4, and 5 are straightforward, so their proof is presented first. On the other hand, condition 2 and condition 3 are quite technical. Their proofs are given in Section 2.7.4 and Section 2.7.5 below.

2.7.1 Condition (4)

This follows by standard limit theorems about the singular values of Wishart matrices (see Appendix 2.9.7, Theorem H.2).

2.7.2 Condition (5)

Recall, $\mathcal{C}_{\boldsymbol{x}}(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \hat{\boldsymbol{\lambda}}_i |b_i|$ for some $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}_{\boldsymbol{x}}^{-1}(\boldsymbol{\lambda}))$, and by definition, $\mathcal{C}_{\boldsymbol{x}}(\boldsymbol{x}) = \mathcal{C}(\boldsymbol{x})$ for all $\boldsymbol{x}$. Since $\widehat{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{C}(\cdot)$ we have $\mathcal{C}(\boldsymbol{\beta}^t) \geq \mathcal{C}(\hat{\boldsymbol{\beta}})$ and by the rearrangement inequality, $\mathcal{C}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}) \geq \mathcal{C}_{\boldsymbol{\beta}^t}(\hat{\boldsymbol{\beta}})$. Therefore, $\mathcal{C}(\boldsymbol{\beta}^t) \geq \mathcal{C}(\hat{\boldsymbol{\beta}}) = \mathcal{C}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}) \geq \mathcal{C}_{\boldsymbol{\beta}^t}(\hat{\boldsymbol{\beta}})$.

2.7.3 Condition (1)

Condition (1) follows, for large enough p, from Lemma 2.7.1, stated below, which proves the asymptotic boundedness of the norms of the AMP estimates $\boldsymbol{\beta}^t$ and the SLOPE estimate $\widehat{\boldsymbol{\beta}}$.

Lemma 2.7.1. *For any parameter vector $\boldsymbol{\lambda} \in \mathbb{R}^p$ defining a SLOPE cost as in* (2.1.2)*, let $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\lambda})$, then for $t \geq 0$,*

$$\operatorname*{plim}_p \frac{1}{p}\|\boldsymbol{\beta}^t\|^2 = \operatorname*{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\eta_p^t(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})\|^2] \leq 2\sigma_\beta^2 + 2\tau_t^2, \tag{2.7.1}$$

for $\eta_p^t(\cdot)$ defined in (2.4.1) *with $\sigma_\beta^2 := \mathbb{E}[B^2] < \infty$ and $\sigma_\beta^2 + \tau_*^2 < \infty$ and*

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2 \leq \mathsf{C}, \tag{2.7.2}$$

where $\mathsf{C} := \mathsf{C}(\delta, \sigma_{\beta}^2, \sigma_w^2, \mathsf{B}_{max}, \mathsf{B}_{min}, \lambda_{min})$ *is a positive constant depending on* $\delta, \sigma_{\beta}^2, \sigma_w^2$, *along with the singular values of* $\boldsymbol{X}$ *through* $\mathsf{B}_{max} \geq \lim_p \sigma_{max}^2(\boldsymbol{X})$, *and* $\mathsf{B}_{min} \leq \lim_p \hat{\sigma}_{min}^2(\boldsymbol{X})$, *and a lower bound on the parameter values* $\lambda_{min} := \lim_p \min(\boldsymbol{\lambda})$.

Proof. The proof is included in Appendix 2.9.4. □

2.7.4 Condition (2)

Condition (2) follows from Lemma 2.7.2 stated below, for ϵ arbitrarily small when t is large enough.

Lemma 2.7.2. *Under the conditions of Theorem 3, for every iteration* t, *there exists a subgradient* $sg(C, \boldsymbol{\beta}^t)$ *of* C *defined in* (2.5.1) *at point* $\boldsymbol{\beta}^t$ *such that almost surely,*

$$\lim_t \operatorname*{plim}_p \frac{1}{p} \|sg(C, \boldsymbol{\beta}^t)\|^2 = 0.$$

The proof is an adaption of [BM11c, Lemma 3.3], though, the subgradient for the SLOPE cost function (studied extensively in Section 2.5.1) is quite different than that of the LASSO cost and our analysis requires handling this carefully. Before we prove Lemma 2.7.2, we state and prove a result which tells us that the asymptotic difference between the AMP output at any two iterations t and $t-1$ goes to zero in ℓ_2 norm as the algorithm runs. This result is crucial to the proof of Lemma 2.7.2.

Lemma 2.7.3. *Under the condition of Theorem 3, the estimates* $\{\boldsymbol{\beta}^t\}_{t \geq 0}$ *and residuals* $\{\boldsymbol{z}^t\}_{t \geq 0}$ *of AMP almost surely satisfy*

$$\lim_t \operatorname*{plim}_p \frac{1}{\delta p} \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}\|^2 = 0, \qquad \textit{and} \qquad \lim_t \operatorname*{plim}_p \frac{1}{n} \|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|^2 = 0$$

Proof of Lemma 2.7.3. This result uses Lemma 2.6.2, which characterizes the large system limit of $\frac{1}{n}\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|^2$ and $\frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2$ as both being equal to $\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2$ where $\Sigma_{t,t-1}$ is the generalized state evolution sequence defined in (2.6.3). Then Lemma 2.9.4 (which is stated and proved in Appendix 2.9.5) shows that $\lim_t (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) = 0$. □

Proof of Lemma 2.7.2. For any vector $\boldsymbol{\nu}^t \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$, note that $\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t)$ is a valid subgradient belonging to the set $\partial\mathcal{C}(\boldsymbol{\beta}^t)$ as defined in Fact 2.5.1. Moreover, by AMP (2.1.3b), $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t = \boldsymbol{z}^t - w^t\boldsymbol{z}^{t-1}$ with $w^t := \frac{1}{\delta p}[\nabla\eta^{t-1}(\boldsymbol{\beta}^{t-1} + \boldsymbol{X}^\top\boldsymbol{z}^{t-1})]$. Therefore we can write,

$$\begin{aligned}
&\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) = \boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{z}^t - w^t\boldsymbol{z}^{t-1}) \\
&= \boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) - (1 - w^t)\boldsymbol{X}^\top\boldsymbol{z}^{t-1} \qquad (2.7.3) \\
&= (\boldsymbol{\nu}^t - \mu_t\boldsymbol{X}^\top\boldsymbol{z}^{t-1}) - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + (\mu_t - (1 - w^t))\boldsymbol{X}^\top\boldsymbol{z}^{t-1},
\end{aligned}$$

where we define $\mu_t := \langle\boldsymbol{\lambda}, \boldsymbol{\theta}_{t-1}\rangle/\|\boldsymbol{\theta}_{t-1}\|^2$ as the ratio of $\boldsymbol{\lambda}$ to $\boldsymbol{\theta}_{t-1}$ so that $\boldsymbol{\lambda} = \mu_t\boldsymbol{\theta}_{t-1}$ (here $\boldsymbol{\theta}_{t-1} := \boldsymbol{\alpha}\tau_{t-1}$ and recall that $\boldsymbol{\alpha}$ is calibrated to be parallel to $\boldsymbol{\lambda}$). It follows that $\partial J_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \mu_t\,\partial J_{\boldsymbol{\theta}_{t-1}}(\boldsymbol{x})$.

Now, by the definition of the proximal operator used in (2.1.3a) and by Fact 2.5.2, we have that $(\boldsymbol{X}^\top\boldsymbol{z}^{t-1} + \boldsymbol{\beta}^{t-1}) - \boldsymbol{\beta}^t \in \partial J_{\boldsymbol{\theta}^{t-1}}(\boldsymbol{\beta}^t)$. Hence we choose $\boldsymbol{\nu}^t$ to be the specific subgradient defined by

$$\boldsymbol{\nu}^t = \mu_t(\boldsymbol{X}^\top\boldsymbol{z}^{t-1} + \boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t) \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t), \qquad (2.7.4)$$

which leads to $\boldsymbol{\nu}^t - \mu_t \boldsymbol{X}^\top \boldsymbol{z}^{t-1} = \mu_t(\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t)$. Plugging into (2.7.3),

$$\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) = \mu_t(\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t) - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + (\mu_t - (1 - w^t))\boldsymbol{X}^\top \boldsymbol{z}^{t-1}. \quad (2.7.5)$$

Then taking the norm, dividing by $\sqrt{p}$, and using the triangular inequality, we have

$$\begin{aligned} &\frac{1}{\sqrt{p}}\|\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t)\| \\ &\leq \frac{\mu_t}{\sqrt{p}}\|\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t\| + \frac{1}{\sqrt{p}}\|\boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1})\| + \frac{(\mu_t - (1 - w^t))}{\sqrt{p}}\|\boldsymbol{X}^\top \boldsymbol{z}^{t-1}\|. \end{aligned}$$

Using Lemma 2.6.2, that $\sigma_{\max}(\boldsymbol{X})$ is almost surely bounded as $p \to \infty$ (cf. Theorem 2), and that $\lim_t \lim_p \mu_t = 1 - \lim_p \frac{1}{\delta p}\mathbb{E}||\operatorname{prox}_{J_{\boldsymbol{A}(p)\tau_*}}(\boldsymbol{B} + \tau_* \boldsymbol{Z})||_0^*$ as in (2.2.11) is finite, the first two terms on the right side of the above → 0. Finally, for the third term, Lemma 2.6.3 gives $\lim_t \operatorname{plim}_p \|z^t\|/\sqrt{p} = \tau_*$, and together with the calibration formula (2.2.11), that $\sigma_{\max}(\boldsymbol{X})$ is almost surely bounded as $p \to \infty$, and the definition of w in the proof of Lemma 2.2.2, we find $\lim_t \lim_p(\mu_t - (1 - w^t)) = 0$, and thus the third term → 0. As $\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) \in \partial\mathcal{C}(\boldsymbol{\beta}^t)$, the proof is complete. □

2.7.5 Condition (3)

We take $\boldsymbol{\nu}^t$ to be the subgradient defined in (2.7.4) and since t is fixed, we drop the superscript t writing $\boldsymbol{\nu} := \boldsymbol{\nu}^t$. Recall the sets $s_t(c_2)$ and $S_t(c_2)$ defined in Condition (3). Then for s' being *any* set of maximal atoms in $[p]$ with $|s'| \leq c_3 p$ and $S' := \{i \in I : I \in s'\}$, we would like to show $\sigma_{min}(\boldsymbol{X}_{S_t(c_2) \cup S'}) \geq c_4$. This holds by Proposition 2.7.4, stated below, whose proof is the main challenge. We state the proposition and then we identify two auxiliary lemmas, Lemma 2.7.5 and 2.7.6, that will be used to ultimately prove Proposition 2.7.4.

Proposition 2.7.4. *There exist constants $c_2 \in (0,1)$, c_3, $c_4 > 0$ and $t_{\min} < \infty$ such that, for any $t \geq t_{\min}$, and set S_t defined in Condition (3)*

$$\min_{s'} \left\{ \sigma_{\min}(\boldsymbol{X}_{S_t(c_2) \cup S'}) \, : S' \subseteq [p] \, , \, |s'| \leq c_3 p \, , \, S' = \{i \in I : I \in s'\} \right\} \geq c_4$$

eventually almost surely as $p \to \infty$.

The proof of Proposition 2.7.4 will use two auxiliary lemmas, Lemma 2.7.5 and 2.7.6, stated below.

Lemma 2.7.5. *Let the set s_t be measurable on the σ-algebra $\mathfrak{S}_t$ generated by $\{\boldsymbol{z}^0, \ldots, \boldsymbol{z}^{t-1}\}$ and $\{\boldsymbol{\beta}^0 + \boldsymbol{X}^* \boldsymbol{z}^0, \ldots, \boldsymbol{\beta}^{t-1} + \boldsymbol{X}^* \boldsymbol{z}^{t-1}\}$ and assume $|s_t| \leq p(\delta - c)$ for some $c > 0$. Define $S_t \subseteq [p]$ as $\{i \in I$ for some $I \in s_t\}$. Then there exists $a_1 = a_1(c) > 0$ (independent of t) and $a_2 = a_2(c,t) > 0$ (depending on t and c) such that*

$$\min_{s'} \left\{ \sigma_{\min}(\boldsymbol{X}_{S_t \cup S'}) \, : S' \subseteq [p] \, , \, |s'| \leq a_1 p \, , \, S' = \{i \in I : I \in s'\} \right\} \geq a_2 \, ,$$

eventually almost surely as $p \to \infty$.

Proof. The proof of Lemma 2.7.5 is given in Appendix 2.9.6. The key difference in SLOPE case (Lemma 2.7.5) and LASSO case (cf. [BM11c, Lemma 3.4]) is the concept of equivalence classes of indices. On a high level, the set s describes some structure in the support space S and such structure restricts the dimension of some linear spaces in the proof of Lemma 2.7.5. □

Lemma 2.7.6. *[BM11c, Lemma 3.5] Fix $\gamma \in (0,1)$ and let the sequence $\{S_t(\gamma)\}_{t\geq 0}$ be defined as before. For any $\xi > 0$ there exists $t_* = t_*(\xi,\gamma) < \infty$ such that, for all $t_2 \geq t_1 \geq t_*$ fixed, we have*

$$\frac{1}{p}|S_{t_2}(\gamma) \setminus S_{t_1}(\gamma)| < \xi \,, \tag{2.7.6}$$

eventually almost surely as $p \to \infty$.

Proof. For LASSO, this result was given in [BM11c, Lemma 3.5], and for SLOPE, the proof stays largely the same so we don't repeat it here. The major difference is that where the work in [BM11c] can appeal to AMP analysis in [BM11a], for SLOPE, we appeal to similar results given in [BMN20] (e.g. Lemma 2.6.1). □

Proof of Proposition 2.7.4. The subgradient in Condition (2) is given by $sg(\mathcal{C}, \boldsymbol{\beta}^t) := \boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t)$ where $\boldsymbol{\nu}^t \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ is the subgradient defined in the Condition (2) proof at Eq. (2.7.4). Recall, $S_t(c_2) = \{i \in I : |\boldsymbol{\nu}_I^t| \succeq \mathcal{P}([\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda})]_I)(1-c_2)\}$. We include a simple visualization for the set $S_t(c_2)$ in Figure 2.4. We have plotted the subgradient $\boldsymbol{\nu}_I^t = (-1,2)$ for (zero) equivalence class $I = \{1,2\}$ when $\boldsymbol{\lambda} = (4,1)$ and $\boldsymbol{\beta}^t = (0,0)$. Then indices of $|\boldsymbol{\nu}_I^t|$, namely $(1,2)$ are in $S_t(c_2)$ unless $c_2 < 0.4$.

We know from the proof of Lemma 2.7.2 Eq. (2.7.4) that $\boldsymbol{\nu}^t = \mu_t(\boldsymbol{X}^\top \boldsymbol{z}^{t-1} + \boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t) \in \mu_t J_{\boldsymbol{\theta}^t}(\boldsymbol{\beta}^t)$ where $\mu_t := \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_{t-1}\rangle / \|\boldsymbol{\theta}_{t-1}\|^2$ and $\boldsymbol{\lambda} = \mu_t \boldsymbol{\theta}^{t-1}$. Therefore, summing over all equivalence classes I,

$$\begin{aligned} |s_t(c_2)| &= \sum_I \mathbb{I}\{|\boldsymbol{\nu}_I^t| \succeq \mathcal{P}([\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda})]_I)(1-c_2)\} \\ &= \sum_I \mathbb{I}\Big\{|\boldsymbol{\beta}^t - [\boldsymbol{X}^\top \boldsymbol{z}^{t-1}] - \boldsymbol{\beta}^{t-1}|_I \succeq \mathcal{P}([\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\theta}^{t-1})]_I)(1-c_2)\Big\}. \end{aligned} \tag{2.7.7}$$

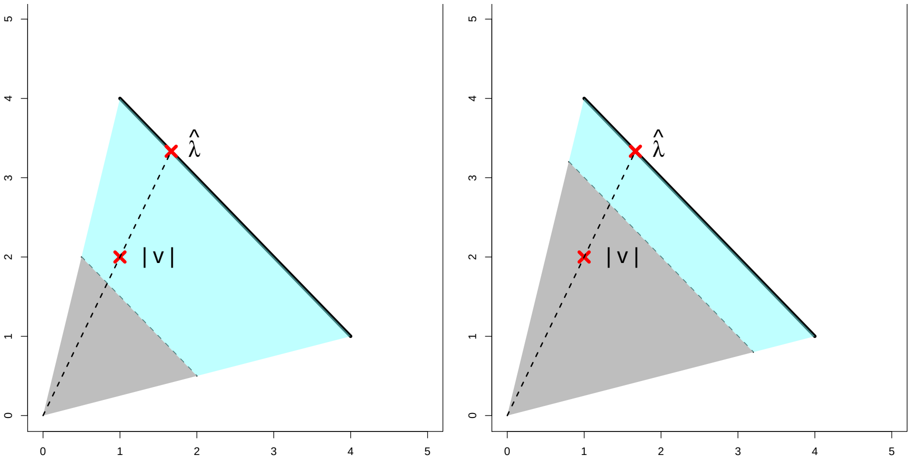


Figure 2.4: Left: $c_2 = 0.5$; Right: $c_2 = 0.2$; Blue area is $\{\boldsymbol{\nu} \in \partial J_{\boldsymbol{\lambda}}(0,0) : |\boldsymbol{\nu}| \succeq (1-c_2)\mathcal{P}(\lambda_1, \lambda_2)\}$ and grey area is complement of blue area in $\partial J_{\boldsymbol{\lambda}}(0,0)$.

As detailed in the proof of Lemma 2.5.5, for non-zero equivalence classes, let $\hat{\boldsymbol{\lambda}}_I = |\boldsymbol{\nu}_I|$, and for the zero equivalence class, let $\hat{\boldsymbol{\lambda}}_I \geq |\boldsymbol{\nu}_I|$, meaning $\hat{\boldsymbol{\lambda}}_I$ is parallel to $|\boldsymbol{\nu}_I|$ for each equivalence class I of $\boldsymbol{\beta}^t$. That such a $\hat{\boldsymbol{\lambda}}$ exists in the set $\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$ follows since $\boldsymbol{\nu}$ is a valid subgradient of $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ (see Fact 2.5.3). We can then simplify the set definitions of $s_t(c_2)$ and $S_t(c_2)$ to be $s_t(c_2) := \{I \subset [p] : |\boldsymbol{\nu}_I| \geq (1-c_2)\hat{\boldsymbol{\lambda}}_I\}$ and $S_t(c_2) := \{i : |\nu_i| \geq (1-c_2)\hat{\lambda}_i\}$, where both use equivalence classes, I, defined for $\boldsymbol{\beta}^t$. Then since $\boldsymbol{\lambda} = \mu_t \boldsymbol{\theta}^{t-1}$, we also let $\hat{\boldsymbol{\theta}}^{t-1}$ be defined such that $\hat{\boldsymbol{\lambda}} = \mu_t \hat{\boldsymbol{\theta}}^{t-1}$.

Therefore, by (2.7.7), $|s_t(c_2)| = \sum_I \mathbb{I}\{|\boldsymbol{\beta}^t - [\boldsymbol{X}^\top \boldsymbol{z}^{t-1}] - \boldsymbol{\beta}^{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\}$. In the notation of (2.9.41), $\boldsymbol{\beta}^t - [\boldsymbol{X}^\top \boldsymbol{z}^{t-1}] - \boldsymbol{\beta}^{t-1} = \boldsymbol{h}^t + \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t) - \boldsymbol{\beta}$ and $\boldsymbol{\beta}^t = \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t)$ and therefore by (2.7.7),

$$|s_t(c_2)| = \sum_I \mathbb{I}\Big\{|\boldsymbol{h}^t + \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t) - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big\}.$$

Now, we note that Lemma 2.6.1 implies weak convergence of the empirical distribution of $\boldsymbol{h}^t$ to $\tau_{t-1}\boldsymbol{Z}_{t-1}$ for $\boldsymbol{Z}_{t-1}$ a vector of i.i.d. standard Gaussian and τ_{t-1} given by the state evolution (2.2.4). Therefore a careful argument using continuous approximations to indicators gives,

$$\begin{aligned}&\operatorname*{plim}_p \frac{1}{p}\sum_I \mathbb{I}\Big\{|\boldsymbol{h}^t + \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t) - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big\}\\ &= \lim_p \mathbb{E}_{\boldsymbol{Z}_{t-1}}\Big\{\frac{1}{p}\sum_I \mathbb{I}\Big\{|\tau_{t-1}\boldsymbol{Z}_{t-1} + \eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}) - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big\}\Big\},\end{aligned} \tag{2.7.8}$$

where in the right side of the above, the equivalence classes I are taken with respect to $\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})$ and $\hat{\boldsymbol{\theta}}_I^{t-1}$ as equal to or larger than $|\tau_{t-1}\boldsymbol{Z}_{t-1} + \eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}) - \boldsymbol{\beta}|_I$ depending on whether I is the zero equivalence class or not. We justify the substitution of $\tau_{t-1}\boldsymbol{Z}_{t-1}$ for $\boldsymbol{h}^t$ by approximating the sum of indicators with a function that counts the number of elements in $\eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t)$ that are strictly greater than its neighbour. Then this function converges to a continuous and bounded function, the function that measures the proportion of η^{t-1} that is non-flat, to which we apply the Portmanteau Theorem (cf. [HL19a], Lemma 1(b) in [BM11a] and Lemma F.3(b) in [BM11c]).

Now, using (2.7.8), we can simplify:

$$\begin{aligned}&\operatorname*{plim}_p \frac{1}{p}|s_t(c_2)|\\ &= \lim_p \frac{1}{p}\sum_I \mathbb{P}_{\boldsymbol{Z}_{t-1}}\Big(|\tau_{t-1}\boldsymbol{Z}_{t-1} - \eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}) - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big),\end{aligned} \tag{2.7.9}$$

and we study the probability on the right side of the above, for a fixed equivalence class I, writing $\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})$ to be η^{t-1}, dropping the input.

$$\begin{aligned}
&\mathbb{P}\left(|\tau_{t-1}\boldsymbol{Z}_{t-1} + \eta^{t-1} - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\right) \\
&= \mathbb{P}\left(|\tau_{t-1}\boldsymbol{Z}_{t-1} + \eta^{t-1} - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2), \eta_1^{t-1} = \mathbf{0}\right) \\
&+ \mathbb{P}\left(|\tau_{t-1}\boldsymbol{Z}_{t-1} + \eta^{t-1} - \boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2), \eta_I^{t-1} \neq \mathbf{0}\right) \\
&\overset{(a)}{=} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_I^{t-1} \geq |\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\right) \\
&+ \mathbb{P}\left(\hat{\boldsymbol{\theta}}_I^{t-1} \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\right)\mathbb{P}(\eta_I^{t-1} \neq \mathbf{0}). \\
&= \mathbb{P}\left(\hat{\boldsymbol{\theta}}_I^{t-1} \geq |\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\right) + \mathbb{P}(\eta_I^{t-1} \neq 0).
\end{aligned} \tag{2.7.10}$$

In the above, step (a) follows when $\eta_I^{t-1} = [\text{prox}_{J_{\boldsymbol{\theta}^{t-1}}}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})]_I = \mathbf{0}$, since we must have $|\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}|_I \leq \hat{\boldsymbol{\theta}}_I^{t-1}$, and when $\eta_I^{t-1} \neq \mathbf{0}$, by Fact 2.5.2 and Fact 2.5.3, we know that $|\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1}) - (\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})|_I \in \mathcal{P}([\hat{\Pi}_{\eta^{t-1}}^{-1}(\boldsymbol{\theta}^{t-1})]_I)$.

It obvious that one can make the first probability arbitrarily small by bringing c_2 to 0. To see this, say $1 \in I$ and notice that $\mathcal{P}([\hat{\Pi}_{\eta^{t-1}}^{-1}(\boldsymbol{\theta}^{t-1})]_I)$ always has Lebesgue measure 0 because it is a subset of the hyperplane $\{\boldsymbol{x} \in \mathbb{R}^p : \sum_{j\in I} x_j = \sum_{j \in I}\theta_j^{t-1}\}$.

On the other hand, notice that

$$\begin{aligned}
&\sum_I \mathbb{P}([\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})]_I \neq \mathbf{0}) = \sum_I \mathbb{E}\{\mathbb{I}([\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})]_I \neq \mathbf{0})\} \\
&= \mathbb{E}_{\boldsymbol{Z}_{t-1}}\|\eta^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|_0^*,
\end{aligned}$$

and that η^{t-1} is asymptotically separable by Lemma 2.3.3. Define $h^{t-1}(x) = h(x; B + \tau_{t-1}Z, \Theta^{t-1})$ with Θ^{t-1} being the distribution to which the empirical distribution of $\boldsymbol{\theta}^{t-1}$ converges, and also define

$$\boldsymbol{W}_{t-1} := \left\{x \;\middle|\; h^{t-1}(x) \neq 0 \text{ and } m\{z \mid |h^{t-1}(z)| = |h^{t-1}(x)|\} = 0\right\}$$

similarly to (2.2.12), where m is the Lebesgue measure. Then,

$$\begin{aligned}
&\lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}_{t-1}}\|\eta^{t-1}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})\|_0^* = \lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}_{t-1}}\|h^{t-1}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})\|_0^* \\
&= \lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}_{t-1}}\sum_{i=1}^{p}\mathbb{I}\left\{(\beta_i-\tau_{t-1}Z_{t-1,i})\in \boldsymbol{W}_{t-1}\right\} \\
&= \lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}_{t-1},\boldsymbol{B}}\|\eta^{t-1}(\boldsymbol{B}-\tau_{t-1}\boldsymbol{Z}_{t-1})\|_0^*,
\end{aligned}$$

where the last equality holds by Lemma 2.3.2.

Then (2.2.10) gives this term is smaller than δ for large t. Hence, by (2.7.9) and (2.7.10),

$$\begin{aligned}
&\operatorname*{plim}_p \frac{1}{p}|s_t(c_2)| = \lim_p \frac{1}{p}\sum_I \mathbb{P}\left(\hat{\boldsymbol{\theta}}_I^{t-1} \geq |\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\right) \\
&+ \lim_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}_{t-1},\boldsymbol{B}}\|\eta^{t-1}(\boldsymbol{B}-\tau_{t-1}\boldsymbol{Z}_{t-1})\|_0^*,
\end{aligned}$$

Therefore, for some $c > 0$, choose $c_2 \in (0,1)$ such that the first term on the right side of the above is arbitrarily small along with $t_{\min,1}(c)$ such that the second term is arbitrarily close to δ, meaning

$$\lim_p \mathbb{P}\left(\frac{1}{p}|s_t(c_2)| < \delta - c\right) = 1,$$

for all fixed t larger than some $t_{\min,1}(c)$.

For any $t \geq t_{\min,1}(c)$ we can apply Lemma 2.7.5 for some $a_1(c)$, $a_2(c,t)$. Note this doesn't immediately give the result we use since the lower bound, a_2, depends on t. To get around this we additionally appeal to Lemma 2.7.6 that tells us after some time t_*, the supports of the AMP estimates don't change appreciably. Now we fix $c > 0$ and consequently $a_1 = a_1(c)$ is fixed. Define $t_{\min} = \max(t_{\min,1}, t_*(a_1/2, c_2))$

with $t_*(\,\cdot\,)$ defined as in Lemma 2.7.6 and let $a_2 = a_2(c, t_{\min})$. Then, by Lemma 2.7.5 and the fact that $a_2(c,t)$ is non-increasing in t,

$$\min\Big\{\sigma_{\min}(\boldsymbol{X}_{S_{t_{\min}}(c_2)\cup S'}) \;:\; \; S' \subseteq [p]\,,\; |s'| \leq a_1 p\Big\} \geq a_2.$$

In addition, by Lemma 2.7.6, $|S_t(c_2) \setminus S_{t_{\min}}(c_2)| \leq pa_1/2$. Both events hold eventually almost surely as $p \to \infty$. The proof completes with $c_3 = a_1(c)/2$ and $c_4 = a_2(c, t_{\min})$, fixed with respect to t. □

2.8 Discussion and Future Work

This work develops and analyzes the dynamics of an approximate message passing (AMP) algorithm with the purpose of solving the SLOPE convex optimization procedure for high-dimensional linear regression. By employing recent theoretical analysis of AMP when the non-linearities used in the algorithm are non-separable [BMN20], as is the case for the SLOPE problem, we provide rigorous proof that the proposed AMP algorithm finds the SLOPE solution asymptotically. Moreover empirical evidence suggests that the AMP estimate is already very close to the SLOPE solution even in few iterations. By leveraging our analysis showing AMP provably solves SLOPE, we provide an exact asymptotic characterization of the ℓ_2 risk of the SLOPE estimator from the underlying truth and insight into other statistical properties of the SLOPE estimator. Though this asymptotic analysis of the SLOPE solution has been demonstrated in other recent work [HL19a] using a different proof strategy, we believe that our AMP-based approach offers a more concrete and algorithmic understanding of the finite-sample behavior of the SLOPE estimator.

A limitation of this approach is that the theory assumes an i.i.d. Gaussian measurement matrix, and moreover, the AMP algorithm can become unstable when the measurement matrix is far from i.i.d., creating the need for heuristic techniques to provide convergence in applications where the measurement matrix is generated by nature (i.e., a real-world experiment or observational study). Additionally, the asymptotical regime studied here, $n/p \to \delta \in (0, \infty)$, requires that the number of columns of the measurement matrix p grow at the same rate as the number of rows n. It is of practical interest to extend the results to high-dimensional settings where p grows faster than n.

2.9 Appendix

2.9.1 State Evolution Analysis

We first prove Theorem 1 and then provide a proof of Proposition 2.2.6.

Proving Theorem 1

Proof of Theorem 1. To begin with, we prove that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ defined in (2.2.8) is concave with respect to τ^2. The proof follows along the same lines as the proof of

[BM11c, Proposition 1.3], however, whereas the proof of [BM11c, Proposition 1.3] proceeds by explicitly expressing the first derivative of the corresponding function F, and then differentiating on the explicit form to get the second derivative, in SLOPE case, because of the averaging that occurs within the proximal operation, it is extremely difficult to similarly derive an explicit form. To work around this, we keep all differentiation implicit. First,

$$\begin{aligned}
\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) &= \frac{\partial}{\partial \tau^2}\Big[\sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z}) - \boldsymbol{B}\|^2\Big] \\
&\overset{(a)}{=} \frac{1}{\delta}\mathbb{E}\Big\{\frac{\partial}{\partial \tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z}) - \boldsymbol{B}\|^2\Big\} \\
&= \frac{2}{\delta p}\sum_{i=1}^{p}\mathbb{E}\Big\{\Big([\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i - B_i\Big)\frac{\partial}{\partial \tau^2}[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i\Big\}.
\end{aligned} \tag{2.9.1}$$

We note that the interchange between the derivative (a limit) and the expectation in step (a) of the above holds due to a dominated convergence argument that relies on the following lemma. First we introduce a bit of notation that will be used throughout the proof. Define an equivalence classes I_i for each index $i = \{1, 2, \ldots, p\}$, defined as

$$I_i := \{j : |[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_j| = |[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i|\}.$$

For any $j \in I_i$, with the above definition, $I_j = I_i$. In general, we use I, without any specific index, to represent an entire equivalence class and let I indicate the collection of unique equivalence classes.

Lemma 2.9.1.

$$\Big|\frac{\partial}{\partial \tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z}) - \boldsymbol{B}\|^2\Big| \leq \frac{1}{p}\sum_{I \in \mathsf{I}}\frac{1}{|I|}\Big(\sum_{i \in I}|\text{sign}(B_i + \tau Z_i)Z_i - \alpha_i|\Big)^2. \tag{2.9.2}$$

Lemma 2.9.1 will be proved below, after we solve $\frac{\partial}{\partial \tau^2}[\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i$.

Now we describe how the bound in Lemma 2.9.1 can be used to produce the dominated convergence result needed in step (a) of (2.9.1). First note,

$$
\begin{aligned}
&\frac{1}{p}\mathbb{E}\Big\{\sum_{I\in\mathsf{I}}\frac{1}{|I|}\Big(\sum_{i\in I}|\mathrm{sign}(B_i+\tau Z_i)Z_i-\alpha_i|\Big)^2\Big\}\\
&\leq \frac{1}{p}\mathbb{E}\Big\{\sum_{I\in\mathsf{I}}\sum_{i\in I}\Big(|\mathrm{sign}(B_i+\tau Z_i)Z_i-\alpha_i|\Big)^2\Big\}\\
&\leq \frac{2}{p}\mathbb{E}\Big\{\sum_{I\in\mathsf{I}}\sum_{i\in I}(Z_i^2+\alpha_i^2)\Big\}=\frac{2}{p}\mathbb{E}\Big\{\sum_{i\in[p]}(Z_i^2+\alpha_i^2)\Big\}=2+2\|\boldsymbol{\alpha}\|^2/p<\infty
\end{aligned}
$$

The first and second inequalities follow from $(\sum_{i=1}^n x_i)^2 \leq n\sum_i x_i^2$. The last inequality comes from entries of $\boldsymbol{\alpha}$ being finite and then $\|\boldsymbol{\alpha}\|^2/p \leq \max_i \alpha_i^2 < \infty$. Therefore we can invoke the dominated convergence theorem that allows the exchange of the derivative and expectation in step (a) of (2.9.1).

Now we want to further simplify (2.9.1). For each $1 \leq i \leq p$, we would like to study $\frac{\partial}{\partial \tau^2}[\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i$. We first note that the mapping $\tau^2 \mapsto [\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i$ can be considered as $f(g(\tau^2))$, where $g:\mathbb{R}\to\mathbb{R}^{2p}$ is defined as $y\mapsto g(y) := (\boldsymbol{B}+\boldsymbol{Z}\sqrt{y}, \boldsymbol{\alpha}\sqrt{y})$ and $f:\mathbb{R}^{2p}\to\mathbb{R}$ is defined as $(\boldsymbol{a},\boldsymbol{b})\mapsto f(\boldsymbol{a},\boldsymbol{b}) := [\mathrm{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})]_i$. Hence,

$$
\begin{aligned}
\frac{\partial}{\partial \tau^2}[\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i &= \boldsymbol{J}_{f\circ g}(\tau^2) \overset{(a)}{=} \boldsymbol{J}_f(g(\tau^2))\boldsymbol{J}_g(\tau^2)\\
&=\Big[\nabla_{\boldsymbol{a}} f(g(\tau^2)), \nabla_{\boldsymbol{b}} f(g(\tau^2))\Big]\Big[\frac{\boldsymbol{Z}}{2\tau},\frac{\boldsymbol{\alpha}}{2\tau}\Big]^\top,
\end{aligned}
\tag{2.9.3}
$$

where $\boldsymbol{J}_h \in \mathbb{R}^{m\times n}$ is the Jacobian matrix of a function $h:\mathbb{R}^n\to\mathbb{R}^m$ and step (a) follows by the chain rule. We denote the proximal operator using a function

$\eta : \mathbb{R}^{2p} \to \mathbb{R}^p$ as $\eta(\boldsymbol{a}, \boldsymbol{b}) := \text{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})$ and consider the partial derivatives of η with respect to its first and second arguments. Denote

$$\begin{aligned} \partial_1\eta(\boldsymbol{a}, \boldsymbol{b}) &:= \text{diag}\Big[\frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}, \ldots, \frac{\partial}{\partial a_p}\Big]\eta(\boldsymbol{a}, \boldsymbol{b}), \\ \text{and} \quad \partial_2\eta(\boldsymbol{a}, \boldsymbol{b}) &:= \text{diag}\Big[\frac{\partial}{\partial b_1}, \frac{\partial}{\partial b_2}, \ldots, \frac{\partial}{\partial b_p}\Big]\eta(\boldsymbol{a}, \boldsymbol{b}). \end{aligned} \tag{2.9.4}$$

Recall that the derivatives computed in $\partial_1\eta(\boldsymbol{a}, \boldsymbol{b})$ are defined in (2.2.2), and by anti-symmetry between two arguments, $\frac{d}{db_j}[\eta(\boldsymbol{a}, \boldsymbol{b})]_i = -\operatorname{sign}([\eta(\boldsymbol{a}, \boldsymbol{b})]_j)\frac{d}{da_j}[\eta(\boldsymbol{a}, \boldsymbol{b})]_i$. Then using the result of (2.2.2):

$$\frac{\partial[\text{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i}{\partial v_j} = \frac{\partial[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_i}{\partial v_j} = \frac{\mathbb{I}\{|[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_i| = |[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_j|\}\operatorname{sign}([\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_i[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_j)}{\#\{1 \leq k \leq p : |[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_k| = |[\eta(\boldsymbol{v}, \boldsymbol{\lambda})]_i|\}}$$

we have

$$\begin{aligned} &\frac{d}{da_j}f(\boldsymbol{a}, \boldsymbol{b}) = \frac{d}{da_j}[\eta(\boldsymbol{a}, \boldsymbol{b})]_i = \mathbb{I}\{|[\eta(\boldsymbol{a}, \boldsymbol{b})]_i| \\ &= |[\eta(\boldsymbol{a}, \boldsymbol{b})]_j|\}\operatorname{sign}([\eta(\boldsymbol{a}, \boldsymbol{b})]_i[\eta(\boldsymbol{a}, \boldsymbol{b})]_j)[\partial_1\eta(\boldsymbol{a}, \boldsymbol{b})]_i, \end{aligned} \tag{2.9.5}$$

and similarly,

$$\begin{aligned} &\frac{d}{db_j}f(\boldsymbol{a}, \boldsymbol{b}) = \frac{d}{db_j}[\eta(\boldsymbol{a}, \boldsymbol{b})]_i = -\mathbb{I}\Big\{|[\eta(\boldsymbol{a}, \boldsymbol{b})]_i| \\ &= |[\eta(\boldsymbol{a}, \boldsymbol{b})]_j|\Big\}\operatorname{sign}\Big([\eta(\boldsymbol{a}, \boldsymbol{b})]_i\Big)\Big[\partial_1\eta(\boldsymbol{a}, \boldsymbol{b})\Big]_i. \end{aligned}$$

Now plugging the above into (2.9.3), we have

$$\begin{aligned} &\frac{\partial}{\partial \tau^2}[\text{prox}_{J_{\alpha\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i \\ &= \frac{1}{2\tau}\Big[\partial_1\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)\Big]_i\operatorname{sign}\Big([\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)]_i\Big) \\ &\sum_{j \in I_i}\Big(\operatorname{sign}([\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)]_j)Z_j - \alpha_j\Big) \end{aligned} \tag{2.9.6}$$

In what follows, we drop the explicit statement of the $\eta(\cdot,\cdot)$ input to save space, writing η_i to mean $[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i$ or $[\partial_1\eta]_i$ to mean $[\partial_1\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i$ for example. Using (2.9.6) in (2.9.1),

$$\begin{aligned}
\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2,\boldsymbol{\alpha}\tau) &= \frac{1}{\delta p\tau}\sum_{i=1}^{p}\sum_{j\in I_i}\mathbb{E}\Big\{(\eta_i-B_i)\,[\partial_1\eta]_i\,\mathrm{sign}(\eta_i)(\mathrm{sign}(\eta_j)Z_j-\alpha_j)\Big\}\\
&= \frac{1}{\delta p}\sum_{i=1}^{p}\sum_{j\in I_i}\mathbb{E}\Big\{([\partial_1\eta]_i)^2+(\eta_i-B_i)[\partial_1^2\eta]_i\Big\} \qquad (2.9.7)\\
&- \frac{1}{\delta p\tau}\sum_{i=1}^{p}\sum_{j\in I_i}\mathbb{E}\Big\{(\eta_i-B_i)\,[\partial_1\eta]_i\,\mathrm{sign}(\eta_i)\alpha_j\Big\}.
\end{aligned}$$

where the second equality follows by Stein's lemma for a fixed i and $j\in I_i$, namely, for standard Gaussian Z we have $\mathbb{E}\{f(Z)Z\}=\mathbb{E}\{f'(Z)\}$ and therefore,

$$\begin{aligned}
&\frac{1}{\tau}\mathbb{E}\Big\{[\partial_1\eta]_i\,\mathrm{sign}(\eta_i)(\eta_i-B_i)\,\mathrm{sign}(\eta_j)Z_j\Big\}\\
&= \mathbb{E}\Big\{\,\mathrm{sign}(\eta_i)\,\mathrm{sign}(\eta_j)\Big[(\eta_i-B_i)\frac{d}{da_j}[\partial_1\eta]_i+[\partial_1\eta]_i\frac{d}{da_j}[\eta]_i\Big]\Big\}\\
&= \mathbb{E}\Big\{(\eta_i-B_i)[\partial_1^2\eta]_i+([\partial_1\eta]_i)^2\Big\}.
\end{aligned}$$

where the last step uses the definition of $\frac{d}{da_j}[\eta(\boldsymbol{a},\boldsymbol{b})]_i$ given in (2.9.5) and the fact that $\frac{d}{da_j}[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})]_i=\mathrm{sign}(\eta_i)\,\mathrm{sign}(\eta_j)[\partial_1^2\eta(\boldsymbol{a},\boldsymbol{b})]_i$.

Therefore, simplifying (2.9.7), we have shown

$$\begin{aligned}
&(\delta p\tau)\times\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)\\
&=\sum_{i=1}^{p}\mathbb{E}\Big\{\tau|I_i|\Big([\partial_1\eta]_i^2+(\eta_i-B_i)[\partial_1^2\eta]_i\Big)-[\partial_1\eta]_i\,\mathrm{sign}(\eta_i)(\eta_i-B_i)\sum_{j\in I_i}\alpha_j\Big\}.
\end{aligned} \qquad (2.9.8)$$

We now have the tools to prove Lemma 2.9.1.

Proof of Lemma 2.9.1. First,

$$\begin{aligned}&\frac{\partial}{\partial \tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2\\&=\frac{2}{p}\sum_{i=1}^{p}\Big([\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i-B_i\Big)\frac{\partial}{\partial \tau^2}[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i.\end{aligned}$$

As in the work above, we denote the proximal operator using a function $\eta:\mathbb{R}^{2p}\to\mathbb{R}^p$ as $\eta(\boldsymbol{a},\boldsymbol{b}):=\text{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})$. Now from (2.9.6), denoting $I_i:=\{j:|[\eta(\boldsymbol{a},\boldsymbol{b})]_j|=|[\eta(\boldsymbol{a},\boldsymbol{b})]_i|\}$, again dropping the explicit statement of the $\eta(\cdot,\cdot)$ input to save space,

$$\frac{\partial}{\partial \tau^2}[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i=\frac{1}{2\tau}[\partial_1\eta]_i\,\text{sign}(\eta_i)\sum_{j\in I_i}(\text{sign}(\eta_j)Z_j-\alpha_j).$$

Therefore,

$$\begin{aligned}&\Big|\frac{\partial}{\partial \tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2\Big|\\&=\frac{1}{\tau p}\Big|\sum_{i=1}^{p}(\eta_i-B_i)\,[\partial_1\eta]_i\,\text{sign}(\eta_i)\sum_{j\in I_i}(\text{sign}(\eta_j)Z_j-\alpha_j)\Big|.\end{aligned}$$

Since the averaging operation reduces the dot product (meaning informally that for a vector $\boldsymbol{v}\in\mathbb{R}^p$, $(\text{mean}(\boldsymbol{v}),...,\text{mean}(\boldsymbol{v}))\cdot\boldsymbol{v}\leq\|\boldsymbol{v}\|^2$), we have for any $i\in\{1,2,\ldots,p\}$ that $[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i-B_i$ can be replaced with $B_i+\tau Z_i-\text{sign}(\eta_i)\alpha_i\tau-B_i$. Using this in the above,

$$\begin{aligned}&\Big|\frac{\partial}{\partial \tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2\Big|\\&\leq\frac{1}{p}\Big|\sum_{i=1}^{p}\sum_{j\in I_i}(Z_i-\text{sign}(\eta_i)\alpha_i)\,[\partial_1\eta]_i\,\text{sign}(\eta_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\Big| \qquad (2.9.9)\\&=\frac{1}{p}\Big|\sum_{i=1}^{p}\sum_{j\in I_i}(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\,[\partial_1\eta]_i\Big|.\end{aligned}$$

Next, using that $0 \leq |[\partial_1 \eta]_i| \leq 1/|I_i|$,

$$\Big|\sum_{i=1}^{p}\sum_{j\in I_i}(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\,[\partial_1\eta]_i\Big|$$
$$\leq \sum_{i=1}^{p}\frac{1}{|I_i|}\sum_{j\in I_i}\Big|(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\Big|.$$

Finally we make the following observation. Any equivalence class I_i is a collection of indices $j \in \{1, 2, \ldots, p\}$ such that $|[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_j| = |[\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})]_i|$, so for any $j \in I_i$, it follows $I_j = I_i$. Recall, I indicates the collection of unique equivalence classes, and we have

$$\sum_{i=1}^{p}\frac{1}{|I_i|}\sum_{j\in I_i}\Big|(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\Big|$$
$$=\sum_{I\in\mathsf{I}}\frac{1}{|I|}\sum_{i,j\in I}\Big|(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\Big|.$$

Now plugging back into (2.9.9),

$$\Big|\frac{\partial}{\partial\tau^2}\frac{1}{p}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2\Big|$$
$$\leq \frac{1}{p}\sum_{I\in\mathsf{I}}\frac{1}{|I|}\sum_{i,j\in I}\Big|(\text{sign}(\eta_i)Z_i-\alpha_i)(\text{sign}(\eta_j)Z_j-\alpha_j)\Big|$$
$$=\frac{1}{p}\sum_{I\in\mathsf{I}}\frac{1}{|I|}\Big(\sum_{j\in I}|\text{sign}(\eta_j)Z_j-\alpha_j|\Big)^2.$$

□

Now considering (2.9.8), for simplicity in our future calculations, we suppress $|I_i|$ to 1 without loss of generality. To see this, recall that $I_i := \{j : |[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_j| = |[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i|\}$ and note that when $|[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_j$ equals $|[\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i$, the terms will remain equal after small changes in τ. Therefore $|I_i|$ is treated as a constant in the derivative and since all operations below preserves linearity, it can safely be assumed to be equal to 1. Note that similarly, $\sum_{j\in I_i}\alpha_j$, will pass through future calculations as a constant. Therefore (2.9.8) becomes

$$\begin{aligned}&(\delta p\tau)\times\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)\\&=\sum_{i=1}^{p}\Big[\mathbb{E}\Big\{\tau([\partial_1\eta]_i)^2+\tau(\eta_i-B_i)[\partial_1^2\eta]_i-\alpha_i\operatorname{sign}(\eta_i)(\eta_i-B_i)[\partial_1\eta]_i\Big\}\Big].\end{aligned}\tag{2.9.10}$$

In what follows we will need to take care with the points $(\boldsymbol{x},\boldsymbol{y})$ such that $[\partial_1^2\eta(\boldsymbol{x},\boldsymbol{y})]_i$ is not equal to 0. We refer to such points as 'kink' points, since these are points where the partial derivative jumps (and the second partial gradient acts like Dirac delta function $\delta(x)$), or in other words the points where the two (sorted, averaged) arguments in η are equal to each other. Informally, define a 'kink' point as an index where the sorted vector $\boldsymbol{x}$ matches the corresponding threshold $\boldsymbol{y}$ exactly. In LASSO, for example, the correspond to the 'kinks' of the soft-thresholding function. We have

$$[\partial_1^2\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i=\delta(B_i+\tau Z_i-\alpha_i\tau)-\delta(B_i+\tau Z_i+\alpha_i\tau)\tag{2.9.11}$$

and

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z},\boldsymbol{B}}\Big\{([\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i-B_i)[\partial_1^2\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i\Big\}\\&=-\mathbb{E}_{\boldsymbol{B}}\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{B}}\Big\{B_i\Big[\delta(B_i+\tau Z_i-\alpha_i\tau)-\delta(B_i+\tau Z_i+\alpha_i\tau)\Big]\Big\}\\&=-\frac{1}{\tau}\mathbb{E}_{\boldsymbol{B}}\Big\{B_i\Big[\phi(\alpha_i-\frac{1}{\tau}B_i)-\phi(-\alpha_i-\frac{1}{\tau}B_i)\Big]\Big\}.\end{aligned}\tag{2.9.12}$$

Therefore, denoting $\odot$ as elementwise multiplication of vectors, by (2.9.10) and

(2.9.12),

$$\begin{aligned}&(\delta p\tau) \times \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)\\ =&\tau\mathbb{E}||\partial_1\eta||^2 - \mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha} - \frac{1}{\tau}\boldsymbol{B}) - \phi(-\boldsymbol{\alpha} - \frac{1}{\tau}\boldsymbol{B})\Big]\Big\}\\ &- \mathbb{E}\Big\{\Big[\boldsymbol{\alpha} \odot \mathrm{sign}(\eta) \odot (\eta - \boldsymbol{B})\Big]^\top \partial_1\eta\Big\}.\end{aligned} \tag{2.9.13}$$

Now we have shown the first derivative, so we consider the second derivative to prove concavity.

Notice, however, that in order to prove concavity of $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ it suffices to show $\frac{\partial}{\partial\tau}[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2, \boldsymbol{\alpha}\tau)] \leq 0$ because $\frac{\partial}{\partial\tau^2}(\frac{\partial \mathsf{F}}{\partial\tau^2}) = \frac{\partial\tau}{\partial\tau^2}[\frac{\partial}{\partial\tau}(\frac{\partial \mathsf{F}}{\partial\tau^2})] = \frac{1}{2\tau}[\frac{\partial}{\partial\tau}(\frac{\partial \mathsf{F}}{\partial\tau^2})]$.

We now show $\frac{\partial}{\partial\tau}[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2, \boldsymbol{\alpha}\tau)] \leq 0$. First,

$$\begin{aligned}&(\delta p) \times \frac{\partial}{\partial\tau}\Big[\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)\Big]\\ =& \frac{\partial}{\partial\tau}\mathbb{E}||\partial_1\eta||^2 - \frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha} - \frac{1}{\tau}\boldsymbol{B}) - \phi(-\boldsymbol{\alpha} - \frac{1}{\tau}\boldsymbol{B})\Big]\Big\}\\ &- \frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}\Big\{\Big[\boldsymbol{\alpha} \odot \mathrm{sign}(\eta) \odot (\eta - \boldsymbol{B})\Big]^\top \partial_1\eta\Big\}.\end{aligned} \tag{2.9.14}$$

To show that (2.9.14) is ≤ 0, we find simplified representations of the three terms on the right side. This requires the same techniques as were used to find the first derivative above and so aren't given in full detail.

The first term on the right side of (2.9.14) can be simplified to the following:

$$\frac{\partial}{\partial\tau}\mathbb{E}||\partial_1\eta||^2 = -\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha} - \frac{1}{\tau}\boldsymbol{B})) - \phi(\boldsymbol{\alpha} + \frac{1}{\tau}\boldsymbol{B}))\Big]\Big\}. \tag{2.9.15}$$

Doing so requires smart uses of the chain rule, a dominated convergence argument, the partials in (2.9.6), and special care for the 'kink' points as discussed above. Similarly, using (2.9.12), one can easily show for the third term on the right side of

(2.9.14),

$$\begin{aligned}&\frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}\Big\{\big[\boldsymbol{\alpha}\odot\operatorname{sign}(\eta)\odot(\eta-\boldsymbol{B})\big]^\top\partial_1\eta\Big\}\\&\geq\frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{[\boldsymbol{\alpha}\odot\boldsymbol{B}^2]^\top[\phi(\boldsymbol{\alpha}+\frac{1}{\tau}\boldsymbol{B})+\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\Big\}.\end{aligned}\tag{2.9.16}$$

Finally, using $\phi'(u)=-u\phi(u)$ and a dominated convergence argument, the second term on the right side of (2.9.14) equals

$$\begin{aligned}&-\frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})\Big]\Big\}\\&=\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})\Big]\Big\}\\&-\frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{(\boldsymbol{B}^2)^\top\Big[(\frac{1}{\tau}\boldsymbol{B}-\boldsymbol{\alpha})\odot\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-(\boldsymbol{\alpha}+\frac{1}{\tau}\boldsymbol{B})\odot\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})\Big]\Big\}.\end{aligned}\tag{2.9.17}$$

Now we plug (2.9.15),(2.9.16), and (2.9.17) back into (2.9.14) to show that $\frac{\partial}{\partial\tau}[\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)]\leq 0$.

$$\begin{aligned}&(\delta p)\times\frac{\partial}{\partial\tau}\Big[\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)\Big]\\&\leq-\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau))-\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau))\Big]\Big\}\\&+\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-\phi(-\boldsymbol{\alpha}-\boldsymbol{B}/\tau)\Big]\Big\}\\&-\frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{(\boldsymbol{B}^2)^\top\Big[(\boldsymbol{B}/\tau-\boldsymbol{\alpha})\odot\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)\odot\phi(-\boldsymbol{\alpha}-\boldsymbol{B}/\tau)\Big]\Big\}\\&-\frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{[\boldsymbol{\alpha}\odot\boldsymbol{B}^2]^\top[\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)+\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)]\Big\}\\&=-\frac{1}{\tau^4}\mathbb{E}_{\boldsymbol{B}}\Big\{[\boldsymbol{B}^3]^\top\Big[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)\Big]\Big\}.\end{aligned}\tag{2.9.18}$$

We justify non-positivity of (2.9.18) by showing that the elementwise term inside the expectation is less than or equal to 0. First assume $B_i\geq 0$, then $\alpha_i-B_i/\tau\leq\alpha_i+B_i/\tau$ and $\phi(\alpha_i-B_i/\tau)\geq\phi(\alpha_i+B_i/\tau)$. The other case $B_i\leq 0$ follows similarly.

Now (2.9.18), implies $\frac{\partial}{\partial \tau}\Big[\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)\Big] \leq 0$ and therefore, we have shown that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ defined in (2.2.8), is concave with respect to τ^2.

Next we show that $\tau^2 \mapsto \mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ is strictly increasing. To do so, it is sufficient to show that $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is positive as $\tau \to \infty$ because the concavity implies that $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is non-increasing. Define $f(\boldsymbol{\alpha}) := \delta \lim_{\tau \to \infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$. First recall that $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is given in (2.9.8). In particular,

$$\begin{aligned}
&\delta \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) \\
&= \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{ |I_i| \Big([\partial_1 \eta]_i^2 + (\eta_i - B_i)[\partial_1^2 \eta]_i \Big) - \frac{1}{\tau}[\partial_1 \eta]_i \operatorname{sign}(\eta_i)(\eta_i - B_i) \sum_{j \in I_i} \alpha_j \Big\},
\end{aligned} \tag{2.9.19}$$

Then taking $\tau \to \infty$ in the above, it is easy to see that $f(\boldsymbol{\alpha})$ is equivalent to setting $\boldsymbol{B} = \mathbf{0}$ in $\eta(\boldsymbol{B} + \tau \boldsymbol{Z}, \boldsymbol{\alpha}\tau)$ and using that $\eta(\tau \boldsymbol{Z}, \boldsymbol{\alpha}\tau) = \tau \eta(\boldsymbol{Z}, \boldsymbol{\alpha})$ (implying that $\partial_1 \eta(\tau \boldsymbol{Z}, \boldsymbol{\alpha}\tau) = \partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha})$). We note that using a simplification of $[\partial_1^2 \eta]_i$ as in (2.9.11)-(2.9.12), means that this term will go to zero as $\tau \to \infty$. Therefore, using $\operatorname{sign}(\eta(\boldsymbol{Z}, \boldsymbol{\alpha})) \odot \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = |\eta(\boldsymbol{Z}, \boldsymbol{\alpha})|$,

$$\begin{aligned}
f(\boldsymbol{\alpha}) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{ &[D(\eta(\boldsymbol{Z}, \boldsymbol{\alpha}))]_i ([\partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i)^2 - [\partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i| \\
&\sum_{j : |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_j| = |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i|} \alpha_j \Big\}.
\end{aligned}$$

In the above we have used the following definition: for a vector $\boldsymbol{v} \in \mathbb{R}^p$, define $\boldsymbol{D}$ elementwise as $[\boldsymbol{D}(\boldsymbol{v})]_i := \#\{j : |v_j| = |v_i|\} = |I_i|$ if $v_i \neq 0$ and ∞ otherwise. Using that $\partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \frac{1}{\boldsymbol{D}(\eta(\boldsymbol{Z}, \boldsymbol{\alpha}))}$,

$$f(\boldsymbol{\alpha}) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{ \Big(1 - |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i| \sum_{j : |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_j| = |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i|} \alpha_j \Big) \frac{1}{[D(\eta(\boldsymbol{Z}, \boldsymbol{\alpha}))]_i} \Big\} \tag{2.9.20}$$

This simplification can be efficiently computed because only $|\eta(\boldsymbol{Z}, \boldsymbol{\alpha})|$ and $\boldsymbol{\alpha}$ need to be memorized.

Now considering (2.9.20), let $\boldsymbol{\alpha} \to \infty$ and note that since $|\boldsymbol{Z}| < \boldsymbol{\alpha}$ almost surely as $\boldsymbol{\alpha} \to \infty$, it follows that $\eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \mathbf{0}$. Therefore $\lim_{\boldsymbol{\alpha} \to \infty} f(\boldsymbol{\alpha}) = 0$. By a very similar argument to the proof of concavity, it is easy to see $f'(\boldsymbol{\alpha}) < 0$, and together these facts imply $f(\boldsymbol{\alpha}) > 0$ for all $\boldsymbol{\alpha}$. The monotonicity of F is now obvious: since F is concave (implying $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is non-increasing) and strictly increasing for τ^2 large enough, it is increasing everywhere. Moreover, the monotonicity of F implies the monotonicity of the sequence $\{\tau_t^2(p)\}_{t \geq 0}$.

Finally we show that there exists a unique τ_* such that $\mathsf{F}(\tau_*^2, \boldsymbol{\alpha}\tau_*) = \tau_*^2$, from which it follows that the monotone sequence $\{\tau_t^2(p)\}_{t \geq 0}$ converges to $\tau_*^2(p)$ as $t \to \infty$. First, by (2.9.20), we know $f(\mathbf{0}) = \mathbb{E}\|\partial_1 \eta(\tau \boldsymbol{Z}, \mathbf{0})\|^2/p = \mathbb{E}\|\mathbf{1}\|^2/p = 1$. This, along with the fact that $f'(\boldsymbol{\alpha}) < 0$, tells us that $0 < f(\boldsymbol{\alpha}) < 1$ for all $\boldsymbol{\alpha}$. Recall the definition of the set $\boldsymbol{A}_{\min}$, namely $\boldsymbol{A}_{\min} := \{\boldsymbol{\alpha} : f(\boldsymbol{\alpha}) = \delta\}$. We know that this set is non-empty since the LASSO case shows $\boldsymbol{\alpha} = (\alpha_{\min}, \cdots, \alpha_{\min})$ belongs to $\boldsymbol{A}_{\min}$ where $\alpha_{\min}$ is the unique non-negative solution of $(1 + \alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \delta/2$. We write $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$ to mean $\boldsymbol{\alpha}$ is larger than at least one element in $\boldsymbol{A}_{\min}$, where we consider one vector $\boldsymbol{v}$ to be larger than another vector $\boldsymbol{u}$ if $v_i \geq u_i$ for all i and $v_j > u_j$ for some j.

To complete the proof, we show that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) > \tau^2$ for small enough τ^2 and $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) < \tau^2$ for large enough τ^2. Therefore, there is at least one τ_* such that

$\mathsf{F}(\tau_*^2, \boldsymbol{\alpha}\tau_*) = \tau_*^2$ since F is continuous in τ. It follows from the concavity of F that the solution is unique and the sequence of iterates $\tau_t^2(p)$ converge to $\tau_*^2(p)$. We first show that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) > \tau^2$ for small enough τ^2. Consider the function $G(\tau^2) := \mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) - \tau^2$. Recalling the definition of $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ in (2.2.8), namely, $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) = \sigma_w^2 + \mathbb{E}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z}) - \boldsymbol{B}\|^2/(\delta p)$, clearly $\mathsf{F}(0, \mathbf{0}) = \sigma_w^2 \geq 0$ and therefore $G(0) = \sigma_w^2 \geq 0$ (with equality only if $\sigma_w^2 = 0$). Now we show that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) < \tau^2$ for large enough τ^2. Since $f(\boldsymbol{\alpha})$ is decreasing in $\boldsymbol{\alpha}$, for $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$, it must be that $f(\boldsymbol{\alpha}) < \delta$. Moreover, $\lim_{\tau\to\infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) = \frac{1}{\delta} f(\boldsymbol{\alpha}) \leq 1$ for $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$. Therefore, $\lim_{\tau\to\infty} \frac{\partial G}{\partial \tau^2}(\tau^2) \leq 0$ meaning G is eventually decreasing (as τ^2 grows) for any $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$. Also, $G(\tau^2)$ is concave and therefore for τ^2 large enough we will have $G(\tau^2) < 0$, in which case $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) < \tau^2$.

Finally, $\left|\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)\right|$ evaluated at at $\tau^2 = \tau_*^2$ is upper bounded by 1 when $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$, as the concavity of F implies that $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is strictly decreasing in τ^2 along with $\lim_{\tau\to\infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) = \frac{1}{\delta} f(\boldsymbol{\alpha}) \leq 1$ when $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$. If this were not the case then there would be multiple fixed points.

□

Proving Proposition 2.2.6

Proof of Proposition 2.2.6. This proof is a generalized result of [BM11c, Proposition 1.4] (originally proved in [DMM11]) and [BM11c, Corollary 1.7]. Here we fixed p and denote $\tau(p)$ as τ.

Recall in the proof of Theorem 1 we have shown the following facts: **(A)** $0 < \lim_{\tau^2\to\infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) < 1$; **(B)** $\tau^2 \mapsto \mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ is concave; **(C)** $\tau^2 \mapsto \mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ is strictly increasing; and **(D)** $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ evaluated at $\tau = \tau_*$, which we denote $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau_*^2, \boldsymbol{\alpha}\tau_*)$ is such that $0 < \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau_*^2, \boldsymbol{\alpha}\tau_*) < 1$.

First we claim $\boldsymbol{\alpha} \mapsto \tau_*^2(\boldsymbol{\alpha})$ is continuously differentiable on $\mathbb{R}_+^p$. This follows from the implicit function theorem on function $G(\boldsymbol{\alpha}, \tau^2) := \tau^2 - \mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ and from Fact **(D)**: G is continuously differentiable and $0 < \frac{\partial G}{\partial \tau^2} < 1$. Hence τ^2 can be written as $\tau^2(\boldsymbol{\alpha})$ which is continuously differentiable. Defining $g(\boldsymbol{\alpha}, \tau^2) := \boldsymbol{\alpha}\tau\Big[1 - \frac{1}{n}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})\|_0^*\Big]$, notice that $\boldsymbol{\lambda}(\boldsymbol{\alpha}) = g(\boldsymbol{\alpha}, \tau_*^2(\boldsymbol{\alpha}))$. Clearly g is continuously differentiable in $\boldsymbol{\alpha}$ and so is $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$.

In the next step, we consider $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}(\delta)$ such that $\boldsymbol{\alpha} \to \boldsymbol{a}_{\min}$ for some $\boldsymbol{a}_{\min} \in \boldsymbol{A}_{\min}(\delta)$ (denote as $\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)$). We claim $\tau_*^2(\boldsymbol{\alpha}) \to +\infty$ as $\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)$. Recall, $f(\boldsymbol{\alpha}) := \delta \lim_{\tau\to\infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ (cf. Theorem 1). Then by concavity of $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ in τ,

$$\begin{aligned}
&\tau_*^2 = \mathsf{F}(\tau_*^2, \boldsymbol{\alpha}\tau_*) \geq \mathsf{F}(0, \mathbf{0}) + \tau_*^2 \lim_{\tau^2\to\infty} \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) = \mathsf{F}(0, \mathbf{0}) + \frac{1}{\delta}\tau_*^2 f(\boldsymbol{\alpha}) \\
&\Rightarrow \quad \tau_*^2 \geq \frac{\mathsf{F}(0, \mathbf{0})}{1 - f(\boldsymbol{\alpha})/\delta}
\end{aligned}$$

Recall $\mathsf{F}(0, \mathbf{0}) = \sigma_w^2$ and $f(\boldsymbol{a}_{\min}) = \delta$ for any $\boldsymbol{a}_{\min} \in \boldsymbol{A}_{\min}(\delta)$. Hence $\tau_*^2(\boldsymbol{\alpha}) \to +\infty$ as $\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)$.

Define $\ell(\boldsymbol{\alpha}) := 1 - \frac{1}{n}\mathbb{E}\| \mathrm{prox}_{J_{\boldsymbol{\alpha}\tau_*}}(\boldsymbol{B} + \tau_*\boldsymbol{Z})\|_0^*$. Then when $\tau_*^2(\boldsymbol{\alpha}) \to +\infty$ as $\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)$,

$$\ell_* := \lim_{\boldsymbol{\alpha}\to\boldsymbol{a}_{\min}} \ell(\boldsymbol{\alpha}) = \lim_{\boldsymbol{\alpha}\to\boldsymbol{a}_{\min}} \left(1 - \frac{1}{n}\mathbb{E}\| \mathrm{prox}_{J_{\boldsymbol{\alpha}\tau_*}}(\tau_*\boldsymbol{Z})\|_0^*\right) = 1 - \frac{1}{n}\mathbb{E}\| \mathrm{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z})\|_0^*.$$

We claim that $\ell_* < 0$. Using the definition of the vector $\boldsymbol{D}$ and the set $\boldsymbol{A}_{\min}(\delta)$ in (2.2.7),

$$\ell_* = 1 - \frac{1}{n}\mathbb{E}\|\operatorname{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z})\|_0^* = 1 - \frac{1}{\delta}\mathbb{E}\Big\langle \frac{1}{\boldsymbol{D}(\operatorname{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z}))}\Big\rangle$$
$$< 1 - \frac{1}{\delta p}\sum_i \mathbb{E}\Big\{\frac{1}{[\boldsymbol{D}(\operatorname{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z}))]_i}\Big(1 - \sum_{j\in I_i}[\boldsymbol{a}_{\min}]_j \cdot |[\operatorname{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z})]_i|\Big)\Big\} = 0,$$

where (writing $\boldsymbol{\eta}$ to mean $\operatorname{prox}_{J_{\boldsymbol{a}_{\min}}}(\boldsymbol{Z})$ and $\boldsymbol{D}$ to mean $\boldsymbol{D}(\eta)$) the inequality in the above uses the fact that

$$\frac{1}{\boldsymbol{D}_i} - \frac{1}{\boldsymbol{D}_i}\Big(1 - \sum_{j\in I_i}[\boldsymbol{\alpha}_{\min}]_j|\boldsymbol{\eta}_i|\Big) = \frac{1}{\boldsymbol{D}_i}\sum_{j\in I_i}[\boldsymbol{\alpha}_{\min}]_j|\boldsymbol{\eta}_i| \geq 0.$$

Notice in the above, the equality only holds when $\boldsymbol{\eta}_i = 0$ but $\boldsymbol{\eta} \neq \mathbf{0}$ almost surely. Therefore, using that $\boldsymbol{\lambda}(\boldsymbol{\alpha}) = g(\boldsymbol{\alpha}, \tau_*^2(\boldsymbol{\alpha})) = \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha})\Big[1 - \frac{1}{n}\mathbb{E}\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha})}}(\boldsymbol{B} + \tau_*(\boldsymbol{\alpha})\boldsymbol{Z})\|_0^*\Big]$,

$$\lim_{\boldsymbol{\alpha}\downarrow\boldsymbol{A}_{\min}(\delta)} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = \ell_* \cdot \lim_{\boldsymbol{\alpha}\downarrow\boldsymbol{A}_{\min}(\delta)} \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha}) = -\infty\,. \tag{2.9.21}$$

Finally we consider the case $\boldsymbol{\alpha} \to \infty$ and observe $\tau_*^2(\boldsymbol{\alpha}) \to \sigma_w^2 + \mathbb{E}\{B^2\}/\delta$. To see this, notice that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) \to \sigma_w^2 + \mathbb{E}\{B^2\}/\delta$ as $\boldsymbol{\alpha} \to \infty$ since $\tau_*^2(\boldsymbol{\alpha}) = \mathsf{F}(\tau_*^2(\boldsymbol{\alpha}), \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha}))$ is bounded above. Moreover, since $\tau_*(\boldsymbol{\alpha})$ is bounded, $\boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha})$ is unbounded as $\boldsymbol{\alpha} \to \infty$ and we have $\lim_{\boldsymbol{\alpha}\to\infty} \ell(\boldsymbol{\alpha}) = 1$ whence

$$\lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = 1 \cdot \lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha}) = \infty\,. \tag{2.9.22}$$

We pause here to summarize that $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ is continuously differentiable on the domain $\{\boldsymbol{\alpha} : \boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}(\delta)\}$ with $\boldsymbol{\lambda}(\boldsymbol{A}_{\min}(\delta)) = -\infty$ and $\lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = +\infty$.

Now to prove the inverse mapping $\boldsymbol{\lambda} \mapsto \boldsymbol{\alpha}(\boldsymbol{\lambda})$ is continuous and non-decreasing when $p \to \infty$, we claim that the invertibility of $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ is sufficient. Precisely, **(1)** invertibility implies strict monotonicity; **(2)** monotonicity plus (2.9.21) and (2.9.22) implies both $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ and $\boldsymbol{\lambda} \mapsto \boldsymbol{\alpha}(\boldsymbol{\lambda})$ are increasing; and **(3)** continuity of $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ implies continuity of $\boldsymbol{\lambda} \mapsto \boldsymbol{\alpha}(\boldsymbol{\lambda})$.

Now we prove the invertibility by contradiction. Assume that there are two distinct such values $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ satisfying $\widetilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\boldsymbol{\alpha}_1) = \boldsymbol{\lambda}(\boldsymbol{\alpha}_2)$. Apply Theorem 3 to both $\boldsymbol{\alpha}(\widetilde{\boldsymbol{\lambda}}) = \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ with $\psi(\boldsymbol{x}, \boldsymbol{y}) = \langle (\boldsymbol{x} - \boldsymbol{y})^2 \rangle$. Then, together with Corollary 2.3.4,

$$\operatorname*{plim}_{p\to\infty} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p = \operatorname*{plim}_{p\to\infty} \mathbb{E}\langle \| \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*}}(\boldsymbol{\beta} + \tau_* \boldsymbol{Z}\,;\, \boldsymbol{\alpha}\tau_*) - \boldsymbol{\beta}\|_2^2 \rangle = \delta(\tau_*^2 - \sigma_w^2)\,.$$

Since $\operatorname{plim}_{p\to\infty} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p$ is independent of $\boldsymbol{\alpha}$, the right side gives $\tau_*(\boldsymbol{\alpha}_1) = \tau_*(\boldsymbol{\alpha}_2)$. Next apply Theorem 3 with $\psi(\boldsymbol{x}, \boldsymbol{y}) = \langle |\boldsymbol{x}| \rangle$, giving $\operatorname{plim}_{p\to\infty} \|\hat{\boldsymbol{\beta}}\|_1/p = \operatorname{plim}_{p\to\infty} \mathbb{E}\langle \| \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*}}(\boldsymbol{\beta} + \tau_* \boldsymbol{Z}\,;\, \boldsymbol{\alpha}\tau_*)\|_1 \rangle$. Obviously, for τ_* and p fixed, $\boldsymbol{\theta} \mapsto \mathbb{E}\langle \| \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_*}}(\boldsymbol{\beta} + \tau_* \boldsymbol{Z}\,;\, \boldsymbol{\theta})\|_1 \rangle$ is strictly decreasing in $\boldsymbol{\theta}$. Therefore $\boldsymbol{\alpha}_1 \tau_*(\boldsymbol{\alpha}_1) = \boldsymbol{\alpha}_2 \tau_*(\boldsymbol{\alpha}_2)$ implying $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$, since $\tau_*(\boldsymbol{\alpha}_1) = \tau_*(\boldsymbol{\alpha}_2)$, which is a contradiction.

□

2.9.2 Verifying Properties (P1) and (P2)

In this appendix we demonstrate that the properties **(P1)** and **(P2)** given in Section 2.4 and relating to the denoiser $\eta_p^t(\cdot)$ defined in (2.4.1) are true.

*Verifying Properties **(P1)** and **(P2)**.* Property **(P1)** follows since $\eta_p^t(\cdot) =$

$\operatorname{prox}_{J_{\alpha\tau_t}}(\cdot)$, as it is easy to show that proximal operators are Lipschitz continuous with Lipschitz constant one. Namely

$$||\eta_p^t(\boldsymbol{v}_1) - \eta_p^t(\boldsymbol{v}_2)|| = ||\operatorname{prox}_{J_{\alpha\tau_t}}(\boldsymbol{v}_1) - \operatorname{prox}_{J_{\alpha\tau_t}}(\boldsymbol{v}_2)|| \leq ||\boldsymbol{v}_1 - \boldsymbol{v}_2||.$$

Next we show that property **(P2)** is true. We restate property **(P2)** for convenience: for any s, t with $(\boldsymbol{Z}, \boldsymbol{Z}')$ a pair of length-p vectors such that (Z_i, Z_i') are i.id. $\sim \mathcal{N}(0, \boldsymbol{\Sigma})$ for $i \in [p]$ where $\boldsymbol{\Sigma}$ is any 2×2 covariance matrix, the following limits exist and are finite.

$$\begin{aligned} &\operatorname*{plim}_{p\to\infty} \frac{1}{p}\|\boldsymbol{\beta}\|, \quad \operatorname*{plim}_{p\to\infty} \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z})], \\ &\operatorname*{plim}_{p\to\infty} \frac{1}{p}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}[\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z})]. \end{aligned} \tag{2.9.23}$$

We first note that the first limit in (2.9.23) exists by Assumption **(A2)** and the strong law of large numbers. We focus on the other two limits. These results follow by [HL19a, Proposition 1] given in Lemma 2.3.3 and the following lemma, which is a classic result in probability theory.

Lemma 2.9.2 (Doob's L^1 maximal inequality, [Doo53] Chapter VII, Theorem 3.4)**.** *Let $X_1, X_2, \ldots, X_p$ be a sequence of nonnegative i.i.d. random variables such that $\mathbb{E}[X_1 \max\{0, \log(X_1)\}] < \infty$. Then,*

$$\mathbb{E}\Big[\sup_{p\geq 1}\Big\{\frac{1}{p}(X_1 + X_2 + \cdots + X_p)\Big\}\Big] \leq \frac{e}{e-1}(1 + \mathbb{E}[X_1 \max\{0, \log(X_1)\}]).$$

Proof. Let $M_p = \frac{1}{p}(X_1 + X_2 + \cdots + X_p)$. Then the sequence $\{M_p\}$ is a submartingale and hence by Doob's maximal inequality,

$$\mathbb{E}\Big[\sup_{p'\geq p\geq 1} M_p\Big] \leq \frac{e}{e-1}(1+\mathbb{E}[M_{p'}\max\{0,\log(M_{p'})\}]).$$

Note the mapping $x \mapsto x\max\{0,\log x\}$ is convex and hence $\mathbb{E}[M_{p'}\max\{0,\log(M_{p'})\}]) \leq \mathbb{E}[X_1\max\{0,\log(X_1)\}]$. The result follows by Fatou's lemma and by noting that $\sup_{p'\geq p\geq 1} M_p \uparrow \sup_{p\geq 1} M_p$ as $p' \to \infty$. □

Before we prove that the second and third limits in (2.9.23) exist and are finite, we state one more result that will be helpful in the proof. This result uses Lemma 2.9.2 along with a Dominated Convergence argument to study expectations taken with respect to $(\boldsymbol{Z},\boldsymbol{Z}')$ like those in (2.9.23).

Lemma 2.9.3. *Consider a function $\psi_p : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that for iterations $s,t \geq 0$,*

$$\frac{1}{p}\Big|\psi_p(\boldsymbol{\beta},\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}),\eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}'))-\psi_p(\boldsymbol{\beta},h^s(\boldsymbol{\beta}+\boldsymbol{Z}),h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big| \to 0, \quad \text{as} \quad p\to\infty, \tag{2.9.24}$$

where h^s, h^t are the unspecified functions of Lemma 2.3.3, and $(\boldsymbol{Z},\boldsymbol{Z}')$ are independent Gaussian vectors having zero-mean and independent entries with finite variance. Assume, for some constant $L > 0$ not depending on p,

$$\begin{aligned}&\frac{1}{p}\Big|\psi_p(\boldsymbol{\beta},\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}),\eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}'))-\psi_p(\boldsymbol{\beta},h^s(\boldsymbol{\beta}+\boldsymbol{Z}),h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big| \\ &\leq L\Big(1+\frac{\|\boldsymbol{\beta}\|^2}{p}+\frac{\|\boldsymbol{Z}\|^2}{p}+\frac{\|\boldsymbol{Z}'\|^2}{p}\Big).\end{aligned} \tag{2.9.25}$$

Then, as $p \to \infty$*,*

$$\frac{1}{p}\Big|\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}))\Big\} - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big\}\Big| \to 0. \tag{2.9.26}$$

Proof. We begin by showing that $\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\sup_{p\geq 1}\frac{1}{p}\Big|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big|\Big\} < \infty$. Using (2.9.25), it is clear that this expectation is finite almost surely if

$$\begin{aligned} &\mathbb{E}\Big[\sup_{p\geq 1}\Big\{\frac{1}{p}\|\boldsymbol{Z}(p)\|^2\Big\}\Big] < \infty, \quad \mathbb{E}\Big[\sup_{p\geq 1}\Big\{\frac{1}{p}\|\boldsymbol{Z}'(p)\|^2\Big\}\Big] < \infty, \\ &\text{and} \quad \mathbb{E}\Big[\sup_{p\geq 1}\Big\{\frac{1}{p}\|\boldsymbol{\beta}(p)\|^2\Big\}\Big] < \infty, \end{aligned}$$

where we have made the dependence of the vectors on the dimension p explicit. But Lemma 2.9.2 immediately implies the above since $\mathbb{E}[B^2 \max\{0, \log B\}] < \infty$ by assumption **(A2)**.

Now by dominated convergence we have,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\operatorname*{plim}_p \frac{1}{p}\Big|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}')) - \psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big|\Big\} \\ &= \operatorname*{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}')) - \psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big| \\ &\geq \operatorname*{plim}_p \frac{1}{p}\Big|\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}))\Big\} \\ &- \mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big\}\Big|. \end{aligned}$$

Then the above implies the desired result (2.9.26) from assumption (2.9.24). □

First consider the second limit in (2.9.23). By Cauchy-Schwarz, (3.3.5) of Lemma 2.3.3 implies that $\Big|\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - \boldsymbol{\beta}^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})\Big|/p \to 0$, as $p \to \infty$. This follows because

$$\left|\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - \boldsymbol{\beta}^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})\right|/p \leq \|\boldsymbol{\beta}\|\|\eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - h^t(\boldsymbol{\beta}+\boldsymbol{Z})\|/p.$$

Then the right side of the above → 0 with growing p because $\|\boldsymbol{\beta}\|/\sqrt{p}$ limits to a constant as justified above (this is the limit in (2.9.23)), and the other term → 0 by (3.3.5) of Lemma 2.3.3. This means that assumption (2.9.24) of Lemma 2.9.3 is satisfied. Assumption (2.9.25) of Lemma 2.9.3 is also satisfied since both η_p^t and h^t are Lipschitz(1), by Cauchy-Schwarz inequality. Therefore Lemma 2.9.3 implies $\left|\mathbb{E}_{\boldsymbol{Z}}\{\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})\} - \mathbb{E}_{\boldsymbol{Z}}\{\boldsymbol{\beta}^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})\}\right|/p \to 0$, as $p \to \infty$. Therefore,

$$\operatorname*{plim}_{p\to\infty} \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})]/p = \operatorname*{plim}_{p\to\infty} \sum_{i=1}^{p} \beta_{0,i}\mathbb{E}_Z\{h^t(\beta_{0,i}+Z_i)\}/p = \mathbb{E}[Bh^t(B+Z)],$$

where B, Z are univariate. By the Cauchy-Schwarz inequality, $\mathbb{E}[Bh^t(B+Z)] < \infty$ if $\mathbb{E}[B^2] < \infty$ and $\mathbb{E}[h^t(B+Z)^2] < \infty$. Since $\mathbb{E}[B^2] = \sigma_{\boldsymbol{\beta}}^2 < \infty$ is given by our assumption, it suffices to show $\mathbb{E}[h^t(B+Z)^2] < \infty$. But this follows from the fact that $h^t(\cdot)$ is Lipschitz(1) and therefore $\mathbb{E}[h^t(B+Z)^2] < \mathbb{E}[(B+Z)^2] \leq \mathbb{E}[B^2] + \mathbb{E}[Z^2] = \sigma_{\boldsymbol{\beta}}^2 + \Sigma_{11} < \infty$.

Finally consider the third limit in (2.9.23). Similarly to the work in studying the second limit in (2.9.23), we will appeal to Lemma 2.9.3. First we will show that

$$\left|\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - h^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})\right|/p \to 0, \quad \text{as} \quad p \to \infty, \tag{2.9.27}$$

meaning that assumption (2.9.24) of Lemma 2.9.3 is satisfied. Then, again, assumption (2.9.25) of Lemma 2.9.3 is satisfied since both $\eta_p^t(\cdot)$ and $h^t(\cdot)$ are Lipschitz(1), using Cauchy-Schwarz.

Now we want to prove (2.9.27). By repeated applications of Cauchy-Schwarz it is not hard to show,

$$\begin{aligned}
&\operatorname*{plim}_p \left|\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z}) - h^s(\boldsymbol{\beta} + \boldsymbol{Z}')^\top h^t(\boldsymbol{\beta} + \boldsymbol{Z})\right|/p \\
&\leq \operatorname*{plim}_p \|h^s(\boldsymbol{\beta} + \boldsymbol{Z}')\| \|\eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z}) - h^t(\boldsymbol{\beta} + \boldsymbol{Z})\|/p \\
&+ \operatorname*{plim}_p \|h^t(\boldsymbol{\beta} + \boldsymbol{Z})\| \|\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}') - h^s(\boldsymbol{\beta} + \boldsymbol{Z}')\|/p \\
&+ \operatorname*{plim}_p \|\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}') - h^s(\boldsymbol{\beta} + \boldsymbol{Z}')\| \|\eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z}) - h^t(\boldsymbol{\beta} + \boldsymbol{Z})\|/p.
\end{aligned}$$

Now, (2.9.27) follows since the right side of the above goes to 0 as p grows. This follows since, by (3.3.5) of Lemma 2.3.3, as $p \to \infty$,

$$\|\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}') - h^s(\boldsymbol{\beta} + \boldsymbol{Z}')\|/\sqrt{p} \to 0 \quad \text{and} \quad \|\eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z}) - h^t(\boldsymbol{\beta} + \boldsymbol{Z})\|/\sqrt{p} \to 0.$$

Moreover, since $h^s(\cdot)$ and $h^t(\cdot)$ are separable, by the Law of Large Numbers,

$$\begin{aligned}
\operatorname*{plim}_p \|h^s(\boldsymbol{\beta} + \boldsymbol{Z}')\|^2/p &= \operatorname*{plim}_p \sum_{i=1}^p [h^s(\beta_i + Z_i')]^2/p = \mathbb{E}[(h^s(B + Z'))^2] < \infty, \\
\operatorname*{plim}_p \|h^t(\boldsymbol{\beta} + \boldsymbol{Z})\|^2/p &= \operatorname*{plim}_p \sum_{i=1}^p [h^t(\beta_i + Z_i)]^2/p = \mathbb{E}[(h^t(B + Z))^2] < \infty,
\end{aligned}$$

where the inequalities follow since $\mathbb{E}[(h^s(B + Z'))^2] \leq \mathbb{E}[(B + Z')^2] \leq \sigma_\beta^2 + \Sigma_{22} < \infty$ and $\mathbb{E}[(h^t(B + Z))^2] \leq \mathbb{E}[(B + Z)^2] \leq \sigma_\beta^2 + \Sigma_{11} < \infty$. This proves (2.9.27) and therefore we can apply Lemma 2.9.3.

Then Lemma 2.9.3 implies,

$$\left|\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\{\eta_p^s(\boldsymbol{\beta} + \boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta} + \boldsymbol{Z})\} - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\{h^s(\boldsymbol{\beta} + \boldsymbol{Z}')^\top h^t(\boldsymbol{\beta} + \boldsymbol{Z})\}\right|/p \to 0, \text{ as } p \to \infty.$$

But now, using the above, we find that

$$\operatorname*{plim}_{p\to\infty} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\{\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})\}/p = \operatorname*{plim}_{p\to\infty} \sum_{i=1}^{p} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\{h^s(\beta_i+Z_i')h^t(\beta_i+Z_i)\}/p$$
$$= \mathbb{E}[h^s(B+Z')h^t(B+Z)],$$

where B, Z', and Z are univariate and $\mathbb{E}[h^s(B+Z')h^t(B+Z)] < \infty$ by Cauchy-Schwarz and the fact that $h^s(\cdot)$ and $h^t(\cdot)$ are Lipschitz(1). Namely, this gives the bound

$$\Big(\mathbb{E}[h^s(B+Z')h^t(B+Z)]\Big)^2$$
$$\leq \mathbb{E}[(h^s(B+Z'))^2]\mathbb{E}[(h^t(B+Z))^2] \leq \mathbb{E}[(B+Z')^2]\mathbb{E}[(B+Z)^2]$$
$$= (\mathbb{E}[B^2]+\mathbb{E}[Z'^2])(\mathbb{E}[B^2]+\mathbb{E}[Z^2]) = (\sigma_{\boldsymbol{\beta}}^2+\Sigma_{22})(\sigma_{\boldsymbol{\beta}}^2+\Sigma_{11}) < \infty.$$

We have now shown that property **(P2)** is true.

□

2.9.3 Proof of Fact 2.2.7

Proof. The fact follows from the asymptotic separability of the proximal operator [HL19a, Proposition 1] (restated in Lemma 2.3.3) and the dominated convergence theorem [Roy68] allowing for interchange of limit and expectation. We sketch the proof of the existence of the limit in (2.2.4) (and the result for the limit in (2.2.11) follows similarly). By Lemma 2.3.3, the weak convergence of $\boldsymbol{\alpha}(p)$ to A, and the Weak Law of Large Numbers, one can argue that

$$\lim_p \|\text{prox}_{J_{\boldsymbol{\alpha}(p)\tau_*}}(\boldsymbol{B}+\tau_*\boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p) = \mathbb{E}\{(h(B+\tau_*Z)-B)^2\}/\delta, \tag{2.9.28}$$

where $h(\cdot) := h(\cdot; B + \tau_* Z, A\tau_*)$ is the unspecified, separable function of Lemma 2.3.3. This is consistent with [Lemma 29, [HL19a]]. The limit in (2.2.4) exists if $\mathbb{E}\{(h(B + \tau_* Z) - B)^2\}/\delta < \infty$ and

$$\mathbb{E}\{(h(B + \tau_* Z) - B)^2\} \leq 2\mathbb{E}\{h(B + \tau_* Z)^2 + B^2\} \leq 2\mathbb{E}\{(B + \tau_* Z)^2 + B^2\}$$
$$\leq 2\mathbb{E}\{2B^2 + 2\tau_*^2 Z^2 + B^2\} = 6\mathbb{E}\{B^2\} + 4\tau_*^2 < \infty.$$

Here the first and third inequalities follow from $(x-y)^2 \leq 2(x^2 + y^2)$ and the second inequality follows from h being Lipschitz(1): $|h(x)| = |h(x) - h(0)| \leq |x - 0| = |x|$. □

2.9.4 Proof of Lemma 2.7.1

Proof. First, the proof of (2.7.1) follows from Theorem 2.4.1. To see this, note that by (2.1.3a), we have $\boldsymbol{\beta}^{t+1} = \text{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t) = \eta_p^t(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t)$, and therefore we apply Theorem 2.4.1 with uniformly pseudo-Lipschitz function $\psi_p(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{z}^t, \boldsymbol{\beta}) = \|\eta_p^t(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{z}^t)\|^2/p$ to get

$$\operatorname*{plim}_p \|\boldsymbol{\beta}^t\|^2/p \overset{p}{=} \operatorname*{plim}_p \ \mathbb{E}_{\boldsymbol{Z}}[\|\eta_p^t(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})\|^2]/p, \tag{2.9.29}$$

for $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$. By the Lipschitz property of η_p^t (Assumption **(A4)**), we have $\mathbb{E}_{\boldsymbol{Z}}[\|\eta_p^t(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})\|^2] \leq \mathbb{E}_{\boldsymbol{Z}}[\|\boldsymbol{\beta} + \tau_t \boldsymbol{Z}\|^2] \leq 2\|\boldsymbol{\beta}\|^2 + 2p\tau_t^2$. Plugging into (2.9.29), we find $\text{plim}_p \|\boldsymbol{\beta}^t\|^2/p \overset{p}{=} 2\,\text{plim}_p \|\boldsymbol{\beta}\|^2/p + 2\tau_t^2 = 2\sigma_{\boldsymbol{\beta}}^2 + 2\tau_t^2$, where the final inequality follows by Assumption **(A2)**.

Now consider the $\widehat{\boldsymbol{\beta}}$ result in (2.7.2). First, note that by definition $\mathcal{C}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{C}(\boldsymbol{0})$

where the cost function $\mathcal{C}(\cdot)$ is defined in (2.1.2). Using that

$$\mathcal{C}(\mathbf{0}) = \frac{1}{2}\|\boldsymbol{y}\|^2 = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{w}\|^2 \leq \|\boldsymbol{X}\boldsymbol{\beta}\|^2+\|\boldsymbol{w}\|^2 \leq \sigma_{\max}^2(\boldsymbol{X})\|\boldsymbol{\beta}\|^2+\|\boldsymbol{w}\|^2, \tag{2.9.30}$$

where $\sigma_{\max}(\boldsymbol{X})$ is the maximum singular value of $\boldsymbol{X}$. We note that this value, $\sigma_{\max}(\boldsymbol{X})$, is bounded almost surely as $p \to \infty$ using standard estimates on the singular values of random matrices since $\boldsymbol{X}$ has i.i.d. Gaussian entries by Assumption **(A1)** (see, for example, [BMN20, Lemma F.2]). Therefore,

$$\operatorname*{plim}_p \mathcal{C}(\widehat{\boldsymbol{\beta}})/p \leq \operatorname*{plim}_p \sigma_{\max}^2(\boldsymbol{X})\|\boldsymbol{\beta}\|^2/p + \operatorname*{plim}_p \|\boldsymbol{w}\|^2/p \leq \mathsf{B}_{max}\sigma_{\beta}^2 + \sigma_w^2, \tag{2.9.31}$$

where we've defined B_{max} to be a bound on the limit of the maximum singular value, i.e. $\lim_p \sigma_{\max}^2(\boldsymbol{X}) \leq \mathsf{B}_{max}$, and the final inequality holds by Assumptions **(A2)** and **(A3)**.

Now we will relate $\frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2$ to $\frac{1}{p}\mathcal{C}(\widehat{\boldsymbol{\beta}})$ and other terms lower-bounded by a constant with high probability. We write $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\perp}+\widehat{\boldsymbol{\beta}}^{||}$ where $\widehat{\boldsymbol{\beta}}^{\perp} \in ker(\boldsymbol{X})^{\perp}$ and $\widehat{\boldsymbol{\beta}}^{||} \in ker(\boldsymbol{X})$. Since $\widehat{\boldsymbol{\beta}}^{||} \in \ker(\boldsymbol{X})$ and $\ker(\boldsymbol{X})$ is a random subspace of size $p - n = p(1-\delta)$, by Kashin Theorem (Theorem H.1.), we have that for some constant $\nu_1 = \nu_1(\delta)$, with high probability

$$\|\widehat{\boldsymbol{\beta}}^{||}\|_2^2 \leq \nu_1\|\widehat{\boldsymbol{\beta}}^{||}\|_1^2/p. \tag{2.9.32}$$

Then we have the following bound

$$\|\widehat{\boldsymbol{\beta}}\|^2 = \|\widehat{\boldsymbol{\beta}}^{||}\|^2 + \|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \overset{(a)}{\leq} \nu_1\|\widehat{\boldsymbol{\beta}}^{||}\|_1^2/p + \|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \overset{(b)}{\leq} 2\nu_1\|\widehat{\boldsymbol{\beta}}\|_1^2/p + (2\nu_1+1)\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2, \tag{2.9.33}$$

where step (a) holds by (2.9.32) and step (a) by the Triangle Inequality and Cauchy-Schwarz as follows

$$\|\widehat{\boldsymbol{\beta}}^{\|}\|_1^2 = \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{\perp}\|_1^2 \leq (\|\widehat{\boldsymbol{\beta}}\|_1 + \|\widehat{\boldsymbol{\beta}}^{\perp}\|_1)^2 \leq 2\|\widehat{\boldsymbol{\beta}}\|_1^2 + 2\|\widehat{\boldsymbol{\beta}}^{\perp}\|_1^2 \leq 2\|\widehat{\boldsymbol{\beta}}\|_1^2 + 2p\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2.$$

Now we bound the second term on the right side of (2.9.33). Define $\hat{\sigma}_{min}(\boldsymbol{X})$ as the minimum non-zero singular value of $\boldsymbol{X}$. By standard results in linear algebra, $\hat{\sigma}_{min}^2(\boldsymbol{X})\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \leq \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\perp}\|^2$. Therefore,

$$\begin{aligned}
&\hat{\sigma}_{min}^2(\boldsymbol{X})\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \leq \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \\
&\leq \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\perp} - \boldsymbol{y} + \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\perp}\|^2 + 2\|\boldsymbol{y}\|^2 \\
&\leq 2\mathcal{C}(\widehat{\boldsymbol{\beta}}) + 2\mathcal{C}(\mathbf{0}) \leq 2\mathcal{C}(\mathbf{0}).
\end{aligned}$$

Therefore, using (2.9.30) and (2.9.31), we have

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \leq \operatorname*{plim}_p \frac{\frac{2}{p}\mathcal{C}(\mathbf{0})}{\hat{\sigma}_{min}^2(\boldsymbol{X})} \leq \frac{2(\mathsf{B}_{max}\sigma_\beta^2 + \sigma_w^2)}{\mathsf{B}_{min}}. \tag{2.9.34}$$

where we've defined B_{min} to be a bound on the limit of the minimum non-zero singular value, i.e. $\lim_p \hat{\sigma}_{min}^2(\boldsymbol{X}) \geq \mathsf{B}_{min}$.

Now we bound the first term on the right side of (2.9.33). Recall the definition of the sort-ed ℓ_1 norm, i.e. $J_{\boldsymbol{\lambda}}(\boldsymbol{b}) = \sum \lambda_i |\boldsymbol{b}|_{(i)}$, then using $\lambda_{min} := \lim_p \min(\boldsymbol{\lambda})$ to lower bound the threshold values,

$$\lambda_{min}\|\widehat{\boldsymbol{\beta}}\|_1 = \sum \lambda_{min}|\widehat{\boldsymbol{\beta}}_i| = \sum \lambda_{min}|\widehat{\boldsymbol{\beta}}|_{(i)} \leq \sum \lambda_i|\widehat{\boldsymbol{\beta}}|_{(i)} = J_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{C}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{C}(\mathbf{0}).$$

Then, using (2.9.30) and (2.9.31), we see

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|_1 \leq \operatorname*{plim}_p \frac{1}{\lambda_{min}}\Big(\frac{1}{p}\mathcal{C}(\mathbf{0})\Big) \leq \frac{1}{\lambda_{min}}(\mathsf{B}_{max}\sigma_\beta^2 + \sigma_w^2). \tag{2.9.35}$$

By (2.9.35), along with the upper bound in (2.9.33), we have

$$\begin{aligned}\operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p} &\leq 2\nu_1 \operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}\|_1^2}{p^2} + (2\nu_1+1)\operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2}{p} \\ &\leq \left[\frac{2\nu_1(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2+\sigma_w^2)}{\lambda_{min}}\right]^2 + \frac{2(2\nu_1+1)(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2+\sigma_w^2)}{\mathsf{B}_{min}}.\end{aligned}$$

□

2.9.5 Proof of Lemma 2.7.3

The proof of Lemma 2.7.3 relies on the following result, Lemma 2.9.4, about the exponential rate of the convergence of the state evolution sequence defined in (2.6.3). We state and prove Lemma 2.9.4, and Lemma 2.7.3 is proved afterward.

Lemma 2.9.4. *Assume $\boldsymbol{\alpha} > \boldsymbol{A}_{\min}(\delta)$ and let $\{\Sigma_{s,t}\}_{s,t\geq 0}$ be defined by the recursion* (2.6.3) *with initial condition* (2.6.2)*. Then there exists constants $B_1, r_1 > 0$ such that for all $t \geq 0$, letting $\tau_* := \lim_t \tau_t$,*

$$|\Sigma_{t,t} - \tau_*^2| \leq B_1 e^{-r_1 t}, \qquad \textit{and} \qquad |\Sigma_{t,t+1} - \tau_*^2| \leq B_1 e^{-r_1 t}.$$

Proof. Throughout the proof, we use the $\{\eta_p^t\}_{p\in\mathbb{N}_{>0}}$ notation introduced in Section 2.4 and defined in (2.4.1) with a slight modification to explicitly state the thresholds. Namely, we consider a sequence of denoisers $\eta_p : \mathbb{R}^{p\times p} \to \mathbb{R}^p$ to be those that apply the proximal operator $\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ defined in (2.1.4), i.e. $\eta_p(\boldsymbol{v}; \boldsymbol{\alpha}\tau_t) := \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{v})$ for a vector $\boldsymbol{v} \in \mathbb{R}^p$.

Then, per the definition in (2.6.3), we have

$$\Sigma_{s+1,t+1} = \sigma_w^2 + \lim_p \mathbb{E}\Big\{[\eta_p(\boldsymbol{B}+\tau_s\boldsymbol{Z}_s;\boldsymbol{\alpha}\tau_s) - \boldsymbol{B}]^\top[\eta_p(\boldsymbol{B}+\tau_t\boldsymbol{Z}_t;\boldsymbol{\alpha}\tau_t) - \boldsymbol{B}]\Big\}/(\delta p),$$

where $\boldsymbol{B} \sim B$ i.i.d. elementwise, independent of length$-p$ jointly Gaussian vectors $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_r$ having $\mathbb{E}[\boldsymbol{Z}_s] = \mathbb{E}[\boldsymbol{Z}_r] = \boldsymbol{0}$, with covariance $\mathbb{E}\{([\boldsymbol{Z}_s]_i)^2\} = \mathbb{E}\{([\boldsymbol{Z}_r]_i)^2\} = 1$ for any element $i \in [p]$, and $\mathbb{E}\{[\boldsymbol{Z}_s]_i[\boldsymbol{Z}_r]_j\} = \frac{\Sigma_{s,r}}{\tau_r \tau_s}\mathbb{I}\{i = j\}$. Recall, $\Sigma_{t,t} = \tau_t^2$ defined in (2.2.4) and by Theorem 1 we know that $\{E_{t,t}\}_{t \geq 0}$ is monotone and converges to τ_*^2 as $t \to \infty$. To prove exponential convergence of $\{E_{t-1,t}\}_{t \geq 0}$ as claimed in the lemma statement, we construct a discrete dynamical system below.

For $t \geq 1$, define the vector $\boldsymbol{y}_t = (y_{t,1}, y_{t,2}, y_{t,3}) \in \mathbb{R}^3$ as

$$y_{t,1} \equiv \Sigma_{t-1,t-1} = \tau_{t-1}^2\,, \quad y_{t,2} \equiv \Sigma_{t,t} = \tau_t^2\,, \quad y_{t,3} \equiv \Sigma_{t-1,t-1} - 2\Sigma_{t,t-1} + \Sigma_{t,t} \tag{2.9.36}$$

A careful argument shows that the vector $\boldsymbol{y}_t = (y_{t,1}, y_{t,2}, y_{t,3})$ belongs to $\mathbb{R}_+^3$. Essentially this requires showing that a matrix $R_T :=$ as in [BM11c, Lemma 5.8] is strictly positive definite. Using the definition of the Σ recursion in (2.6.3), it is immediate to see that this sequence is updated according to the mapping $\boldsymbol{y}_{t+1} = G(\boldsymbol{y}_t)$ where

$$G_1(\boldsymbol{y}_t) \equiv y_{t,2}\,, \tag{2.9.37}$$

$$G_2(\boldsymbol{y}_t) \equiv \sigma_w^2 + \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \boldsymbol{B}\|^2\Big\}/(\delta p), \tag{2.9.38}$$

$$G_3(\boldsymbol{y}_t) \equiv \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \eta_p(\boldsymbol{B} + \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\sqrt{y_{t,1}})\|^2\Big\}/(\delta p), \tag{2.9.39}$$

where $(\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1})$ are length$-p$ jointly Gaussian vectors, independent of $\boldsymbol{B} \sim B$ i.i.d. elementwise, having $\mathbb{E}[\boldsymbol{Z}_t] = \mathbb{E}[\boldsymbol{Z}_{t-1}] = \boldsymbol{0}$ and with covariance $\mathbb{E}\{([\boldsymbol{Z}_t]_i)^2\} = \mathbb{E}\{([\boldsymbol{Z}_{t-1}]_i)^2\} = 1$ for any element $i \in [p]$, and $\mathbb{E}\{[\boldsymbol{Z}_t]_i[\boldsymbol{Z}_{t-1}]_j\} = \frac{\Sigma_{t,t-1}}{\tau_t \tau_{t-1}}\mathbb{I}\{i = j\}$. Notice that $\mathbb{E}\{\|\sqrt{y_{t,2}}\boldsymbol{Z}_t - \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}\|^2\} = y_{t,3}$, where we emphasize that $G_3(\boldsymbol{y}_t)$

depends on $y_{t,3}$ through the covariance of $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-1}$. Moreover, if $\sigma_w^2 > 0$, then $y_{t,1}$ and $y_{t,2}$ are both strictly positive and by the map defined above it is easy to see that $y_{t,3}$ for all $t \geq 0$. This mapping is defined for $y_{t,3} \leq 2(y_{t,1} + y_{t,2})$.

In the following, we will show by induction on t, for $t \geq 1$, that the stronger inequality $y_{t,3} < (y_{t,1} + y_{t,2})$ holds. The initial condition implied by Eq. (2.6.2) is

$$y_{1,1} = \sigma_w^2 + \mathbb{E}[B^2]/\delta, \qquad y_{1,2} = \sigma_w^2 + \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0; \boldsymbol{\alpha}\tau_0) - \boldsymbol{B}\|^2\Big\}/(\delta p),$$
$$y_{1,3} = \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0; \boldsymbol{\alpha}\tau_0)\|^2\Big\}/(\delta p),$$

It follows that

$$\begin{aligned} y_{1,1} + y_{1,2} - y_{1,3} &= 2\sigma_w^2 + 2\lim_p \mathbb{E}\Big\{\boldsymbol{B}^\top\Big(\boldsymbol{B} - \eta_p(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0; \boldsymbol{\alpha}\tau_0)\Big)\Big\}/(\delta p) \\ &= 2\sigma_w^2 + 2\lim_p \mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big(\boldsymbol{B} - \mathbb{E}_{\boldsymbol{Z}_0}\{\eta_p(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0; \boldsymbol{\alpha}\tau_0)\}\Big)\Big\}/(\delta p). \end{aligned}$$

Using the above, it is easy to show $y_{1,3} < y_{1,1} + y_{1,2}$. This follows since $\mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big(\boldsymbol{B} - \mathbb{E}_{\boldsymbol{Z}_0}\{\eta_p^0(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0)\}\Big)\Big\}$ is asymptotically separable using Lemma 2.3.3 and because the function $x \mapsto x - \mathbb{E}_Z h^0(x + \tau_0 Z)$ is monotone increasing. It follows that $\lim_p \mathbb{E}_{\boldsymbol{B}}\Big\{\boldsymbol{B}^\top\Big(\boldsymbol{B} - \mathbb{E}_{\boldsymbol{Z}_0}\{\eta_p^0(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0)\}\Big)\Big\}/(\delta p) > 0$.

Suppose that $y_{t,3} < y_{t,1} + y_{t,2}$, we want to show $y_{t+1,3} < y_{t+1,1} + y_{t+1,2}$. By the induction hypothesis, $\mathbb{E}\{[\boldsymbol{Z}_t]_i[\boldsymbol{Z}_{t-1}]_i\} = \frac{y_{t,1}+y_{t,2}-y_{t,3}}{2\sqrt{y_{t,1}y_{t,2}}} > 0$, so elementwise $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-1}$

are positively correlated.

$$
\begin{aligned}
&y_{t+1,1} + y_{t+1,2} - y_{t+1,3}\\
&= 2\sigma_w^2\\
&+ \lim_p 2\mathbb{E}\Big\{[\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \boldsymbol{B}]^\top[\eta_p(\boldsymbol{B} + \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\sqrt{y_{t,1}}) - \boldsymbol{B}]\Big\}/(\delta p).
\end{aligned}
\tag{2.9.40}
$$

Notice that $x \mapsto \eta(b + c \cdot x\,; \theta) - b$ is monotone for any constants b and $c > 0$ and consider the following result: for g, a monotone function, and X_1 and X_2, two positively correlated standard Gaussians, $\mathbb{E}[g(X_1)g(X_2)] \geq 0$. This is a special case of a theorem in [Pit82], which shows $\mathbb{E}[g(X_1)g(X_2)] \geq \mathbb{E}[g(X_1)]\mathbb{E}[g(X_2)] = (\mathbb{E}[g(X_1)])^2 > 0$. Then since $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-1}$ are positively correlated, $\mathbb{E}\Big\{[\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \boldsymbol{B}]^\top[\eta_p(\boldsymbol{B} + \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\sqrt{y_{t,1}}) - \boldsymbol{B}]\Big\} \geq 0$, which yields $y_{t+1,3} < (y_{t+1,1} + y_{t+1,2})$.

We can hereafter therefore assume $y_{t,3} < y_{t,1} + y_{t,2}$ for all t.

We will consider the above iteration for arbitrary initialization y_0 (satisfying $y_{0,3} < y_{0,1} + y_{0,2}$) and will show the following three facts:

Fact (i). $y_{t,1}, y_{t,2} \to \tau_*^2$ as $t \to \infty$. Further the convergence is monotone.

Fact (ii). If $y_{0,1} = y_{0,2} = \tau_*^2$ and $y_{0,3} \leq 2\tau_*^2$, then $y_{t,1} = y_{t,2} = \tau_*^2$ for all t and $y_{t,3} \to 0$.

Fact (iii). The Jacobian $J = J_G(y_*)$ of G at $y_* = (\tau_*^2, \tau_*^2, 0)$ has spectral radius $\sigma(J) < 1$.

By simple compactness arguments, Facts (i) and (ii) imply $y_t \to y_*$ as $t \to \infty$. (Notice that $y_{t,3}$ remains bounded since $y_{t,3} \le (y_{t,1} + y_{t,2})$ and by the convergence of $y_{t,1}, y_{t,2}$.) Fact (iii) implies that convergence is exponentially fast.

Proof of Fact (i). Notice that $y_{t,2}$ evolves independently by $y_{t+1,2} = G_2(y_t) = F(y_{2,t}, \boldsymbol{\alpha}\sqrt{y_{2,t}})$, with $F(\,\cdot\,,\,\cdot\,)$ the state evolution mapping introduced in (2.2.8). It follows from Proposition 1.3 that $y_{t,2} \to \tau_*^2$ monotonically for any initial condition. Since $y_{t+1,1} = y_{t,2}$, the same happens for $y_{t,1}$.

Proof of Fact (ii). Consider the function

$$G_*(x) = G_3(\tau_*^2, \tau_*^2, x) = \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B} + \tau_* \boldsymbol{Z}_t; \boldsymbol{\alpha}\tau_*) - \eta_p(\boldsymbol{B} + \tau_* \boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\tau_*)\|^2\Big\}/(\delta p),$$

where

$$\mathbb{E}\{[\boldsymbol{Z}_t]_i[\boldsymbol{Z}_{t-1}]_i\} = \frac{y_{t,1} + y_{t,2} - y_{t,3}}{2\sqrt{y_{t,1}y_{t,2}}} = \frac{2\tau_*^2 - x}{2\tau_*^2}$$

is no longer time-dependent. This function is defined for $x \in [0, 2\tau_*^2]$. Further G_* can be represented as follows in terms of the independent random vectors $\boldsymbol{Z}$, $\boldsymbol{W} \sim N(0, \mathbb{I})$:

$$\begin{aligned} G_*(x) = \lim\nolimits_p \tfrac{1}{\delta p}\mathbb{E}\Big\{&\|\eta_p(\boldsymbol{B} + \boldsymbol{Z}\sqrt{\tau_*^2 - \tfrac{1}{4}x} + \boldsymbol{W}(\tfrac{1}{2}\sqrt{x}); \boldsymbol{\alpha}\tau_*) \\ &-\eta_p(\boldsymbol{B} + \boldsymbol{Z}\sqrt{\tau_*^2 - \tfrac{1}{4}x} - \boldsymbol{W}(\tfrac{1}{2}\sqrt{x}); \boldsymbol{\alpha}\tau_*)\|^2\Big\}, \end{aligned}$$

where

$$(\tau_*\boldsymbol{Z}_{t-1}, \tau_*\boldsymbol{Z}_t) \overset{d}{=} \Big(\boldsymbol{Z}\sqrt{\tau_*^2 - \frac{1}{4}x} - \boldsymbol{W}(\frac{1}{2}\sqrt{x}), \boldsymbol{Z}\sqrt{\tau_*^2 - \frac{1}{4}x} + \boldsymbol{W}(\frac{1}{2}\sqrt{x})\Big).$$

Obviously $G_*(0) = 0$. A simple Taylor expansion about the first argument around

$\boldsymbol{B}$ yields (recall higher derivatives of η are 0 almost everywhere)

$$\begin{aligned}G_*(x) &= \lim_p \mathbb{E}\Big\{\|\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*) + \Big(\boldsymbol{Z}\sqrt{\tau_*^2 - \frac{1}{4}x} + \boldsymbol{W}(\frac{1}{2}\sqrt{x})\Big) \odot \partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*) \\ &\quad - \eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*) - \Big(\boldsymbol{Z}\sqrt{\tau_*^2 - \frac{1}{4}x} - \boldsymbol{W}(\frac{1}{2}\sqrt{x})\Big) \odot \partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}/(\delta p) \\ &= \lim_p x\mathbb{E}\Big\{\|\boldsymbol{W} \odot \partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}/(\delta p) = \lim_p x\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}/(\delta p).\end{aligned}$$

Using the above, we study $G_*'(x)$. First, we can exchange the limit and differentiation because $f_p(x) := x\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}/(\delta p)$ converges uniformly to $f(x) := \lim_p x\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}/(\delta p)$. To see this, notice f_p, f are linear in x and defined on $[0, 2\tau_*^2]$. Hence for every $\epsilon > 0$, there exists p_0 such that

$$\begin{aligned}|f_{p_0}(x) - f(x)| &= x\Big|\frac{1}{\delta p_0}\mathbb{E}\Big\{\|\partial_1\eta_{p_0}(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\} - \lim_p \frac{1}{\delta p}\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}\Big| \\ &\leq 2\tau_*^2\Big|\frac{1}{\delta p_0}\mathbb{E}\Big\{\|\partial_1\eta_{p_0}(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\} - \lim_p \frac{1}{\delta p}\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\}\Big| < \epsilon.\end{aligned}$$

By uniform convergence we have,

$$G_*'(x) = \lim_p \frac{1}{\delta p}\mathbb{E}\Big\{\|\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]\|^2\Big\} = G_*'(0) \leq \lim_p \frac{1}{\delta p}\sum_{i=1}^p \mathbb{E}\Big\{[\partial_1\eta_p(\boldsymbol{B};\boldsymbol{\alpha}\tau_*)]_i\Big\}.$$

Hence $G_*'(0) < 1$, using (2.2.10) since $\boldsymbol{\lambda} > \boldsymbol{0}$. Then $y_{t,3} = [G_*'(0)]^t y_{0,3} \to 0$ as $t \to \infty$ as claimed.

Proof of Fact (iii). By the definition of G, the Jacobian is given by

$$J_G(y_*) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \mathsf{F}'(\tau_*^2) & 0 \\ a & G_*'(0) & b \end{pmatrix}$$

denoting $\mathsf{F}'(\tau_*^2) \equiv \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ evaluated at $\tau^2 = \tau_*^2$ with a and b constants whose values are not important to the proof. Computing the eigenvalues of the Jacobian, we get $\sigma(J) = \max\left\{ \mathsf{F}'(\tau_*^2) \,,\, G_*'(0) \right\}$. Since $G_*'(0) < 1$ proved above and $\mathsf{F}(\tau_*^2) < 1$ by Theorem 1, the claim follows. □

Proof of Lemma 2.7.3. We show that Lemma 2.7.3 follows by Lemmas 2.9.4 and 2.6.2. By Lemma 2.6.2,

$$\operatorname*{plim}_n \Big(\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|^2/n - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) \Big) = 0,$$

$$\operatorname*{plim}_p \Big(\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2/(\delta p) - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) \Big) = 0,$$

and so it is sufficient to show that $\lim_t(\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) = 0$. Note that this follows from Lemma 2.9.4 since $\tau_t^2 = \Sigma_{t,t}$ and $\tau_{t-1}^2 = \Sigma_{t-1,t-1}$ both converge to τ_*^2 as does $\Sigma_{t,t-1}$.

□

2.9.6 Technical Details for the Condition (3) Proof

We first introduce some notation and ideas that will be used throughout the proof. The proof is similar to [BM11c, Section 5.3], with the key difference being the concept of equivalence classes as described in Section 2.5.1.

We now introduce a more general recursion than the AMP algorithm in (2.1.3a)-(2.1.3b). Given $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, define the column vectors $\boldsymbol{h}^{t+1}, \boldsymbol{q}^{t+1} \in \mathbb{R}^p$ and $\boldsymbol{b}^t, \boldsymbol{m}^t \in \mathbb{R}^n$, recursively, for $t \geq 0$ as follows, starting with initial condition $\boldsymbol{\beta}^0 = 0$

and $\boldsymbol{z}^0 = \boldsymbol{y}$.

$$\boldsymbol{h}^{t+1} = \boldsymbol{\beta} - (\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t), \qquad \boldsymbol{q}^t = \boldsymbol{\beta}^t - \boldsymbol{\beta}, \qquad \boldsymbol{b}^t = \boldsymbol{w} - \boldsymbol{z}^t, \qquad \boldsymbol{m}^t = -\boldsymbol{z}^t. \tag{2.9.41}$$

Note that these definitions of $\boldsymbol{h}^t$ and $\boldsymbol{m}^t$ match those used in Section 2.6.

Denoting $[\boldsymbol{u}|\boldsymbol{v}]$ to mean the matrix of concatenating vectors $\boldsymbol{u}, \boldsymbol{v}$ horizontally, we define

$$\begin{aligned} \underbrace{[\boldsymbol{h}^1 + \boldsymbol{q}^0 | \cdots | \boldsymbol{h}^t + \boldsymbol{q}^{t-1}]}_{\boldsymbol{A}_t} &= \boldsymbol{X}^\top \underbrace{[\boldsymbol{m}^0 | \cdots | \boldsymbol{m}^{t-1}]}_{\boldsymbol{M}_t}, \\ \underbrace{[\boldsymbol{b}^0 | \boldsymbol{b}^1 + \kappa_1 \boldsymbol{m}^0 | \cdots | \boldsymbol{b}^{t-1} + \kappa_{t-1} \boldsymbol{m}^{t-2}]}_{\boldsymbol{Y}_t} &= \boldsymbol{X} \underbrace{[\boldsymbol{q}^0 | \cdots | \boldsymbol{q}^{t-1}]}_{\boldsymbol{Q}_t}, \end{aligned} \tag{2.9.42}$$

where the scalars κ_t are defined as $\kappa_t := -[\nabla \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^{t-1})]/n$.

Define the σ-algebra generated by $\boldsymbol{b}^0, \cdots, \boldsymbol{b}^{t-1}, \boldsymbol{m}^0, \cdots, \boldsymbol{m}^{t-1}, \boldsymbol{h}^1, \cdots, \boldsymbol{h}^t$, $\boldsymbol{q}^0, \cdots, \boldsymbol{q}^t$ as $\mathfrak{S}_t$. Then [BM11a; BMN20], says that the conditional distribution of the random matrix $\boldsymbol{X}$ given $\mathfrak{S}_t$ is

$$\boldsymbol{X}|_{\mathfrak{S}_t} \stackrel{d}{=} \boldsymbol{E}_t + \boldsymbol{P}^\perp_{\boldsymbol{M}_t} \tilde{\boldsymbol{X}} \boldsymbol{P}^\perp_{\boldsymbol{Q}_t}, \tag{2.9.43}$$

where $\tilde{\boldsymbol{X}} \stackrel{d}{=} \boldsymbol{X}$ is independent of the conditioning sigma-algebra $\mathfrak{S}_t$ and $\boldsymbol{E}_t = \mathbb{E}(\boldsymbol{X}|\mathfrak{S}_t)$ is given by:

$$\boldsymbol{E}_t := \boldsymbol{Y}_t(\boldsymbol{Q}_t^\top \boldsymbol{Q}_t)^{-1}\boldsymbol{Q}_t^\top + \boldsymbol{M}_t(\boldsymbol{M}_t^\top \boldsymbol{M}_t)^{-1}\boldsymbol{A}_t^\top + \boldsymbol{M}_t(\boldsymbol{M}_t^\top \boldsymbol{M}_t)^{-1}\boldsymbol{M}_t^\top \boldsymbol{Y}_t(\boldsymbol{Q}_t^\top \boldsymbol{Q}_t)^{-1}\boldsymbol{Q}_t^\top.$$

In (2.9.43), we use the notation $\boldsymbol{P}^\perp_{\boldsymbol{M}_t} = \mathbb{I} - \boldsymbol{P}_{\boldsymbol{M}_t}$ and$\boldsymbol{P}^\perp_{\boldsymbol{Q}_t} = \mathbb{I} - \boldsymbol{P}_{\boldsymbol{Q}_t}$ where $\boldsymbol{P}_{\boldsymbol{Q}_t}$ and $\boldsymbol{P}_{\boldsymbol{M}_t}$ are orthogonal projectors onto column spaces of $\boldsymbol{Q}_t, \boldsymbol{M}_t$ respectively. From now on, since t is fixed, we will drop the subscript t when it is clear. A proof of

(2.9.43) can be found in [BM11a, Lemma 11]. We note that there are no differences in this conditional distribution in the nonseparable case, since the analysis (in both cases) is just that of an i.i.d. Gaussian matrix conditional on linear constraints.

Given the above notations, we claim that Lemma 2.7.5 is implied by the following statement.

Lemma 2.9.5. *Let s be a set of maximal atoms in $[p]$ such that $|s| \leq p(\delta - \gamma)$, for some $\gamma > 0$. Then there exists $\alpha_1 = \alpha_1(\gamma) > 0$ (independent of t) and $\alpha_2 = \alpha_2(\gamma, t) > 0$ (depending on t and γ) with*

$$\mathbb{P}\left\{ \min_{\|\boldsymbol{v}\|=1,\, \mathrm{supp}^*(\boldsymbol{v}) \subseteq s} \left\| \boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp} \tilde{\boldsymbol{X}} \boldsymbol{P}_{\boldsymbol{Q}}^{\perp} \boldsymbol{v} \right\| \leq \alpha_2 \,\middle|\, \mathfrak{S}_t \right\} \leq e^{-p\alpha_1},$$

eventually almost surely as $p \to \infty$, with $\boldsymbol{E}\boldsymbol{v} = \boldsymbol{Y}(\boldsymbol{Q}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}^*\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v} + \boldsymbol{M}(\boldsymbol{M}^*\boldsymbol{M})^{-1}\boldsymbol{X}^*\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}$.

We prove such implication in the next section now.

Proof of Lemma 2.7.5. The proof is adapted from [BM11c, Section 5.3.1]. First note that by Borel-Cantelli, it is sufficient to show that, for s measurable on $\mathfrak{S}_t$ and $|s| \leq p(\delta - c)$ there exist $a_1 = a_1(c) > 0$ and $a_2 = a_2(c, t) > 0$, such that

$$\mathbb{P}\left\{ \min_{|s'| \leq a_1 p} \; \min_{\|\boldsymbol{v}\|=1,\, \mathrm{supp}^*(\boldsymbol{v}) \subseteq s \cup s'} \|\boldsymbol{X}\boldsymbol{v}\| < a_2 \right\} \leq 1/p^2,$$

for all p large enough, using $\sigma_{\min}(\boldsymbol{X}_{S_t \cup S'}) = \min_{\|\boldsymbol{v}\|=1,\, \mathrm{supp}^*(\boldsymbol{v}) \subseteq s \cup s'} \|\boldsymbol{X}\boldsymbol{v}\|$. To shorten notation, the set $\{\|\boldsymbol{v}\| = 1,\ \mathrm{supp}^*(\boldsymbol{v}) \subseteq s \cup s'\}$ is denoted $\boldsymbol{v}(s')$. Now, conditioning on $\mathfrak{S}_t$, by a union bound,

$$\mathbb{P}\{\min_{|s'|\le a_1 p}\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\}\le\sum_{|s'|\le a_1 p}\mathbb{P}\{\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\}$$
$$\le\Big[\sum_{k=1}^{a_1 p}\binom{p}{k}\Big]\max_{|s'|\le pa_1}\mathbb{P}\{\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\}\le e^{ph(a_1)}\max_{|s'|\le a_1 p}\mathbb{P}\{\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\}\,, \tag{2.9.44}$$

where $h(a)=-a\log a-(1-a)\log(1-a)$ is the binary entropy function (cf. [MS77, Chapter 10, Corollary 9]). Therefore, using iterated expectation and (2.9.44),

$$\mathbb{P}\Big\{\min_{|s'|\le a_1 p}\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big\}=\mathbb{E}\Big\{\mathbb{P}\Big\{\min_{|s'|\le a_1 p}\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\Big\}\Big\}$$
$$\le e^{ph(a_1)}\mathbb{E}\Big\{\max_{|s'|\le a_1 p}\mathbb{P}\Big\{\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big|\mathfrak{S}_t\Big\}\Big\}\,,$$

Now, we fix $a_1<c/2$ in such a way that $h(a_1)\le\frac{1}{2}\alpha_1(\frac{c}{2})$ and let $a_2=\frac{1}{2}\alpha_2(\frac{c}{2},t)$ where α_1 and α_2 are defined by Lemma 2.9.5. Then,

$$\mathbb{P}\Big\{\min_{|s'|\le a_1 p}\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big\}$$
$$\le e^{\frac{1}{2}p\alpha_1(\frac{c}{2})}\mathbb{E}\Big\{\max_{|s'|\le a_1 p}\mathbb{P}\Big\{\min_{\|\boldsymbol{v}\|=1,\ \mathrm{supp}^*(\boldsymbol{v})\subseteq s\cup s'}\|\boldsymbol{X}\boldsymbol{v}\|<\frac{1}{2}\alpha_2(\frac{c}{2},t)\Big|\mathfrak{S}_t\Big\}\Big\}$$
$$\le e^{\frac{1}{2}p\alpha_1(\frac{c}{2})}\,\mathbb{E}\Big\{\max_{|s''|\le p(\delta-\frac{c}{2})}\mathbb{P}\Big\{\min_{\|\boldsymbol{v}\|=1,\ \mathrm{supp}^*(\boldsymbol{v})\subseteq s''}\|\boldsymbol{X}\boldsymbol{v}\|<\frac{1}{2}\alpha_2(\frac{c}{2},t)\Big|\mathfrak{S}_t\Big\}\Big\}\,.$$

Finally, using (cf. [BM11c, Lemma 5.1]),

$$\boldsymbol{X}\boldsymbol{v}|_{\mathfrak{S}}\overset{d}{=}\boldsymbol{Y}(\boldsymbol{Q}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}^*\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}+\boldsymbol{M}(\boldsymbol{M}^*\boldsymbol{M})^{-1}\boldsymbol{X}^*\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}+\boldsymbol{P}_{\boldsymbol{M}}^{\perp}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\,. \tag{2.9.45}$$

to estimate $\boldsymbol{X}\boldsymbol{v}$ and applying Lemma 2.9.5, we get, for all p large enough,

$$\mathbb{P}\Big\{\min_{|s'|\le a_1 p}\min_{\boldsymbol{v}(s')}\|\boldsymbol{X}\boldsymbol{v}\|<a_2\Big\}\le e^{\frac{1}{2}p\alpha_1}\,\mathbb{E}\Big\{\max_{|s''|\le p(\delta-\frac{c}{2})}e^{-p\alpha_1}\Big\}\le 1/p^2\,.$$

□

Now we prove Lemma 2.9.5, using a proof that is similar to that of [BM11c, Section 5.3.2]. We first state some lemmas that will be used in the proof, but we will not migrate the full proofs from [BM11c] for the sake of brevity. Instead, we describe the key points of proofs with an emphasis on the technical differences for the SLOPE problem and provide pointers to the original proofs.

The concept of maximal atoms are reflected in these lemmas via the sets s and correspondingly $\boldsymbol{P}_s$, where $\boldsymbol{P}_s$ is the $p \times p$ projector matrix onto the subspace of vectors whose supp* equals s. In the LASSO case where supp$^* \equiv$ supp and $s \equiv S$, the projector is orthogonal, but in general, we must define $\boldsymbol{P}_s[\cdot, j] = \frac{1}{|I|}\sum_{i \in I} \boldsymbol{e}_i$ for $j \in I$ where $\boldsymbol{P}_s[\cdot, j]$ is the j^{th} column of $\boldsymbol{P}_s$ for $1 \leq j \leq p$ and $\boldsymbol{e}_i$ is the i^{th} vector of the standard basis. For example, when $p = 4$ and $s = \{\{1\}, \{2, 4\}\}$,

$$\boldsymbol{P}_s = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}.$$

Such a projector is not necessarily orthogonal and its rank is described via $|s|$ (the number of equivalence classes), not via $|S|$ (the number of non-zero elements) as for the LASSO. We may view this projector as an orthogonal projector onto the subspace of maximal atoms: for a maximal atom $I \in s$, the projector maps elements whose indices belong to I onto their average value.

We begin with the auxiliary lemmas.

Lemma 2.9.6. *[Adapted from [BM11c, Lemma 5.4]] Let s be a set of maximal atoms in $[p]$ such that $|s| \le p(\delta-\gamma)$, for some $\gamma > 0$. Recall that $\boldsymbol{E}\boldsymbol{v} = \boldsymbol{Y}(\boldsymbol{Q}^\top\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v} + \boldsymbol{M}(\boldsymbol{M}^\top\boldsymbol{M})^{-1}\boldsymbol{A}^\top\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}$ and consider the event*

$$\varepsilon_1 := \Big\{\Big\|\boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^\perp\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2 \ge \frac{\gamma}{4\delta}\Big\|\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2 + \frac{\gamma}{4\delta}\Big\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2 \ \forall \boldsymbol{v} \\ \textit{s.t. } \|\boldsymbol{v}\| = 1 \textit{ and } \operatorname{supp}^*(\boldsymbol{v}) \subseteq s\Big\}.$$

Then there exists $a = a(\gamma) > 0$ such that $\mathbb{P}\{\varepsilon_1|\mathfrak{S}_t\} \ge 1 - e^{-pa}$.

Sketch proof. Define an event $\widetilde{\varepsilon}_1$ as follows:

$$\widetilde{\varepsilon}_1 = \Big\{|(\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v})^\top(\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v})| \le \Big(1 - \tfrac{\gamma}{2\delta}\Big)^{1/2}\|\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\|\,\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\|\Big\}, \tag{2.9.46}$$

where the event $\widetilde{\varepsilon}_1$ is meant to hold for all $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$ and $\operatorname{supp}^*(\boldsymbol{v}) \subseteq s$. We claim that $\mathbb{P}\{\widetilde{\varepsilon}_1|\mathfrak{S}_t\} \ge 1 - e^{-pa}$. To prove the claim, we use that for any $\boldsymbol{v}$, the unit vector $\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}/\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\|$ belongs to the random linear space $\operatorname{im}(\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{P}_s)$ with dimension at most $p(\delta-\gamma)$. Also, $\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}$ belongs to space spanned by the column space of the matrices $\boldsymbol{M}$ and of $\boldsymbol{B}$ where $\boldsymbol{B}_t = [\boldsymbol{b}^0|\ldots|\boldsymbol{b}^{t-1}]$ defined in (2.9.41) and (2.9.42), having dimension at most $2t$. Applying Proposition 2.9.9 using $m = n, m\lambda = p(\delta-\gamma), d = 2t$ and $\varepsilon = (1-\frac{\gamma}{2\delta})^{1/2}(1-\frac{\gamma}{\delta})^{1/2}$ gives that the event

$$\left(\frac{\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}}{\|\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\|}\right)^\top \frac{\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}}{\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\|} \le \sqrt{\lambda} + \varepsilon = \left(1 - \frac{\gamma}{2\delta}\right)^{1/2},$$

holds with the desired probability, proving the claim. Conditional on event (2.9.46), one can show

$$\Big\|\boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^\perp\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2 \ge \left(1 - \left(1 - \frac{\gamma}{2\delta}\right)^{1/2}\right)\Big\{\Big\|\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2 + \Big\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^\perp\boldsymbol{v}\Big\|^2\Big\}.$$

Finally observe that $1-(1-\frac{\gamma}{2\delta})^{1/2}\geq\frac{\gamma}{4\delta}$ and therefore since event $\widetilde{\varepsilon}_1$ occurring implies ε_1 occurs, giving the desired probability of ε_1 as well. □

Next we estimate the term $\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|^2$ in the above lower bound.

Lemma 2.9.7. *[Adapted from [BM11c, Lemma 5.5]] Let s be a set of maximal atoms in $[p]$ such that $|s|\leq p(\delta-\gamma)$, for some $\gamma>0$. Then there exists constant $c_1=c_1(\gamma)$, $c_2=c_2(\gamma)$ such that the event*

$$\varepsilon_2:=\Big\{\left\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\right\|\geq c_1(\gamma)\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|\ \forall\,\boldsymbol{v}\ \textit{ such that }\ \mathrm{supp}^*(\boldsymbol{v})\subseteq s\Big\}$$

holds with probability $\mathbb{P}\{\varepsilon_2|\mathfrak{S}_t\}\geq 1-e^{-pc_2}$.

Sketch proof. Let V be the linear space $V=\mathrm{im}(\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{P}_s)$ having dimension at most $p(\delta-\gamma)$. For all $\boldsymbol{v}$ with $\mathrm{supp}^*(\boldsymbol{v})\subseteq s$,

$$\left\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\right\|\geq\sigma_{\min}(\tilde{\boldsymbol{X}}|_V)\left\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\right\|, \tag{2.9.47}$$

where $\tilde{\boldsymbol{X}}|_V$ refers to the restriction of $\tilde{\boldsymbol{X}}$ to V. Then $\sigma_{\min}(\tilde{\boldsymbol{X}}|_V)$ is distributed as the minimum singular value of a Gaussian matrix of dimensions $p\delta\times\dim(V)$, which is almost surely bounded away from 0 as $p\to\infty$ (see Theorem G. 2). Large deviation estimates [Lit+05] imply that the probability that $\sigma_{\min}$ is smaller than a constant $c_1(\gamma)$ is exponentially small. □

In the next step we estimate the norm $\boldsymbol{E}\boldsymbol{v}$ by quoting the following result.

Lemma 2.9.8. *[BM11c, Lemma 5.6] There exists a constant $c=c(t)>0$ such that, defining the event,*

$$\mathcal{E}_3:=\Big\{\|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\|\geq c(t)\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\|\,,\|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|\leq c(t)^{-1}\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|,\ \textit{ for all }\ \boldsymbol{v}\in\mathbb{R}^p\Big\}, \tag{2.9.48}$$

we have that $\mathcal{E}_3$ holds eventually almost surely as $p \to \infty$.

Finally, we can now prove Lemma 2.9.5 with the ingredients given in Lemmas 2.9.6-2.9.8. We restate the proof from [BM11c, Lemma 5.3] with minor changes.

Proof of Lemma 2.9.5. We start with Lemma 2.9.8 by which we assume that event $\mathcal{E}_3$ holds for some function $c = c(t)$ (without loss of generality $c < 1/2$). For $\alpha_2(t) > 0$ small enough, let $\mathcal{E}$ be the event

$$\mathcal{E} := \left\{ \min_{\|\boldsymbol{v}\|=1,\, \mathrm{supp}^*(\boldsymbol{v}) \subseteq s} \left\| \boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp} \tilde{\boldsymbol{X}} \boldsymbol{P}_{\boldsymbol{Q}}^{\perp} \boldsymbol{v} \right\| \leq \alpha_2(t) \right\}. \tag{2.9.49}$$

First assume $\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \leq c^2/10$, from which it follows,

$$\begin{aligned} &\|\boldsymbol{E}\boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq \|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\| - \|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| - \|\boldsymbol{P}_{\boldsymbol{M}}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \\ \geq & c\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\| - (c^{-1} + \|\tilde{\boldsymbol{X}}\|_2)\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq \frac{c}{2} - \frac{c}{10} - \|\tilde{\boldsymbol{X}}\|_2\frac{c^2}{10} = \frac{2c}{5} - \|\tilde{\boldsymbol{X}}\|_2\frac{c^2}{10}, \end{aligned}$$

where the last inequality uses $\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\| = \sqrt{1 - \|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|^2} \geq 1/2$ under the assumption $\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \leq c^2/10$. Therefore, using Lemma 2.9.6, we get

$$\mathbb{P}\{\mathcal{E}|\mathfrak{S}_t\} \leq \mathbb{P}\left\{ \frac{2c}{5} - \|\tilde{\boldsymbol{X}}\|_2\frac{c^2}{10} \leq \left(\frac{4\delta}{\gamma}\right)^{1/2} \alpha_2(t) \middle| \mathfrak{S}_t \right\} + e^{-pa},$$

and the thesis follows from large deviation bounds on the norm $\|\tilde{\boldsymbol{X}}\|_2$ (see [Led01]) by first taking c small enough, and then choosing $\alpha_2(t) < \frac{c}{5}\sqrt{\frac{\gamma}{4\delta}}$.

Next assume $\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq c^2/10$. By Lemma 2.9.6 and 2.9.7, we can assume events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Therefore $\left\|\boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\right\| \geq (\frac{\gamma}{4\delta})^{1/2}\left\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\right\| \geq (\frac{\gamma}{4\delta})^{1/2}c_1(\gamma)\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|$, proving our thesis. □

2.9.7 Some Useful Auxiliary Material

We collect some auxiliary results that are necessary in our proof. Most of these are results that were initially stated in [BM11c] that we repeat here for the reader.

The following proposition is used in the proof of Lemma 2.9.6. The proof is identical to that of [BM11c, Proposition E.1] and it follows from a standard concentration of measure argument in [Led01]. For this reason, we don't repeat it here.

Proposition 2.9.9. *Let $V \subseteq \mathbb{R}^m$ a uniformly random linear space of dimension d. For $\lambda \in (0,1)$, let $\boldsymbol{P}_\lambda$ denote the projector onto the first $m\lambda$ maximal atoms in $[m]$: assume that $s = \{I_1, ..., I_d\}$, is the set of maximal atoms, then the j^{th} column, $\boldsymbol{P}_\lambda[:,j] = \frac{1}{|I_r|}\sum_{i\in I_r} \boldsymbol{e}_i$ if $j \in I_r$ for some $r \leq m\lambda$; otherwise $\boldsymbol{P}_\lambda[:,j] = \mathbf{0}$. Define $Z(\lambda) := \sup\{\|\boldsymbol{P}_\lambda \boldsymbol{v}\| \,:\, \boldsymbol{v} \in V,\ \|\boldsymbol{v}\| = 1\}$. Then, for any $\varepsilon > 0$ there exists $c(\varepsilon) > 0$ such that, for all m large enough (and d fixed) $\mathbb{P}\{|Z(\kappa) - \sqrt{\lambda}| \geq \varepsilon\} \leq e^{-m\, c(\varepsilon)}$.*

We next state a result due to Kashin [Kas77] relating to the equivalence of ℓ^2 and ℓ^1 norms on random vector spaces (cf. also [BM11c, Theorem F.1]).

Theorem G.1. *[Kas77] For any positive number υ there exist a universal constant c_υ such that for any $n \geq 1$, with probability at least $1 - 2^{-n}$, for a uniformly random subspace $V_{n,\upsilon}$ of dimension $\lfloor n(1-\upsilon)\rfloor$, for all $x \in V_{n,\upsilon}$, we have $c_\upsilon \|x\|_2 \leq \|x\|_1/\sqrt{n}$.*

Finally we state a general result about the limit behavior of extreme singular values of random matrices, as proved in [BY08] (cf. also [BM11c, Theorem F.2]).

Theorem G.2. *[BY08] Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ have i.i.d. entries with $\mathbb{E}\{A_{ij}\} = 0$, $\mathbb{E}\{A_{ij}^2\} = 1/n$, and $n/p = \delta$. Let $\sigma_{\max}(\boldsymbol{A})$ be it largest singular value, and $\hat{\sigma}_{\min}(\boldsymbol{A})$ be its smallest non-zero singular value. Then,*

$$\lim_{p\to\infty} \sigma_{\max}(\boldsymbol{A}) \overset{a.s.}{=} 1/\sqrt{\delta} + 1, \qquad \textit{and} \qquad \lim_{p\to\infty} \hat{\sigma}_{\min}(\boldsymbol{A}) \overset{a.s.}{=} 1/\sqrt{\delta} - 1.$$

Chapter 3

Characterizing the SLOPE Trade-off: A Variational Perspective and the Donoho-Tanner Limit

This chapter is based on "Zhiqi Bu, Jason Klusowski, Cynthia Rush, and Weijie J. Su. "Characterizing the SLOPE Trade-off: A Variational Perspective and the Donoho-Tanner Limit." arXiv preprint arXiv:2105.13302 (2021).".

3.1 Introduction

Reconstructing the signal from noisy linear measurements is vital in many disciplines, including statistical learning, signal processing, and biomedical imaging. In many modern applications where the number of explanatory variables often exceeds the number of measurements, the signal is often believed—or, wished—to be sparse in the sense that most of its entries are zero or approximately zero. Put differently, this means that a majority of the explanatory variables are simply irrelevant to the response of interest.

Accordingly, a host of methods have been developed to tackle these problems by leveraging the sparsity of signals in high-dimensional linear regression. These methods often rely on, among others, the concept of *regularization* to constrain the search space of the unknown signals. Perhaps the most influential instantiation of this concept is ℓ_1 regularization, which gives rise to the Lasso method [Tib96a]. The optimal amount of regularization, however, hinges on the sparsity level of the signal. Intuitively speaking, if the sparsity level is low, then more regularization should be imposed, and vice versa (see, for example, [Abr+06]).

This intuition necessitates the development of a regularization technique that is adaptive to the sparsity level of signals, which is typically unknown in practical problems. To achieve this desired adaptivity, [Bog+15a] introduced *sorted ℓ_1 regularization*. This new regularization technique turns into a method called SLOPE in the setting of a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tag{3.1.1}$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, and $\boldsymbol{w} \in \mathbb{R}^n$ is the noise term. Explicitly, SLOPE estimates the coefficients by solving the convex programming problem

$$\arg\min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \tag{3.1.2}$$

where $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$ are the order statistics in absolute value of $\boldsymbol{b} = (b_1, \ldots, b_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ (with at least one strict inequality) are the regularization parameters. The sorted ℓ_1 penalty, $\sum_{i=1}^{p} \lambda_i |b|_{(i)}$, is a norm, and the optimization problem for SLOPE is, therefore, convex (see also [FN16]). As an important feature, the sorted ℓ_1 norm penalizes larger entries more heavily than smaller ones. Indeed, this regularization technique is shown to be adaptive to the degree of sparsity level and enables SLOPE to obtain optimal estimation performance for certain problems [SC16]. Notably, in the special case $\lambda_1 = \cdots = \lambda_p$, the sorted ℓ_1 norm reduces to the usual ℓ_1 norm. Thus, the Lasso can be regarded as a special instance of SLOPE.

A fundamental question, yet to be better addressed, is how to quantitatively characterize the benefits of using the sorted ℓ_1 regularization. To explore this question, Figure 3.1 compares the model selection performance of SLOPE and the Lasso in terms of the *false discovery proportion* (FDP) and *true positive proportion*

(TPP) or, equivalently, between measures of type I error and power. Needless to say, a model is preferred if its FDP is small while its TPP is large. As the first impression conveyed by this figure, both methods seem to undergo a trade-off between the FDP and TPP when the TPP is below a certain limit. More interestingly, while *nowhere* on the Lasso path is the TPP above a limit, which is about 0.5707 in the left plot of Figure 3.1 and 0.4343 in the right, SLOPE is able to pass the limit toward achieving full power. To be sure, these contrasting patterns persist even for an arbitrarily large signal-to-noise ratio. This distinction must be attributed to the flexibility of the SLOPE regularization sequence $(\lambda_1, \ldots, \lambda_p)$ compared to a single value as in the Lasso case. Recognizing this message, we are tempted to ask (1) *why* the use of sorted ℓ_1 regularization brings a significant benefit over ℓ_1 regularization in the high TPP regime and, equally importantly, (2) *why* SLOPE exhibits a trade-off between the FDP and TPP just as the Lasso does in the low TPP regime.

3.1.1 A peek at our results

To address these two questions, in this paper we characterize the optimal trade-off of SLOPE between the TPP and FDP, uncovering several intriguing findings of sorted ℓ_1 regularization. Assuming TPP $\approx u$ for $0 \le u \le 1$, loosely speaking, the trade-off curve gives the smallest possible value of the FDP of SLOPE using any regularization sequence in the large system limit. To prepare for a rough description of our contributions, in brief, we work in the setting where the design

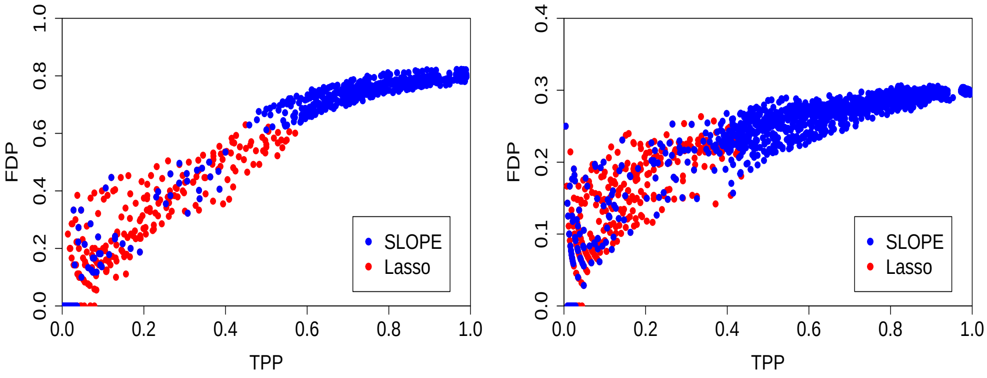


Figure 3.1: Comparison between SLOPE and the Lasso in terms of the TPP–FDP trade-off. Given an estimate $\widehat{\boldsymbol{\beta}}$, define its FDP $= \frac{|\{j:\beta_j=0 \text{ and } \widehat{\beta}_j\neq 0\}|}{|\{j:\widehat{\beta}_j\neq 0\}|}$ and TPP $= \frac{|\{j:\beta_j\neq 0 \text{ and } \widehat{\beta}_j\neq 0\}|}{|\{j:\beta_j\neq 0\}|}$. The SLOPE regularization sequence $\boldsymbol{\lambda}_{\lambda,r\lambda,w}$ is defined in (3.2.4), with varying $0 < r < 1$ and $\lambda > 0$, and $w = 0.2$ in the left plot and $w = 0.3$ in the right plot. The results of the Lasso are taken over its entire solution path, and its highest TPP is about 0.5707 in the left plot and 0.4343 in the right plot. Left: $(n, p) = (300, 1000), |\{j : \beta_j \neq 0\}|/p = 0.2$, and $\boldsymbol{w} = \mathbf{0}$ (noiseless); right: $(n, p) = (400, 1000), |\{j : \beta_j \neq 0\}|/p = 0.7$, and $\boldsymbol{w} = \mathbf{0}$. On both plots, non-zero entries of $\boldsymbol{\beta}$ are i.i.d. draws from the standard normal distribution. More specifications of the setup are detailed in Section 3.2. The result presents 10 independent trials.

has i.i.d. Gaussian entries and the regression coefficients $\beta_1, \ldots, \beta_p$ are i.i.d. draws from a distribution that takes non-zero values with a certain probability. Notably, it is generally nontrivial to define false discoveries in high dimensions [GHT13], which is not an issue however in the case of independent regressors. The assumption on the signal prior corresponds to the *linear sparsity* regime. In addition, we assume that both $n, p \to \infty$ and the sampling ratio n/p converges to a constant (see more detailed assumptions in Section 3.2). From a technical viewpoint, these assumptions allow us to make use of tools from approximate message passing (AMP) theory [DMM09a; BM11a].

Breaking the Donoho–Tanner power limit To explain the contrasting results presented in Figure 3.1, we prove that under the aforementioned assumptions, SLOPE can achieve an arbitrarily high TPP. Moving from sorted ℓ_1 regularization to ℓ_1 regularization, in stark contrast, the Lasso exhibits the Donoho–Tanner (DT) power limit when $n < p$ and the sparsity is above a certain threshold [Don06; Don05]. Informally, the DT power limit is the largest possible power that any estimate along the Lasso path can achieve in the large system limit. For example, in the setting of Figure 3.1 this power limit is about 0.5676 in the left plot and 0.4401 in the right plot. For SLOPE and a certain choice of the regularization sequence, interestingly, we show that the asymptotic TPP-FDP trade-off of SLOPE beyond the DT power limit is given by a simple Möbius transformation, which is shown by the blue curve in Figure 3.2. This Möbius transformation naturally serves as an upper bound on the (optimal) SLOPE trade-off curve above the DT power limit.

Lower bound via convex optimization Next, we address the second question by lower bounding the optimal trade-off for SLOPE, followed by a comparison between the trade-offs for the two methods in the low TPP regime. To put it into perspective, the Lasso trade-off obtained by [SBC17] is plotted as the green solid curve in Figure 3.2. Apart from the simple fact that the SLOPE trade-off is better than or equal to the Lasso counterpart, however, it requires new tools to take into account the structure of sorted ℓ_1 regularization. To this end, we develop a technique based on a class of infinite-dimensional convex optimization problems. The resulting lower bound is shown in red in Figure 3.2. It is worth noting that the development of this technique presents several novel ideas that might be of independent interest for other regularization schemes.

Instance superiority of SLOPE The results illustrated so far are taken from an optimal-case viewpoint. Moving to a more practical standpoint, we are interested in comparing the two methods on a specific problem instance and, in particular, wish to find a SLOPE regularization sequence that allows SLOPE to outperform the Lasso with any given penalty parameter in terms of, for example, the TPP, the FDP, or the ℓ_2 estimation risk. Surprisingly, we prove that on any problem instance, SLOPE can dominate the Lasso according to these three indicators simultaneously. This comparison conveys the message that the flexibility of the sorted ℓ_1 regularization can turn into appreciable benefits. This result is formally stated in Theorem 6.

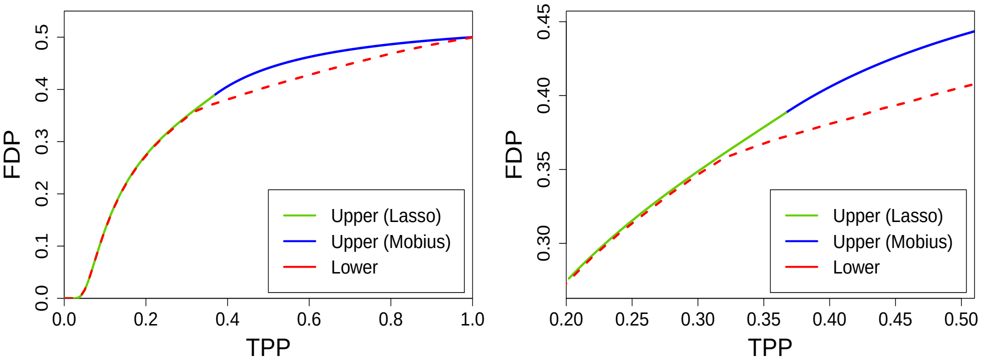


Figure 3.2: Illustration of the upper bound $q^\star$ and lower bound $q_\star$ for the SLOPE TPP–FDP trade-off. The right plot is the zoom-in of the left. Here $n/p = 0.3$ and $|\{j : \beta_j \neq 0\}|/p = 0.5$ (see more details in the working assumptions in Section 3.2). The Lasso trade-off curve shown in green is truncated at the DT power limit about 0.3669 [SBC17]. The optimal SLOPE trade-off curve must lie between the two curves. Notably, the two bounds agree at TPP $= 1$.

3.1.2 Organization

The remainder of this paper is structured as follows. In Section 3.2, we present the main results of this paper. Next, Section 3.3 introduces the AMP machinery at a minimal level as a preparation for the proofs of our main results. In Section 3.4, we detail the derivation of the lower bound based on variational calculus and infinite-dimensional convex optimization. In Section 3.5, we specify the upper bound, especially the part given by a Möbius transformation above the DT power limit. We conclude this paper in Section 3.6 by proposing several future research directions. Omitted proofs are relegated to the appendix.

3.2 Main results

Throughout this paper, we make the following working assumptions to specify the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, and noise $\boldsymbol{w} \in \mathbb{R}^n$ in the linear model (3.1.1), as well as the SLOPE regularization sequence $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$. To obviate any ambiguity, we consider a sequence of problems indexed by (n, p) with both n, p tending to infinity.

(A1) The matrix $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries. The sampling ratio n/p converges to a constant $\delta > 0$.

(A2) The entries of $\boldsymbol{\beta}$ are i.i.d. copies of a random variable Π satisfying $\mathbb{P}(\Pi \neq 0) = \epsilon$ for a constant $0 < \epsilon < 1$ and $\mathbb{E}(\Pi^2 \max\{0, \log \Pi\}) < \infty$. The noise vector $\boldsymbol{w}$ consists of i.i.d. copies of a random variable W with bounded second moment $\sigma^2 := \mathbb{E}(W^2) < \infty$.

(A3) The SLOPE regularization sequence $\boldsymbol{\lambda}(p) = (\lambda_1, \ldots, \lambda_p)$ is the order statistics of p i.i.d. realizations of a (nontrivial) non-negative random variable Λ.

Moreover, we assume that $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{w}$ are independent. Notice that the sparsity level of $\boldsymbol{\beta}$ is about ϵp and that each column of $\boldsymbol{X}$ has approximately a unit ℓ_2 norm. The noise variance σ^2 can equal 0, meaning that our results apply to both noisy and noiseless settings. In (A3), by “nontrivial” we mean that Λ is not always equal to 0. As an aside, SLOPE is reduced to the Lasso if the distribution of Λ is a unit probability mass at some positive value.

The working assumptions are mainly driven by their necessity in AMP theory [DMM09a; BM11a], which enables the use of the recent analysis of an AMP algorithm when applied to solve SLOPE [HL19a; Bu+20a]. Regarding (A2), the condition $\mathbb{P}(\Pi \neq 0) = \epsilon$, which implies linear sparsity of the regression coefficients, is not required for AMP theory. Rather, this condition is only made so that the TPP and FDP are well-defined. Besides, the merit of the linear sparsity regime has been increasingly recognized in the high-dimensional literature [MMB+18b; WMZ18; Su18; SCC19; WWM19].

3.2.1 Bounds on the SLOPE trade-off

Our main result is the characterization of a trade-off curve that teases apart asymptotically achievable TPP and FDP pairs from the asymptotically unachievable pairs for SLOPE [1]. For any estimate $\widehat{\boldsymbol{\beta}}$, recall that its FDP and TPP are defined as

$$\mathrm{FDP} = \frac{|\{j : \beta_j = 0 \text{ and } \widehat{\beta}_j \neq 0\}|}{|\{j : \widehat{\beta}_j \neq 0\}|}, \quad \mathrm{TPP} = \frac{|\{j : \beta_j \neq 0 \text{ and } \widehat{\beta}_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|},$$

with the convention $0/0 = 0$. When it comes to the SLOPE estimator, we use $\mathrm{TPP}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ and $\mathrm{FDP}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ to denote its TPP and FDP, respectively.

[1] R code to reproduce the results, e.g., to calculate $q_\star$ and $q^\star$, is available at `https://github.com/woodyx218/SLOPE_AMP`.

Likewise, we define the thresholded FDP and TPP, namely,

$$\mathrm{FDP}_\xi = \frac{|\{j : \beta_j = 0 \text{ and } |\widehat{\beta}_j| > \xi\}|}{|\{j : |\widehat{\beta}_j| > \xi\}|}, \quad \mathrm{TPP}_\xi = \frac{|\{j : \beta_j \neq 0 \text{ and } |\widehat{\beta}_j| > \xi\}|}{|\{j : \beta_j \neq 0\}|},$$

which reduce to FDP and TPP when $\xi = 0$. These thresholded versions of FDP and TPP are introduced purely for technical reasons, and have been used in previous work including [WWM17]. Specifically, the SLOPE estimator is known to possibly have many elements that are very close to 0, but not strictly 0, thereby causing FDP and TPP not to converge as expected. We refer interested readers to [HL19a, Example 3 and Figure 3] for a concrete example that illustrates such a phenomenon. In order for FDP and TPP to converge, we consider ξ in the set

$$\Xi := \{\xi : \mathbb{P}(\widehat{\Pi}(\Pi, \Lambda) = \xi) = 0\}, \tag{3.2.1}$$

where $\widehat{\Pi}$ is the limiting distribution of $\widehat{\beta}_j$ that will be defined in (3.3.1)..

Our main results are stated in the following two theorems, which give lower and upper bounds on the optimal SLOPE trade-off. Taken together, they demonstrate a fundamental separation between asymptotically achievable TPP–FDP pairs and the unachievable pairs over all signal priors Π and SLOPE regularization sequences $\boldsymbol{\lambda}$. Note that both the upper bound $q^\star$ and lower bound $q_\star$ are defined on $[0, 1]$ and completely determined by ϵ and δ. The expression for $q^\star$ is given in (3.2.8), while $q_\star$ is detailed in Section 3.4.

Theorem 4 (Lower bound)**.** *Under the working assumptions, namely (A1), (A2), and (A3), for $\xi \in \Xi$ in* (3.2.1)*, the following inequality holds with probability tending* *to one:*

$$\mathrm{FDP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \geq q_\star\left(\mathrm{TPP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}); \delta, \epsilon\right) - c_\xi,$$

for some positive constant c_ξ which tends to 0 as $\xi \to 0$.

Theorem 5 (Upper bound)**.** *Under the working assumptions, namely (A1), (A2), and (A3), for any $0 \leq u \leq 1$ and $\xi \in \Xi$ in (3.2.1), there exist a signal prior Π and a SLOPE regularization prior Λ such that the following inequalities hold with probability tending to one:*

$$\mathrm{FDP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \leq q^\star\left(\mathrm{TPP}_\xi(\boldsymbol{\beta}, \boldsymbol{\lambda}); \delta, \epsilon\right) + c_\xi \quad \textit{and} \quad |\underset{\xi}{\mathrm{TPP}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) - u| \leq c_\xi.$$

for some positive constant c_ξ which tends to 0 as $\xi \to 0$.

Remark 3.2.1. Above, 0.0001 can be replaced by an arbitrarily small positive constant. The probability is taken with respect to the randomness in the design matrix, regression coefficients, noise, and SLOPE regularization sequence in the large system limit $n, p \to \infty$. In relating to the assumptions made previously, this theorem holds even for $\sigma^2 = 0$, the noiseless case.

The proofs of Theorem 4 and Theorem 5 are given in Section 3.4 and Section 3.5, respectively. Most notably, our proof of Theorem 4 starts by formulating the problem of finding a tight lower bound as a calculus of variations problem. Relying on several novel elements, we further reduce this problem to a class of infinite-dimensional convex programs.

On the one hand, Theorem 4 says that it is impossible to achieve high power and a low FDP simultaneously using any sorted ℓ_1 regularization sequences, and this trade-off is specified by $q_\star$. On the other hand, Theorem 5 demonstrates that SLOPE can achieve at least the same trade-off as that given by $q^\star$ by specifying a prior Π and a regularization sequence $\boldsymbol{\lambda}$. Indeed, the proof of this theorem is constructive in that we will show that SLOPE can come arbitrarily close to any point on the curve $q^\star$ in Section 3.5. Another important observation from Theorem 5 is that SLOPE can achieve any power levels, which is not necessarily the case for ℓ_1 regularization-based methods as we show in Section 3.2.2.

Informally, let q_{SLOPE} denote the optimal SLOPE trade-off curve. That is, $q_{\text{SLOPE}}(u)$ is asymptotically the minimum possible value of the FDP under the constraint that the TPP is about u, over all possible SLOPE regularization sequences (see formal definition in Section 3.3). Combining the two theorems above, we readily see that the optimal SLOPE trade-off must be sandwiched between $q^\star$ and $q_\star$:

$$q_\star(u) \le q_{\text{SLOPE}}(u) \le q^\star(u)$$

for all $0 \le u \le 1$. Consequently, the sharpness of the approximation to the SLOPE trade-off rests on the gap between the two curves, and throughout the paper, we refer to the gap as the function $u \mapsto q^\star(u) - q_\star(u)$. Figure 3.3 illustrates several examples of the two curves for various pairs of ϵ, δ. Importantly, the plots show that the two bounds are very close to each other, thereby demonstrating tightness of our bounds. In fact, the gap between $q_\star$ and $q^\star$ is an upper bound of the gap between the analytical $q^\star$ and the true trade-off q_{SLOPE}. Furthermore, a closer look at the plots reveals that the two curves seem to coincide exactly when the TPP is below a certain value. In this regard, the SLOPE trade-off might have been uncovered exactly in this regime of TPP. Future investigation is required to obtain a fine-grained comparison between the two curves.

Looking at Figure 3.3, the reader may wonder where the non-monotonicity in ϵ of the trade-off curves originates from. We argue that this is due to the DT phase transition. In the case of the Lasso, for fixed δ, the trade-off curves are monotonically increasing in ϵ: in other words, $q_{\text{Lasso}}(u; \delta, \epsilon_1) > q_{\text{Lasso}}(u; \delta, \epsilon_2)$ whenever $\epsilon_1 > \epsilon_2$. However, in some settings, we empirically observe that TPP $= 1$ is achieved with a dense SLOPE estimator. When this occurs, $q_{\text{SLOPE}}(1) = 1 - \epsilon$ and thus $q_{\text{SLOPE}}(1; \delta, \epsilon_1) < q_{\text{SLOPE}}(1; \delta, \epsilon_2)$. In words, the SLOPE trade-off at TPP $= 1$ is monotonically *decreasing* in ϵ. Therefore, the patterns may not be monotone between the TPP upper limit $u^\star_{\text{DT}}$ and 1, shifting from increasing in ϵ to decreasing in ϵ at the extreme. In short, the regime beyond DT phase limit is different for SLOPE and when SLOPE enters this regime, breaking the monotonicity in ϵ may occur.

To be complete, we remark that the message conveyed by these two theorems does not contradict earlier results established for FDR control of SLOPE [Bog+13a; Bog+15a; Brz+19; KB20]. The crucial difference between the two sides arises from the linear sparsity assumed in the present paper, which is a clear departure from the much lower sparsity level considered in the literature. In this regard, our results complement the literature by extending our understanding of the inferential properties of the SLOPE method to an unchartered regime.

3.2.2 Breaking the Donoho–Tanner power limit

To better appreciate the trade-off results presented in Theorem 5 for SLOPE, it is instructive to compare them with the TPP and FDP trade-off for the Lasso, which is arguably the most popular method leveraging ℓ_1 regularization.

To put it into perspective, first recall some results concerning the optimal trade-off between the TPP and FDP for the Lasso. A surprising fact is that under the working assumptions,[2] the Lasso cannot achieve full power even with an arbitrarily large signal-to-noise ratio when $\delta < 1$ (that is, $\boldsymbol{X}$ is "fat") and the sparsity ratio ϵ is above a threshold, which we denote by $\epsilon^\star(\delta)$. The dependence of this value on δ is specified by the parametric equations

$$\delta = \frac{2\phi(s)}{2\phi(s) + s(2\Phi(s) - 1)}, \qquad \epsilon^\star = \frac{2\phi(s) - 2s\Phi(-s)}{2\phi(s) + s(2\Phi(s) - 1)} \tag{3.2.2}$$

for $s > 0$.[3] For simplicity, henceforth (δ, ϵ) is said to be in the *supercritical* regime if $\delta < 1, \epsilon > \epsilon^\star(\delta)$. Otherwise, it is in the *subcritical* regime when $\delta < 1, \epsilon \le \epsilon^\star(\delta)$, or $\delta \ge 1$ (that is, $\boldsymbol{X}$ is "thin"). In the supercritical regime, [SBC17] proved that the

[2] Note that, in the case of the Lasso, (A3) is replaced by the assumption that $\lambda > 0$ is a constant.

[3] In the compressed sensing literature, $\epsilon^\star$ corresponds to the sparsity level where the Donoho–Tanner phase transition occurs [DT09b; DT09a].

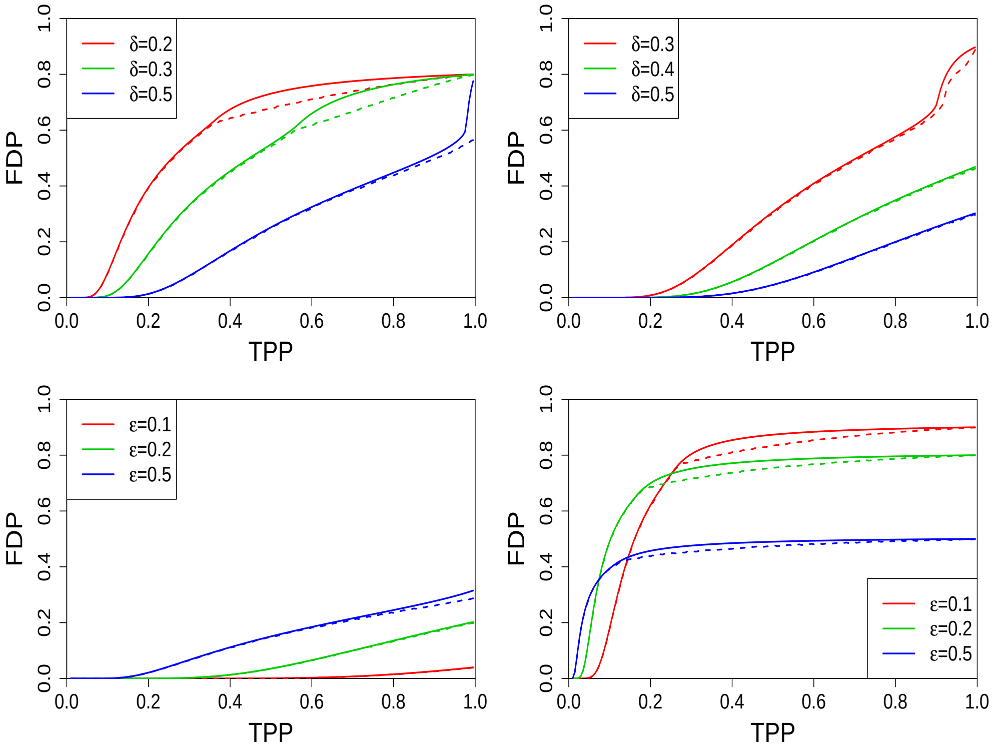


Figure 3.3: Examples of the SLOPE trade-off bounds $q^\star$ and $q_\star$ for different (δ, ϵ) pairs. Top-left: $\epsilon = 0.2$; top-right: $\epsilon = 0.1$; bottom-left: $\delta = 0.9$; bottom-right: $\delta = 0.1$. For a given δ, note that the trade-off for SLOPE is not monotone with respect to ϵ, which is a departure from the Lasso counterpart (see [SBC17, Figure 4]). Numerically, the upper and lower bounds seem to coincide when the TPP is below a threshold (see more details in Figure 3.5). To give more details, in one regime with $\delta = 0.1, \epsilon = 0.5$, the maximum gap between the upper and lower bounds $\max_u[q^\star(u) - q_\star(u)]$ is less than 0.0235; whereas in another regime with $\delta = 0.5, \epsilon = 0.1$, the maximum gap is always less than 0.0056.

highest achievable TPP of the Lasso, denoted $u^\star_{\mathrm{DT}}$, takes the form

$$u^\star_{\mathrm{DT}}(\delta,\epsilon) := 1 - \frac{(1-\delta)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)} < 1. \tag{3.2.3}$$

Throughout the paper, $u^\star_{\mathrm{DT}}$ is referred to as the *DT power limit*. For completeness, in the subcritical regime the Lasso can achieve any power level. As such, we formally set $u^\star_{\mathrm{DT}}(\delta,\epsilon) = 1$ when $\delta < 1, \epsilon \le \epsilon^\star(\delta)$, or $\delta \ge 1$.

This existing result, in conjunction with Theorem 5, immediately gives the following contrasting result concerning the Lasso and SLOPE. We use $\mathrm{TPP}_{\mathrm{Lasso}}(\boldsymbol{\beta},\lambda)$ and $\mathrm{FDP}_{\mathrm{Lasso}}(\boldsymbol{\beta},\lambda)$ to denote, respectively, the TPP and FDP of the Lasso with penalty parameter λ. Likewise, we use $\mathrm{TPP}_{\mathrm{SLOPE}}(\boldsymbol{\beta},\boldsymbol{\lambda})$ and $\mathrm{FDP}_{\mathrm{SLOPE}}(\boldsymbol{\beta},\boldsymbol{\lambda})$ to denote those of SLOPE as $\xi \to 0$.

Corollary 3.2.2 (SLOPE breaks the DT power limit)**.** *In the supercritical regime, the following conclusions hold under the working assumptions:*

(a) The power of the Lasso satisfies $\mathrm{TPP}_{\mathrm{Lasso}}(\boldsymbol{\beta},\lambda) < u^\star_{\mathrm{DT}}$ *with probability tending to one.*

(b) For any $0 \le u < 1$*, there exist a SLOPE regularization prior* Λ *and a signal prior* Π *such that* $\mathrm{TPP}_{\mathrm{SLOPE}}(\boldsymbol{\beta},\boldsymbol{\lambda}) > u$ *with probability tending to one.*

For illustration, Figure 3.1 in the introduction reflects this distinction between SLOPE and the Lasso with $u^\star_{\mathrm{DT}}(0.4, 0.7) = 0.4401$ in the left plot. Another illustration is Figure 3.1 right plot and Figure 3.4, which is vertically truncated at $u^\star_{\mathrm{DT}}(0.3, 0.2) = 0.5676$. Notice that the SLOPE breaks the DT power limit, i.e.

$u^{\star}_{\mathrm{DT}} < \mathrm{TPP}_{\mathrm{SLOPE}} < 1$, but still preserves non-trivial FDP, i.e. $\mathrm{FDP}_{\mathrm{SLOPE}} < 1 - \epsilon$, where $1-\epsilon$ is the FDP associated with the trivial procedure that selects all predictors.

Corollary 3.2.2 highlights the benefit of using sorted ℓ_1 regularization over the less flexible ℓ_1 regularization in terms of power. This sharp distinction persists no matter how large the effect sizes are and, therefore, it must be attributed to the flexibility of the SLOPE regularization sequence. As is well-known, the Lasso selects no more than n variables. Worse, a significant proportion of false variables are always interspersed on the Lasso path in the linear sparsity regime and, therefore, even though the Lasso can select up to $n > k$ variables, it would always miss a fraction of true variables, thereby imposing a limit on the power. In contrast, SLOPE does not bear the constraint that $\|\widehat{\boldsymbol{\beta}}\|_0 \leq n$ owing to the flexibility of its regularization sequence. In fact, the corresponding constraint for SLOPE is that the number of *unique* non-zero entries is no more than n [SC16]. This flexibility allows SLOPE to have arbitrarily high power regardless of the regime (δ, ϵ) belongs to.

Moving forward, we ask which regularization prior Λ and signal prior Π are "flexible" enough to enable SLOPE to break the DT power limit. To achieve desired flexibility, interestingly, it only requires a simple two-level regularization sequence for SLOPE. Consider the following *two-level* SLOPE regularization prior: given constants $a > b \geq 0$ and $0 < w < 1$, let $\Lambda_{a,b,w} = a$ with probability w and otherwise $\Lambda_{a,b,w} = b$. The SLOPE regularization sequence drawn from this prior takes the form

$$\boldsymbol{\lambda}_{a,b,w} := \Big(\underbrace{a, a, \cdots, a}_{\text{about } wp}, \underbrace{b, b, \cdots, b}_{\text{about } (1-w)p} \Big). \tag{3.2.4}$$

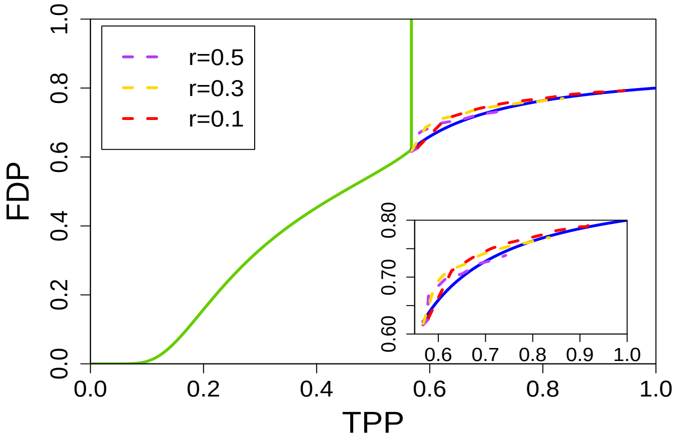


Figure 3.4: The Möbius part of the SLOPE trade-off upper bound $q^\star$. The solid curve denotes the upper bound specified by $(\delta, \epsilon) = (0.3, 0.2)$. The green line is the Lasso part of $q^\star$ and the blue one is the Möbius part. The numerical pairs of the TPP and FDP are obtained from experiments that are specified by the following parameters: $n = 300, p = 1000, \sigma^2 = 0$, signal prior $\Pi_M(\epsilon^\star/\epsilon)$ with $M = 10000$ in (3.2.5) (note that $\epsilon^\star(0.3) = 0.087$), and regularization prior $\boldsymbol{\lambda}_{\sqrt{M}, r\sqrt{M}, w}$ in (3.2.4) with varying w. Each pair is averaged over 50 independent trials.

Next, for any $M > 0$ and $0 \le \epsilon' \le 1$, define the following signal prior:

$$\Pi_M(\epsilon') := \begin{cases} M, & \text{w.p.} \quad \epsilon\epsilon' \\ M^{-1}, & \text{w.p.} \quad \epsilon - \epsilon\epsilon' \\ 0, & \text{w.p.} \quad 1 - \epsilon. \end{cases} \tag{3.2.5}$$

Henceforth in this paper, denote by $\boldsymbol{\beta}_M(\epsilon')$ the regression coefficients sampled from $\Pi_M(\epsilon')$.

Now we are ready to state the following result, which shows that SLOPE with the two-level regularization sequence can approach any point on the Möbius transformation (3.2.6) arbitrarily close. This result also partially specifies the upper bound $q^\star$ in Theorem 5 in the supercritical regime:

$$q^\star(u; \delta, \epsilon) = \frac{\epsilon(1-\epsilon)u - \epsilon^\star(1-\epsilon)}{\epsilon(1-\epsilon^\star)u - \epsilon^\star(1-\epsilon)} \tag{3.2.6}$$

for $u^\star_{\mathrm{DT}} \le u \le 1$ (above the DT power limit). Note that this function takes the form of a *Möbius transformation*. Notably, taking $u = 1$ gives $q^\star(1; \delta, \epsilon) = \dfrac{(\epsilon - \epsilon^\star)(1-\epsilon)}{\epsilon(1-\epsilon^\star) - \epsilon^\star(1-\epsilon)} = 1 - \epsilon$, which is the FDP achieved by the trivial procedure that simply selects all predictors..

Proposition 3.2.3. *For any $u^\star_{\mathrm{DT}} \le u \le 1$ in the supercritical regime, there exist w such that $\boldsymbol{\lambda}_{a,b,w}$ and $\boldsymbol{\beta}_M(\epsilon^\star/\epsilon)$ make SLOPE approach the point $(u, q^\star(u))$ in the sense*

$$\lim_{M\to\infty} \lim_{\xi\to 0} \lim_{n,p\to\infty} \left(\mathrm{TPP}_\xi(\boldsymbol{\beta}_M(\epsilon^\star/\epsilon), \boldsymbol{\lambda}_{a,b,w}), \mathrm{FDP}_\xi(\boldsymbol{\beta}_M(\epsilon^\star/\epsilon), \boldsymbol{\lambda}_{a,b,w})\right) \to (u, q^\star(u)),$$

where $a = \sqrt{M}, b = r\sqrt{M}$ for a certain value $0 \le r \le 1$.

Figure 3.4 provides a numerical example that corroborates this proposition.

This result in fact implies Theorem 5 for $u^\star_{\mathrm{DT}} \le u \le 1$ in the supercritical regime. Note that the first limit $\lim_{n,p\to\infty}$ is taken in the sense of convergence in probability. See more details in its proof in Section 3.5.1. It is worthwhile to mention that the three-component mixture (3.2.5) is considered in [SBC17] for the construction of favorable priors under sparsity constraint (see a generalization in [WYS20]). This mixture prior is used to ensure that the effect sizes are either very strong or very weak. In particular, Proposition 3.2.3 remains true if M and $1/M$ are replaced by any value diverging to infinity and any value converging to 0, respectively.

3.2.3 Below the Donoho–Tanner power limit

Next, we continue to interpret Theorem 4 and Theorem 5, but with a focus on the regime below the DT power limit.

First of all, the two right plots of Figure 3.5 show that the lower bound and the upper bound are very close to each other when $0 \le \mathrm{TPP} \le u^\star_{\mathrm{DT}}$ (recall that $u^\star_{\mathrm{DT}} = 1$ in the subcritical regime). As a matter of fact, the upper bound in this regime is given by [SBC17], which showed that under the working assumptions, there exists a function $q^\star_{\mathrm{Lasso}}(\cdot; \delta, \epsilon)$ such that

$$\mathrm{FDP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda) \ge q^\star_{\mathrm{Lasso}}(\mathrm{TPP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda); \delta, \epsilon) - 0.0001$$

holds with probability tending to one as $n, p \to \infty$. Here 0.0001 can be replaced by any arbitrarily small positive constant. Moreover, $q^\star_{\mathrm{Lasso}}$ is tight in the sense that the

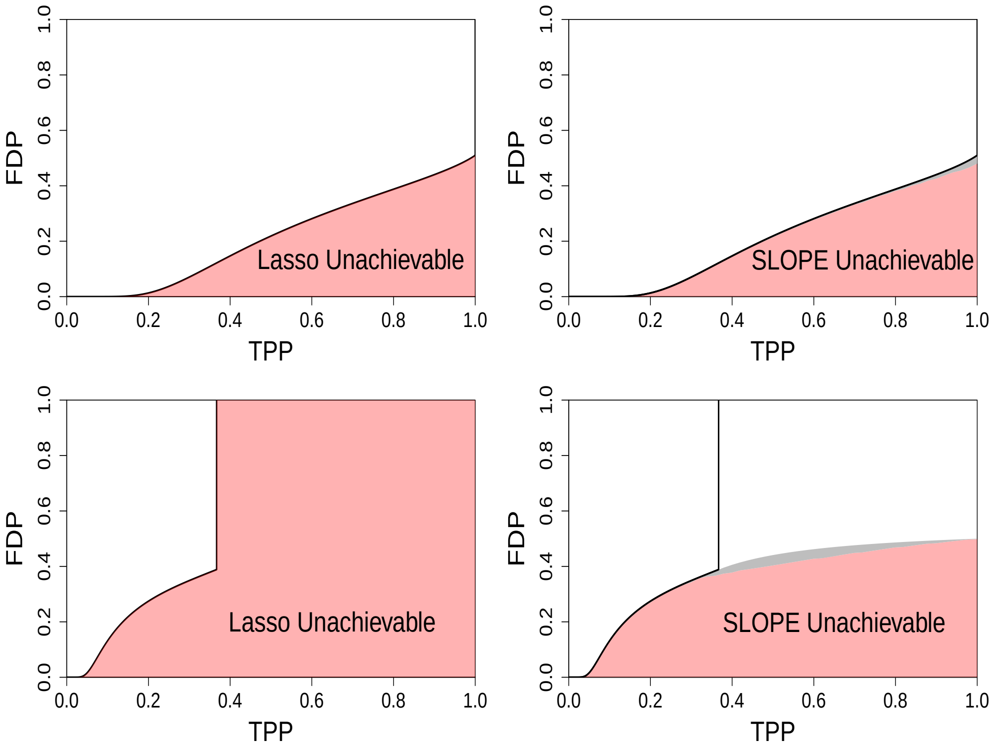


Figure 3.5: Examples of the TPP–FDP trade-off curve, with $(\delta, \epsilon) = (0.3, 0.2)$ on the top panel and $(0.3, 0.5)$ on the bottom. The left plot is the Lasso trade-off curve and the right plot describes the SLOPE trade-off gain. Neither the Lasso nor SLOPE can approach the red regions. The gray regions are sandwiched by the upper and lower bounds on the SLOPE trade-off.

Lasso can come arbitrarily close to any point on this curve by specifying a prior and a penalty parameter (see refined results in [WYS20]). Recognizing that the Lasso is an instance of SLOPE, the tightness of $q^\star_{\text{Lasso}}$ allows us to set $q^\star(u) = q^\star_{\text{Lasso}}(u)$ for $0 \le u \le u^\star_{\text{DT}}$. For information, letting $t^\star(u)$ be the largest positive root of the equation

$$\frac{2(1-\epsilon)\left[(1+x^2)\Phi(-x)-x\phi(x)\right]+\epsilon(1+x^2)-\delta}{\epsilon\left[(1+x^2)(1-2\Phi(-x))+2x\phi(x)\right]}=\frac{1-u}{1-2\Phi(-x)}, \tag{3.2.7}$$

we have

$$q^\star(u;\delta,\epsilon)=\begin{cases} q^\star_{\text{Lasso}}(u;\delta,\epsilon)=\dfrac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u))+\epsilon u}, & \text{if } u\le u^\star_{\text{DT}}(\delta,\epsilon),\\ \\ \dfrac{\epsilon(1-\epsilon)u-\epsilon^\star(1-\epsilon)}{\epsilon(1-\epsilon^\star)u-\epsilon^\star(1-\epsilon)}, & \text{if } u> u^\star_{\text{DT}}(\delta,\epsilon). \end{cases} \tag{3.2.8}$$

Above, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

Returning to the lower bound, in stark contrast, the situation becomes much more challenging. To be sure, to obtain a lower bound requires a good understanding of the superiority of sorted ℓ_1 regularization over its usual ℓ_1 counterpart. From a theoretical viewpoint, a major difficulty in the analysis of SLOPE arises from the *non-separability* of sorted ℓ_1 regularization. Note that the non-separability results from the sorting operation in the penalty term $\sum_{i=1}^p \lambda_i |b|_{(i)}$ in the SLOPE optimization program (3.1.2). To tackle this technical issue, in this paper we formulate the SLOPE trade-off as a calculus of variations problem and further cast it into infinite-dimensional convex optimization problems (see more details in Section 3.4).

In a nutshell, the flexibility of the SLOPE regularization sequence seems to only bring up limited improvement on the trade-off between the TPP and FDP below the DT power limit. However, the two right plots of Figure 3.5 present a noticeable departure between the two bounds when the TPP is slightly below $u^\star_{\mathrm{DT}}$. This departure is not an artifact of our analysis. Indeed, in Section 3.5.3 we provide a problem instance whose asymptotic TPP and FDP trade-off falls strictly between the upper bound and the lower bound:

$$q_\star(u) + 0.0001 < \mathrm{FDP} < q^\star(u) - 0.0001$$

and TPP $\approx u < u^\star_{\mathrm{DT}}$ with probability tending to one.

In passing, it is worthwhile mentioning that the performance in the high power regime is likely to carry more weight. In this sense, SLOPE overall outperforms the Lasso in terms of the trade-off between the TPP and FDP.

3.2.4 Instance-superiority of SLOPE

An important but less-emphasized point is that the above-mentioned comparison between the two methods is over the *lower envelope* of all the instance-specific problems. In this regard, it would be too quick to conclude that the flexibility of the penalty sequence does not gain any benefits for SLOPE, even at points where $q_\star(u)$ may be very close to $q^\star_{\mathrm{Lasso}}(u)$. Under the working hypotheses, indeed, we can formally prove that SLOPE is superior to the Lasso in the sense that we can always find a SLOPE regularization prior that strictly improves the Lasso on the same linear regression problem in terms of both model selection and estimation. Below, we let $\widehat{\boldsymbol{\beta}}$ denote the SLOPE or the Lasso estimate, and use the subscript to distinguish between the two methods.

Theorem 6. *Under the working assumptions, namely (A1), (A2), and (A3), given any bounded signal prior* Π *and any Lasso regularization parameter* $\lambda > 0$*, there exists a SLOPE regularization* Λ *such that the following inequalities hold simultaneously with probability tending to one:*

(a) $\mathrm{TPP}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) > \mathrm{TPP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda)$*;*

(b) $\mathrm{FDP}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) < \mathrm{FDP}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda)$*;*

(c) $\|\widehat{\boldsymbol{\beta}}_{\mathrm{SLOPE}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) - \boldsymbol{\beta}\|^2 < \|\widehat{\boldsymbol{\beta}}_{\mathrm{Lasso}}(\boldsymbol{\beta}, \lambda) - \boldsymbol{\beta}\|^2$*.*

This theorem shows that SLOPE can outperform the Lasso from both the model selection and the estimation viewpoints. The proof strategy of this theorem leverages a simple form of SLOPE regularization sequences that admits two distinct values (see (3.2.4)). Due to space constraints, we relegate the proof of this theorem to Section 3.7.1. It is somewhat surprising that such a simple two-level sequence can already exploit the benefits of using SLOPE over the Lasso. Having said this, from a practical standpoint it is not entirely clear how to select the optimal SLOPE penalty sequence to outperform the Lasso. We leave this important direction for future research.

As an aside, we remark that SLOPE has been shown to achieve the asymptotically exact minimax estimation when the sparsity level is much lower than considered in the present paper, largely owing to the adaptivity of sorted ℓ_1 regularization [SC16]. When it comes to the Lasso, however, cross validation is needed to select a penalty parameter that enables the Lasso to achieve similar estimation performance, which however is not exact as the constant is not sharp [BLT18].

3.3 Preliminaries for Proofs

In this section, we collect some preliminary results about SLOPE and AMP theory that allow us to get analytic expression of the TPP and FDP asymptotically. Informally speaking, the AMP theory given in [Bu+20a, Theorem 3] and [HL19a, Theorem 1] characterizes the *asymptotic* joint distribution of the SLOPE estimator $\widehat{\boldsymbol{\beta}}$ and the true regression coefficients $\boldsymbol{\beta}$. Notably, since $\widehat{\boldsymbol{\beta}}$ depends on $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, when studying asymptotic properties of $\widehat{\boldsymbol{\beta}}$, we will work with their asymptotic distributions (Π, Λ). In this way, we drop the dependence on finite-sample quantities like n, p and the sparsity level $|\{j : \beta_j \neq 0\}|$ and instead work with asymptotic quantities such as (δ, ϵ) henceforth.

To be specific, under pseudo-Lipschitz functions (see [Bu+20a, Definition 3.1]) on $(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, the asymptotic distribution of the SLOPE (including the Lasso) estimator $\widehat{\boldsymbol{\beta}}$, which we denote as $\widehat{\Pi}$, can be described as

$$\widehat{\Pi} \overset{\mathcal{D}}{=} \eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z), \tag{3.3.1}$$

where Z is an independent standard normal and the superscript $\mathcal{D}$ means "in distribution". We will refer to η (to be introduced in (3.3.5)) as the *limiting scalar function*, and (τ, A) is the unique solution to the *state evolution* and the *calibration* equations

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left(\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z) - \Pi\right)^2, \tag{3.3.2}$$

$$\Lambda \overset{\mathcal{D}}{=} \mathrm{A}\tau\left(1 - \frac{1}{\delta}\mathbb{E}\Big(\eta'_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z)\Big)\right). \tag{3.3.3}$$

In order to discuss properties of the limiting scalar function η, we first introduce the SLOPE proximal operator on $(\boldsymbol{y}, \boldsymbol{\theta}) \in \mathbb{R}^p \times \mathbb{R}^p$, where $\boldsymbol{\theta}$ is proportional to $\boldsymbol{\lambda}$ and $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \geq 0$ with at least one inequality. We define the proximal operator as

$$\mathrm{prox}_J(\boldsymbol{y}; \boldsymbol{\theta}) := \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \Big\{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{b}\|^2 + J_{\boldsymbol{\theta}}(\boldsymbol{b})\Big\}, \tag{3.3.4}$$

where $J_{\boldsymbol{\theta}}(\boldsymbol{b}) = \sum_{i=1}^p \theta_i |b|_{(i)}$. In the Lasso case when the penalty parameter is a constant, the proximal operator reduces to the soft-thresholding function:

$$\mathrm{prox}_J(\boldsymbol{y}; \theta) = \eta_{\mathrm{soft}}(\boldsymbol{y}; \theta) := \mathrm{sign}(\boldsymbol{y}) \cdot \max\{|\boldsymbol{y}| - \theta, 0\}.$$

Generally speaking, the SLOPE proximal operator in (3.3.4) is *adaptive* and *non-separable*, in the sense that an element of the output generally will depend on all elements of the input. As a concrete example, we obtain via Algorithm 2 that

$$\begin{aligned}&\mathrm{prox}_J([20, 13, 10, 6, 4]; [12, 10, 5, 5, 5])\\ =&\eta_{\mathrm{soft}}([20, 13, 10, 6, 4]; [12, 9, 6, 5, 5]) = [8, 4, 4, 1, 0].\end{aligned}$$

On the one hand, the adaptivity arises from the fact that larger penalties are applied to larger elements of the input. On the other hand, for example, two elements of input $[13, 10]$ are not directly thresholded by the penalty $[10, 5]$, but rather an averaging step is triggered by the existence of the other inputs, which gives an effective threshold of $[9, 6]$.

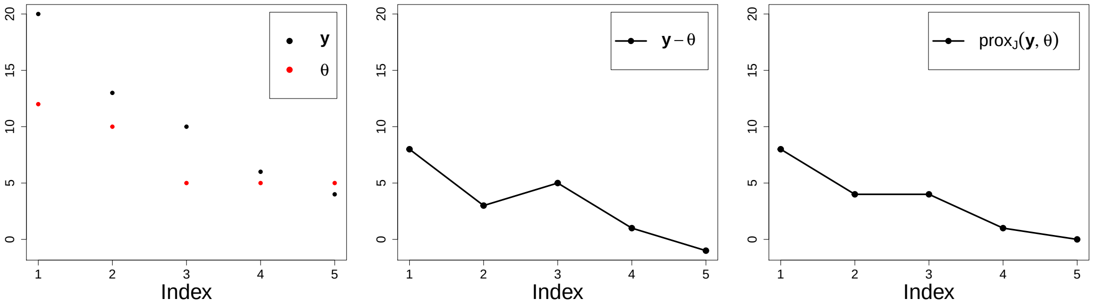


Figure 3.6: Illustration of how the SLOPE proximal operator can be interpreted as using an effective threshold. The leftmost figure plots two vectors $\boldsymbol{y}$ and $\boldsymbol{\theta}$. The middle image plots their difference $\boldsymbol{y} - \boldsymbol{\theta}$ and the rightmost image plots the output of the proximal operator $\text{prox}_J(\boldsymbol{y}; \boldsymbol{\theta})$.

Although the SLOPE proximal operator given in (3.3.4) is non-separable, nevertheless, as introduced in [HL19a, Proposition 1], the SLOPE proximal operator is *asymptotically separable*: for sequences $\{\boldsymbol{\theta}(p)\}$ and $\{\boldsymbol{v}(p)\}$ growing in p with empirical distributions that weakly converge to distributions Θ and V, respectively, there exists a limiting scalar function η (determined by Θ and V) such that as $p \to \infty$,

$$\frac{1}{p}\|\text{prox}_J(\boldsymbol{v}(p); \boldsymbol{\theta}(p)) - \eta_{V,\Theta}(\boldsymbol{v}(p))\|^2 \to 0. \tag{3.3.5}$$

The work in [HL19a] discusses many properties of this limiting scalar function, η. Indeed, it can be shown to be odd, increasing, Lipschitz continuous with constant

1 and that it applies coordinate-wise to $\boldsymbol{v}(p)$ (hence it is separable; see [HL19a, Proposition 2]). In more details, $\eta_{V,\Theta}(x)$ takes a scalar input, x, and performs soft-thresholding with a penalty adaptive to x in a way that depends on V and Θ, meaning the input-dependent penalty $\lambda_{V,\Theta}(x)$ such that $\eta_{V,\Theta}(x) = \eta_{\text{soft}}(x; \lambda_{V,\Theta}(x))$. More details on the adaptive penalty function that relates the SLOPE proximal operator to the soft-thresholding function can be found in Section 3.7.3.

We now discuss in more detail the so-called state evolution and calibration equations given in (3.3.2) and (3.3.3). We refer to A, which is defined implicitly via (3.3.3), as the *normalized penalty* distribution. Notice that A only differs from the original penalty distribution Λ by a constant factor. In fact, there exists a one-to-one mapping between A and Λ by [Bu+20a, Proposition 2.6], allowing one to analyze in either regime flexibly. In particular, when it is clear from the context, we will use (Π, Λ) and (π, A) interchangeably since there exists a bijective calibration between the original problem instance and the normalized one. Moreover, for a fixed Π, the quantity $\tau(\mathrm{A})$ can be uniquely derived from (3.3.2) and, as shown in [Bu+20a, Corollary 3.4], it can be used to characterize the estimation error via $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/p \to \delta(\tau^2 - \sigma^2)$. In this work, we will use τ as a factor to define the *normalized prior*,

$$\pi(\Pi, \mathrm{A}) := \Pi/\tau(\Pi, \Lambda).$$

and, in particular, when it is clear from the context, we will use (Π, Λ) and (π, A) interchangeably since there exists a bijective calibration between the original problem instance and the normalized one, provided by fixed point recursion for the state evolution and the calibration mappings, (3.3.2) and (3.3.3). We refer the interested readers to Section 3.7.2 for a discussion of many nice properties of this fixed point recursion, such as the explicit form of the divergence η'.

Under the characterization of the asymptotic SLOPE distribution of (3.3.1), we define $\mathrm{FDP}^{\infty}(\Pi,\Lambda)$ and $\mathrm{TPP}^{\infty}(\Pi,\Lambda)$ as the large system limits of FDP and TPP. The proof of convergence in probability is given in the next lemma.

Lemma 3.3.1. *Under the working assumptions, namely (A1), (A2), and (A3), for* $\xi \in \Xi$ *in* (3.2.1)*, the SLOPE estimator* $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ *with the penalty sequence* $\boldsymbol{\lambda}$ *satisfies*

$$\mathrm{FDP}_{\xi}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j = 0\}|}{|\{j : |\widehat{\beta}_j| > \xi\}|} \xrightarrow{P} \mathrm{FDP}^{\infty}_{\xi}(\Pi,\Lambda) := \frac{(1-\epsilon)\,\mathbb{P}\left(\left|\eta_{\pi+Z,\mathrm{A}}(Z)\right| > \xi\right)}{\mathbb{P}\left(\left|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)\right| > \xi\right)},$$

$$\mathrm{TPP}_{\xi}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|} \xrightarrow{P} \mathrm{TPP}^{\infty}_{\xi}(\Pi,\Lambda) := \mathbb{P}\left(\left|\eta_{\pi+Z,\mathrm{A}}(\pi^{\star}+Z)\right| > \xi\right),$$

where superscript P denotes convergence in probability, Z is a standard normal independent of Π*, and* (τ,A) *is the unique solution to the state evolution* (3.3.2) *and calibration* (3.3.3)*. Furthermore,* $\Pi^{\star} := (\Pi|\Pi \neq 0)$ *is the signal prior distribution of the non-zero elements.*

By continuity of probability measure, we obtain

$$\begin{aligned} \lim_{\xi\to 0}\mathrm{FDP}^{\infty}_{\xi}(\Pi,\Lambda) = \mathrm{FDP}^{\infty}(\Pi,\Lambda) &:= \frac{(1-\epsilon)\,\mathbb{P}\left(\eta_{\pi+Z,\mathrm{A}}(Z) \neq 0\right)}{\mathbb{P}\left(\eta_{\pi+Z,\mathrm{A}}(\pi+Z) \neq 0\right)}, \\ \lim_{\xi\to 0}\mathrm{TPP}^{\infty}_{\xi}(\Pi,\Lambda) = \mathrm{TPP}^{\infty}(\Pi,\Lambda) &:= \mathbb{P}\left(\eta_{\pi+Z,\mathrm{A}}(\pi^{\star}+Z) \neq 0\right). \end{aligned} \tag{3.3.6}$$

Here, $\pi = \Pi/\tau$ is the normalized prior distribution and $\pi^\star := \Pi^\star/\tau$. We give the proof of Lemma 3.3.1 in Section 3.7.4 by extending [Bog+13a, Theorem B.1].

Following the notions of FDP^∞ and TPP^∞ given in Lemma 3.3.1, we mathematically define the SLOPE trade-off curve as the envelope of all achievable SLOPE $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ pairs:

$$q_{\text{SLOPE}}(u;\delta,\epsilon) := \inf_{(\Pi,\Lambda)\colon \mathrm{TPP}^\infty(\Pi,\Lambda)=u} \mathrm{FDP}^\infty(\Pi,\Lambda).$$

To study the SLOPE trade-off, we will make use of a critical concept, the *zero-threshold* $\alpha(\Pi,\Lambda)$, which will be defined in Definition 3.4.1. Using the zero threshold, the limiting values in (3.3.6) can be simplified to

$$\begin{aligned}
\mathrm{TPP}^\infty(\Pi,\Lambda) &= \mathbb{P}(|\pi^\star + Z| > \alpha(\Pi,\Lambda)),\\
\mathrm{FDP}^\infty(\Pi,\Lambda) &= \frac{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda))}{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda)) + \epsilon\cdot \mathrm{TPP}^\infty(\Pi,\Lambda)}.
\end{aligned} \tag{3.3.7}$$

Note from the equations above that for fixed $\mathrm{TPP}^\infty = u$, the formula of FDP^∞ is decreasing in α. Therefore we consider the maximum of feasible zero-thresholds,

$$\alpha^\star(u) := \sup_{(\Pi,\Lambda)\colon \mathrm{TPP}^\infty=u} \alpha(\Pi,\Lambda),$$

in order to derive the minimum FDP^∞ on the SLOPE trade-off

$$q_{\text{SLOPE}}(u;\delta,\epsilon) := \frac{2(1-\epsilon)\Phi(-\alpha^\star(u))}{2(1-\epsilon)\Phi(-\alpha^\star(u)) + \epsilon u}. \tag{3.3.8}$$

3.4 Lower bound of SLOPE trade-off

The main purpose of this section is to provide a lower bound $q_\star$ for q_{SLOPE}. We accomplish this by (equivalently) giving an upper bound for $\alpha^\star(u)$ for *fixed* u, which we denote as $t_\star(u)$. As we shall see, in contrast to Lasso, our derivation for SLOPE requires non-standard tools from the calculus of variations and quadratic programming.

To construct the upper bound $t_\star(u)$, we examine the state evolution (3.3.2), which gives

$$\tau^2 \geq \frac{1}{\delta}\mathbb{E}\left(\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z)-\Pi\right)^2 = \frac{\tau^2}{\delta}\mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)-\pi\right)^2.$$

Rearranging the above inequality yields the state evolution condition

$$E(\Pi,\Lambda) := \mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)-\pi\right)^2 \leq \delta. \tag{3.4.1}$$

Here the quantity $E(\Pi,\Lambda)$ can be viewed as the asymptotic mean squared error between the SLOPE estimator and the truth, scaled by $1/\tau^2$, since $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|/p \to \tau^2 E(\Pi,\Lambda)$ in probability by [Bu+20a, Corollary 3.4].

Before we proceed, we first introduce an important (scalar) quantity that governs the sparsity, the TPP, and the FDP of the SLOPE estimator and will be used throughout the paper.

Definition 3.4.1. *Let* (Π,Λ) *be a pair of prior and penalty distributions (or, equivalently, the normalized* (π,A)*) and suppose* $\alpha(\Pi,\Lambda)$ *is a positive number such that* $\eta_{\pi+Z,\mathrm{A}}(x)=0$ *if and only if* $|x|\leq\alpha(\Pi,\Lambda)$. *Then we say that* $\alpha=\alpha(\Pi,\Lambda)$ *is the* zero-threshold.

Intuitively, the zero-threshold is a positive threshold, below which, the input is mapped to zero. Note that the necessary condition (3.4.1) sets the feasible domain of (π, A) pairs and thus prescribes limits to the zero-threshold α. In the Lasso case, the zero-threshold is indeed equivalent to the normalized penalty scalar A; but in SLOPE, it is a quantity derived from the normalized penalty distribution A in a highly nontrivial manner (see Proposition 3.7.5 for details).

Next, we state another useful definition. Recall from Section 3.3 that the limiting scalar function η of SLOPE is separable and assigns a different penalty to different inputs. We therefore define the *effective penalty function* accordingly.

Definition 3.4.2. *Given a normalized pair of prior and penalty* (π, A)*, the effective penalty function* $\widehat{\mathrm{A}}_{\text{eff}} : \mathbb{R} \to \mathbb{R}_+$ *is a function such that*

$$\eta_{\text{soft}}(x; \widehat{\mathrm{A}}_{\text{eff}}(x)) = \eta_{\pi+Z,\mathrm{A}}(x).$$

It is not hard to show that $\widehat{\mathrm{A}}_{\text{eff}}$ is well-defined. In fact, given $\eta_{\pi+Z,\mathrm{A}}$, we can represent $\widehat{\mathrm{A}}_{\text{eff}}$ via the zero-threshold from Definition 3.4.1, namely,

$$\widehat{\mathrm{A}}_{\text{eff}}(x) = \begin{cases} x - \eta_{\pi+Z,\mathrm{A}}(x) & \text{if } x > \alpha(\pi, \mathrm{A}), \\ -x + \eta_{\pi+Z,\mathrm{A}}(x) & \text{if } x < -\alpha(\pi, \mathrm{A}), \\ \alpha(\pi, \mathrm{A}) & \text{if } |x| < \alpha(\pi, \mathrm{A}). \end{cases}$$

Equipped with this effective penalty function, we can rewrite the state evolution condition (3.4.1) as

$$F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}] := \mathbb{E}\left(\eta_{\text{soft}}(\pi + Z; \widehat{\mathrm{A}}_{\text{eff}}(\pi + Z)) - \pi\right)^2 \le \delta,$$

in which the functional objective F_α is defined on the effective penalty function $\widehat{\mathrm{A}}_{\text{eff}}$ as well as the probability density function of $\pi^\star$. Note here that $\pi^\star$ and π determine each other uniquely since $\pi^\star := \pi | \pi \neq 0$. We provide an explicit expression for $F_\alpha[\widehat{\mathrm{A}}_{\mathrm{eff}}, p_{\pi^\star}]$ in (3.7.28).

Since the constraint (3.3.2) remains the same if π is replaced by $|\pi|$, we assume $\pi \geq 0$ without loss of generality. We minimize $F_\alpha[\widehat{\mathrm{A}}_{\mathrm{eff}}, p_{\pi^\star}]$ over the functional space of $(\widehat{\mathrm{A}}_{\mathrm{eff}}, p_{\pi^\star})$ through a relaxed variational problem:

$$\begin{aligned} &\min_{\mathrm{A}_{\mathrm{eff}}, \rho \geq 0} \quad F_\alpha[\mathrm{A}_{\mathrm{eff}}, \rho] \\ &\text{s.t.} \quad \mathrm{A}_{\mathrm{eff}}(\alpha) \geq \alpha, \mathrm{A}_{\mathrm{eff}}'(z) \geq 0 \text{ for all } z \geq \alpha, \\ &\int_0^\infty \rho(t) dt = 1, \int_0^\infty [\Phi(t-\alpha) + \Phi(-t-\alpha)]\rho(t) dt = u. \end{aligned} \tag{3.4.2}$$

Here the function $\mathrm{A}_{\mathrm{eff}}$ is implicitly defined on $[\alpha, \infty)$ as $\mathrm{A}_{\mathrm{eff}}(z) = \alpha$ for $0 \leq z < \alpha$ and ρ is a probability measure defined on $[0, \infty)$. We remark that the constraints for $\mathrm{A}_{\mathrm{eff}}$ in problem (3.4.2) are derived from the properties of $\widehat{\mathrm{A}}_{\mathrm{eff}}$ in Section 3.7.3, i.e. $\mathrm{A}_{\mathrm{eff}}' \geq 0$ comes from Fact 3.7.3 and the boundary condition $\mathrm{A}_{\mathrm{eff}}(\alpha) \geq \alpha$ comes from Proposition 3.7.5. Because some additional properties of $\widehat{\mathrm{A}}_{\mathrm{eff}}$ may have been excluded in the relaxation, these constraints are only necessary and may not be sufficient. Therefore,

$$\min_{(\widehat{\mathrm{A}}_{\mathrm{eff}}, p_{\pi^\star})} F_\alpha[\widehat{\mathrm{A}}_{\mathrm{eff}}, p_{\pi^\star}] \geq \min_{(\mathrm{A}_{\mathrm{eff}}, \rho)} F_\alpha[\mathrm{A}_{\mathrm{eff}}, \rho],$$

with the inequality possibly being strict, provided the first optimization problem (3.4.1) is solved subject to (i) $\widehat{\mathrm{A}}_{\mathrm{eff}}$ corresponds to the effective penalty in the limiting scalar function; and (ii) $p_{\pi^\star}$ is a probability density function such that $\mathrm{TPP}^\infty = \mathbb{P}(|\pi^\star + Z| > \alpha) = u$.

Leveraging the above relaxation (3.4.2), in order to lower bound q_{SLOPE} in (3.3.8), we can analogously define the maximum feasible zero-threshold $\alpha^\star(u)$ and upper bound it with $t_\star(u)$ as follows:

$$\alpha^\star(u) := \sup\left\{\alpha : \min_{(\Pi,\Lambda)} F_\alpha[\widehat{\mathrm{A}}_{\text{eff}}, p_{\pi^\star}] \le \delta\right\} \le t_\star(u) := \sup\left\{\alpha : \min_{(\mathrm{A}_{\text{eff}},\rho)} F_\alpha[\mathrm{A}_{\text{eff}}, \rho] \le \delta\right\}. \tag{3.4.3}$$

With these definitions in place, we are now in a position to describe the procedure to find the optimal prior and the optimal penalty in problem (3.4.2), given $\mathrm{TPP}^\infty = u$ and $\alpha(\Pi, \Lambda) = \alpha$.

3.4.1 Optimal prior is three-point prior

To solve problem (3.4.2), we must search over all possible distributions $\pi^\star$, which is generally infeasible. To overcome this obstacle, we use the concept of extreme points (i.e. points that do not lie on the line connecting any other two points of the same set) to show that the optimal $\pi^\star$ for problem (3.4.2) is a two-point distribution, having probability mass at only two non-negative (and possibly infinite) values (t_1, t_2). In doing so, we significantly reduce the search domain, from infinite dimensional to two-dimensional. Because π has an additional point mass at 0, the optimal prior π (that can achieve minimum FDP when accompanied with the properly chosen penalty) is a three-point prior taking values at $(0, t_1, t_2)$. We recall that the two-point $\pi^\star$ is consistent to the Lasso result in [SBC17, Section 2.5], where the optimal $\pi^\star$ is the infinity-or-nothing distribution with $t_1 = 0^+, t_2 = \infty$.

To see that $\pi^\star$ admits a two-point form, suppose that $(\mathrm{A}^*_{\text{eff}}, \rho^*)$ is the global minimum of problem (3.4.2). Then clearly ρ^* is also the global minimum of the following linear problem (3.4.4) with linear constraints.

$$\begin{aligned} \min_{\rho \geq 0} \quad & F_\alpha[\mathrm{A}^*_{\text{eff}}, \rho] \\ \text{s.t.} \quad & \int_0^\infty \rho(t)dt = 1, \int_0^\infty [\Phi(t-\alpha) + \Phi(-t-\alpha)]\rho(t)dt = u. \end{aligned} \tag{3.4.4}$$

Intuitively, since there are two constraints, we need two parameters (which will be t_1, t_2) to characterize the minimum. We formalize this intuition in the next lemma (proved in Appendix 3.7.7) and show that ρ^* indeed takes the form of a sum of two Dirac delta functions.

Lemma 3.4.3. *If ρ^* is a global minimum of problem* (3.4.4)*, then*

$$\rho^*(t) = p_1\delta(t-t_1) + p_2\delta(t-t_2)$$

for some constants p_1, p_2, t_1, t_2, and $p_1 + p_2 = 1$, $p_1, p_2 \geq 0$.

The above specific form of the optimal ρ^* allows us to search over all (t_1, t_2), each pair of which uniquely corresponds to either a single-point prior $\rho(t; t_1, t_2) = \delta(t-t_1)$ if $t_1 = t_2$, or a two-point prior by

$$\begin{aligned} \rho(t; t_1, t_2) &= p_1\delta(t-t_1) + p_2\delta(t-t_2), \\ p_1(t_1, t_2) &= \frac{u - [\Phi(t_2-\alpha) + \Phi(-t_2-\alpha)]}{[\Phi(t_1-\alpha) + \Phi(-t_1-\alpha)] - [\Phi(t_2-\alpha) + \Phi(-t_2-\alpha)]}, \\ p_2(t_1, t_2) &= 1 - p_1(t_1, t_2), \end{aligned} \tag{3.4.5}$$

where the last two equations come from the constraints in problem (3.4.4).

In light of Lemma 3.4.3, each pair (t_1, t_2) forms a different instantiation of problem (3.4.2), which will be problem (3.4.6) and whose optimal penalty is denoted by $\mathrm{A}^*_{\mathrm{eff}}(\cdot; t_1, t_2)$ so as to be explicitly dependent on (t_1, t_2). Before we proceed to optimize the penalty $\mathrm{A}_{\mathrm{eff}}(\cdot; t_1, t_2)$, we assure the skeptical reader that, doing a grid search on (t_1, t_2) and considering the minimal value of all programs (3.4.6) parameterized by (t_1, t_2) to be equivalent to the minimal value of problem (3.4.2), is indeed a valid approach. This claim is theoretically grounded by noting that $F_\alpha[\mathrm{A}^*_{\mathrm{eff}}(\cdot; t_1, t_2), \rho(\cdot; t_1, t_2)]$ is continuous in (t_1, t_2). Continuity can be seen from a perturbation analysis of the optimal value in problem (3.4.6). In our case, the perturbation analysis is not hard since the constraint is independent of (t_1, t_2) and F_α depends on $\mathrm{A}^*_{\mathrm{eff}}$ in a strongly-convex manner: a small perturbation in (t_1, t_2) only results in a small perturbation in $\mathrm{A}^*_{\mathrm{eff}}$ and thus in $F_\alpha[\mathrm{A}^*_{\mathrm{eff}}(\cdot; t_1, t_2), \rho(\cdot; t_1, t_2)]$. We refer the curious reader to a line of perturbation analysis for such optimization problems in [BS13; Sha92; BS98].

3.4.2 Characterizing optimal penalty analytically

By Lemma 3.4.3, we reduce the multivariate non-convex problem (3.4.2) to a set of univariate convex problems (3.4.6) over $\mathrm{A}_{\mathrm{eff}}$. In this section, we describe the optimal penalty function $\mathrm{A}^*_{\mathrm{eff}}(\cdot; t_1, t_2)$, which is the solution to the problem below:

$$\begin{aligned} \min_{\mathrm{A}_{\mathrm{eff}}} \quad & F_\alpha[\mathrm{A}_{\mathrm{eff}}, \rho(\cdot; t_1, t_2)] \\ \text{s.t.} \quad & \mathrm{A}_{\mathrm{eff}}(\alpha) \ge \alpha, \quad \mathrm{A}'_{\mathrm{eff}}(z) \ge 0 \text{ for all } z \ge \alpha. \end{aligned} \tag{3.4.6}$$

This is a quadratic problem with a non-holonomic constraint. To see this, we can expand the objective functional F_α from (3.7.28) and split it into a functional integral that involves A_{eff} and other terms which do not, i.e.

$$\begin{aligned} F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)] = \int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz + \epsilon p_1 t_1^2 \Big[\Phi(\alpha - t_1) - \Phi(-\alpha - t_1) \Big] \\ + \epsilon p_2 t_2^2 \Big[\Phi(\alpha - t_2) - \Phi(-\alpha - t_2) \Big]. \end{aligned}$$

This split changes our objective functional from $F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)]$ to the new functional $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ with

$$\begin{aligned} L(z, \mathrm{A}_{\text{eff}}) :&= 2(1-\epsilon)(z - \mathrm{A}_{\text{eff}}(z))^2 \phi(z) \\ &+ \epsilon p_1 \left(\Big(z - t_1 - \mathrm{A}_{\text{eff}}(z) \Big)^2 \phi(z - t_1) + \Big(-z - t_1 + \mathrm{A}_{\text{eff}}(z) \Big)^2 \phi(-z - t_1) \right) \\ &+ \epsilon p_2 \left(\Big(z - t_2 - \mathrm{A}_{\text{eff}}(z) \Big)^2 \phi(z - t_2) + \Big(-z - t_2 + \mathrm{A}_{\text{eff}}(z) \Big)^2 \phi(-z - t_2) \right). \end{aligned} \tag{3.4.7}$$

We will numerically optimize the functional $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ together with the constraints in problem (3.4.6). In addition, although we cannot derive the analytic form of $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$ from problem (3.4.6), we can still analytically characterize it at points z where the monotonicity constraint is non-binding (that is, when $\mathrm{A}^*_{\text{eff}}(\cdot; t_1, t_2)$ is strictly increasing in a neighborhood of z), as shown in Section 3.7.5.

3.4.3 Searching over optimal penalty numerically

To solve the functional optimization problem (3.4.6), we approximate it by a discrete optimization problem via Euler's finite difference method. Specifically, we approximate the function $L(z, \mathrm{A}_{\text{eff}})$ (and hence F_α) on a discretized uniform grid of z and solve the resulting quadratic programming problem with linear constraints.

To this end, we denote vectors $\boldsymbol{z} = [\alpha, \alpha+\Delta z, \alpha+2\Delta z, \cdots, \alpha+m\Delta z]$ and $\mathbf{A} = [\mathrm{A}_{\text{eff}}(\alpha), \mathrm{A}_{\text{eff}}(\alpha+\Delta z), \cdots, \mathrm{A}_{\text{eff}}(\alpha+m\Delta z)]$ for some small Δz and large m. Then problem (3.4.6) is discretized into the convex quadratic program

$$\begin{aligned} \min_{\mathbf{A}_{\text{eff}}} \quad & \bar{F}_\alpha(\mathbf{A}; t_1, t_2) \\ \text{s.t.} \quad & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \mathbf{A} \geq \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \end{aligned} \tag{3.4.8}$$

in which the new objective $\bar{F}_\alpha(\mathbf{A}; t_1, t_2)$ (derived in (3.7.29) and also presented below) is the discretized objective of $F_\alpha[\mathrm{A}_{\text{eff}}, \rho(\cdot; t_1, t_2)]$ from problem (3.4.6).

As $\Delta z \to 0$ and $m \to \infty$, problem (3.4.8) recovers problem (3.4.6) by well-known convergence theory for Euler's finite difference method. To simplify the exposition, we write the objective of problem (3.4.8) in matrix and vector notation as follows:

$$\mathbf{Q} = \operatorname{diag}\left(2(1-\epsilon)\phi(\boldsymbol{z}) + \epsilon \sum_{j=1,2} p_j \Big[\phi(\boldsymbol{z} - t_j) + \phi(-\boldsymbol{z} - t_j)\Big]\right),$$

$$\mathbf{d} = 2(1-\epsilon)\boldsymbol{z}\phi(\boldsymbol{z}) + \epsilon \sum_{j=1,2} p_j \Big[(\boldsymbol{z} - t_j)\phi(\boldsymbol{z} - t_j) + (\boldsymbol{z} + t_j)\phi(\boldsymbol{z} + t_j)\Big],$$

and observe that

$$\bar{F}_\alpha(\mathbf{A}; t_1, t_2) = (\mathbf{A}^\top \mathbf{Q} \mathbf{A} - 2\mathbf{A}^\top \mathbf{d})\Delta z + \epsilon \sum_{j=1,2} p_j t_j^2 \Big[\Phi(\alpha - t_j) - \Phi(-\alpha - t_j)\Big].$$

The discretized problem (3.4.8) is equivalent to a standard quadratic programming problem, whose objective is the discrete version of $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}})dz$ in (3.4.7),

$$\begin{aligned} \min_{\mathbf{A}} \quad & \frac{1}{2}\mathbf{A}^\top \mathbf{Q}\mathbf{A} - \mathbf{A}^\top \mathbf{d} \\ \text{s.t.} \quad & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \mathbf{A} \geq \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned} \tag{3.4.9}$$

3.4.4 Solving the quadratic program

Here we briefly discuss our numerical approach to solving the quadratic program (3.4.9). Generally speaking, quadratic programming problems do not admit closed-form solutions. However, they can be efficiently solved by classical numerical methods, including the interior point method [Dik67; SNW12], active set method [MY88; Fer+14] and other dual methods [GI83; FW56]. In this work, we use the dual method in [GI83], as implemented in the R library `quadprog`, to solve (3.4.9).

We remark that problem (3.4.9) is not the only way to discretize problem (3.4.6) and we now mention other approaches, which can result in better discretization accuracy. The discretization of problem (3.4.6) contains two parts: (i) a numerical integration to approximate the objective and (ii) a numerical differentiation to approximate the constraints.

When formulating the quadratic programming problem (3.4.9), we chose to apply the left endpoint rule to approximate the objective integral $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ in (3.4.7) by $(\mathbf{A}^\top \mathbf{Q}\mathbf{A} - 2\mathbf{A}^\top \mathbf{d})\Delta z$, as well as the backward finite difference (with first-order accuracy) to describe the constraint $\mathrm{A}'_{\text{eff}}(z) \geq 0$. Alternatively, one can use different numerical quadratures to approximate the integral $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$ or use a change of variable to approximate a different integral. We can also apply different finite differences to discretize the monotonicity constraint in problem (3.4.6).

Numerical integration to approximate the objective

More specifically, for the approximation of the objective in problem (3.4.6), we can alternatively apply numerical quadratures such as the trapezoid rule, Simpson's rule, or Gauss-Laguerre quadrature [SZ49] to improve the numerical integration for $\int_\alpha^\infty L(z, \mathrm{A}_{\text{eff}}) dz$. On the other hand, we may use a change of variable $z = \frac{x}{1-x} + \alpha$ to transform the integral $\int_\alpha^\infty L(z) dz$ over an infinite interval $[\alpha, \infty)$ to the integral $\int_0^1 L\left(\frac{x}{1-x} + \alpha\right) \frac{dx}{(1-x)^2}$ over a finite interval $[0, 1]$. This new integral can then be approximated by the same left endpoint rule (or other rules) but with different $\mathbf{Q}$ and $\mathbf{d}$.

Numerical differentiation to approximate the constraints

As for the monotonicity constraint $\mathrm{A}'_{\text{eff}}(z) \geq 0$, we may alternatively use other difference methods, e.g. the central difference, or higher-order accuracies. Doing so will result in a different matrix that left-multiplies $\mathbf{A}$ in the constraint of (3.4.9).

In conclusion, different numerical integration and differentiation schemes will lead to other formulations of the quadratic programming that are different from (3.4.9). We do not pursue these additional numerical aspects in the present work.

3.4.5 Summary

To summarize everything so far, the procedure of finding the lower bound $q_\star(u)$ involves the following steps: fixing $\mathrm{TPP}^\infty = u$, we search over a line of zero-thresholds $\{\alpha\}$; for each α, we search over a two-dimensional finite grid of (t_1, t_2), each pair defining a standard quadratic programming problem (3.4.9); we then solve the quadratic problem and reject (t_1, t_2) if the minimal value of the equivalent problem (3.4.8) is larger than δ; if all (t_1, t_2) are rejected, then the current zero-threshold α is too large to be valid. We set the largest valid zero-threshold as $t_\star(u)$ in (3.4.3) and write the lower bound of the FDP^∞ as $q_\star(u) = \frac{2(1-\epsilon)\Phi(-t_\star(u))}{2(1-\epsilon)\Phi(-t_\star(u))+\epsilon u}$. Note that $q_\star(u) > 0$ for any possible $t_\star(u)$.

We note that, in addition to minimizing FDP at a fixed TPP over all penalty-prior pairs, our quadratic programming approach also works when the prior Π is fixed. The fixed prior scenario has been extensively studied in [HL19a], who optimize over the limiting scalar function η while we are optimizing over the penalty function $\mathrm{A}_{\mathrm{eff}}$. Our approach adds a new angle that can be algorithmically more efficient. We defer the details of the procedure to Section 3.7.9.

3.4.6 Differences between SLOPE and Lasso

We end this section by discussing why deriving the SLOPE trade-off is fundamentally more complicated than the Lasso case. We highlight that the variational problem (3.4.2) is *non-convex*, even though it is convex with respect to each variable A_{eff} and ρ (i.e. it is bi-convex but non-convex). Generally speaking, approximate solutions to non-convex problems are not accompanied by theoretical guarantees, except for some special cases. Our bi-convex problem (3.4.2) cannot be solved by alternating descent, namely, fixing one variable, optimizing over the other and then alternating. Furthermore, our constraints only add another layer of complexity to the problem: in particular, the monotonicity constraint of A_{eff} is non-holonomic (i.e. the constraint $\mathrm{A}_{\text{eff}}' \geq 0$ does not depend explicitly on A_{eff}).

More precisely, the difficulty in directly solving the problem (3.4.2) is two-fold. The first difficulty lies in the search for the optimal penalty. For the Lasso case, the penalty distribution A and the penalty function $\widehat{\mathrm{A}}_{\text{eff}}$ are not adaptive to the input and hence they both equal the zero-threshold α. Therefore, we can perform a grid search on $\mathrm{A} \in \mathbb{R}$ and simply optimize over ρ. However, for SLOPE, the penalty $\widehat{\mathrm{A}}_{\text{eff}}$ is a *function* and hence it is intractable to search over the SLOPE penalty function space. The functional form of the penalty is the reason we must rely on the calculus of variations to study the associated optimization problem.

To demonstrate the second difficulty, we again consider the convex problem (3.4.4), which is over the probability density function ρ, assuming the optimal penalty $\mathrm{A}^*_{\mathrm{eff}}$ has been obtained. In the Lasso case, it was shown in [SBC17, Equation (C.2)] that the optimal $\pi^\star$ is the infinity-or-nothing distribution: $\mathbb{P}(\pi^\star = 0) = 1 - \epsilon'$ and $\mathbb{P}(\pi^\star = \infty) = \epsilon'$. In other words, given $\mathrm{A}^*_{\mathrm{eff}}$, we can easily derive the optimal ρ. However, a key concavity result in [SBC17, Lemma C.1], which holds for Lasso and determines the optimal $\pi^\star$, unfortunately breaks in SLOPE. Therefore, the optimal form of $\pi^\star$ is inaccessible for SLOPE with existing tools, even if the optimal penalty $\mathrm{A}^*_{\mathrm{eff}}$ is known.

3.5 Upper bound of SLOPE trade-off

In this section, we rigorously analyze the SLOPE trade-off upper boundary curve $q^\star$ (defined in (3.2.8)). As stated in Theorem 5, $q^\star$ takes two forms: below the DT power limit, i.e. when $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$ for $u^\star_{\mathrm{DT}}$ defined in (3.2.3), we have $q^\star = q^\star_{\mathrm{Lasso}}$, and beyond the DT power limit, $q^\star$ is a Möbius curve.

We start by giving some intuition for why the domain of $q^\star$ is the entire interval $[0, 1]$, whereas, the Lasso trade-off curve is only defined on $[0, u^\star_{\mathrm{DT}})$. Intuitively, SLOPE is capable of overcoming the DT power limit and achieving 100% TPP since it is possible for SLOPE estimators to select all p features, hence, by the definition of TPP (see Section 3.2.1), one can find a completely dense SLOPE estimator whose TPP is automatically 1. This is not true for the Lasso, since it can select *at most* n out of p features. The corresponding constraint for the SLOPE estimator follows from the AMP calibration in (3.3.3) (discussed in detail in Section 3.7.2), namely it says that the number of *unique absolute values* in the entries of the SLOPE estimator is at most n out of p. However, this does not directly constrain the sparsity of SLOPE estimator, and thus it can still be dense. In other words, the SLOPE estimator always satisfies

$$\text{number of unique non-zero magnitude } |\widehat{\beta}_i| \text{ in } \widehat{\boldsymbol{\beta}}(p) \leq n. \tag{3.5.1}$$

Notice that, in the Lasso sub-case, the above implies a direct sparsity constraint $|\{i : \widehat{\beta}_i \neq 0\}| \leq n$ as just discussed, since all non-zero entries in Lasso have unique magnitudes.

With this intuition, we are prepared to prove Theorem 5 and show that $q^\star$ indeed serves as an upper bound of q_{SLOPE}. Following Proposition 3.2.3, we have the tightness of $q^\star$ when $u \geq u^\star_{\text{DT}}$. We will further discuss the proof of Proposition 3.2.3 in Section 3.5.1, but leave the full details for Section 3.7.4. The tightness of $q^\star$ when $u < u^\star_{\text{DT}}$ follows from the existing tightness result on the Lasso trade-off (see [SBC17, Section 2.5]), since the Lasso is a sub-case of SLOPE and $q^\star$ matches the Lasso trade-off curve for $u < u^\star_{\text{DT}}$. Hence, we have the corollary below.

Corollary 3.5.1. *For any $0 \leq u \leq 1$, there exists an $\epsilon' \in [0, \epsilon^\star/\epsilon]$, and values $r(u) \in [0,1]$ and $w(u) \in [0,1]$, both depending on u, such that the penalty $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\sqrt{M}, r(u)\sqrt{M}, w(u)}$ (defined in (3.2.4)) and the prior $\boldsymbol{\beta}_M(\epsilon')$ (defined in (3.2.5)) make SLOPE approach the point $(u, q^\star(u))$ in the sense*

$$\lim_{M\to\infty} \lim_{\xi\to 0} \lim_{n,p\to\infty} (\text{TPP}_\xi(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda}), \text{FDP}_\xi(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda})) \to (u, q^\star(u)).$$

Moreover, when $u < u^\star_{\mathrm{DT}}$*, we can set* $r(u) = 1$*, without specifying* $w(u)$*, and* $\epsilon' = \epsilon'(u)$ *will also depend on* u*. When* $u \geq u^\star_{\mathrm{DT}}$*, we fix* $\epsilon' = \epsilon^\star/\epsilon$ *and set* $r(u)$ *via* (3.5.2) *and* $w(u)$ *via* (3.5.3) *below.*

An interesting aspect of this result is that there are two different strategies for attaining $q^\star(u)$, depending on whether $\mathrm{TPP}^\infty = u$ is above or below the DT power limit. In both cases, we use a two-level penalty $\boldsymbol{\lambda}_{\sqrt{M},r(u)\sqrt{M},w(u)}$ and a sparse prior (see (3.2.5)) with very small and very large non-zeros. However, when $\mathrm{TPP}^\infty = u < u^\star_{\mathrm{DT}}$, the strategy for attaining $q^\star(u)$ is to vary the proportion of strong signals (which equals $\epsilon\epsilon'$ and ϵ' varies with u), but when $u \geq u^\star_{\mathrm{DT}}$, sharpness in the Möbius part of $q^\star$ is attained by keeping the sequence of priors fixed and instead tuning the ratio between strong and weak penalties.

The sharpness result of Corollary 3.5.1 shows that over the entire domain, $q^\star$ is arbitrarily closely achievable, thus, $q^\star(u)$ must serve as the upper bound of the minimum FDP^∞, $q_{\mathrm{SLOPE}}(u)$, hence we have completed the proof of Theorem 5.

3.5.1 Möbius upper bound is achievable

In this section, we will sketch the proof of Proposition 3.2.3, which is used to prove Corollary 3.5.1 in the regime $u \geq u^\star_{\mathrm{DT}}$. To complement Proposition 3.2.3 and Corollary 3.5.1, for concreteness, we give a specific prior and penalty pair in (3.2.6) that approaches $q^\star(u)$ when $u \geq u^\star_{\mathrm{DT}}$. The fully rigorous proof of Proposition 3.2.3, together with the derivation of (r, w), is given in Section 3.7.4.

Before we sketch the proof, we will provide some intuition for what makes the specific choice of priors and penalties behave effectively in terms of reducing the FDP^∞ while still driving TPP^∞ to 1, in order that we are able to approach $q^\star(u)$ for all $u \geq u^\star_{\mathrm{DT}}$. We remind the reader that, because there is a one-to-one correspondence between original instance (Π, Λ) and the normalized (π, A), we will use the two notations interchangeably.

First, for fixed $\mathrm{TPP}^\infty = u$, we can reduce the FDP^∞ through a smart use of the priors defined in (3.2.5), where many elements equal 0 exactly, while some non-zero elements are small (equal to $1/M$) and others large (equal to M) with M tending to ∞. This is the same strategy as was used for demonstrating the achievability of the Lasso curve in [SBC17], and the intuition that we present here is based on this analysis. Mathematically speaking, for the Lasso, [SBC17, Lemma C.1] revealed a concave relationship in Π between the normalized estimation error $E(\Pi, \Lambda) = \mathbb{E}(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) - \pi)^2$ in (3.4.1) and the sparsity $\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) \neq 0)$, which also depends on the pair (Π, Λ). The idea is that minimizing FDP^∞ corresponds to minimizing the sparsity (this can be seen, for example, by the relationship in (3.5.11) where $\kappa(\Pi, \Lambda)$ denotes the sparsity). Therefore, to find a prior Π that satisfies the state evolution condition (3.4.1), while minimizing the sparsity, the optimal (normalized) distribution for the non-zero elements, $\pi^\star$, for the Lasso case has probability masses concentrated at the endpoints of the domain, namely 0^+ and ∞. In this way, the form of the signal prior Π contributes to reducing the FDP^∞ by mixing the weak effects β_i with the zero effects.

Combining the priors discussed above, with a special subset of the possible penalties, namely the two-level penalties defined in (3.2.4), we are able to reduce the FDP^∞ while still increasing the TPP^∞ to its maximum value of 1, hence attaining $q^\star(u)$ for all $u \geq u^\star_{\text{DT}}$. Interestingly, the fact that SLOPE can do this, is through its penalty, which mixes the weak predictors $\widehat{\beta}_i$ and the zero predictors (see Figure 3.10). This mix-up is in fact triggered by the averaging step in the SLOPE proximal operator (see Algorithm 2; the averaging is determined by the sorted ℓ_1 norm in the SLOPE problem), which creates non-zero magnitudes that are shared by some predictors and hence maintains the quota of unique magnitudes in (3.5.1). As a consequence, the SLOPE estimator can overcome the DT power limit (and reach higher TPP^∞) without violating the uniqueness constraint (3.5.1) on its magnitudes.

When constructing the two-level penalties just discussed, we must choose a pair (r, w) that, respectively, defines the downweighting of the $\sqrt{M}$ used for the smaller penalty and the proportion of penalties getting each value. Concretely speaking, in Proposition 3.2.3 and Corollary 3.5.1, we set

$$r(u) = \Phi^{-1}\left(\frac{2\epsilon - \epsilon^\star - \epsilon u}{2(\epsilon - \epsilon^\star)}\right) / t^\star(u^\star_{\text{DT}}). \tag{3.5.2}$$

where $\epsilon^\star$ and $u^\star_{\text{DT}}$ define the DT power limit and are given in (3.2.2)-(3.2.3) and $t^\star$ is defined in (3.2.7). Moreover,

$$w(u) = \epsilon^\star + \tfrac{2(1-\epsilon^\star)}{1-r}\left[\Phi(-t^\star(u^\star_{\text{DT}})) - r\Phi(-rt^\star(u^\star_{\text{DT}})) - \tfrac{\phi(-t^\star(u^\star_{\text{DT}})) - \phi(-rt^\star(u^\star_{\text{DT}}))}{t^\star(u^\star_{\text{DT}})}\right], \tag{3.5.3}$$

where r in the above is shorthand for the $r(u)$ from (3.5.2).

Without going into details, the key reason for choosing such pair (r, w) is so that the sequence of two-level penalties have two different penalization effects: for one, the SLOPE estimator $\eta_{\pi+Z,A}(\pi + Z)$ is equivalent to a Lasso estimator $\eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}}))$ in the sense of (3.5.4); for the other, the SLOPE estimator is equivalent to a different Lasso estimator $\eta_{\text{soft}}(\pi + Z; rt^\star(u^\star_{\text{DT}}))$ in the sense of (3.5.5).

To be precise, it can be shown that

$$\eta_{\pi+Z,\mathrm{A}}(\pi + Z) \overset{P}{=} \eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}})),$$

and

$$\mathbb{E}(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) - \pi)^2 = \mathbb{E}(\eta_{\text{soft}}(\pi + Z; t^\star(u^\star_{\text{DT}})) - \pi)^2, \tag{3.5.4}$$

so when considering the asymptotic magnitude of the elements of the SLOPE estimator, or its asymptotic estimation error (3.4.1), we can analyze the limiting scalar function instead using a soft-thresholding function with threshold given by $t^\star(u^\star_{\text{DT}})$. Moreover, this implies that SLOPE satisfies the state evolution constraint (3.4.1) in a similar way to how the Lasso satisfies its corresponding state evolution constraint.

However, analysis of the asymptotic sparsity of the SLOPE estimator or of its asymptotic TPP and FDP, relies on the fact that one can prove

$$\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) \neq 0) = \mathbb{P}(\eta_{\text{soft}}(\pi + Z; rt^\star(u^\star_{\text{DT}})) \neq 0), \tag{3.5.5}$$

Hence, again, instead of analyzing the limiting scalar function one can analyze a soft-thresholding function, but now with a smaller threshold given by $rt^\star(u^\star_{\mathrm{DT}})$ for some $0 \le r \le 1$ defined in (3.5.2). Reducing the threshold in this way functions to improve the attainable TPP–FDP over the comparable Lasso problem by allowing more elements in the estimate with non-zero values. We visualize the above claims in Figure 3.10(d).

Essentially, the state evolution condition (3.4.1) must always hold, but it uses the larger pseudo zero-threshold $t^\star(u^\star_{\mathrm{DT}})$, while inference is conducted on the true, but smaller, zero-threshold $rt^\star(u^\star_{\mathrm{DT}})$. In this way, we can extend attainability of $q^\star_{\mathrm{Lasso}}$ to attainability $q^\star$, while still working within the state evolution constraint (3.4.1).

3.5.2 Infinity-or-nothing prior has FDP above upper bound

The goal of this section is to provide some intuition for the Möbius form of the curve $q^\star(u)$ when u is larger than the DT power limit. This will be done by demonstrating that, in the case of infinity-or-nothing priors, with a special subset of penalties, the SLOPE FDP^∞ is always above $q^\star$ in Proposition 3.5.2. This also motivates the achievability results of Section 3.5.1, as the proof given in Section 3.5.1 essentially tries to construct prior penalty pairs such that the inequality in Proposition 3.5.2 becomes an equality. While we only consider infinity-or-nothing priors here, we remark that in the Lasso case these are actually the *optimal* priors (ses [SBC17, Section 2.5]), meaning that they achieve the minimum FDP^{∞} given TPP^{∞}.

Proposition 3.5.2. *Under the working assumptions, namely (A1), (A2), and (A3), for $\xi \in \Xi$ in (3.2.1), assuming that $\boldsymbol{\beta}$ is sampled i.i.d. from (3.2.5) for any $\epsilon' \in [0,1]$, $M \to \infty$, and that $\boldsymbol{\lambda}$ is the order statistics of i.i.d. realization of a non-negative Λ with $\mathbb{P}(\Lambda = \max \Lambda) \geq \epsilon\epsilon'$, the following inequality holds with probability tending to one:*

$$\mathrm{FDP}_{\xi}(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda}) \geq q^{\star}\left(\mathrm{TPP}_{\xi}(\boldsymbol{\beta}_M(\epsilon'), \boldsymbol{\lambda}); \delta, \epsilon\right) - c_{\xi}.$$

for some positive constant c_{ξ} which tends to 0 as $\xi \to 0$.

Proof of Proposition 3.5.2. As in Section 3.4, we assume $\pi \geq 0$ without loss of generality since the analysis holds if we replace π by $|\pi|$. Consider a *subset of priors*, namely the infinity-or-nothing priors: for some $\epsilon' \in [0,1]$,

$$\pi_{\infty}(\epsilon') = \begin{cases} \infty & \text{w.p. } \epsilon\epsilon', \\ \\ 0 & \text{w.p. } 1 - \epsilon\epsilon'. \end{cases} \tag{3.5.6}$$

Although the infinity-or-nothing prior in (3.5.6) does not satisfy the assumption (A2) that $\mathbb{P}(\Pi \neq 0) = \mathbb{P}(\pi \neq 0) = \epsilon$, this does not affect our discussion[4].

In fact, as demonstrated by Lemma 3.5.3 below, for infinity-or-nothing priors, the state evolution constraint (3.4.1) guarantees that $\epsilon' \leq \epsilon^{\star}/\epsilon$. Since $\epsilon^{\star}$ is the same for the Lasso and SLOPE, this means that the maximum proportion of ∞ signals in the infinity-or-nothing prior is the same for both as well.

[4] The infinity-or-nothing prior can be approximated arbitrarily closely by a sequence of priors that satisfy the assumption. For example, let $M \to \infty$ and consider $\pi_M(\epsilon')$ defined in (3.2.5).

Lemma 3.5.3. *Under assumptions in Proposition 3.5.2, we must have* $\epsilon' \in [0, \epsilon^\star/\epsilon]$.

The proof of Lemma 3.5.3 is given in Section 3.7.4. It turns out that the DT threshold $\epsilon^\star$ plays an important role in understanding the relationship between the sparsity and TPP^∞. Before illustrating this relationship, we introduce the concept of *sparsity*. In a finite dimension, the sparsity of SLOPE estimator is $|\{j : \widehat{\beta}_j \neq 0\}|$. However, as $p \to \infty$, the count of non-zeros will also go to infinity, meaning a quantity like $\lim_p |\{j : \widehat{\beta}_j \neq 0\}|$ is not well-defined. Therefore we introduce the *asymptotic sparsity* of the SLOPE estimator via the distributional characterization in (3.3.1), denoting the limit in probability by plim,

$$\kappa(\Pi, \Lambda) := \mathbb{P}\left(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) \neq 0\right) = \mathbb{P}\left(\widehat{\Pi} \neq 0\right) = \operatorname{plim} |\{j : \widehat{\beta}_j \neq 0\}|/p. \quad (3.5.7)$$

Making use of the DT threshold $\epsilon^\star(\delta)$, we show in Lemma 3.5.4 that the sparsity $\kappa(\Pi, \Lambda)$ sets an upper bound on achievable TPP^∞.

Lemma 3.5.4. *Consider SLOPE based on the pair* (Π, Λ) *with* Π *from* (3.2.5) *and set* $M \to \infty$*. Then with the asymptotic sparsity* $0 \le \kappa(\Pi, \Lambda) < 1$*, we have* $\mathrm{TPP}^\infty(\Pi, \Lambda) \le u^\star(\kappa(\Pi, \Lambda); \epsilon, \delta)$ *where*

$$u^\star(\kappa; \epsilon, \delta) := \begin{cases} 1 - \frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}, & \text{if } \delta < 1 \text{ and } \epsilon > \epsilon^\star(\delta), \\ 1, & \text{otherwise.} \end{cases} \quad (3.5.8)$$

Proof of Lemma 3.5.4. We will only prove $\mathrm{TPP}^\infty(\Pi, \Lambda) \le 1 - \frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}$ when $\delta < 1$ and $\epsilon > \epsilon^\star(\delta)$. We note that the bound on $u^\star$ given in (3.5.8) when $\delta \geq 1$ or $\epsilon \le \epsilon^\star(\delta)$ is trivial since, by definition, $\mathrm{TPP}^\infty(\Pi, \Lambda) \le 1$.

As $M \to \infty$ in (3.2.5), the prior π converges to the infinity-or-nothing priors $\pi_\infty(\epsilon')$ in (3.5.6). In addition, $\pi^\star = \pi_\infty(\epsilon'/\epsilon)$. By the intermediate value theorem, there must exist some $\epsilon' \in [0,1]$ such that

$$\begin{aligned}\mathrm{TPP}^\infty(\Pi,\Lambda) = \mathbb{P}(|\pi^\star + Z| > \alpha) &= (1-\epsilon')\,\mathbb{P}(|Z| > \alpha) + \epsilon'\,\mathbb{P}(|\infty + Z| > \alpha)\\ &= 2(1-\epsilon')\Phi(-\alpha) + \epsilon'.\end{aligned}$$

Here the first equality is given by (3.3.7) and $\alpha \equiv \alpha(\Pi,\Lambda)$ is the zero-threshold in Definition 3.4.1. The second equality follows from substituting the infinity-or-nothing $\pi^\star$. Therefore, the asymptotic sparsity in (3.5.7) is

$$\kappa(\Pi,\Lambda) = \mathbb{P}(|\pi + Z| > \alpha) = (1-\epsilon)\,\mathbb{P}(|Z| > \alpha) + \epsilon\,\mathrm{TPP}^\infty = 2(1-\epsilon\epsilon')\Phi(-\alpha) + \epsilon\epsilon',$$

where the first equality follows by the definition of the zero-threshold in Definition 3.4.1, the second uses that $\mathrm{TPP}^\infty(\Pi,\Lambda) = \mathbb{P}(|\pi^\star + Z| > \alpha)$, and the third is the result from the previous equation.

Some rearrangement gives

$$\Phi(-\alpha) = \frac{\kappa(\Pi,\Lambda) - \epsilon\epsilon'}{2(1-\epsilon\epsilon')}, \quad \text{and} \quad \mathrm{TPP}^\infty(\Pi,\Lambda) = \frac{(1-\epsilon')(\kappa(\Pi,\Lambda) - \epsilon\epsilon')}{1-\epsilon\epsilon'} + \epsilon'. \tag{3.5.9}$$

Simple calculus shows that the $\mathrm{TPP}^\infty(\Pi,\Lambda)$ in (3.5.9) is an increasing function of ϵ'. To see this, notice that the derivative is $\frac{(1-\epsilon)(1-\kappa)}{(1-\epsilon\epsilon')^2} \geq 0$. Given that $\epsilon' \leq \epsilon^\star/\epsilon$ by Lemma 3.5.3, we have

$$\mathrm{TPP}^\infty(\Pi,\Lambda) \leq \frac{(1-\frac{\epsilon^\star}{\epsilon})(\kappa(\Pi,\Lambda) - \epsilon\cdot\frac{\epsilon^\star}{\epsilon})}{1-\epsilon\cdot\frac{\epsilon^\star}{\epsilon}} + \frac{\epsilon^\star}{\epsilon} = 1 - \frac{(1-\kappa)(\epsilon-\epsilon^\star)}{\epsilon(1-\epsilon^\star)}.$$

□

In fact, Lemma 3.5.4 is an extension of [SBC17, Lemma C.2] (restated in Corollary 3.2.2(a)), which claims that, in the Lasso case, for all priors including those are not infinity-or-nothing, $\text{TPP}^\infty \leq u^\star(\delta;\epsilon,\delta)$. In particular, we remark that $u^\star(\delta;\epsilon,\delta)$ is equivalent to $u^\star_{\text{DT}}(\delta,\epsilon)$, since any Lasso estimator has an asymptotic sparsity no larger than δ.

As an immediate consequence of Lemma 3.5.4, we can reversely set a lower bound on the sparsity $\kappa(\Pi,\Lambda)$ given $\text{TPP}^\infty(\Pi,\Lambda)$. This is achieved by inverting the mapping in (3.5.8) and setting $u^\star = \text{TPP}^\infty$:

$$\kappa(\Pi,\Lambda) \geq 1 - \frac{\epsilon(1-\text{TPP}^\infty(\Pi,\Lambda))(1-\epsilon^\star)}{\epsilon-\epsilon^\star}. \tag{3.5.10}$$

Finally, leveraging the lower bound on the sparsity, we can minimize the FDP^∞ by minimizing the sparsity $\kappa(\Pi,\Lambda)$, since by definition

$$\text{FDP}^\infty(\Pi,\Lambda) = 1 - \frac{\epsilon\cdot\text{TPP}^\infty(\Pi,\Lambda)}{\kappa(\Pi,\Lambda)}. \tag{3.5.11}$$

Plugging (3.5.10) into (3.5.11), we finish the proof that $\text{FDP}^\infty \geq q^\star(\text{TPP}^\infty)$ for the SLOPE when we restrict the priors to be infinity-or-nothing: with $\text{TPP}^\infty = u$,

$$\text{FDP}^\infty(\Pi,\Lambda) \geq q^\star(u;\delta,\epsilon) := 1 - \frac{\epsilon u}{1-\frac{\epsilon(1-u)(1-\epsilon^\star)}{\epsilon-\epsilon^\star}} = \frac{\epsilon u(1-\epsilon)-\epsilon^\star(1-\epsilon)}{\epsilon u(1-\epsilon^\star)-\epsilon^\star(1-\epsilon)}.$$

□

3.5.3 Gap between upper and lower bounds

Considering Figure 3.2, we observe that the upper and lower boundary curves, $q_\star$ and $q^\star$, can be visually and numerically close to each other, especially when $\text{TPP}^\infty < u^\star_{\text{DT}}$.

One may wonder whether these boundaries actually coincide below the DT power limit. We answer this question in the negative and show analytically that there may exist pairs of $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ with the FDP^∞ strictly below $q^\star(\mathrm{TPP}^\infty)$ when $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$. In other words, there are instances where $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ points lie between the boundary curves $q_\star$ and $q^\star$.

Proposition 3.5.5. *For some (δ, ϵ), there exists $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}(\delta, \epsilon)$ defined in (3.2.3) such that*

$$q_\star(\mathrm{TPP}^\infty) < \mathrm{FDP}^\infty < q^\star(\mathrm{TPP}^\infty).$$

In the following, we prove Proposition 3.5.5 by constructing a specific problem instance (Π, Λ) which has FDP^∞ falling between the bounds. By showing that the gap between $q^\star(u)$ and $q_\star(u)$ indeed exists, we rigorously demonstrate a gap between $q^\star(u)$ and the unknown SLOPE trade-off q_{SLOPE}. We note that, for the Lasso trade-off at $(u, q^\star(u))$, the zero-threshold $\alpha(\Pi, \lambda) = t^\star(u)$ (defined in (3.2.7)) exactly and the state evolution constraint (3.4.1) is binding, i.e. $E(\Pi, \lambda) = \delta$ (see [SBC17, Lemma C.4, Lemma C.5]).

Fixing $\mathrm{TPP}^\infty = u$, our strategy (detailed in Section 3.7.5) is to construct (π, A) for SLOPE such that $\alpha(\pi, \mathrm{A}) = t^\star(u)$ as well but the state evolution constraint (3.4.1) is not binding, i.e. $E(\Pi, \Lambda) < \delta$. If such a construction succeeds, we can use a strictly larger zero-threshold than $t^\star(u)$ that can increase until $E(\Pi, \Lambda) > \delta$. Then, by using a larger zero-threshold, the SLOPE FDP^∞ is guaranteed to be strictly smaller than $q^\star(\mathrm{TPP}^\infty)$ by (3.3.7). Thus we will complete the proof that

$q_\star(u) < q^\star(u)$ for some $u < u^\star_{\text{DT}}$.

To construct (π, A) satisfying $\alpha(\pi, \mathrm{A}) = t^\star(u)$ with $E(\Pi, \Lambda) < \delta$, we leverage our empirical observation that the optimal priors $\pi^\star$, in the sense of problem (3.4.2), which achieves the lower bound $q_\star$, are oftentimes either infinity-or-nothing or constant. This motivates us to consider constant priors $\pi^\star = t_1$, for some constant t_1 (i.e. $p_1 = 1, t_1 = t_2$ in (3.4.5)), and hence

$$\pi = \begin{cases} t_1 & \text{w.p. } \epsilon, \\ 0 & \text{w.p. } 1-\epsilon. \end{cases}$$

In fact, conditioning on $\alpha(\Pi, \Lambda) = t^\star$ and $\text{TPP}^\infty = u$, the constant $t_1(u)$ is uniquely determined by (3.3.7):

$$\mathbb{P}\left(|t_1 + Z| > t^\star(u)\right) = u,$$

where Z is a standard normal.

Next, we use a common tool in the calculus of variations, known as the Euler-Lagrange equation (detailed in Section 3.7.5), to construct an effective penalty function $\mathrm{A}_{\text{eff}}(z)$ analytically on the interval $[0, \infty)$. The explicit form of $\mathrm{A}_{\text{eff}}(z)$ is defined in (3.7.23) with $\alpha = t^\star$. We emphasize that the constructed A_{eff} may not be a feasible SLOPE penalty function in the sense that it may violate the constraints in problem (3.4.6); however, if A_{eff} is increasing, then the optimal SLOPE effective penalty must be A_{eff}, as it is the minimizer of the unconstrained version of problem (3.4.6) and clearly satisfies the constraints. In the case that A_{eff} is feasible, we compare $E(\Pi, \Lambda) = F_{t^\star(u)}[\mathrm{A}_{\text{eff}}, p_{t_1}]$ with δ to determine whether $q^\star(u) > q_\star(u)$.

We now give a concrete example, which is elaborated in Section 3.7.5. When $\delta = 0.3, \epsilon = 0.2, \Pi^\star = 4.9006, \mathrm{TPP}^\infty = u^\star_{\mathrm{DT}} = 0.5676$, the maximum Lasso zero-threshold $t^\star(u^\star_{\mathrm{DT}}) = 1.1924$ and the minimum Lasso $\mathrm{FDP}^\infty = 0.6216$. We can construct the SLOPE penalty $\mathrm{A}_{\mathrm{eff}}$ that has the same zero-threshold and achieves $E(\Pi, \Lambda) = 0.2773 < \delta$. We can further construct the SLOPE penalty with larger zero-threshold, up to 1.2567, eventually have the SLOPE $\mathrm{FDP}^\infty = 0.5954$, which is much smaller than the minimum Lasso FDP^∞. In fact, our method can construct SLOPE penalty that outperforms the Lasso trade-off for any $\mathrm{TPP}^\infty \in (0.5283, 1]$, as shown in Figure 3.14.

3.6 Discussion

In this paper, we have investigated the possible advantages of employing sorted ℓ_1 regularization in model selection instead of the usual ℓ_1 regularization. Focusing on SLOPE, which instantiates sorted ℓ_1 regularization, our main results are presented by lower and upper bounds on the trade-off between false and true positive rates. On the one hand, the two tight bounds together demonstrate that type I and type II errors cannot both be small simultaneously using the SLOPE method with any regularization sequences, no matter how large the effect sizes are. This is the same situation as the Lasso [SBC17], which instantiates ℓ_1 regularization. More importantly, our results on the other hand highlight several benefits of using sorted ℓ_1 regularization. First, SLOPE is shown to be capable of achieving arbitrarily high power, thereby breaking the DT power limit. For comparison, the Lasso cannot pass the DT power limit in the supercritical regime, no matter how strong the effect sizes are. Second, moving to the regime below the DT power limit, we provide a problem instance where the SLOPE TPP and FDP trade-off is strictly better than the Lasso. Third, we introduce a comparison theorem which shows that any solution along the Lasso path can be dominated by a certain SLOPE estimate in terms of both the TPP and FDP and the estimation risk. In other words, the flexibility of sorted ℓ_1 regularization can always improve on the usual ℓ_1 regularization in the instance-specific setting.

The assumptions underlying the above-mentioned results include the random designs that have independent Gaussian entries and linear sparsity. In the venerable literature on high-dimensional regression, however, a more common sparsity regime is sublinear regimes where k/p tends to zero. As such, it is crucial to keep in mind the distinction in the sparsity regime when interpreting the results in this paper. From a technical viewpoint, our assumptions here enable the use of tools from AMP theory and in particular a very recent technique for tackling non-separable penalties. To obtain the lower bound, moreover, we have introduced several novel elements that might be useful in establishing trade-offs for estimators using other penalties.

In closing, we propose several directions for future research. Perhaps the most pressing question is to obtain the exact optimal trade-off for SLOPE. Regarding this question, a closer look at Figure 3.3 and Figure 3.5 suggests that our lower and upper bounds seem to coincide exactly when the TPP is small. If so, part of the optimal trade-off would already be specified. Having shown the advantage of SLOPE over the Lasso, a question of practical importance is to develop an approach to selecting regularization sequences for SLOPE to realize these benefits. Next, we would welcome extensions of our results to other methods using sorted ℓ_1 regularization, such as the group SLOPE [Brz+19]. For this purpose, our optimization-based technique for the variational calculus problems would likely serve as an effective tool. Recognizing that we have made heavy use of the two-level regularization sequences in many of our results, one is tempted to examine the possible benefits of using multi-level sequences for SLOPE [ZB21]. Finally, a challenging question is to investigate the SLOPE trade-off under correlated design matrices; the recent development by [CMW20] in AMP theory can be a stepping stone for this highly desirable generalization.

3.7 Appendix

3.7.1 When does SLOPE outperform Lasso?

When studying the SLOPE tradeoff curve, we consider results that hold true for all combinations of signal prior distribution and penalty distribution, (Π, Λ). In this section, we will instead look at instances of *fixed* bounded signal prior distributions.

Although the SLOPE trade-off upper bound $q^\star$ is no better than the Lasso one

$q^{\star}_{\text{Lasso}}$ when $\text{TPP}^{\infty} < u^{\star}_{\text{DT}}(\delta)$, as has been studied extensively in the previous sections of the paper, it is still possible that for a *fixed* prior distribution Π, the SLOPE can outperform Lasso using a smart choice of penalty vector. We emphasize that such cases are important, since in the real-world, the ground truth prior of the signal is indeed unknown but fixed. In fact, we will demonstrate that the SLOPE can always outperform Lasso in terms of the TPP, the FDP, the mean squared error (MSE).

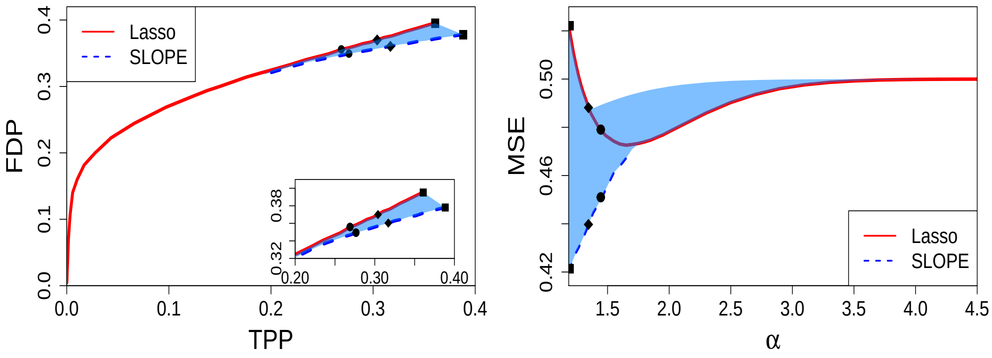


Figure 3.7: SLOPE outperforms the Lasso below the DT power limit. The red line is the Lasso paths when Π is Bernoulli(ϵ), $\delta = 0.3, \epsilon = 0.5$, and $\sigma = 0$. The blue region is the SLOPE $(\text{TPP}^{\infty}, \text{FDP}^{\infty})$, produced by $\Theta(\ell, \alpha_L, 0.1)$ where $\alpha_L \in (\alpha_0, \infty)$ is the zero-threshold shared by the Lasso and the SLOPE for all $\ell \geq \alpha_L$. The blue dashed line is the boundary of blue region. The black dots on the red line are specific $(\text{TPP}^{\infty}, \text{FDP}^{\infty}$, MSE) by the Lasso, while the dots on the blue dashed line correspond to the Lasso dots by shape.

The proofs we provide only consider the two-level SLOPE penalty sequences of the form $\boldsymbol{\lambda} = \boldsymbol{\theta}_{\lambda_1,\lambda_2,w}$ as in (3.2.4). Despite the simplicity of the penalty sequence, we are already able to leverage the advantages of the flexibility of the SLOPE penalty relative to the Lasso. We moreover believe that the advantages of the SLOPE over the Lasso could be even greater when more general SLOPE penalty sequences are considered, though we leave this to future work.

To be specific, we consider a Lasso pair (Π_L, Λ_L) and aim to construct a corresponding SLOPE pair (Π_S, Λ_S) that outperforms the Lasso, under the requirement that $\Pi_L = \Pi_S = \Pi$. We will demonstrate that, for any fixed bounded Π, each Lasso penalty Λ_L can be dominated by some two-level SLOPE penalty distributions Λ_S, in the sense that the SLOPE produces strictly better $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ and MSE. We further demonstrate a method to search for such dominating SLOPE penalties Λ_S and then we reinforce these ideas with simulation results.

The theoretical result of this section can be found in Theorem 6. In specific, we demonstrate that switching from Lasso to the simple two-level SLOPE can achieve better TPP^∞, better FDP^∞ and better MSE at the same time. The full proof is in Appendix 3.7.6 and we discuss the ideas of proof here.

In the following, we work in the normalized A or α regime (given by the AMP calibration (3.3.3); see also the interpretations below that equation) instead of the Λ or λ regime. The minimum $\alpha \in \mathbb{R}_+$ such that the corresponding the Lasso penalty $\lambda(\alpha)$ is non-negative, is denoted α_0. We denote the normalized prior $\pi_L := \Pi/\tau_L < \Pi/\tau_S := \pi_S$ and their non-zero conditional distribution as $\pi_L^\star, \pi_S^\star$ respectively. Here τ_L, τ_S are computed from the state evolution (3.3.2) of the Lasso and the SLOPE.

The high-level idea of the proof is, for any Lasso penalty A_L, to find a SLOPE penalty distribution A_S which has

(1) the same zero-threshold $\alpha(\pi_S, \mathrm{A}_S) = \alpha(\pi_L, \mathrm{A}_L)$ (defined in Definition 3.4.1);

(2) a smaller τ_S than the Lasso τ_L (from Equation (3.4.1));

(3) a larger sparsity $\kappa(\Pi, \mathrm{A}_S)$ than $\kappa(\Pi, \mathrm{A}_L)$ (defined in Equation (3.5.7)).

To see why such A_S is dominating, we can show (2) together with [Bu+20a, Corollary 3.4], restated in (3.7.24), implies that the SLOPE MSE is strictly smaller than the Lasso MSE.

Further results follow from the definitions of TPP and FDP: by (3.3.7), we get

$$\mathrm{TPP}^{\infty}(\Pi, \Lambda_S) = \mathbb{P}(|\pi_S^{\star} + Z| > \mathrm{A}_L) > \mathbb{P}(|\pi_L^{\star} + Z| > \mathrm{A}_L) = \mathrm{TPP}^{\infty}(\Pi, \Lambda_L),$$

where we have used the equal zero-threshold condition (1). Finally, we finish the proof for the FDP^{∞} result by using (3.5.11) as well as the sparsity condition (3).

Using the above conditions as the searching criteria, we have designed an algorithm that, for any fixed prior Π and for each Lasso penalty $\mathrm{A}_L = \alpha_L$, finds a superior two-level SLOPE penalty $\mathrm{A}_S = \Theta_{\ell, \alpha_L, w}$ by searching over (ℓ, w). As presented in Figure 3.7, the SLOPE $(\mathrm{TPP}^{\infty}, \mathrm{FDP}^{\infty})$ with two-level penalties outperforms the Lasso path.

3.7.2 Detailed preliminary results of SLOPE AMP

In this section, we introduce the proximal operator of SLOPE, its limiting form (known as the limiting scalar function, on which the SLOPE AMP algorithm is based), and the SLOPE AMP theory relating to the state evolution and calibration equations.

SLOPE proximal operator

We start with the definition of the proximal operator. For input $\boldsymbol{y} \in \mathbb{R}^p$, define the **proximal operator** of a function $f : \mathbb{R}^p \to \mathbb{R}$ as

$$\text{prox}_f(\boldsymbol{y}) := \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{b}\|^2 + f(\boldsymbol{b}) \right\}.$$

For SLOPE, the proximal operator uses $f(\boldsymbol{b}) = J_{\boldsymbol{\theta}}(\boldsymbol{b}) := \sum_{i=1}^p \theta_i |b|_{(i)}$ for some penalty vector $\boldsymbol{\theta}$ and as discussed in [Bog+15a], the SLOPE proximal operator can be computed by Algorithm 2[5].

For the Lasso, the relevant proximal operator uses $f(\boldsymbol{b}) = \theta\|\boldsymbol{b}\|_1$ and is known as the soft-thresholding function, which we will denote as $\eta_{\text{soft}} : \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}^p$. Namely, for any index $i \in \{1, 2, \ldots, p\}$, the soft-thresholding function is defined as

$$[\text{prox}_{\theta\|\cdot\|_1}(\boldsymbol{y})]_i = [\eta_{\text{soft}}(\boldsymbol{y};\theta)]_i := \begin{cases} y_i - \theta, & \text{if } y_i > \theta, \\ y_i + \theta, & \text{if } y_i < -\theta, \\ 0, & \text{otherwise.} \end{cases}$$

[5] The SLOPE proximal operator can be computed by R library `SLOPE`.

Algorithm 2 Solving $\text{prox}_J(\boldsymbol{s}; \boldsymbol{\theta})$ by [Bog+15a, Algorithm 3]

```
(1). Sorting: Sort $|\boldsymbol{s}|$ in decreasing order, returning $\text{sort}(|\boldsymbol{s}|)$;
(2). Differencing: Calculate a difference sequence, $\boldsymbol{S}$, defined as
         $\boldsymbol{S} = \text{sort}(|\boldsymbol{s}|) - \boldsymbol{\theta};$
(3). Averaging: Repeatedly average out strictly increasing subsequences in $\boldsymbol{S}$ until none remains. We refer to the decreasing sequence after all the averaging as $\text{AVE}(\boldsymbol{S})$, and reassign
         $\boldsymbol{S} = \text{AVE}(\boldsymbol{S});$
(4). Truncating: Set negative values in the difference sequence to 0 and reassign
         $\boldsymbol{S} = \max(\boldsymbol{S}, 0);$
(5). Restoring: Restore the order and the sign of $\boldsymbol{s}$ from step (1) to $\boldsymbol{S}$. Now $\boldsymbol{S}$ with the restored order and sign is the final output.
```

Note that the Lasso proximal operator is indeed *separable*, meaning that any element of its output depends only on the corresponding element of its input. This generally does not hold for the SLOPE proximal operator, which renders the analysis of SLOPE much more difficult. Nevertheless, the SLOPE proximal operator is an *asymptotically separable* function (as discussed in (3.3.5)) and enables the analysis of the input-dependent penalty, which is detailed in Section 3.7.3.

In what follows, we denote $\text{prox}_J(\boldsymbol{v}; \boldsymbol{\lambda})$ as the SLOPE proximal operator

$\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})$.

SLOPE AMP algorithm

Under the working assumptions (see $(A1) \sim (A3)$ in Section 3.2) and using the SLOPE proximal operator, the SLOPE optimization problem (3.1.2) can be solved by the following AMP algorithm with any intial conditions ([Bu+20a]):

$$\begin{aligned}
\boldsymbol{\beta}^{t+1} &= \operatorname{prox}_J(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t; \boldsymbol{\alpha}\tau_t), \\
\boldsymbol{z}^{t+1} &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{t+1} + \frac{\boldsymbol{z}^t}{\delta}\left[\nabla \operatorname{prox}_J(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t; \boldsymbol{\alpha}\tau_t)\right],
\end{aligned}$$

where ∇ denotes the divergence and $(\boldsymbol{\alpha}, \tau_t)$ is defined in the equations known as the state evolution and the calibration, which we will describe shortly.

It has been shown in [Bu+20a, Theorem 2] that asymptotically $\boldsymbol{\beta}^t$ converges to the true minimizer $\widehat{\boldsymbol{\beta}}$. In addition, for uniformly pseudo-Lipschitz sequence functions ϕ_p, we have from [Bu+20a, Theorem 3] that

$$\operatorname*{plim}_p \phi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_t \operatorname*{plim}_p \mathbb{E}_Z[\phi_p(\operatorname{prox}_J(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}; \boldsymbol{\alpha}(p)\tau_t), \boldsymbol{\beta})].$$

Loosely speaking, AMP theory characterizes the SLOPE estimator by

$$\widehat{\boldsymbol{\beta}} \overset{\mathcal{D}}{\approx} \operatorname{prox}_J(\boldsymbol{\beta} + \tau \boldsymbol{Z}; \boldsymbol{\alpha}\tau),$$

whose empirical distribution weakly converges to $\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z)$ and we describe (A, τ) below.

State evolution of SLOPE AMP

Rigorously speaking, the **state evolution** for SLOPE is

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big(\eta_{\Pi+\tau Z;\mathrm{A}\tau}(\Pi+\tau Z)-\Pi\Big)^2 = \sigma^2 + \frac{\tau^2}{\delta}\mathbb{E}\Big(\eta_{\pi+Z;\mathrm{A}}(\pi+Z)-\pi\Big)^2, \tag{3.7.1}$$

which can be solved iteratively via

$$\tau_{t+1}^2 = F(\tau_t, \mathrm{A}\tau_t) = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big(\eta_{\Pi+\tau_t Z;\mathrm{A}\tau_t}(\Pi+\tau_t Z)-\Pi\Big)^2,$$

From the algorithmic perspective of AMP, we use the finite approximation of the state evolution,

$$\tau^2 = F(\tau, \boldsymbol{\alpha}\tau) := \sigma^2 + \frac{1}{\delta}\mathbb{E}\left\langle[\mathrm{prox}_J(\boldsymbol{\beta}+\tau\boldsymbol{Z};\boldsymbol{\alpha}\tau)-\boldsymbol{\beta}]^2\right\rangle, \tag{3.7.2}$$

which can be recursively solved from the fixed point recursion $\tau_{t+1}^2(p) = F(\tau_t(p), \boldsymbol{\alpha}\tau_t(p))$ for each vector $\boldsymbol{\alpha} \in \mathbb{R}^p$. Here $\langle\boldsymbol{u}\rangle = \sum_{i=1}^p u_i/p$. Furthermore, this state evolution enjoys nice convergence properties: it is shown in [Bu+20a, Theorem 1] that $\{\tau_t\}$ converges monotonically to a unique fixed point τ, under any initial condition.

Calibration of SLOPE AMP

For finite p, we have seen that the state evolution term τ depends on Π and $\boldsymbol{\alpha}$. Therefore fixing Π, we can view $\tau(\boldsymbol{\alpha})$ as a function of $\boldsymbol{\alpha}$ and then gives the **calibration** mapping of the SLOPE penalty $\boldsymbol{\lambda} \in \mathbb{R}^p$ through [Bu+20a, Lemma 2.2],

$$\boldsymbol{\lambda}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}\tau\left(1-\frac{1}{\delta p}\mathbb{E}(\nabla\,\mathrm{prox}_J(\boldsymbol{\Pi}+\tau\boldsymbol{Z};\boldsymbol{\alpha}\tau))\right), \tag{3.7.3}$$

where the divergence of the proximal operator is defined as

$$\nabla \operatorname{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}\tau) := \operatorname{diag}\left(\frac{\partial}{\partial v_1}, \frac{\partial}{\partial v_2}, ...\right) \cdot \operatorname{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}\tau).$$

There are certain critical observations given by [Bu+20a, Lemma 2.1] and [SC16, Proofs of Fact 3.2 and 3.3] that explain the divergence term as

$$\nabla \operatorname{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}\tau) = \| \operatorname{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}\tau)\|_0^*,$$

where $\|\boldsymbol{x}\|_0^*$ counts the *unique* non-zero magnitudes in the vector $\boldsymbol{x}$. E.g. $\|(2, -1, 1, 0)\|_0^* = 2$. This norm reduces to ℓ_0 norm in the Lasso case, where all non-zero elements in $\operatorname{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}\tau)$ have unique magnitudes. Therefore we can express (3.7.3) as

$$\boldsymbol{\lambda}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}\tau\Big(1 - \frac{1}{n}\mathbb{E}\| \operatorname{prox}_J(\boldsymbol{\Pi} + \tau\boldsymbol{Z}; \boldsymbol{\alpha}\tau)\|_0^*\Big). \tag{3.7.4}$$

In the asymptotic case, the calibration lies between the original penalty distribution Λ and the normalized penalty A, to which the empirical distributions of $\{\boldsymbol{\lambda}\}$ and $\{\boldsymbol{\alpha}\}$ converge weakly:

$$\begin{aligned}\Lambda &\overset{\mathcal{D}}{=} \mathrm{A}\tau\left(1 - \frac{1}{\delta}\lim_p \frac{1}{p}\mathbb{E}\| \operatorname{prox}_J(\boldsymbol{\Pi} + \tau\boldsymbol{Z}; \boldsymbol{A}(p)\tau)\|_0^*\right) \\ &= \mathrm{A}\tau\left(1 - \frac{1}{\delta}\,\mathbb{P}\left(\eta_{\pi+Z;\mathrm{A}}(\pi + Z) \in \mathcal{U}(\eta_{\pi+Z;\mathrm{A}}(\pi+Z))\right)\right),\end{aligned} \tag{3.7.5}$$

where $\mathcal{U}(\cdot)$ is defined by

$$\mathcal{U}(\eta) := \Big\{h \in \mathbb{R}_+ : \mathbb{P}(|\eta| = h) = 0\Big\}.$$

This quantity represents the portion of the probability space on which $|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|$ has zero probability mass. In addition, $\mathbb{P}(\eta \in \mathcal{U}(\eta))$ is the asymptotic proportion of *unique* non-zeros in the SLOPE estimator η. For example, if η follows a Bernoulli-Gaussian distribution with 30% probability being zero, then $\mathcal{U} = (0, \infty)$, since the only point mass is concentrated at 0 and $\mathbb{P}(\eta \in \mathcal{U}(\eta)) = 0.7$.

3.7.3 Bridging SLOPE and soft-thresholding

In this section we describe a connection between the SLOPE proximal operator and the Lasso proximal operator, i.e. the soft-thresholding function. This connection is built on top of the concept of **effective penalty** in Definition 3.4.2, which allows one to **reduce** the SLOPE proximal operator to the soft-thresholding function with an input-dependent penalty. In analyzing both bounds of the SLOPE trade-off, $q_\star$ and $q^\star$, we use this technique so that we can study the much more amenable soft-thresholding function in place of the SLOPE proximal operator.

Here we use '$\mathrm{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})$' to denote the SLOPE proximal operator and η_{soft} to denote the soft-thresholding function, both defined in Section 3.3. Note that unlike the soft-thresholding function, the SLOPE proximal operator does not have an explicit formula (nor does its limiting form given by the limiting scalar function η), however it can be efficiently computed by Algorithm 2. Recall that the SLOPE penalty vector $\boldsymbol{\alpha}$ is decreasing and non-negative. The first result we present in this section says that in finite dimension, we can always design an effective penalty

$\widehat{\boldsymbol{\alpha}}(\boldsymbol{v}, \boldsymbol{\alpha})$, such that applying prox_J on penalty $\boldsymbol{\alpha}$ is equivalent to applying elementwise soft-thresholding on $\widehat{\boldsymbol{\alpha}}$.

Fact 3.7.1. *For any $\boldsymbol{\alpha}, \boldsymbol{v} \in \mathbb{R}^p$, there exists $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ such that*

$$\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}) = \eta_{\text{soft}}(\boldsymbol{v}; \widehat{\boldsymbol{\alpha}}).$$

Proof of Fact 3.7.1. For $\boldsymbol{v} \geq 0$, the soft-thresholding operator is $\eta_{\text{soft}}(\boldsymbol{v}; \widehat{\boldsymbol{\alpha}}) = \max(\boldsymbol{v} - \widehat{\boldsymbol{\alpha}}, 0)$. Note that $\boldsymbol{v} \geq 0$ implies $\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}) \geq 0$ and $[\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})]_i \leq v_i$ for every i. Then we can simply design $\widehat{\boldsymbol{\alpha}}$ by setting $\widehat{\boldsymbol{\alpha}} = \boldsymbol{v} - \text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})$. More generally, for any $\boldsymbol{v}$, we can set $\widehat{\boldsymbol{\alpha}} = |\boldsymbol{v}| - \text{prox}_J(|\boldsymbol{v}|; \boldsymbol{\alpha})$ (c.f. [HL19a, Proposition 2]). □

We notice that there are possibly multiple valid designs of $\widehat{\boldsymbol{\alpha}}$. An example would be

$$\text{prox}_J([6, 5, 3, 2, 1]; [5, 2, 2, 2, 2]) = [2, 2, 1, 0, 0] = \eta_{\text{soft}}([6, 5, 3, 2, 1]; \widehat{\boldsymbol{\alpha}}), \tag{3.7.6}$$

and both $\widehat{\boldsymbol{\alpha}} = [4, 3, 2, 2, 1]$ or $[4, 3, 2, 2, 2]$ give the desired result.

We remark that the asymptotic version of the above fact is established in [HL19a, Proposition 1 and Algorithm 1]. However, we emphasize that although the construction of $\widehat{\boldsymbol{\alpha}}$ is trivial once $\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})$ is known beforehand, it is difficult to derive $\widehat{\boldsymbol{\alpha}}$ in general: $\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})$ has no explicit form and its computation is complicated, as can be seen in Algorithm 2. Nevertheless, certain useful properties of the effective penalty $\widehat{\boldsymbol{\alpha}}$ can be extracted.

Fact 3.7.2. *Suppose $\boldsymbol{v}$ is sorted in decreasing absolute values, then $\widehat{\boldsymbol{\alpha}}$ agrees with $\boldsymbol{\alpha}$ at the non-zero entries of $\operatorname{prox}_J(\boldsymbol{v};\boldsymbol{\alpha})$ where the proximal operation takes no averaging.*

Proof of Fact 3.7.2. From Algorithm 2, for each entry of $\boldsymbol{v}$, one may think of prox_J as either applying a soft-thresholding or applying a soft-thresholding followed by an averaging. □

In the example given in (3.7.6), the subsequence $[3,2,1]$ of $\boldsymbol{v}$ experiences the soft-thresholding with respect to the penalty subsequence $[2,2,2]$; on the other hand, the subsequence $[6,5]$ experiences the soft-thresholding with respect to the penalty subsequence $[5,2]$ (resulting in $[1,3]$) then the averaging (resulting in $[2,2]$); this output is equivalent to $[6,5]$ experiencing the soft-thresholding with respect to the effective penalty subsequence $[4,3]$ instead of the actual penalty subsequence $[5,2]$.

In other words, if v_i is indeed penalized by α_i without averaging, then the effective penalty $\widehat{\alpha}_i$ agrees with the actual penalty α_i.

The above result generally does not hold when v is not sorted in decreasing magnitudes. For instance,

$$\operatorname{prox}_J([3,5,-6];[5,2,2]) = [1,2,-2] = \eta_{\text{soft}}([3,5,-6];[2,3,4]).$$

Nevertheless, we show that larger input (in magnitude) matches with larger penalties.

Fact 3.7.3. *Suppose $\boldsymbol{v}$ is sorted in decreasing absolute values, so is $\widehat{\boldsymbol{\alpha}}$. Then larger input will have larger effective penalty.*

Proof of Fact 3.7.3. For the simplicity of discussion, we assume $\boldsymbol{v} \geq 0$. Then we have $\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha}) = \max(\text{AVE}(\boldsymbol{v} - \boldsymbol{\alpha}), 0)$ where $\text{AVE}(\cdot)$ is the averaging operator in Algorithm 2. For indices where the averaging does not take place on the sequence $\boldsymbol{v} - \boldsymbol{\alpha}$, we have $\widehat{\alpha}_i = \alpha_i$ from Fact 3.7.2. Clearly $\widehat{\boldsymbol{\alpha}}$ is decreasing on these indices as $\boldsymbol{\alpha}$ is decreasing by the definition of the sorted ℓ_1 norm. For indices where the averaging does take place, say the averaged magnitude is $c := [\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})]_I$ for some set of indices I, then $\widehat{\alpha}_i = v_i - [\text{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})]_i = v_i - c$ (by Fact 3.7.1), which is decreasing in i since $\boldsymbol{v}_I$ is a decreasing subsequence. □

Now that we have derived some properties of the sequence $\widehat{\boldsymbol{\alpha}}$ as a whole, we will focus on a particular point of the sequence. Before we move on, we introduce a quantile-related concept.

Definition 3.7.4. *For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we denote the k-th largest element in absolute values as $\boldsymbol{v}_{(k)}$. For a distribution V, we denote $V_{(\mathtt{k})}$ as the* ***upper* $\mathtt{k}$*-quantile*** *with* $\mathtt{k} \in [0, 1]$*:*

$$\mathbb{P}(|V| \geq V_{(\mathtt{k})}) = \mathtt{k}.$$

For example, $V_{(0.25)}$ is the upper quartile of $|V|$; $V_{(0.5)}, V_{(0)}, V_{(1)}$ are the median, maximum and minimum of $|V|$ respectively.

We show an asymptotic result that $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}$ agree at a specific point closely related to the zero-threshold defined in Definition 3.4.1.

Proposition 3.7.5. *Suppose $\boldsymbol{v}(p), \boldsymbol{\alpha}(p), \widehat{\boldsymbol{\alpha}}(p)$ converge weakly to distributions $V, \mathrm{A}, \widehat{\mathrm{A}}$ respectively, with V being a continuous distribution whose support contains*

0. Then

$$\widehat{\mathrm{A}}_{(\kappa)} = \mathrm{A}_{(\kappa)} = |V|_{(\kappa)},$$

where κ is the asymptotic sparsity of the SLOPE estimator, defined in (3.5.7)*.*

To see how this asymptotic result relates to the zero-threshold $\alpha(\pi, \mathrm{A})$ in Definition 3.4.1, it is helpful to consider $V = \pi + Z$ (which is continuous even if π is discrete), since the SLOPE estimator's distribution is $\widehat{\Pi} = \eta_{\pi+Z,\mathrm{A}}(\pi + Z)$.

Proof of Proposition 3.7.5. Then the asymptotic sparsity is

$$\kappa := \mathrm{plim}\, |\{i : [\mathrm{prox}_J(\boldsymbol{v}; \boldsymbol{\alpha})]_i \neq 0\}|/p = \mathbb{P}\left(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) \neq 0\right) = \mathbb{P}\left(|\pi + Z| > \alpha(\pi, \mathrm{A})\right).$$

On the other hand, from the soft-thresholding effect of $\eta_{V,\mathrm{A}}$, we have

$$\eta_{V,\mathrm{A}}(V) = \eta_{\mathrm{soft}}(V; \widehat{\mathrm{A}}),$$

and equivalently

$$\kappa = \mathbb{P}\left(|\eta_{V,\mathrm{A}}(V)| \neq 0\right) = \mathbb{P}\left(\eta_{\mathrm{soft}}(|V|; \widehat{\mathrm{A}}) \neq 0\right) = \mathbb{P}\left(|V| > \widehat{\mathrm{A}}\right),$$

which indicates

$$|V|_{(\kappa)} = \widehat{\mathrm{A}}_{(\kappa)} = \alpha(\pi, \mathrm{A}).$$

From the proof of Fact 3.7.1 (also from [HL19a, Proposition 2]), we know $\widehat{\mathrm{A}}_{(\kappa)} = |V|_{(\kappa)} - \eta_{V,\mathrm{A}}(|V|_{(\kappa)})$. Together with the above, it holds that $\eta_{V,\mathrm{A}}(|V|_{(\kappa)}) = 0$.

Notice that $\eta_{V,\mathrm{A}}$ is continuous, thus there must exist some interval $[|V|_{(\kappa)}, x]$ where $\eta_{V,\mathrm{A}}$ is not constant (i.e. penalties are not averaged), because $\eta_{V,\mathrm{A}}(|V|_{(\kappa)}) = 0$ but $\eta_{V,\mathrm{A}}(x) > 0$. Hence by Fact 3.7.2, we obtain $\widehat{\mathrm{A}}_{(\kappa)} = \mathrm{A}_{(\kappa)}$. □

To summarize, we can reduce the non-separable SLOPE proximal operator to some separable soft-thresholding, asymptotically. In this way, we can alternatively study the effective penalty used in the soft-thresholding, instead of the implicit SLOPE proximal operator. We emphasize that Section 3.7.3 is the key to study the SLOPE TPP-FDP trade-off bounds $q_\star$ and $q^\star$ in Section 3.4 and Section 3.5.

3.7.4 SLOPE trade-off and Möbius upper bound

In this section we provide some useful results that describe the SLOPE TPP–FDP trade-off curve beyond the Lasso phase transition. In particular, we show that the SLOPE state evolution and calibration constraints can be translated to analogous constraints based on the soft-thresholding function.

Using AMP to characterize the asymptotic TPP and FDP

In this section, we give a sketch of the proof of Lemma 3.3.1, which consists of justifying the use of AMP to characterize the FDP and TPP of SLOPE asymptotically.

It has been rigorously proven in [Bu+20a, Theorem 3] that $\frac{1}{n}\sum_{i=1}^n \psi([\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\boldsymbol{\beta}+\tau Z)]_i, \beta_i)$ is asymptotically equal in distribution to that of $\frac{1}{n}\sum_{i=1}^n \psi(\widehat{\beta}_i, \beta_i)$, when $\psi:\mathbb{R}^2\to\mathbb{R}$ is a pseudo-Lipschitz continuous function. We would like to use this result to analyze the FDP and TPP, where from Lemma 3.3.1 we see that

$$\mathrm{FDP}_\xi(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{|\{j:|\widehat{\beta}_j|>\xi,\beta_j=0\}|}{|\{j:|\widehat{\beta}_j|>\xi\}|} = \frac{\sum_j \varphi_{V,\xi}(\widehat{\beta}_j,\beta_j)}{\sum_j \varphi_{V,\xi}(\widehat{\beta}_j,\beta_j)+\sum_j \varphi_{T,\xi}(\widehat{\beta}_j,\beta_j)}, \quad (3.7.7)$$

and

$$\mathrm{TPP}_\xi(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|} = \frac{\sum_j \varphi_{T,\xi}(\widehat{\beta}_j, \beta_j)}{\sum_j 1(\beta_j \neq 0)}, \tag{3.7.8}$$

are determined by sums of discontinuous functions, $\varphi_{V,\xi}(x,y) = 1(|x| > \xi)1(y = 0)$ and $\varphi_{T,\xi}(x,y) = 1(|x| > \xi)1(y \neq 0)$, and not pseudo-Lipschitz functions. Therefore [Bu+20a, Theorem 3] does not apply directly. Nevertheless, we are still able to use the characterization given by AMP, as is demonstrated in Lemma 3.7.6. The proof of Lemma 3.7.6 is an extension of the analogous result for the Lasso case given in [SBC17, Lemma A.1]. We notice that the result of Lemma 3.3.1 is just that given in (3.7.11).

Lemma 3.7.6. *Under the working assumptions, namely (A1), (A2), and (A3), for ξ such that $\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) = \xi) = 0$, the SLOPE estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ obeys*

$$\frac{V_\xi(\boldsymbol{\lambda})}{p} := \sum_j \frac{\varphi_{V,\xi}(\widehat{\beta}_j, \beta_j)}{p} = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j = 0\}|}{p} \xrightarrow{P} \mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi + Z)| > \xi, \pi = 0), \tag{3.7.9}$$

$$\frac{T_\xi(\boldsymbol{\lambda})}{p} := \sum_j \frac{\varphi_{T,\xi}(\widehat{\beta}_j, \beta_j)}{p} = \frac{|\{j : |\widehat{\beta}_j| > \xi, \beta_j \neq 0\}|}{p} \xrightarrow{P} \mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi + Z)| > \xi, \pi \neq 0), \tag{3.7.10}$$

where Z is a standard normal independent of Π, (τ, A) is the unique solution to the state evolution (3.3.2) *and the calibration* (3.3.3)*, and $\pi = \Pi/\tau$. Consequently, we have using the representations in* (3.7.7) *and* (3.7.8) *and the definitions of $V_\xi(\boldsymbol{\lambda})$*

and $T_\xi(\boldsymbol{\lambda})$ *above, that*

$$\mathrm{FDP}_\xi = \frac{V_\xi(\boldsymbol{\lambda})}{V_\xi(\boldsymbol{\lambda}) + T_\xi(\boldsymbol{\lambda})} \xrightarrow{P} \mathrm{FDP}_\xi^\infty, \qquad \textit{and} \qquad \mathrm{TPP}_\xi = \frac{T_\xi(\boldsymbol{\lambda})}{\sum_j 1(\beta_j \neq 0)} \xrightarrow{P} \mathrm{TPP}_\xi^\infty. \tag{3.7.11}$$

Proof of Lemma 3.7.6. The analogous result for when $\widehat{\boldsymbol{\beta}}$ is a Lasso solution is proven rigorously in [Bog+13b]. Here we adapt their proof for SLOPE. The high level idea for the proof of (3.7.9) and (3.7.10) is to construct two series of pseudo-Lipschitz continuous functions

$$\varphi_{V,\xi,h}(x,y) = (1 - R_h(x))Q_h(y) \quad \text{and} \quad \varphi_{T,\xi,h}(x,y) = (1 - R_h(x))(1 - Q_h(y)),$$

that approach $\varphi_{V,\xi}, \varphi_{T,\xi}$ as $h \to 0^+$. Here $Q_h(y) = \max\{1 - |y/h|, 0\}$ and

$$R_h(x) = \begin{cases} 0 & \text{if } |x| > \xi + h \\ \frac{\xi + h - |x|}{2h} & \text{if } \xi - h < |x| < \xi + h \\ 1 & \text{if } |x| < \xi - h \end{cases} \cdot$$

Since for small h,

$$|\varphi_{V,\xi,h}(x,y) - \varphi_{V,\xi}(x,y)| \leq 1(\xi - h < |x| < \xi + h) + 1(0 < |y| < h),$$

for any $c > 0$,

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{1}{p}\sum_{i=1}^p \varphi_{V,\xi}\left(\hat{\beta}_i, \beta_i\right) - \frac{1}{p}\sum_{i=1}^p \varphi_{V,\xi,h}\left(\hat{\beta}_i, \beta_i\right)\right| > c\right) \\ \leq &\mathbb{P}\left(\frac{1}{p}\sum_{i=1}^p 1\left(\xi - h < \left|\hat{\beta}_i\right| < \xi + h\right) > \frac{c}{2}\right) + \mathbb{P}\left(\frac{1}{p}\sum_{i=1}^p 1\left(0 < |\beta_i| < h\right) > \frac{c}{2}\right) \end{aligned}$$

We will show that both terms on the right hand side converge to zero as $p \to \infty$ and then $h \to 0$. The second term converges to zero by the weak Law of Large Numbers. To deal with the first term, we introduce another pseudo-Lipschitz continuous function

$$G_h(x) = \begin{cases} 1 & \text{if } \xi - h < x < \xi + h \\ 0 & \text{if } x > \xi + 2h \text{ or } x < \xi - 2h \\ \frac{x-(\xi-2h)}{h} & \text{if } \xi - 2h < x < \xi - h \\ \frac{(\xi+2h)-x}{h} & \text{if } \xi + h < x < \xi + 2h \end{cases}$$

which upper bounds the function $1\left(< \xi - h < |x| < \xi + h\right)$. Then the AMP theory in [Bu+20a, Theorem 3] gives

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} 1\left(\xi - h < \left|\hat{\beta}_i\right| < \xi + h\right) \le \mathbb{P}\left(\xi - h < \left|\hat{\Pi}\right| < \xi + h\right) \to 0$$

as $h \to 0$, where $\hat{\Pi}$ is defined in (3.3.1). Hence, one can then argue

$$\begin{aligned} \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \varphi_{V,\xi}(\hat{\beta}_i, \beta_i) &\overset{P}{=} \lim_{h\to 0}\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \varphi_{V,\xi,h}(\hat{\beta}_i, \beta_i) \\ &\overset{P}{=} \lim_{h\to 0} \mathbb{E}\varphi_{V,\xi,h}\left(\eta_{\Pi+\tau Z, \mathrm{A}\tau}(\Pi + \tau Z), \Pi\right) \\ &= \mathbb{E}\varphi_{V,\xi}\left(\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi + \tau Z), \Pi\right), \end{aligned}$$

where the second equality in the above employs the AMP results for the pseudo-Lipschitz continuous function $\varphi_{V,\xi,h}(\cdot,\cdot)$. The technical aspects of the proof involve making this argument rigorous. The final result follows by noticing that

$$\begin{aligned}\mathbb{E}\varphi_{V,\xi}\left(\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z),\Pi\right) &= \mathbb{E}\left[1\left(|\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z)|>\xi\right)1\left(\Pi=0\right)\right] \\ &= \mathbb{P}\left(|\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z)|>\xi, \Pi=0\right) \\ &= \mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|>\xi, \pi=0).\end{aligned}$$

Now leveraging results (3.7.9) and (3.7.10), that give

$$V_\xi(\boldsymbol{\lambda})/p \xrightarrow{\mathbb{P}} \mathbb{P}\left(|\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z)|>\xi, \Pi=0\right) = \mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|>\xi, \pi=0),$$

$$T_\xi(\boldsymbol{\lambda})/p \xrightarrow{\mathbb{P}} \mathbb{P}\left(|\eta_{\Pi+\tau Z,\mathrm{A}\tau}(\Pi+\tau Z)|>\xi, \Pi\neq 0\right) = \mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|>\xi, \pi\neq 0),$$

and using that

$$\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|>\xi, \pi=0) = (1-\epsilon)\,\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(Z)|>\xi),$$

and

$$\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi+Z)|>\xi, \pi\neq 0) = \epsilon\,\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(\pi^\star+Z)|>\xi),$$

where we recall $\pi^\star$ is the distribution of the non-zero part of π, we finally obtain

$$\begin{aligned}\mathrm{FDP}^\infty_\xi(\Pi,\Lambda) &= \operatorname{plim}\frac{V_\xi(\boldsymbol{\lambda})}{V_\xi(\boldsymbol{\lambda})+T_\xi(\boldsymbol{\lambda})} \\ &= \frac{(1-\epsilon)\,\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(Z)|>\xi)}{(1-\epsilon)\,\mathbb{P}(|\eta_{\pi+Z,\mathrm{A}}(Z)|>\xi)+\epsilon\,\mathbb{P}\left(|\eta_{\pi+Z,\mathrm{A}}(\pi^\star+Z)|>\xi\right)}, \\ \mathrm{TPP}^\infty_\xi(\Pi,\Lambda) &= \operatorname{plim}\frac{T_\xi(\boldsymbol{\lambda})}{|\{j:\beta_j\neq 0\}|} = \mathbb{P}\left(|\eta_{\pi+Z,\mathrm{A}}(\pi^\star+Z)|>\xi\right).\end{aligned}$$

□

The result of Lemma 3.3.1 (and Lemma 3.7.6) implies that when studying the trade-off between the FDP and TPP asymptotically, we can work with the explicit and amenable quantities $\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(Z) \neq 0)$ and $\mathbb{P}(\eta_{\pi+Z,\mathrm{A}}(\pi^\star + Z) \neq 0)$, by considering $\xi \to 0$.

A better understanding the Donoho-Tanner threshold

In this section, we introduce an equivalent definition of the DT threshold $\epsilon^\star$, originally defined in (3.2.2), from a non-parametric viewpoint. This definition is necessary for our analysis of the SLOPE trade-off upper bound $q^\star$ discussed in Section 3.5.

To specify the threshold $\epsilon^\star$ when $\delta < 1$, we consider the equation

$$2(1-\epsilon)[(1+x^2)\Phi(-x) - x\phi(x)] + \epsilon(1+x^2) = \delta \tag{3.7.12}$$

in $x > 0$. Above, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution, respectively. We demonstrate the properties of (3.7.12) can be found in Figure 3.8 and Figure 3.9.

The key point we will use is that this equation has a unique positive root in x if and only if $0 < \epsilon < 1$ takes a certain value $\epsilon^\star(\delta)$ that depends only on δ. This unique root is $x := t^\star(u^\star_{\mathrm{DT}}(\delta))$, as given by [SBC17, Appendix C]. Furthermore, (3.7.12) has two roots when $\epsilon \leq \epsilon^\star$ and no root otherwise. In fact, (3.7.12) originates from the state evolution (3.4.1) for the Lasso when we consider the infinity-or-nothing priors defined in (3.5.6), and it can also be found in [SBC17, Equation (C.5)].

In summary, (3.7.12) gives an equivalent representation of $\epsilon^\star(\delta)$ that we will find useful in the upcoming proofs. Namely, $\epsilon^\star(\delta)$ is the specific value of $0 < \epsilon < 1$ such that (3.7.12) has a unique root.

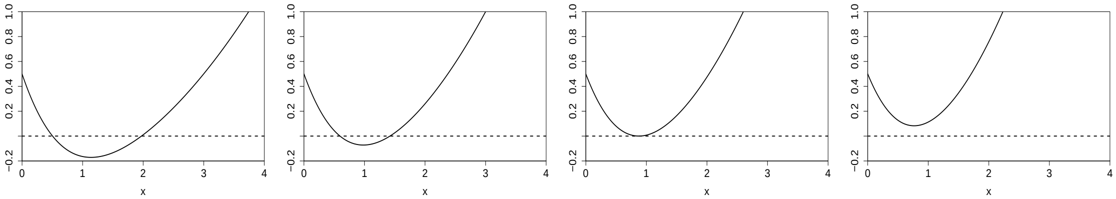


Figure 3.8: A plot of the function $f(x) = 2(1-\epsilon)[(1+x^2)\Phi(-x) - x\phi(x)] + \epsilon(1+x^2) - \delta$ defined in (3.7.12) for $\delta = 0.5$ and $\epsilon \in \{0.1, 0.15, \epsilon^\star(\delta) = 0.1928, 0.25\}$ (from left to right).

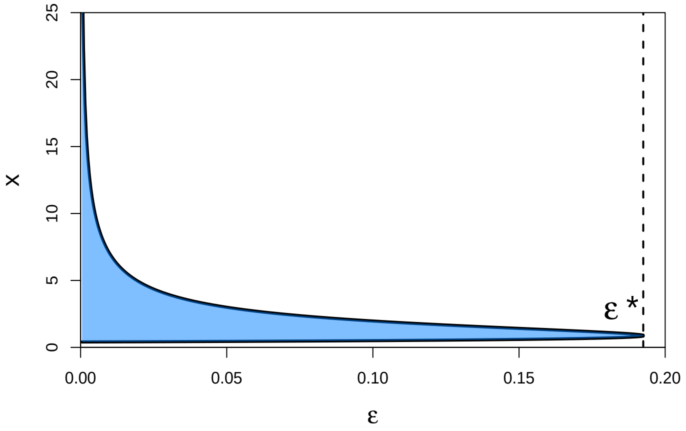


Figure 3.9: A plot to demonstration the roots of (3.7.12) with $\delta = 0.5$. Here $\epsilon^\star = 0.1928$ is the dashed line and the blue area contains valid x for which the inequality in (3.7.12) holds for each ϵ. The black solid lines are the roots. Notice that the blue area corresponding to $\epsilon = 0.1$ corresponds to the area under the dashed line in leftmost plot of Figure 3.8.

Proof of Lemma 3.5.3

Proof of Lemma 3.5.3. For infinity-or-nothing priors where $\pi = \infty$ with probability $\epsilon\epsilon'$ or $\pi = 0$ with probability $1 - \epsilon\epsilon'$, the state evolution constraint (3.4.1) gives,

$$\begin{aligned}\delta &\geq \mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi + Z) - \pi\right)^2 \\ &= \mathbb{P}(\pi = \infty)\mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi^\star + Z) - \pi^\star\right)^2 + \mathbb{P}(\pi \neq \infty)\mathbb{E}\eta_{\pi+Z,\mathrm{A}}(Z)^2 \qquad (3.7.13) \\ &= \epsilon\epsilon'\mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi^\star + Z) - \pi^\star\right)^2 + (1-\epsilon\epsilon')\mathbb{E}\eta_{\pi+Z,\mathrm{A}}(Z)^2.\end{aligned}$$

Using the effective penalty function $\widehat{\mathrm{A}}_{\mathrm{eff}}$ defined in Definition 3.4.2, we can write the above as

$$\delta \geq \epsilon\epsilon'\mathbb{E}\left(\eta_{\mathrm{soft}}(\pi^\star + Z; \widehat{\mathrm{A}}_{\mathrm{eff}}(\pi^\star + Z)) - \pi^\star\right)^2 + (1-\epsilon\epsilon')\mathbb{E}\eta_{\mathrm{soft}}(Z; \widehat{\mathrm{A}}_{\mathrm{eff}}(Z))^2.$$

Now, we denote the distribution $\widehat{\mathrm{A}} :\overset{\mathcal{D}}{=} \widehat{\mathrm{A}}_{\mathrm{eff}}(\pi + Z)$ and in what follows we study the distribution of $\widehat{\mathrm{A}}_{\mathrm{eff}}(\pi^\star + Z)$ in more detail. Using the fact that $\pi^\star + Z$ is almost surely larger than Z (since $\pi^\star = \infty$) and Fact 3.7.3, which states SLOPE assigns larger effective penalty to larger input, we conclude

$$\widehat{\mathrm{A}}_{\mathrm{eff}}(\pi^\star + Z) \overset{\mathcal{D}}{=} \widehat{\mathrm{A}}\Big|\widehat{\mathrm{A}} > \widehat{\mathrm{A}}_{(\epsilon\epsilon')}.$$

which, we will shortly show, is a constant. In the above, the quantity with a subscript, $\widehat{\mathrm{A}}_{(\epsilon\epsilon')}$, is a quantile-related scalar such that $\mathbb{P}(\widehat{\mathrm{A}} > \widehat{\mathrm{A}}_{(\epsilon\epsilon')}) = \epsilon\epsilon'$, defined in Definition 3.7.4. In words, the larger part of $\widehat{\mathrm{A}}$ $\left(\text{i.e. } \widehat{\mathrm{A}}\Big|\widehat{\mathrm{A}} > \widehat{\mathrm{A}}_{(\epsilon\epsilon')}\right)$ is assigned to the larger part of the input $\pi^\star + Z$; and $\widehat{\mathrm{A}}\Big|\widehat{\mathrm{A}} \leq \widehat{\mathrm{A}}_{(\epsilon\epsilon')}$ is assigned to the input Z.

Furthermore, using the assumption that

$$\epsilon\epsilon' \leq \mathbb{P}(\Lambda = \max \Lambda) = \mathbb{P}(\mathrm{A} = \max \mathrm{A}), \tag{3.7.14}$$

where the final equality follows since Λ and A only differ by a constant (see the calibration equation (3.3.3)), we get

$$\widehat{\mathrm{A}}\Big|\widehat{\mathrm{A}} \geq \widehat{\mathrm{A}}_{(\epsilon\epsilon')} \overset{\mathcal{D}}{=} \mathrm{A}\Big|\mathrm{A} \geq \mathrm{A}_{(\epsilon\epsilon')} = \mathrm{A}_{(\epsilon\epsilon')} \in \mathbb{R}.$$

In the above, the first equality comes from the fact that the upper $\epsilon\epsilon'$ quantile of A is Lasso-like, following from (3.7.14) (hence, there is no averaging in the SLOPE proximal operator and Fact 3.7.2 applies) and the second equality also follows from (3.7.14) as well.

Therefore, using that $\widehat{\mathrm{A}}_{\text{eff}}(\pi^\star + Z)$ is a constant equal to $\mathrm{A}_{(\epsilon\epsilon')}$, the state evolution constraint becomes

$$\begin{aligned}
\delta &\geq \epsilon\epsilon'\mathbb{E}\left(\eta_{\text{soft}}(\pi^\star + Z; \mathrm{A}_{(\epsilon\epsilon')}) - \pi^\star\right)^2 + (1-\epsilon\epsilon')\mathbb{E}\eta_{\text{soft}}(Z; \widehat{\mathrm{A}}_{\text{eff}}(Z))^2 \\
&= \epsilon\epsilon'\mathbb{E}(\mathrm{A}_{(\epsilon\epsilon')} - Z)^2 + (1-\epsilon\epsilon')\mathbb{E}\eta_{\text{soft}}(Z; \widehat{\mathrm{A}}_{\text{eff}}(Z))^2 \\
&= \epsilon\epsilon'(1 + \mathrm{A}^2_{(\epsilon\epsilon')}) + (1-\epsilon\epsilon')\mathbb{E}\eta_{\text{soft}}(Z; \widehat{\mathrm{A}}_{\text{eff}}(Z))^2,
\end{aligned} \tag{3.7.15}$$

where the first equality follows by the definition of the soft-thresholding function and the fact that $\mathrm{A}_{(\epsilon\epsilon')}$ is constant and the second from the fact that $Z \sim \mathcal{N}(0,1)$.

Notice that, again by Fact 3.7.3, $\widehat{\mathrm{A}}_{\text{eff}}(z)$ is increasing in absolute value of z, hence

$$\widehat{\mathrm{A}}_{\text{eff}}(Z) \leq \sup\left(\widehat{\mathrm{A}}\Big|\widehat{\mathrm{A}} \leq \widehat{\mathrm{A}}_{(\epsilon\epsilon')}\right) = \widehat{\mathrm{A}}_{(\epsilon\epsilon')} = \mathrm{A}_{(\epsilon\epsilon')},$$

in which the last equality holds from Fact 3.7.2, as a consequence of

$$\lim_{x \nearrow \epsilon\epsilon'} \left|\eta_{\pi+Z,\mathrm{A}}(\pi + Z)\right|_{(x)} = \infty > \sup_{Z} \left|\eta_{\pi+Z,\mathrm{A}}(Z)\right| = \lim_{x \searrow \epsilon\epsilon'} \left|\eta_{\pi+Z,\mathrm{A}}(\pi + Z)\right|_{(x)},$$

i.e. no averaging takes place at the quantile $\epsilon\epsilon'$ (here the limits are one-sided limits). Additionally, we observe that $\eta_{\text{soft}}(z;x)^2$ is decreasing in the scalar x. Therefore, we get

$$\mathbb{E}\eta_{\text{soft}}(Z;\widehat{\mathrm{A}}_{\text{eff}}(Z))^2 \geq \mathbb{E}\eta_{\text{soft}}(Z;\mathrm{A}_{(\epsilon\epsilon')})^2.$$

Applying the above bound into (3.7.15) and then using some simple algebra to express the soft-thresholding function, we find

$$\begin{aligned}\delta &\geq \epsilon\epsilon'(1+\mathrm{A}^2_{(\epsilon\epsilon')}) + (1-\epsilon\epsilon')\mathbb{E}\eta_{\text{soft}}(Z;\mathrm{A}_{(\epsilon\epsilon')})^2 \\ &= \epsilon\epsilon'(1+\mathrm{A}^2_{(\epsilon\epsilon')}) + 2(1-\epsilon\epsilon')\left[(1+\mathrm{A}^2_{(\epsilon\epsilon')})\Phi(-\mathrm{A}_{(\epsilon\epsilon')}) - \mathrm{A}_{(\epsilon\epsilon')}\phi(\mathrm{A}_{(\epsilon\epsilon')})\right].\end{aligned} \tag{3.7.16}$$

Following the discussion around (3.7.12), the above inequality can only possibly hold when $\epsilon\epsilon' \leq \epsilon^\star$, or when $\epsilon' \in [0, \epsilon^\star/\epsilon]$ as desired. □

Achieving the Möbius curve of $q^\star$

In this section we prove Proposition 3.2.3, or in other words, we show that with the special design of a two-level SLOPE penalty and infinity-or-nothing prior, we can approach the Möbius part of $q^\star$ arbitrarily close.

Proof of Proposition 3.2.3. To give the proof, we consider a specific prior $\Pi_M(\epsilon^\star/\epsilon)$ as in (3.2.5) and let $M \to \infty$. Here $\epsilon^\star$ is defined in (3.2.2). This is equivalent to setting the normalized prior π to the infinity-or-nothing prior $\pi_\infty(\epsilon^\star/\epsilon)$, defined in (3.5.6) as:

$$\pi_\infty(\epsilon^\star/\epsilon) = \begin{cases} \infty & w.p. \quad \epsilon^\star, \\ 0 & w.p. \quad 1-\epsilon^\star. \end{cases} \tag{3.7.17}$$

As for the SLOPE penalty, we consider a sub-class of two-level penalty distributions Λ that satisfy $\mathbb{P}(\Lambda = \max \Lambda) \geq \epsilon^\star$, or in the notation of (3.2.4) we will have $w > \epsilon^\star$. By setting the penalty as such, we satisfy the assumption in Proposition 3.5.2 and consequently we can apply the results in Lemma 3.5.3 and Lemma 3.5.4.

Now we are ready to present the proof. For any $\mathrm{TPP}^\infty = u \geq u^\star_{\mathrm{DT}}(\delta)$, we recall from (3.5.9) in the proof of Lemma 3.5.4 that the asymptotic sparsity $\kappa(\Pi, \Lambda)$ (defined in (3.5.7)) satisfies

$$\kappa(\Pi, \Lambda) = 1 - \frac{\epsilon(1-u)(1-\epsilon\epsilon')}{\epsilon - \epsilon\epsilon'}. \tag{3.7.18}$$

From (3.5.11), minimizing $\mathrm{FDP}^\infty(\Pi, \Lambda)$ is equivalent to minimizing $\kappa(\Pi, \Lambda)$, which from (3.7.18) we see is further equivalent to maximizing ϵ'. Since Lemma 3.5.3 states that $\epsilon' \leq \epsilon^\star/\epsilon$, we aim to achieve a sparsity with $\epsilon' = \epsilon^\star/\epsilon$, namely a sparsity of

$$\kappa(\Pi, \Lambda) = 1 - \frac{\epsilon(1-u)(1-\epsilon^\star)}{\epsilon - \epsilon^\star}, \tag{3.7.19}$$

which is given in (3.5.10) as the smallest sparsity for which $\mathrm{TPP}^\infty \geq u$ is possible.

Therefore, we consider a specific prior $\Pi_M(\epsilon^\star/\epsilon)$ as in (3.2.5) and let $M \to \infty$. Then the limiting normalized prior π is the infinity-or-nothing prior defined in (3.7.17). Next, we seek the penalty Λ that can result in the desired sparsity $\kappa(\pi, \mathrm{A})$ in (3.7.19), or equivalently, we seek the normalized version of Λ given by A, defined via the calibration equation (3.3.3).

To find such a penalty Λ, we consider the state evolution constraint (3.4.1), and emphasize that when achieving the desired sparsity, (3.4.1) must be satisfied by the pair (π, A). We use the result of (3.7.16) and more generally, the proof of Lemma 3.5.3 in Section 3.7.4, to give for $\epsilon' = \epsilon^\star/\epsilon$,

$$(1-\epsilon^\star)\mathbb{E}\eta_{\text{soft}}(Z;\mathrm{A}_{(\epsilon^\star)})^2 + \epsilon^\star(1+\mathrm{A}_{(\epsilon^\star)}^2) \le \mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)-\pi\right)^2 \le \delta, \quad (3.7.20)$$

where again $\mathrm{A}_{(\epsilon^\star)}$ is a scalar defined in Definition 3.7.4, i.e., it is chosen such that $\mathbb{P}(\mathrm{A} > \mathrm{A}_{(\epsilon^\star)}) = \epsilon^\star$. In particular, the first inequality above only holds with equality when $\eta_{\pi+Z,\mathrm{A}}(Z) \overset{\mathcal{D}}{=} \eta_{\text{soft}}(Z;\mathrm{A}_{(\epsilon^\star)})$, which can be seen by comparing the bounds in (3.7.16) and (3.7.13).

From another direction, by the alternative definition of $\epsilon^\star$ in (3.7.12), we have

$$(1-\epsilon^\star)\mathbb{E}\eta_{\text{soft}}(Z;x)^2 + \epsilon^\star(1+x^2) \ge \delta, \quad (3.7.21)$$

for all $x > 0$, with the equality holding only when $x = t^\star(u^\star_{\text{DT}}(\delta))$, as has been discussed in Section 3.7.4. We notice that (3.7.21) equals (3.7.12) since $\mathbb{E}\eta_{\text{soft}}(Z;x)^2 = 2[(1+x^2)\Phi(-x) - x\phi(x)]$. Combining (3.7.20) and (3.7.21), we obtain

$$\delta \overset{(a)}{\le} (1-\epsilon^\star)\mathbb{E}\eta_{\text{soft}}(Z;\mathrm{A}_{(\epsilon^\star)})^2 + \epsilon^\star(1+\mathrm{A}_{(\epsilon^\star)}^2) \overset{(b)}{\le} \mathbb{E}\left(\eta_{\pi+Z,\mathrm{A}}(\pi+Z)-\pi\right)^2 \overset{(c)}{\le} \delta,$$

which is only valid when we meet the equality conditions for all the inequalities above. I.e, the penalty distribution A must be chosen to satisfy

$$(a) \quad t^\star(u^\star_{\text{DT}}) = \mathrm{A}_{(\epsilon^\star)};$$

$$(b) \quad \eta_{\pi+Z,\mathrm{A}}(Z) \overset{\mathcal{D}}{=} \eta_{\text{soft}}(Z;t^\star(u^\star_{\text{DT}}));$$

Notice that the condition (c) is automatically satisfied when the condition (a) is satisfied.

To design such A, it suffices to set a two-level penalty distribution $\mathrm{A} = \mathrm{A}_{t^\star(u^\star_{\mathrm{DT}}), rt^\star(u^\star_{\mathrm{DT}}), w}$ in (3.2.4) for carefully chosen $r(u)$ and $w(u)$, with $w(u) > \epsilon^\star$. Then the condition $w(u) > \epsilon^\star$ gives $\mathrm{A}_{(\epsilon^\star)} = t^\star(u^\star_{\mathrm{DT}})$ by design, thus we satisfy the condition (a). In words, the infinite input $\pi^\star + Z$ is assigned to match with the first level of the two-level SLOPE penalty A.

We now turn to the more difficult condition (b) and explicitly choose $r(u)$ and $w(u)$ so that it is satisfied. Before giving the exact values of $r(u)$ and $w(u)$ and showing how they lead to satisfying condition (b), we take some time to further investigate the sparsity of the SLOPE estimator, $\kappa(\pi, \mathrm{A})$. Recall that the zero-threshold, defined in Definition 3.4.1, is the value $\alpha(\pi, \mathrm{A})$ such that $\eta_{\pi+Z,\mathrm{A}}(x) = 0$ if and only if $|x| \leq \alpha(\pi, \mathrm{A})$ and by Proposition 3.7.5, we know that the zero-threshold must be equal to one of the two levels of SLOPE penalty.

Now, when w is small, few input values are subjected to the larger level of the penalty and of those inputs, all will correspond to infinite signal prior elements. Thus, the zero-threshold will be the smaller level of the penalty, namely it equals $rt^\star(u^\star_{\mathrm{DT}})$ (visualized in Figure 3.10(a)(b)). In more details, for small $w(u)$, the value $r(u)$ controls the sparsity in the sense that

$$\kappa(\pi, \mathrm{A}) = \mathbb{P}\left(|\pi + Z| > rt^\star(u^\star_{\mathrm{DT}})\right) = \epsilon^\star + (1 - \epsilon^\star)\,\mathbb{P}(|Z| > rt^\star(u^\star_{\mathrm{DT}})).$$

Following the above equation, there exists an one-to-one map between $\mathrm{TPP}^\infty = u$

and $r(u)$ to achieve the desired sparsity of (3.7.19):

$$
\begin{aligned}
1-\frac{\epsilon(1-u)(1-\epsilon^\star)}{\epsilon-\epsilon^\star}=\kappa(\pi,\mathrm{A})&=\epsilon^\star+(1-\epsilon^\star)\,\mathbb{P}\left(|Z|>rt^\star(u^\star_{\mathrm{DT}})\right)\\
&=\epsilon^\star+2(1-\epsilon^\star)\Phi\Big(-rt^\star(u^\star_{\mathrm{DT}})\Big).
\end{aligned}
$$

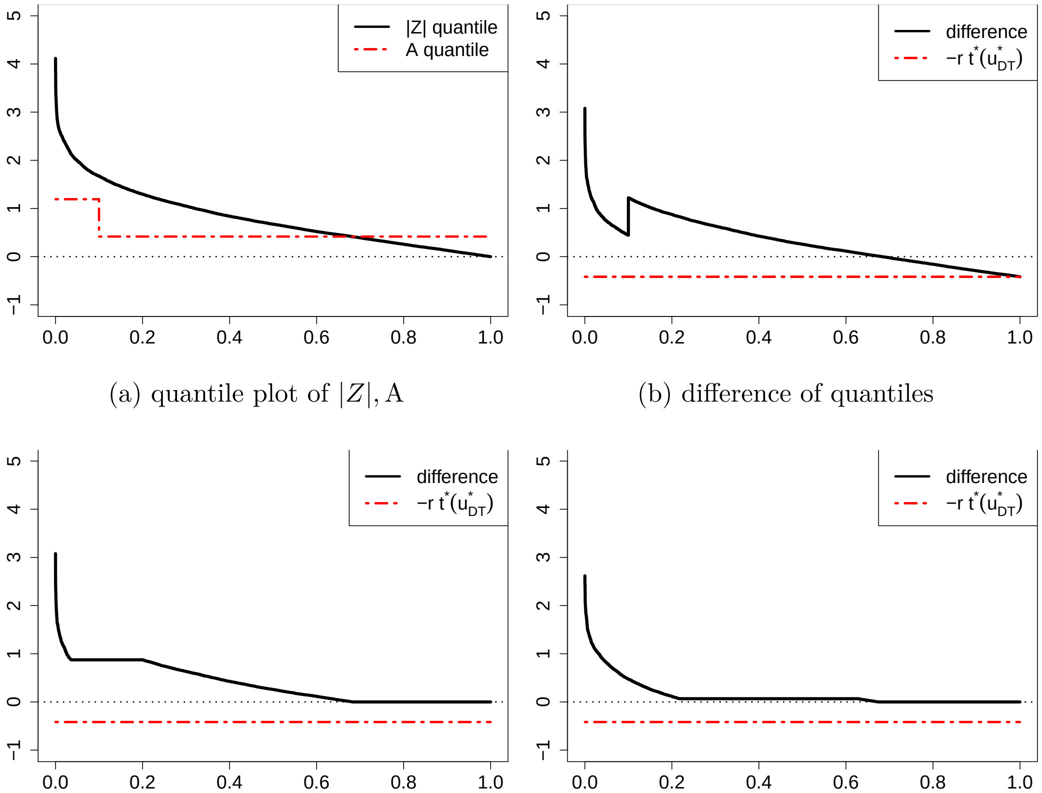
Explicitly, by rearranging the above, we conclude that the sparsity condition (3.7.19) is satisfied if one sets

$$
r(u)=\Phi^{-1}\left(\frac{2\epsilon-\epsilon^\star-\epsilon u}{2(\epsilon-\epsilon^\star)}\right)/t^\star(u^\star_{\mathrm{DT}}),
$$

given that $rt^\star(u^\star_{\mathrm{DT}})$ is the zero-threshold. In what follows, we always aim to keep the zero-threshold at $rt^\star(u^\star_{\mathrm{DT}})$.

As w increases, more and more input values are subjected to the larger level of the penalty. Thus, the zero-threshold and sparsity will remain the same, taking as values the second level of the penalty, $rt^\star(u^\star_{\mathrm{DT}})$, and that in (3.7.19), respectively, until w moves above a certain bound and forces the zero-threshold to increase to the larger level of A (again by Proposition 3.7.5 the zero-threshold can only take these two levels).

Moreover, as w increases to this bound, we observe that $\eta_{\pi+Z,A}(Z)$ becomes more similar to $\eta_{\mathrm{soft}}(Z;t^\star(u^\star_{\mathrm{DT}}))$, as demonstrated in Figure 3.10, and hence $\mathbb{E}\eta_{\pi+Z,A}(Z)^2$ becomes more similar to $\mathbb{E}\eta_{\mathrm{soft}}(Z;t^\star(u^\star_{\mathrm{DT}}))^2$. To observe this similarity property rigorously, notice that the quantile function of $\pi+Z$ has a sharp drop at $\epsilon^\star$ since $\mathbb{P}(\pi=\infty)=\epsilon^\star$, which splits the quantiles corresponding to $\infty+Z$ and to Z. Accordingly, Fact 3.7.3 says that for the input value corresponding to the 'infinity'



(a) quantile plot of $|Z|, \mathrm{A}$

(b) difference of quantiles

(c) averaging and truncating

(d) larger w

Figure 3.10: Fixing r and varying w within some range remains the sparsity of the SLOPE estimator (around 0.6) but forces the SLOPE proximal operator to approach the soft-thresholding. In other words, as w increases, the flat (averaged) quantile of $\eta_{\pi+Z,\mathrm{A}}(\pi + Z)$ takes a magnitude converging to zero. This implies that $\eta_{Z,\mathrm{A}^*}(Z) \to \eta_{\mathrm{soft}}(Z; t^\star(u^\star_{\mathrm{DT}}))$. Here $\pi = 0, \mathrm{A} = \mathrm{A}_{t^\star(u^\star_{\mathrm{DT}}),rt^\star(u^\star_{\mathrm{DT}}),w}, \delta = 0.3, \epsilon = 0.2, \epsilon^\star = 0.087, t^\star(u^\star_{\mathrm{DT}}) = 1.1924, u = 0.8176, r(u) = 0.3500, w(u) = 0.4819$.

part of the signal, $\infty + Z$, since $w > \epsilon^\star$, the SLOPE assigns a penalty given by the upper $\epsilon^\star$ quantiles of A, namely $\left(\mathrm{A}|\mathrm{A} \geq \mathrm{A}_{(\epsilon^\star)}\right) = \mathrm{A}_{(\epsilon^\star)} = t^\star(u^\star_{\mathrm{DT}})$ and for the inputs corresponding to the ‘nothing’ part of the signal, Z, SLOPE assigns a penalty given by the lower $1-\epsilon^\star$ quantiles of A, denoted by

$$\mathrm{A}^*(w) := \Big(\mathrm{A}|\mathrm{A} \le \mathrm{A}_{(\epsilon^\star)}\Big) = \mathrm{A}_{t^\star(u^\star_{\mathrm{DT}}), rt^\star(u^\star_{\mathrm{DT}}), \frac{w-\epsilon^\star}{1-\epsilon^\star}}.$$

Notice that what the above says is that a fraction, $\frac{w-\epsilon^\star}{1-\epsilon^\star}$, of the 'nothing' signals Z match with the large penalty $t^\star(u^\star_{\mathrm{DT}})$ and the remaining fraction match with the smaller penalty $rt^\star(u^\star_{\mathrm{DT}})$. In this way, when considering just the 'nothing' part of the signal, we can write $\eta_{\pi+Z,\mathrm{A}}(Z) \overset{\mathcal{D}}{=} \eta_{Z,\mathrm{A}^*(w)}(Z)$[6].

We now determine the exact $w(u)$ such that $\eta_{Z,\mathrm{A}^*(w)}(Z) \overset{\mathcal{D}}{=} \eta_{\mathrm{soft}}(Z; t^\star(u^\star_{\mathrm{DT}}))$, so as to satisfy condition (b). Our strategy is to select a $w(u)$ such that we are able to divide the output of $\eta_{Z,\mathrm{A}^*(w)}(Z)$ into clearly non-zero, arbitrarily close to zero and zero parts, so as to look like the soft-thresholding function as desired. That we can do this is visualized in Figure 3.10 and follows from the fact that with the two-level penalty, there will be only one flat averaged region in the output of $\eta_{Z,\mathrm{A}^*(w)}(Z)$, which we want to suppress to almost zero. Denoting

$$P_1 := \mathbb{P}(|Z| > t^\star(u^\star_{\mathrm{DT}})), \quad P_2 := \frac{w-\epsilon^\star}{1-\epsilon^\star}, \quad \text{and} \quad P_3 := \mathbb{P}(|Z| > rt^\star(u^\star_{\mathrm{DT}})), \tag{3.7.22}$$

we quantitatively define these three parts (clearly non-zero, close to zero, and zero) as the quantiles of $\eta_{Z,\mathrm{A}^*(w)}(Z)$ on the probability intervals $(0, P_1)$, (P_1, P_3) and $(P_3, 1)$

[6] Notice that, because no averaging takes place at the wp-th position, $\mathrm{prox}_J(\boldsymbol{\pi} + \boldsymbol{Z}; \boldsymbol{\alpha}) = [\mathrm{prox}_J(\boldsymbol{\pi}^\star + \boldsymbol{Z}; (\alpha_1, \cdots, \alpha_{wp})), \mathrm{prox}_J(\boldsymbol{Z}; (\alpha_{wp+1}, \cdots, \alpha_p))]$, in which $[\cdot]$ means concatenation.

respectively: i.e. we want $w(u)$ such that

$$|\eta_{Z,\mathrm{A}^*(w)}(Z)| \overset{\mathcal{D}}{=} \begin{cases} \eta_{\mathrm{soft}}(|Z|\Big||Z| > t^\star(u^\star_{\mathrm{DT}}); t^\star(u^\star_{\mathrm{DT}})) & \text{w.p. } P_1 \\ 0.0001 & \text{w.p. } P_3 - P_1 \\ 0 & \text{w.p. } 1 - P_3 \end{cases}$$

Here 0.0001 can be an arbitrarily small positive constant, which tends to 0 as $w \nearrow w(u)$. By such a construction, we have met our goal: we have determined $w(u)$ such that $\eta_{Z,\mathrm{A}^*(w)}(Z) \overset{\mathcal{D}}{=} \eta_{\mathrm{soft}}(Z; t^\star(u^\star_{\mathrm{DT}}))$. For example, in Figure 3.10(d), $P_1 \approx 0.23$ and $P_3 \approx 0.68$. Given that the averaged sub-interval between P_1 and P_3 is arbitrarily close to zero, we can write the scaled conditional expectation of $|\eta_{Z,\mathrm{A}^*(w)}(Z)|$ being on the flat region as an integral of the quantile function:

$$\int_{P_1}^{P_2} (|Z|_{(x)} - t^\star(u^\star_{\mathrm{DT}}))dx + \int_{P_2}^{P_3} (|Z|_{(x)} - rt^\star(u^\star_{\mathrm{DT}}))dx$$
$$= \int_{P_1}^{P_3} \eta_{Z,\mathrm{A}^*(w)}(|Z|_{(x)})dx = 0.0001(P_3 - P_1).$$

Setting $w = w(u)$ and thus the right hand side to 0, and rearranging the equation,

$$\int_{P_1}^{P_3} |Z|_{(x)}dx = t^\star(u^\star_{\mathrm{DT}})(P_2 - P_1) + rt^\star(u^\star_{\mathrm{DT}})(P_3 - P_2) = t^\star(u^\star_{\mathrm{DT}})[(P_2 - P_1) + r(P_3 - P_2)],$$

where the left hand side is the scaled conditional expectation of the random variable $|Z|$ given $rt^\star(u^\star_{\mathrm{DT}}) < |Z| < t^\star(u^\star_{\mathrm{DT}})$, with an explicit form as

$$\begin{aligned}\int_{P_1}^{P_3} |Z|_{(x)}dx &= \mathbb{E}\left(|Z|\Big|rt^\star(u^\star_{\mathrm{DT}}) < |Z| < t^\star(u^\star_{\mathrm{DT}})\right)\mathbb{P}\left(rt^\star(u^\star_{\mathrm{DT}}) < |Z| < t^\star(u^\star_{\mathrm{DT}})\right) \\ &= \mathbb{E}\left(Z\Big|rt^\star(u^\star_{\mathrm{DT}}) < Z < t^\star(u^\star_{\mathrm{DT}})\right)\mathbb{P}\left(rt^\star(u^\star_{\mathrm{DT}}) < |Z| < t^\star(u^\star_{\mathrm{DT}})\right) \\ &= 2\phi(rt^\star(u^\star_{\mathrm{DT}})) - 2\phi(t^\star(u^\star_{\mathrm{DT}})),\end{aligned}$$

in which the last equality holds from a direct calculation of the expection of a two-sided truncated normal distribution. Hence, we have,

$$2\phi(rt^\star(u^\star_{\text{DT}})) - 2\phi(t^\star(u^\star_{\text{DT}})) = t^\star(u^\star_{\text{DT}})[(P_2 - P_1) + r(P_3 - P_2)],$$

which, upon rearrangement, gives

$$P_2 = \frac{P_1 - rP_3}{1-r} - \frac{2}{(1-r)}\left[\frac{\phi(t^\star(u^\star_{\text{DT}})) - \phi(rt^\star(u^\star_{\text{DT}}))}{t^\star(u^\star_{\text{DT}})}\right].$$

Then, plugging in the values in (3.7.22), the above simplifies to

$$w(u) = \epsilon^\star + \tfrac{2(1-\epsilon^\star)}{1-r}\left[\Phi(-t^\star(u^\star_{\text{DT}})) - r\Phi(-rt^\star(u^\star_{\text{DT}})) - \tfrac{\phi(-t^\star(u^\star_{\text{DT}}))-\phi(-rt^\star(u^\star_{\text{DT}}))}{t^\star(u^\star_{\text{DT}})}\right].$$

We claim $w(u)$ can be uniquely determined by $r(u)$, and $w(u)$ is clearly larger than $\epsilon^\star$ as the second term is positive. To see this, we study the term in the bracket and claim that its derivative over $t^\star$ is $(\phi(-t^\star) - \phi(-rt^\star))/(t^\star)^2$, which is negative and hence the term is larger than when $t^\star = \infty$, i.e. 0.

Combining (3.5.2) for designing $r(u)$ and (3.5.3) for designing $w(u)$, we can design the two-level SLOPE penalty that, together with infinity-or-nothing prior $\pi_\infty(\epsilon^\star/\epsilon)$, approaches $(u, q^\star(u))$ arbitrarily close. □

On a side note, if the w is larger than the specific choice in (3.5.3), i.e. when the flat quantile in Figure 3.10 drops below zero, the SLOPE proximal operator has the same effect as soft-thresholding and the analysis for the Lasso case follows. Consequently, TPP^∞ reduces to the interval $[0, u^\star_{\text{DT}})$. Graphically speaking, when one fixes r and increases w from 0 to 1 (similar to Figure 3.4), the SLOPE TPP^∞

will increase from $u^\star_{\mathrm{DT}}$ to above, until $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ touches the Möbius curve. Then TPP^∞ will suddenly jump below $u^\star_{\mathrm{DT}}$, once w is larger than (3.5.3), and then remain constant afterwards.

Achievable TPP–FDP region by SLOPE

The trade-off boundary curves $q_\star$ and $q^\star$ only provide information that splits the entire TPP–FDP region into two parts: the possibly achievable $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ and the unachievable ones. See the red and non-red regions in Figure 3.5. Although we have shown the achievability of the upper boundary $q^\star$ via Proposition 3.2.3, such achievability of the curve does not directly distinguish the achievability of the regions, until the recent work on Lasso by [Wan+20] which gives a complete Lasso TPP–FDP diagram.

Here we leverage the homotopy result in [Wan+20, Lemma 3.8] to bridge from the achievability of the boundary curve to the achievability of the region. Thus we establish the actually achievable region by SLOPE.

The idea of homotopy is quite intuitive: suppose there are two curves, Curve A (our upper boundary curve $q^\star$) and Curve B (the horizontal line $\mathrm{FDP}^\infty = 1 - \epsilon$), and a continuous transformation f moving from Curve A to B. During the movement, f sweeps out a region whose boundaries include Curve A and B, where every single point in this region is passed by the transforming curve during the transformation. Formally, we have a homotopy lemma below.

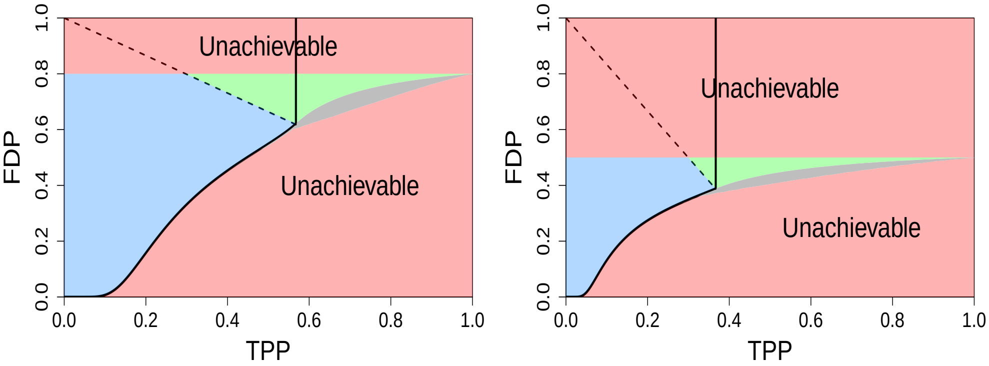


Figure 3.11: SLOPE TPP–FDP diagram by Proposition 3.7.8. Left: $(\delta, \epsilon) = (0.3, 0.2)$. Right: $(\delta, \epsilon) = (0.3, 0.5)$. The red regions are $(\mathrm{TPP}^\infty, \mathrm{FDP}^\infty)$ pairs not achievable by SLOPE nor by the Lasso, regardless of the prior distribution or the penalty tuning. The blue regions are achievable by both the SLOPE and by the Lasso. The green region is achievable only by the SLOPE but not by the Lasso. We note that the boundary between the blue region and the green one is a line segment connection (0,1) and $(u^\star_{\mathrm{DT}}, q^\star(u^\star_{\mathrm{DT}}))$, same as given by [Wan+20] for the Lasso case. The gray region is where the SLOPE trade-off lies in, and thus is possibly achievable by the SLOPE but not by the Lasso.

Lemma 3.7.7 (Lemma 3.7, [Wan+20])**.** *If a continuous curve is parameterized by* $f : [0,1] \times [0,1] \to \mathbb{R}^2$ *and if the four curves*

- $\mathcal{C}_1 = \{f(u,0) : 0 \leq u \leq 1\}$,
- $\mathcal{C}_2 = \{f(u,1) : 0 \leq u \leq 1\}$,
- $\mathcal{C}_3 = \{f(0,s) : 0 \leq s \leq 1\}$,
- $\mathcal{C}_4 = \{f(1,s) : 0 \leq s \leq 1\}$,

join together as a simple closed curve, $\mathcal{C} := \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4$*, then* $\mathcal{C}$ *encloses an interior area* $\mathcal{D}$*, and* $\forall (x,y) \in \mathcal{D}, \exists (u,s) \in [0,1] \times [0,1]$ *such that* $f(u,s) = (x,y)$*. In other words, every point inside the region* $\mathcal{D}$ *enclosed by curve* $\mathcal{C}$ *is realizable by some* $f(u,s)$*.*

Now we can show a region $\mathcal{D}_{\epsilon,\delta}$ defined below is indeed asymptotically achievable. This directly give the SLOPE TPP–FDP diagram in Figure 3.11.

Proposition 3.7.8. *Any* $(\text{TPP}^\infty, \text{FDP}^\infty)$ *in* $\mathcal{D}_{\epsilon,\delta}$ *is asymptotically achievable by the SLOPE. Here* $\delta < 1, \epsilon > \epsilon^\star(\delta)$ *and* $\mathcal{D}_{\epsilon,\delta}$ *is enclosed by the four curves:* $\text{FDP}^\infty = 1 - \epsilon, \text{FDP}^\infty = q^\star(\text{TPP}^\infty)$*,* $\text{TPP}^\infty = 0$ *and* $\text{TPP}^\infty = 1$*.*

Proof of Proposition 3.7.8. Note that $(\text{TPP}^\infty, \text{FDP}^\infty)$ is a function of Λ, σ and Π and hence we can denote every TPP–FDP point in $[0,1] \times [0,1]$ as

$$g = \left(\text{TPP}^\infty(\Lambda, \sigma, \Pi), \text{FDP}^\infty(\Lambda, \sigma, \Pi)\right).$$

To characterize the boundary of the achievable region, i.e. $\mathcal{C}_1$, we parameterize the (two-level) penalty distribution $\Lambda_*(u)$ and the (infinity-or-nothing) prior distribution $\Pi_*(u)$, empowered by the achievability result Corollary 3.5.1 (which holds for finite noise, including the noiseless case), such that

$$\mathrm{TPP}^\infty(\Pi_*(u), \Lambda_*(u)) = u,$$

$$\mathrm{FDP}^\infty(\Pi_*(u), \Lambda_*(u)) = q^\star(u).$$

Leveraging this parameterization, we define the transformation

$$f(u,s) = g(\Lambda_*(u), \tan(\frac{\pi s}{2}), \Pi_*(u)),$$

that is employed in Lemma 3.7.7. Therefore, $\mathcal{C}_1$ is the curve described by $q^\star$.

When the noise $\sigma = \infty$, clearly $\mathrm{FDP}^\infty(\Lambda, \infty, \Pi) = 1 - \epsilon$. It follows that $\mathcal{C}_2$ is $\mathrm{FDP}^\infty = 1 - \epsilon$. When $\mathrm{TPP}^\infty = u = 0$, this is the case that the penalty $\Lambda = \infty$ and we get $\mathcal{C}_3$ is $\mathrm{TPP}^\infty = 0$. When $u = 1$, we have $f(1, s) = (1, 1 - \epsilon)$. This is the case that the penalty $\Lambda = 0$ and $\mathcal{C}_4$ is $\mathrm{TPP}^\infty = 1$.

We notice that $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$ indeed composes a closed curve. Therefore, f sweeps from $\mathcal{C}_1$ to $\mathcal{C}_2$ and each point in $\mathcal{D}_{\epsilon,\delta}$ is achievable by some $f(u, s)$ by the homotopy lemma in Lemma 3.7.7. □

3.7.5 Lower bound not equal to upper bound

To complement Section 3.5.3, we give concrete examples that the upper bound $q^\star$ does not equal the lower bound $q_\star$. Visually, in Figure 3.2, it is not difficult to distinguish the two bounds when $\mathrm{TPP}^\infty \geq u^\star_{\mathrm{DT}}$. However, when $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$, the difference can be rather small (see Figure 3.12), but we assert that, at least for some $\mathrm{TPP}^\infty < u^\star_{\mathrm{DT}}$, the difference indeed exists and is not a result of numerical errors.

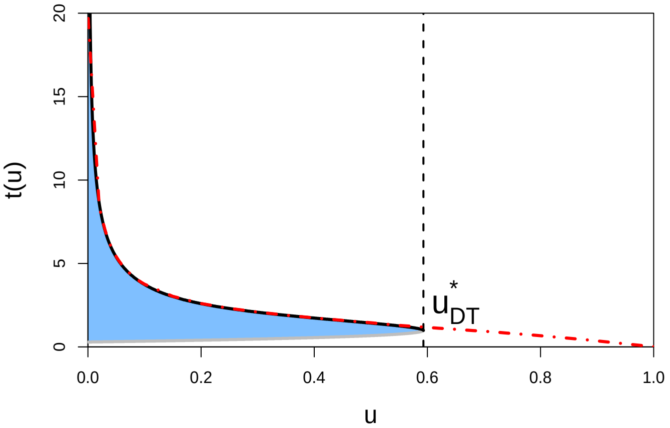


Figure 3.12: Demonstration of $t^\star(u)$ and $t_\star(u)$ with $\delta = 0.3, \epsilon = 0.2$. The blue area is valid x if (3.2.7) is an inequality with the left hand side being smaller than the right one; the black solid line is the larger root of (3.2.7), i.e. $t^\star(u)$, with the support $[0, u^\star_{\mathrm{DT}})$; the gray line is the smaller root. The red dotted line is $t_\star(u)$, with support $[0, 1]$. Note that if $t^\star(u) > t_\star(u)$ at some u, then $q^\star(u) > q_\star(u)$.

Characterizing the analytic SLOPE penalty

In order to characterize the optimal SLOPE penalty analytically, we discuss the complementary slackness condition on the monotonicity constraint in problem (3.4.6). We start with the case when the monotonicity constraint is **not binding** (i.e. when $\mathrm{A}'_{\mathrm{eff}}(z) > 0$ for all $z \geq \alpha$). We apply the Euler-Lagrange Multilplier Theorem to derive the following Euler-Lagrange equation on $L(z, \mathrm{A}_{\text{eff}})$ defined in (3.4.7):

$$\frac{\partial L}{\partial \mathrm{A}_{\text{eff}}} - \frac{d}{dz}\frac{\partial L}{\partial \mathrm{A}'_{\text{eff}}} = 0.$$

This is a necessary condition of the optimal SLOPE penalty function. Since L does not explicitly depend on A'_{eff}, the Euler-Lagrange equation gives, for the optimal SLOPE penalty function $\mathrm{A}^*_{\text{eff}}$ of problem (3.4.6),

$$\begin{aligned}&\frac{\partial L}{\partial \mathrm{A}_{\text{eff}}} = 4(1-\epsilon)(\mathrm{A}^*_{\text{eff}}(z) - z)\phi(z)\\ &+2\epsilon \sum_{j=1,2} p_j \left[\Big(\mathrm{A}^*_{\text{eff}}(z) - (z - t_j)\Big)\phi(z-t_j) + \Big(\mathrm{A}^*_{\text{eff}}(z) - (z + t_j)\Big)\phi(z+t_j)\right] = 0.\end{aligned}$$

This equation can be significantly simplified: if we denote a function

$$H(z) := 4(1-\epsilon)\phi(z) + 2\epsilon \sum_{j=1,2} p_j \left[\phi(z-t_j) + \phi(z+t_j)\right],$$

then the Euler-Lagrange equation above claims that

$$H(z)\mathrm{A}^*_{\text{eff}}(z) + H'(z) = 0,$$

which is equivalent to

$$\mathrm{A}^*_{\text{eff}}(z) = -H'(z)/H(z).$$

On the other hand, when the monotonicity constraint is **binding** (i.e. when $\mathrm{A}'_{\text{eff}}(z) = 0$ for all $z \geq \alpha$), clearly the penalty function $\mathrm{A}^*_{\text{eff}}$ is a constant. In short, the optimal penalty function $\mathrm{A}^*_{\text{eff}}$ coincides with the function $-H'/H$ in the interval (α, ∞) when $\mathrm{A}^*_{\text{eff}}$ is strictly increasing and stays (piecewise) constant elsewhere; in particular, $\mathrm{A}^*_{\text{eff}}(z) = \alpha$ on $[0, \alpha]$.

Unfortunately, the conclusion so far only gives the necessary but not sufficient condition for any penalty function to be optimal. Putting differently, the condition is not specific enough to uniquely determine $\mathrm{A}^*_{\mathrm{eff}}$ and thus we have to rely on the numerical approach to find $\mathrm{A}^*_{\mathrm{eff}}$, as shown in Section 3.4. Nevertheless, the condition we derived above will serve as an essential tool to build up analytic SLOPE penalty in the following sections.

Analytic SLOPE penalty for two-point prior

Here we derive the optimal SLOPE penalty analytically for a special two-point prior, which can be used to prove $q_\star(u) < q^\star(u)$ for some u, including those below the DT power limit. We review what is known for the Lasso trade-off: fixing δ, ϵ and $\mathrm{TPP}^\infty = u$, the maximum Lasso zero-threshold $t^\star(u)$ satisfies (3.2.7) and the minimum Lasso FDP^∞ is achieved at such threshold (see (3.2.8)). If for SLOPE we can find a larger zero-threshold α than $t^\star(u)$, then by the definition in (3.3.7):

$$\mathrm{FDP}^\infty(\Pi, \Lambda) = \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon u}$$

for the SLOPE must be smaller than the minimum Lasso FDP^∞.

We first determine the prior that we want to study. We focus on a zero-or-constant prior

$$\pi = \begin{cases} t_1 & \text{w.p. } \epsilon \\ 0 & \text{w.p. } 1-\epsilon, \end{cases}$$

whose probability density function is $p_\pi(t) = (1-\epsilon)\delta(t) + \epsilon\delta(t-t_1)$ and clearly $\pi^\star = t_1$. For the Lasso, from (3.3.7), we see that $\text{TPP}^\infty = u$ defines a unique t_1 by

$$\mathbb{P}(|t_1 + Z| > t^\star(u)) = \Phi(t_1 - t^\star) + \Phi(-t_1 - t^\star) = u,$$

i.e. $t_1(u, \delta, \epsilon)$ only depends on u, δ and ϵ.

Now that we have determined the prior, we seek a feasible SLOPE penalty function A_S which allows a larger zero-threshold α with this prior: let

$$\mathrm{A}_S(z) = \begin{cases} \alpha & \text{if } |z| < \alpha \\ \\ \max(\alpha, -H'_{t_1}(z)/H_{t_1}(z)) & \text{if } |z| \geq \alpha, \end{cases} \tag{3.7.23}$$

where $H_{t_1}(z)$ is the function $H(z)$ in the Section 3.7.5 but specific to our new prior, i.e. $p_1 = 1, p_2 = 0, t_1 = t_2$ in (3.4.5): we get

$$H_{t_1}(z) = 4(1-\epsilon)\phi(z) + 2\epsilon[\phi(z - t_1) + \phi(z + t_1)].$$

We remark that the SLOPE penalty A_S is clearly feasible for problems (3.4.2) and (3.4.6) if it is monotonically increasing. Furthermore, this monotonicity condition indeed holds true for some t_1 and α (such that A_S is increasing in z; we will give examples shortly), for which we can show $q_\star(u) < q^\star(u)$.

In summary, fixing (u, δ, ϵ), we can uniquely determine $t_1 \in \mathbb{R}$ for the two-point zero-or-constant prior and the maximum Lasso penalty $t^\star \in \mathbb{R}_+$. Looking at t_1, we can construct A_S using (3.7.23) on the interval $(\alpha = t^\star, \infty)$. If furthermore A_S is increasing, then this non-constant penalty A_S is feasible and must outperform the constant penalty of the Lasso (which is $t^\star$), based on the Euler-Lagrange equation discussed in Section 3.7.5. In consequence, the SLOPE allows strictly larger zero-threshold α than the maximum Lasso zero-threshold $t^\star$, until for some α we saturate the state evolution condition (3.4.1) by having $F_\alpha[\mathrm{A}_S, p_{\pi^\star}] = \delta$.

We give an example as follows for the framework described above.

An example of SLOPE FDP below the Lasso trade-off

As a concrete example of SLOPE FDP$^\infty$ being smaller than the minimum Lasso FDP$^\infty$, i.e. $q^\star_{\text{Lasso}}$, we use $\delta = 0.3, \epsilon = 0.2, \sigma = 1, u = u^\star_{\text{DT}}(\delta, \epsilon) = 0.56760$ by (3.2.3). Then the maximum Lasso zero-threshold $t^\star(u)$ (or equivalently the Lasso penalty A_L) equals 1.19241 by (3.2.7). In this case, the Lasso FDP$^\infty = 0.62160$ by (3.3.7). We can compute $t_1(u, \delta, \epsilon) = 1.34864$ by (3.3.7).

One can check that the function $-H'_{t_1}/H_{t_1}$ as well as the penalty function A_S in (3.7.23) (with α set as $t^\star$) are indeed increasing. Hence A_S is the unique optimal SLOPE penalty that satisfies the Euler-Lagrange equation. We can analytically compute the state evolution condition in problem (3.4.2) (see also (3.7.28) for the formula):

$$
\begin{aligned}
F_{t^\star}[\mathrm{A}_S, p_{\pi^\star}] &= 2(1-\epsilon)\int_{t^\star}^{\infty}(z - \mathrm{A}_S(z))^2\phi(z)dz + \epsilon\Big[t^2(\Phi(t^\star - t_1) - \Phi(-t^\star - t_1)) \\
&\quad + \int_{t^\star}^{\infty}\Big(\Big(z - t_1 - \mathrm{A}_S(z)\Big)^2\phi(z - t_1) + \Big(-z - t_1 + \mathrm{A}_S(z)\Big)^2\phi(-z - t_1)\Big)\,dz\Big] \\
&= \int_{t^\star}^{\infty}\left[\frac{1}{2}H_t(z)\mathrm{A}_S(z)^2 + H'_t(z)\mathrm{A}_S(z)\right]dz + \epsilon {t_1}^2(\Phi(t^\star - t_1) - \Phi(-t^\star - t_1)) \\
&\quad + 2(1-\epsilon)\int_{t^\star}^{\infty}z^2\phi(z)dz + \epsilon\int_{t^\star}^{\infty}(z - t_1)^2\phi(z - t_1)dz + \epsilon\int_{t^\star}^{\infty}(z + t_1)^2\phi(z + t_1)dz.
\end{aligned}
$$

Using the facts that $\ddot{\phi}(z) = (z^2 - 1)\phi(z)$ and $\dot{\phi}(z) = -z\phi(z)$, we get

$\frac{d}{dz}\left(\Phi(z) - z\phi(z)\right) = z^2\phi(z)$, which can be used to simplify the last three integrals:

$$F_{t^\star}[\mathrm{A}_S, p_{\pi^\star}] = \int_{t^\star}^{\infty} \left[\frac{1}{2}H_t(z)\mathrm{A}_S(z)^2 + H_t'(z)\mathrm{A}_S(z)\right] dz$$

$$+ \epsilon {t_1}^2(\Phi(t^\star - t_1) - \Phi(-t^\star - t)) + 2(1-\epsilon)\left[t^\star\phi(t^\star) + \Phi(-t^\star)\right]$$

$$+ \epsilon\left[(t^\star - t_1)\phi(t^\star - t_1) + \Phi(-t^\star + t_1)\right] + \epsilon\left[(t^\star + t_1)\phi(t^\star + t_1) + \Phi(-t^\star - t_1)\right].$$

Together with $\mathrm{A}_S(z) = \max\left(t^\star, -H_{t_1}'(z)/H_{t_1}(z)\right)$ on $(t^\star, \infty)$, this analytic quantity can be calculated by numerical integration to arbitrary precision, and it gives $E(\pi, \mathrm{A}_S) = 0.27727 < \delta$. In words, at the Lasso maximum zero-threshold $t^\star$, the SLOPE and the Lasso have the same FDP^∞, but the SLOPE has a smaller normalized estimation error E in the state evolution condition (3.4.1). Hence, this leaves a margin to further reduce the FDP^∞ before we use up the margin.

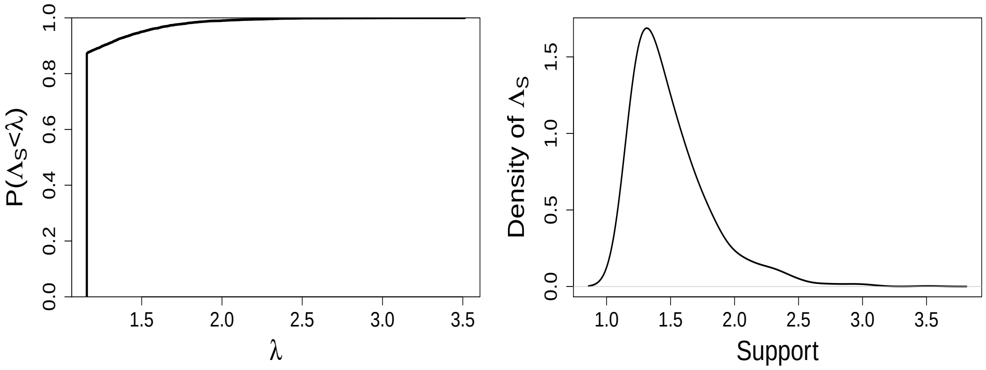


Figure 3.13: Cumulative distribution function of the optimal SLOPE penalty Λ_S and probability density of its non-constant component, at $(\text{TPP}^\infty, \text{FDP}^\infty) = (0.5676, 0.6216)$. Here $\delta = 0.3, \epsilon = 0.2, \sigma = 1, \Pi^\star = 4.9006$.

Up until now, we are working in the normalized regime on (π, A) and we want to determine the original prior and penalty (Π, Λ_S). To do so, we use the state evolution (3.7.1) to compute $\tau = \sqrt{\frac{\sigma^2}{1-E(\pi,\mathrm{A}_S)/\delta}} = 3.6337$, which uniquely defines the two-point prior via $\Pi^\star = t_1 \cdot \tau = 4.9006$. We then apply the calibration (3.7.5) to derive Λ_S and visualize the distribution in Figure 3.13.

To saturate the state evolution condition (3.4.1) so that $E(\pi, \mathrm{A}_S) = \delta$, while still fixing $\mathrm{TPP}^\infty = 0.5676$, we can increase the zero-threshold α from $t^\star$ (which is 1.19241) to 1.25672 and derive $t_1 = 1.41748$ via (3.3.7):

$$\mathbb{P}(|t_1 + Z| > \alpha) = \Phi(t_1 - \alpha) + \Phi(-t_1 - \alpha) = u.$$

Again, A_S constructed by (3.7.23) is increasing and optimal. This new SLOPE zero-threshold α implies $\mathrm{FDP}^\infty = 0.5954$ which is strictly smaller than the Lasso minimum $\mathrm{FDP}^\infty = 0.6216$.

A new TPP threshold $u^\dagger$

In this section, we find the minimum TPP^∞ such that we can leverage Section 3.7.5 to construct SLOPE FDP^∞ below the Lasso trade-off $q^\star_{\text{Lasso}}(\mathrm{TPP}^\infty)$.

When $\mathrm{TPP}^\infty = u$ and the zero-threshold α equals $t^\star(u)$ defined in (3.2.7), the SLOPE penalty may have a normalized estimation error $E(\pi, \mathrm{A}_S) < \delta$. In the above example, we increase the zero-threshold α until the state evolution constraint is binding: $E(\pi, \mathrm{A}_S) = \delta$, thus obtaining smaller FDP^∞. From a different angle, we can decrease u (and change $q^\star(u)$ and $t^\star(u)$ consequently) until $E(\pi, \mathrm{A}_S) = \delta$.

To be specific, we test a general $\mathrm{TPP}^\infty = u$ and set the zero-threshold α at $t^\star(u)$. Then the single point $\pi^\star = t_1$ can be computed via $\mathbb{P}(|t_1 + Z| > t^\star(u)) = u$ and the SLOPE penalty function A_S is determined via (3.7.23). Lastly, we compute the normalized estimation error $E(\pi, \mathrm{A}_S)$ if A_S is increasing.

We define the smallest u such that $E \le \delta$ as our new TPP threshold $u^\dagger$:

$$u^\dagger(\delta,\epsilon) := \inf\{u \text{ s.t. } F_{t^\star(u)}[\max(t^\star(u), -H'_{t_1}/H_{t_1}), \rho_{t_1}] \le \delta \\ \text{and } \max(t^\star(u), -H'_{t_1}/H_{t_1}) \text{ is increasing}\}.$$

Here the function $\rho_{t_1}(t) = \delta(t - t_1)$ is the probability density function of $\pi^\star = t_1$ and the functional F is defined in (3.7.28).

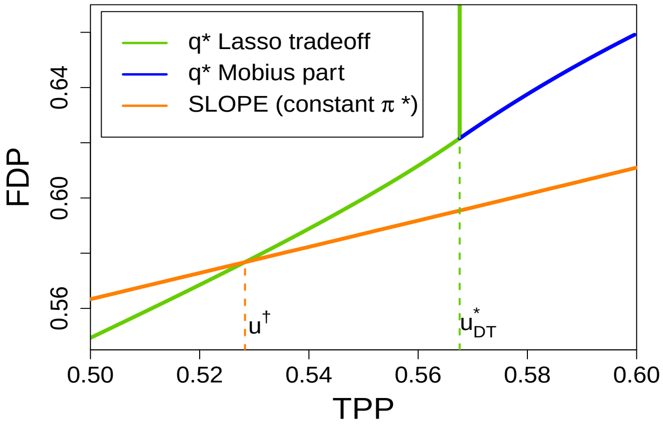


Figure 3.14: SLOPE TPP$^\infty$–FDP$^\infty$ of constant $\pi^\star$ and the new TPP threshold $u^\dagger$ when $(\delta,\epsilon) = (0.3, 0.2)$. The green line is the Lasso trade-off $q^\star$ and the blue line is the Möbius part of $q^\star$. The orange line is the TPP$^\infty$–FDP$^\infty$ realized by constant $\pi^\star$ and A_{eff} for the maximum zero-threshold α. In this example, $u^\dagger = 0.5283 < u^\star_{\text{DT}} = 0.5676$.

Under $\delta = 0.3, \epsilon = 0.2$, we find that $u^\dagger = 0.5283 < u^\star_{\text{DT}} = 0.5676$ (visualized in Figure 3.14). This indicates that we can show $q_\star < q^\star$ for a range of u smaller than the DT power limit. We observe that below $u^\dagger$, the Lasso penalty and the infinity-

or-nothing prior achieve smaller FDP$^\infty$ than our SLOPE penalty and constant prior, and vice versa. We further offer graphical demonstration of the difference between $u^\dagger$ and $u^\star_{\text{DT}}$ in Figure 3.15.

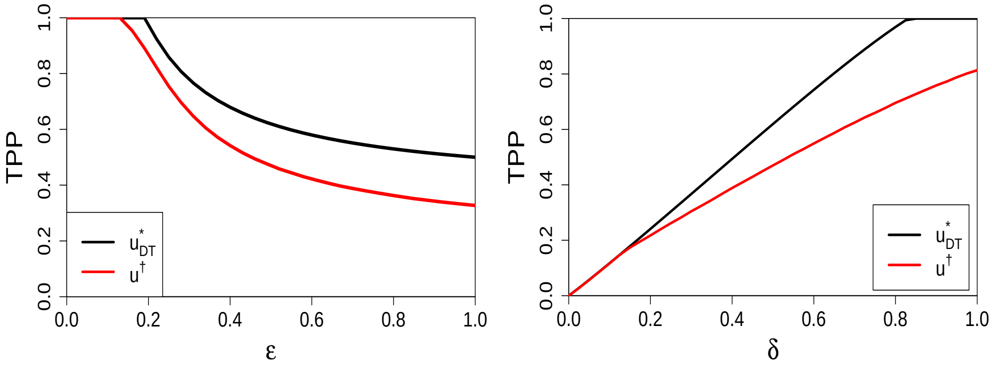


Figure 3.15: Comparison of $u^\dagger$ and $u^\star_{\text{DT}}$, fixing $\delta = 0.5$ (left) or $\epsilon = 0.5$ (right). The difference can be as large as 0.173 on the left plot and 0.282 on the right one.

3.7.6 Proving SLOPE outperforms the Lasso for fixed prior

Proof of Theorem 6. The proof is broken down into three pieces: we start with the MSE, then the asymptotic TPP and lastly the asymptotic FDP [7].

SLOPE has smaller MSE Fixing any bounded prior Π and any scalar Lasso penalty λ, we can derive the corresponding (τ_L, α_L) from the calibration (3.7.5) and the state evolution (3.7.1), so as to work in the normalized regime (π_L, A_L). The

[7] Theorem 6 can be generalized to further include certain unbounded signal prior Π as long as Equation (3.7.27) is satisfied. For example, for any Gaussian or Exponential Π, the SLOPE can outperform the Lasso.

quantity τ relates to the MSE by [Bu+20a, Corollary 3.4]:

$$\operatorname{plim} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p = \delta(\tau^2 - \sigma^2). \tag{3.7.24}$$

Obviously, the SLOPE estimator has a smaller MSE than the Lasso one if and only if $\tau_S < \tau_L$, where (τ_S, A_S) is the solution to the SLOPE calibration (3.7.5) and the state evolution (3.7.1).

We now illustrate that this is always feasible by carefully designing the SLOPE penalty vector $\boldsymbol{\alpha}_S(p)$, or in the asymptotic sense, the SLOPE penalty distribution A_S. We directly work with the SLOPE AMP state evolution (3.7.1) instead of the Lasso AMP, since SLOPE covers the Lasso as a sub-case. In particular, we consider the two-level SLOPE of the form $\mathrm{A}_{\ell,\alpha_L,w}$ defined in (3.2.4).

Our goal is to show that for any Lasso penalty $\mathrm{A}_L = \alpha_L = \mathrm{A}_{\alpha_L,\alpha_L,w}$, we can find a SLOPE penalty $\mathrm{A}_S = \mathrm{A}_{\ell,\alpha_L,w}$ for some sufficiently small w and $\ell > \alpha_l$ such that $\tau_S < \tau_L$. In other words, among all the two-level SLOPE penalties $\mathrm{A}_{\ell,\alpha_L,w}$ with a zero-threshold α_L, we show the optimal MSE is not achieved at $\ell = \alpha_L$.

To present a clear proof, we simplify the notation of $\operatorname{prox}_J(\boldsymbol{a}; \boldsymbol{b})$ by using $\boldsymbol{\eta}(\boldsymbol{a}, \boldsymbol{b})$, or simply $\boldsymbol{\eta}$, where $\boldsymbol{a} := \boldsymbol{\Pi} + \tau \boldsymbol{Z}$ and $\boldsymbol{b} := \boldsymbol{\alpha}\tau$, where $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\ell,\alpha_L,w}$ (defined in (3.2.4)) with $\ell \geq \alpha_L$. On convergence of the state evolution (3.7.1), we can differentiate both sides of

$$\tau^2 = \sigma^2 + \frac{1}{\delta} \lim_{p\to\infty} \mathbb{E}\langle[\boldsymbol{\eta}(\boldsymbol{\Pi} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau) - \boldsymbol{\Pi}]^2\rangle = \sigma^2 + \frac{1}{\delta} \lim_{p\to\infty} \mathbb{E}\langle[\boldsymbol{\eta}(\boldsymbol{a}, \boldsymbol{b}) - \boldsymbol{\Pi}]^2\rangle,$$

with respect to $\ell \in \mathbb{R}$. Denoting $\tau' = \frac{\partial \tau}{\partial \ell}$, we obtain

$$2\tau\tau' = \frac{\partial}{\partial \ell}\left(\sigma^2 + \frac{1}{\delta}\lim_{p\to\infty}\mathbb{E}\langle[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) - \boldsymbol{\Pi}]^2\rangle\right) = \frac{1}{\delta}\lim_{p\to\infty}\mathbb{E}\frac{\partial}{\partial \ell}\langle[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) - \boldsymbol{\Pi}]^2\rangle.$$

Then the chain rule leads to

$$\tau\tau' = \lim_{p\to\infty}\frac{1}{n}\sum_j \mathbb{E}(\eta_j - \Pi_j)\frac{\partial \eta_j}{\partial \ell} = \lim_{p\to\infty}\frac{1}{n}\sum_j \mathbb{E}(\eta_j - \Pi_j)\sum_k\left[\frac{d\eta_j}{da_k}Z_k\frac{\partial \tau}{\partial \ell} + \frac{d\eta_j}{db_k}\frac{\partial b_k}{\partial \ell}\right]. \tag{3.7.25}$$

To investigate the derivative terms, we copy some important facts in [Bu+20a, Appendix A] here for reader's convenience:

$$\begin{aligned}
&\frac{d}{da_k}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = \mathbb{I}\{|\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_j = |\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_k\}\,\mathrm{sign}([\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_k)[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})]_j,\\
&\frac{d}{db_k}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = -\mathbb{I}\Big\{|\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_j = |\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_{o(k)}\Big\}\,\mathrm{sign}\left([\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j\right)\Big[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})\Big]_j.
\end{aligned}$$

where the permutation $o : i \to j$ finds the index of the i-th largest magnitude, i.e. $|\eta|_{o(i)} := |\eta|_{(i)} = |\eta|_j$, and its inverse function is the rank of the magnitudes. In the above notation, we have used (again from [Bu+20a, Appendix A])

$$[\partial_1\eta]_j = \frac{\partial \eta_j}{\partial a_j} = \frac{1}{\Big|\{1 \le k \le p : |\eta_k| = |\eta_j|\}\Big|},$$

which converges to 1 as $p \to \infty$ if η_j is unique in the magnitudes of $\boldsymbol{\eta}$ and to 0 otherwise. In the Lasso case (i.e. $\ell = \alpha_L$), each non-zero entry in $|\boldsymbol{\eta}|$ is indeed unique, and hence we can simplify $[\partial_1\eta]_j$ to $\mathbb{I}\{|\eta|_j \neq 0\}$. We now rewrite (4.6.6) as

$$\begin{aligned}
\tau\tau' &= \lim_{p\to\infty} \frac{1}{n} \sum_j \mathbb{E}(\eta_j - \Pi_j) \left[[\partial_1\eta]_j Z_j \tau' - \text{sign}(\eta_j)[\partial_1\eta]_j \frac{\partial b_{o^{-1}(j)}}{\partial \ell} \right] \\
&= \lim_{p\to\infty} \frac{1}{n} \sum_{j:\eta_j\neq 0} \mathbb{E}(\eta_j - \Pi_j) \left[Z_j \tau' - \text{sign}(\eta_j) \frac{\partial b_{o^{-1}(j)}}{\partial \ell} \right] \\
&= \lim_p \frac{1}{n} \sum_{j:\eta_j\neq 0} \mathbb{E}(\eta_j - \Pi_j) \Big[Z_j \tau' - \text{sign}(\eta_j)(\alpha_{o^{-1}(j)}\tau' + \mathbb{I}\{o^{-1}(j) \leq wp\}\tau) \Big] \\
&= \lim_p \frac{1}{n} \sum_{j:\eta_j\neq 0} \mathbb{E}(\eta_j - \Pi_j) \left[\left(Z_j - \text{sign}(\eta_j)\alpha_L\right) \tau' - \text{sign}(\eta_j)\mathbb{I}\{o^{-1}(j) \leq wp\}\tau \right].
\end{aligned}$$

To summarize, we derive that

$$\frac{\partial \tau}{\partial \ell} = \frac{\lim_p \frac{1}{n} \sum_{j:o^{-1}(j)\leq wp} \mathbb{E}(\eta_j - \Pi_j)\,\text{sign}(\eta_j)\tau}{\lim_p \frac{1}{n} \sum_{j:\eta_j\neq 0} \mathbb{E}(\eta_j - \Pi_j)\left(Z_j - \text{sign}(\eta_j)\alpha_L\right) - \tau}. \tag{3.7.26}$$

The rest of the proof contains two statements: (1) We will show that the numerator term is positive for sufficiently small w; (2) We also show that the denominator term is always negative for any Lasso penalty α_L.

To show that the numerator in (3.7.26) is positive for small w, we write

$$\begin{aligned}
\lim_p \frac{\tau}{n} \sum_{j:o^{-1}(j)\leq wp} \mathbb{E}(\eta_j - \Pi_j)\,\text{sign}(\eta_j) =& \frac{w\tau}{\delta} \mathbb{E}\left[(\eta_j - \Pi_j)\,\text{sign}(\eta_j) \Big| o^{-1}(j) \leq wp \right] \\
=& \frac{w\tau^2}{\delta} \mathbb{E}\left[(\eta_{\text{soft}} - \pi)\,\text{sign}(\eta_{\text{soft}}) \Big| |\eta_{\text{soft}}| \geq q_w \right],
\end{aligned}$$

in which we slightly abuse the notation for the distribution $\eta \overset{\mathcal{D}}{:=} \eta_{\text{soft}}(\pi + Z; \alpha_L)$ and define q_w as the w-quantile of $|\eta_{\text{soft}}(\pi + Z; \alpha_L)|$ such that $\mathbb{P}\left(|\eta| \geq q_w\right) = w$.

Next, simple substitution gives that it is equivalent to show

$$\mathbb{E}\left(Z\,\text{sign}(\eta) \Big| |\eta| \geq q_w \right) > \alpha_L. \tag{3.7.27}$$

We notice that as $w \to 0$, $q_w \to \infty$. Hence we can always consider w small enough that the desired inequality above holds. The full proof of this fact is referred to Section 3.7.7.

The next step is to show that the denominator in (3.7.26) is negative, similar to the proof in [ZB21]. By multiplying with the positive τ, the denominator becomes

$$\begin{aligned}
&\lim_p \frac{1}{n} \sum_{j:\eta_j \neq 0} \mathbb{E}(\eta_j - \Pi_j)\left[Z_j - \operatorname{sign}(\eta_j)\alpha_L\right] - \tau \\
\propto &\lim_p \frac{1}{n} \sum_{j:\eta_j \neq 0} \mathbb{E}(\eta_j - \Pi_j)\left[\tau Z_j - \tau \operatorname{sign}(\eta_j)\alpha_L\right] - \tau^2 \\
= &\lim_p \frac{1}{n} \sum_{j:\eta_j \neq 0} \mathbb{E}(\eta_j - \Pi_j)^2 - \tau^2 \\
= &\lim_p \frac{1}{n} \sum_{j:\eta_j} \mathbb{E}(\eta_j - \Pi_j)^2 - \lim_p \frac{1}{n} \sum_{j:\eta_j = 0} \mathbb{E}(\eta_j - \Pi_j)^2 - \tau^2 \\
= &- \lim_p \frac{1}{n} \sum_{j:\eta_j = 0} \mathbb{E}(\eta_j - \Pi_j)^2 - \sigma^2 < 0,
\end{aligned}$$

where the last equality follows from (3.7.24).

All in all, we finish the proof that $\frac{\partial \tau}{\partial \ell}$ in (3.7.26) is negative for $\mathrm{A}_{\ell,\alpha_L,w}$ at $\ell = \alpha_L$ and small w. Along this negative gradient $\frac{\partial \tau}{\partial \ell}$, increasing the first argument of $\mathrm{A}_{\ell,\alpha_L,w}$ from $\ell = \alpha_L$ (the Lasso case) leads to a SLOPE penalty A_S and reduces τ_L to a smaller τ_S. Equivalently, the SLOPE MSE is strictly smaller than the Lasso MSE.

SLOPE has higher TPP To prove the TPP result, we need the SLOPE to have smaller MSE (as shown previsouly) as well as the same zero-threshold as the Lasso. To achieve this, we claim that, for sufficiently small w and some

$\ell > \alpha_L$, the SLOPE zero-threshold $\alpha(\Pi, \Lambda_S)$ is the same as the Lasso zero-threshold $\alpha(\Pi, \Lambda_L) = \alpha_L$.

In fact, the two-level SLOPE $\mathrm{A}_{\ell,\alpha_L,w}$ by its levels must have the zero-threshold as either ℓ or α_L (see Proposition 3.7.5), and the zero-threshold will be α_L if and only if the sparsity $\kappa(\Pi, \Lambda_S) > w$ (see Fact 3.7.2). Therefore it suffices to guarantee $\kappa(\Pi, \Lambda_S) > w$. From $\mathbb{P}(|\Pi/\tau_S + Z| > \ell) \leq \kappa(\Pi, \Lambda_S) \leq \mathbb{P}(|\Pi/\tau_S + Z| > \alpha_L)$, it is not hard to obtain that the sparsity κ is continuous in ℓ. Hence for any $w < \kappa(\Pi, \Lambda_L)$, there exists some $\ell > \alpha_L$ but close to α_L so that the SLOPE sparsity $\kappa(\Pi, \Lambda_S) > w$.

Now that we have $\tau_S < \tau_L$ and $\alpha(\Pi, \Lambda_S) = \alpha(\Pi, \Lambda_L)$, we can finish the proof by the definition of TPP^∞: intuitively, $\pi_S := \Pi/\tau_S > \pi_L := \Pi/\tau_L$ and SLOPE $\mathrm{TPP}^\infty = \mathbb{P}(|\pi_S^\star + Z| > \alpha_L) > \mathbb{P}(|\pi_L^\star + Z| > \alpha_L) =$ the Lasso TPP^∞; formally, we show by Equation (3.3.7) that

$$\mathrm{TPP}^\infty(\Pi, \Lambda_S) = \mathbb{P}(|\pi_S^\star + Z| > \alpha(\Pi, \Lambda_S)) = \int_{-\infty}^{\infty} \mathbb{P}(|t/\tau_S + Z| > \alpha_L) p_{\Pi^\star}(t) dt$$
$$> \int_{-\infty}^{\infty} \mathbb{P}(|t/\tau_L + Z| > \alpha_L) p_{\Pi^\star}(t) dt = \mathbb{P}(|\pi_L^\star + Z| > \alpha(\Pi, \Lambda_L)) = \mathrm{TPP}^\infty(\Pi, \Lambda_L).$$

SLOPE has lower FDP To prove the FDP result, we again use the fact that the SLOPE shares the same zero-threshold as the Lasso but has larger

TPP. By Equation (3.3.7),

$$\begin{aligned}\mathrm{FDP}^{\infty}(\Pi,\Lambda_S) =&\frac{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda_S))}{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda_S))+\epsilon\,\mathrm{TPP}^{\infty}(\Pi,\Lambda_S)}\\ <&\frac{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda_L))}{2(1-\epsilon)\Phi(-\alpha(\Pi,\Lambda_L))+\epsilon\,\mathrm{TPP}^{\infty}(\Pi,\Lambda_L)} = \mathrm{FDP}^{\infty}(\Pi,\Lambda_L).\end{aligned}$$

□

3.7.7 Auxiliary proofs

Here we give some technical proofs that have been used in this work.

Derivation of F_α in Section 3.4

We are ready to give the explicit form of the functional F_α given that the zero-threshold is $\alpha \in \mathbb{R}_+$. Expanding the term in the integral form, we get

$$\begin{aligned}F_\alpha[\mathrm{A}_{\mathrm{eff}}, p_{\pi^\star}] &:= \mathbb{E}\left(\eta_{\mathrm{soft}}(\pi+Z;\mathrm{A}_{\mathrm{eff}}(\pi+Z))-\pi\right)^2\\ &=\int_0^{\infty}\int_{-\infty}^{\infty}\Big(\eta_{\mathrm{soft}}\Big(t+z;\mathrm{A}_{\mathrm{eff}}(t+z)\Big)-t\Big)^2\phi(z)dzp_\pi(t)dt,\end{aligned}$$

where p_π is the probability density function of the normalized prior π, which is uniquely determined by $p_{\pi^\star}$ in a way to be explained shortly.

We further expand the quadratic term in the inner integral, by using the fact that $\eta_{\pi+Z,\mathrm{A}}(t+z) = \eta_{\mathrm{soft}}(t+z;\mathrm{A}_{\mathrm{eff}}(t+z)) = 0$ for $-\alpha \le t+z \le \alpha$, since α is the zero-threshold in Definition 3.4.1. We obtain

$$\begin{aligned}F_\alpha[\mathrm{A}_{\text{eff}}, p_{\pi^\star}] =&\int_0^\infty \Big[\int_{-\alpha-t}^{\alpha-t} t^2\phi(z)dz + \int_{\alpha-t}^{\infty}\Big(z-\mathrm{A}_{\text{eff}}(t+z)\Big)^2\phi(z)dz \\ &+\int_{-\infty}^{-\alpha-t}\Big(z+\mathrm{A}_{\text{eff}}(t+z)\Big)^2\phi(z)dz\Big]p_\pi(t)dt \\ =&\int_0^\infty \Big[t^2(\Phi(\alpha-t)-\Phi(-\alpha-t)) + \int_\alpha^\infty \Big(z-t-\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(z-t)dz \\ &+\int_\alpha^\infty \Big(-z-t+\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(-z-t)dz\Big]p_\pi(t)dt.\end{aligned}$$

By the definition of $\pi^\star$ in Lemma 3.3.1, we use $p_\pi(t) = (1-\epsilon)\delta(t) + \epsilon p_{\pi^\star}(t)$, in which $p_{\pi^\star}(t)$ is the probability density function of $\pi^\star$, to write

$$\begin{aligned}&F_\alpha[\mathrm{A}_{\text{eff}}, p_{\pi^\star}] \\ =&2(1-\epsilon)\int_\alpha^\infty (z-\mathrm{A}_{\text{eff}}(z))^2\phi(z)dz + \epsilon\int_0^\infty \Big[t^2(\Phi(\alpha-t)-\Phi(-\alpha-t)) \\ &+\int_\alpha^\infty \Big(z-t-\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(z-t)dz \\ &+\int_\alpha^\infty \Big(-z-t+\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(-z-t)dz\Big]p_{\pi^\star}(t)dt \qquad (3.7.28) \\ =&2(1-\epsilon)\int_\alpha^\infty (z-\mathrm{A}_{\text{eff}}(z))^2\phi(z)dz + \epsilon\int_0^\infty \Big[t^2(\Phi(\alpha-t)-\Phi(-\alpha-t)) \\ &+\int_\alpha^\infty \Big(\Big(z-t-\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(z-t) \\ &+\Big(-z-t+\mathrm{A}_{\text{eff}}(z)\Big)^2\phi(-z-t)\Big)dz\Big]p_{\pi^\star}(t)dt.\end{aligned}$$

Since Lemma 3.4.3 states that the optimal $p_{\pi^\star}(t)$ takes the form $\rho(t;t_1,t_2) =$

$p_1\delta(t-t_1)+p_2\delta(t-t_2)$ in (3.4.5), the above functional turns into

$$\begin{aligned}
&F_\alpha[\mathrm{A}_{\mathrm{eff}},\rho(\cdot;t_1,t_2)]\\
=&\int_\alpha^\infty \Big[2(1-\epsilon)(z-\mathrm{A}_{\mathrm{eff}}(z))^2\phi(z)\\
&+\epsilon p_1\Big(\Big(z-t_1-\mathrm{A}_{\mathrm{eff}}(z)\Big)^2\phi(z-t_1)+\Big(-z-t_1+\mathrm{A}_{\mathrm{eff}}(z)\Big)^2\phi(-z-t_1)\Big)\\
&+\epsilon p_2\Big(\Big(z-t_2-\mathrm{A}_{\mathrm{eff}}(z)\Big)^2\phi(z-t_2)+\Big(-z-t_2+\mathrm{A}_{\mathrm{eff}}(z)\Big)^2\phi(-z-t_2)\Big)\Big]dz\\
&+\epsilon p_1 t_1^2\Big[\Phi(\alpha-t_1)-\Phi(-\alpha-t_1)\Big]+\epsilon p_2 t_2^2\Big[\Phi(\alpha-t_2)-\Phi(-\alpha-t_2)\Big].
\end{aligned}$$

To construct the quadratic programming in problem (3.4.6), we can apply the left endpoint rule and approximate $F_\alpha[\mathrm{A}_{\mathrm{eff}},\rho(\cdot;t_1,t_2)]$ by

$$\begin{aligned}
\bar{F}_\alpha(\mathbf{A};t_1,t_2)=&\,2(1-\epsilon)\sum_{i=1}^m(z_i-\mathrm{A}_i)^2\phi(z_i)\Delta z\\
&+\epsilon p_1\sum_{i=1}^m\Big((z_i-t_1-\mathrm{A}_i)^2\phi(z_i-t_1)+(-z_i-t_1+\mathrm{A}_i)^2\phi(-z_i-t_1)\Big)\Delta z\\
&+\epsilon p_2\sum_{i=1}^m\Big((z_i-t_2-\mathrm{A}_i)^2\phi(z_i-t_2)+\Big(-z_i-t_2+\mathrm{A}_i)^2\phi(-z_i-t_2)\Big)\Delta z\\
&+\epsilon p_1 t_1^2\Big[\Phi(\alpha-t_1)-\Phi(-\alpha-t_1)\Big]+\epsilon p_2 t_2^2\Big[\Phi(\alpha-t_2)-\Phi(-\alpha-t_2)\Big].
\end{aligned}\tag{3.7.29}$$

Proof of Lemma 3.4.3

Proof. In general, ρ^* can always be approximated by a sum of Dirac delta functions, $\rho^*(t)=\sum_{i=1}^m p_i\delta(t-t_i)$. In particular, since ρ^* is a probability density function, we require $0<p_i<1$: otherwise if for some i, $p_i=1$, then $m=1$ and we are done.

We now show $m<3$ by contradiction. The vertex principle of linear programming states that the minimum value of the linear objective function occurs at the vertices of the feasible region. Hence it suffices to show that all vertices are two-point Dirac delta functions.

The constraints in problem (3.4.4) lead to

$$\sum_i p_i = 1, \quad \sum_i p_i[\Phi(t_i - \alpha) + \Phi(-t_i - \alpha)] = u.$$

Suppose $m \geq 3$, then there always exists $\rho'(t) = \sum_{i=1}^m p'_i \delta(t - t_i)$ such that

- $p_i = p'_i$ for $i > 3$;
- $p_1 + p_2 + p_3 = p'_1 + p'_2 + p'_3$;
- $p_1 h_\alpha(t_1) + p_2 h_\alpha(t_2) + p_3 h_\alpha(t_3) = p'_1 h_\alpha(t_1) + p'_2 h_\alpha(t_2) + p'_3 h_\alpha(t_3)$;

where we denote $h_\alpha(t_i) := \Phi(t_i - \alpha) + \Phi(-t_i - \alpha)$. In other words, we can find ρ' such that

$$\mathbf{0} = \begin{pmatrix} 1 & 1 & 1 \\ h_\alpha(t_1) & h_\alpha(t_2) & h_\alpha(t_3) \end{pmatrix} \left[\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} - \begin{pmatrix} p'_1 \\ p'_2 \\ p'_3 \end{pmatrix} \right].$$

Since there are only two equations involving the three unknown variables p'_1, p'_2 and p'_3, in the generic case, we can represent the infinitely many ρ' with one degree of freedom,

$$\begin{pmatrix} p'_1 \\ p'_2 \\ p'_3 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + s \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix},$$

using the null vector $(c_1, c_2, c_3)^\top$ of the above matrix.

As $0 < p_i' < 1$ for $i = 1, 2, 3$, we claim that for all $s \in (-s_0, s_0)$ with some $s_0 > 0$, $(p_1', p_2', p_3')^\top$ defined above is feasible for problem (3.4.4). In other words, suppose we explicitly define

$$\rho_s(t) = \sum_{i=1}^{3}(p_i + c_i s)\delta(t - t_i) + \sum_{i>3} p_i \delta(t - t_i),$$

then there exists a range of $s \in \mathbb{R}$ such that ρ_s is feasible. However, one can easily check that

$$\rho^* = \frac{1}{2}\left(\rho_s + \rho_{-s}\right)$$

is also a feasible solution. Hence ρ^* is not a vertex. Contradiction. □

Proof of Equation (3.7.27)

Proof. To see that for large enough q_w, $\mathbb{E}\left(Z\,\mathrm{sign}(\eta)\middle||\eta| > q_w\right) \geq \alpha_L$, we write $\tilde{q}_w := q_w + \alpha_L$ and study

$$\mathbb{E}\left(Z\,\mathrm{sign}(\eta)\middle||\eta| \geq q_w\right) = \mathbb{E}\left(Z\,\mathrm{sign}(\eta)\middle||\pi + Z| \geq \tilde{q}_w\right).$$

We have

$$\begin{aligned}
&\mathbb{E}\left(Z\,\mathrm{sign}(\eta)\middle||\pi + Z| \geq \tilde{q}_w\right)\\
=&\mathbb{E}\left(Z\middle|\pi + Z \geq \tilde{q}_w\right)\frac{\mathbb{P}(\pi + Z \geq \tilde{q}_w)}{\mathbb{P}(|\pi + Z| \geq \tilde{q}_w)} - \mathbb{E}\left(Z\middle|\pi + Z \leq -\tilde{q}_w\right)\frac{\mathbb{P}(\pi + Z \leq -\tilde{q}_w)}{\mathbb{P}(|\pi + Z| \geq \tilde{q}_w)}\\
=&\frac{\int_{-\infty}^{\infty}\left[\phi\left(t - \tilde{q}_w\right) + \phi\left(-\tilde{q}_w - t\right)\right]p_\pi(t)dt}{\int_{-\infty}^{\infty}\left[\Phi\left(t - \tilde{q}_w\right) + \Phi\left(-\tilde{q}_w - t\right)\right]p_\pi(t)dt},
\end{aligned}$$

in which $p_\pi(t)$ is the unknown but fixed probability density function of π.

Now we show cases where the above ratio of integrals goes to ∞ as $q_w \to \infty$ or equivalently $\tilde{q}_w \to \infty$. For bounded π (in fact for priors with bounded essential infimum and essential supreme), denoting the minimum and maximum as $\pi_{\min}$ and $\pi_{\max}$, then the ratio is

$$\frac{\int_{\pi_{\min}}^{\pi_{\max}} \left[\phi\left(t-\tilde{q}_w\right)+\phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}{\int_{\pi_{\min}}^{\pi_{\max}} \left[\Phi\left(t-\tilde{q}_w\right)+\Phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}.$$

Using the fact that $\phi(x)+x\Phi(x)>0$, we have

$$\begin{aligned}
&\mathbb{E}\left(Z\operatorname{sign}(\eta)\Big| |\pi+Z|\geq\tilde{q}_w\right)\\
>&\frac{\int_{\pi_{\min}}^{\pi_{\max}} \left[(\tilde{q}_w-t)\Phi\left(t-\tilde{q}_w\right)+(\tilde{q}_w+t)\Phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}{\int_{\pi_{\min}}^{\pi_{\max}} \left[\Phi\left(t-\tilde{q}_w\right)+\Phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}\\
>&\frac{\int_{\pi_{\min}}^{\pi_{\max}} \left[(\tilde{q}_w-\pi_{\max})\Phi\left(t-\tilde{q}_w\right)+(\tilde{q}_w+\pi_{\min})\Phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}{\int_{\pi_{\min}}^{\pi_{\max}} \left[\Phi\left(t-\tilde{q}_w\right)+\Phi\left(-\tilde{q}_w-t\right)\right] p_\pi(t)dt}\\
>&\min\{\tilde{q}_w-\pi_{\max},\tilde{q}_w+\pi_{\min}\}.
\end{aligned}$$

In summary, when $w\to 0$, all $q_w, \tilde{q}_w$ and $\mathbb{E}(Z\operatorname{sign}(\eta)\Big||\eta|\geq q_w)\to\infty$. Therefore we have $\mathbb{E}(Z\operatorname{sign}(\eta)\Big||\eta|\geq q_w)>\alpha_L$ for sufficiently small w. □

3.7.8 Computation of SLOPE AMP quantities

In order to compute $\boldsymbol{\alpha}$ and τ, e.g. for the AMP calibration or for computing the estimation error, we need to estimate the SLOPE proximal operator in the state evolution (3.7.2). Despite that the Monte Carlo method is easy to implement, it is often unstable nor efficient for high-dimensional SLOPE problems, say when p is in the order of thousands. Here we demonstrate how to approximate the normalized estimation error $E(\Pi, \Lambda)$ in a way that matches the truth asymptotically and has satisfactory approximation error in the finite dimension (see bottom-right plot in Figure 3.17).

Notice in this section, the prior distribution Π is general and does not necessarily satisfy the sparsity assumption $\mathbb{P}(\Pi \neq 0) = \epsilon$.

Approximating state evolution and calibration with quantiles

In state evolution (3.7.2), the expectation term can be difficult to evaluate because of the ordering and the non-separability of the sorted norm. In addition, the convolution between Π and Z also makes the estimation difficult.

We propose the following method for estimation: denote q_D as discretized quantile function of distribution D at $\{\frac{1}{p}, \frac{2}{p}, ..., \frac{p-1}{p}\}$. Denote Π as the true distribution of $\boldsymbol{\beta}$, $\boldsymbol{\Pi}_p$ as p-variate Π with i.i.d. entries; π as Π/τ and $\boldsymbol{\pi}_p$ as p-variate π with i.i.d. entries accordingly. Similarly denote standard normals Z and $\boldsymbol{Z}_p$. Assume $\boldsymbol{\alpha} \in \mathbb{R}^p$ is p-variate A with i.i.d. entries (in decreasing order). We can decompose

$$\begin{aligned}
&\mathbb{E}\langle[\text{prox}_J(\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p; \boldsymbol{\alpha}\tau) - \boldsymbol{\Pi}_p]^2\rangle\\
=&\mathbb{E}\langle[\text{prox}_J(\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p; \boldsymbol{\alpha}\tau)]^2\rangle + \mathbb{E}\langle\boldsymbol{\Pi}_p^2\rangle - \frac{2}{p}\mathbb{E}[\text{prox}_J(\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p; \boldsymbol{\alpha}\tau)^\top\boldsymbol{\Pi}_p]\\
=&\mathbb{E}\langle[\text{prox}_J(\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p; \boldsymbol{\alpha}\tau)]^2\rangle + \mathbb{E}\Pi^2 - \frac{2}{p}\mathbb{E}[\text{prox}_J(\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p; \boldsymbol{\alpha}\tau)^\top\boldsymbol{\Pi}_p]\\
\approx&\langle[\text{prox}_J(q_{\Pi+\tau Z}; q_{\text{A}\tau})]^2\rangle + \frac{1}{p}q_\Pi^\top q_\Pi - \frac{2}{p}\,\text{prox}_J(q_{\Pi+\tau Z}; q_{\text{A}\tau})^\top\mathbb{E}[\boldsymbol{\Pi}_p|\boldsymbol{\Pi}_p + \tau\boldsymbol{Z}_p = q_{\Pi+\tau Z}].
\end{aligned} \tag{3.7.30}$$

Such approximation for (3.7.30) is consistent and can be visualized in Figure 3.17.

This is due to the fact that the ordering and the signs do not affect the sum of squares and the property of Riemann Stieltjes integral (see [JP12; KF75; Rud+76]).

Fact 3.7.9 (Existence of Riemann Stieltjes integral)**.** *Suppose f is continuous and g is of bounded variation. For every $\epsilon > 0$, there exists $\delta > 0$ such that for every partition $P := (a = x_0 < x_1 < \cdots < x_p = b)$ with mesh$(P) < \delta$, and for every choice of points c_i in $[x_i, x_{i+1}]$, we have*

$$\left| S(P, f, g) - \int_a^b f(x) dg(x) \right| < \epsilon,$$

where $S(P, f, g) := \sum_{i=1}^{p-1} f(c_i)(g(x_{i+1}) - g(x_i))$.

We start with the simplest second term in (3.7.30). Set $f(x) = x^2$ and $g(x)$ to be cumulative distribution function of Π. Denote q_i as $\frac{i}{p}$-th quantile. Setting $c_i = q_i$ and $P = (q_1, ..., q_{p-1})$, then we have the approximate sum as

$$\begin{aligned} \frac{1}{p} q_\Pi^\top q_\Pi &= S(P, f, g) = q_1^2 \, \mathbb{P}(q_1 < \Pi < q_2) + \cdots + q_{p-1}^2 \, \mathbb{P}(q_{p-1} < \Pi < q_p) \\ &= \sum_i \int_{q_i}^{q_{i+1}} q_i^2 dg(x) \to \int_{-\infty}^{\infty} x^2 dg(x) = \mathbb{E}\Pi^2. \end{aligned}$$

Similarly, for the first term in (3.7.30), denote the distribution to which the empirical distribution of $\text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{\alpha}\tau)$ converges as $\widehat{\Pi}$. By approximating the Riemann Stieltjes integral twice, we have

$$\langle [\text{prox}_J(q_{\Pi+\tau Z}; q_{\mathrm{A}\tau})]^2 \rangle \to \mathbb{E}\widehat{\Pi}^2 = \text{plim}\, \mathbb{E}\langle [\text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{\alpha}\tau)]^2 \rangle.$$

For the last term in (3.7.30), we transform the term via the law of total expectation,

$$
\begin{aligned}
&\operatorname{plim}\frac{1}{p}\mathbb{E}[\operatorname{prox}_J(\mathbf{\Pi}+\tau\boldsymbol{Z};\boldsymbol{\alpha}\tau)^\top\mathbf{\Pi}]\\
=&\lim\frac{1}{p}\mathbb{E}[\operatorname{prox}_J(\mathbf{\Pi}+\tau\boldsymbol{Z};q_{\mathrm{A}\tau})^\top\mathbf{\Pi}]\\
=&\lim\frac{1}{p}\mathbb{E}_{\Pi+\tau Z}[\mathbb{E}_{\Pi,Z}[\operatorname{prox}_J(\mathbf{\Pi}+\tau\boldsymbol{Z};q_{\mathrm{A}\tau})^\top\mathbf{\Pi}\Big|\mathbf{\Pi}+\tau\boldsymbol{Z}]]\\
=&\lim\frac{1}{p}\mathbb{E}_{\Pi,Z}[\operatorname{prox}_J(q_{\Pi+\tau Z};q_{\mathrm{A}\tau})^\top\mathbf{\Pi}\Big|\mathbf{\Pi}+\tau\boldsymbol{Z}=q_{\Pi+\tau Z}]\\
=&\lim\frac{1}{p}\left(\operatorname{prox}_J(q_{\Pi+\tau Z};q_{\mathrm{A}\tau})^\top\mathbb{E}_{\Pi,Z}[\mathbf{\Pi}\Big|\mathbf{\Pi}+\tau\boldsymbol{Z}=q_{\Pi+\tau Z}]\right).
\end{aligned}
$$

Before we move on to the next section where we look at the conditional expectation term above, we pause to remark that the approximation via quantiles can be also used for computing the calibration (3.7.4): the calculation of $\mathbb{E}\|\operatorname{prox}_J(\mathbf{\Pi}+\tau\boldsymbol{Z};\boldsymbol{A}\tau)\|_0^*$ can be approximated by the number of unique values in $|\operatorname{prox}_J(q_{\Pi+\tau Z};q_{\mathrm{A}\tau})|$.

Closed-form of conditional expectation

The challenge remains on computing the vector $\mathbb{E}[\mathbf{\Pi}|\mathbf{\Pi}+\tau\boldsymbol{Z}=q_{\Pi+\tau Z}]$. We will derive its closed-form by applying the inverse transform sampling on each entry. The effect of approximation using our explicit form is demonstrated in Figure 3.16.

For any $q\in\mathbb{R}$, denoting the support of Π as supp and the probability density function as p_Π, we get

$$
\mathbb{E}[\Pi|\Pi+\tau Z=q]=\frac{\int_{\mathrm{supp}(\Pi)}x\cdot\frac{1}{\sqrt{2\pi}\tau}\exp(-\frac{(q-x)^2}{2\tau^2})p_\Pi(x)dx}{\int_{\mathrm{supp}(\Pi)}\frac{1}{\sqrt{2\pi}\tau}\exp(-\frac{(q-x)^2}{2\tau^2})p_\Pi(x)dx}.
$$

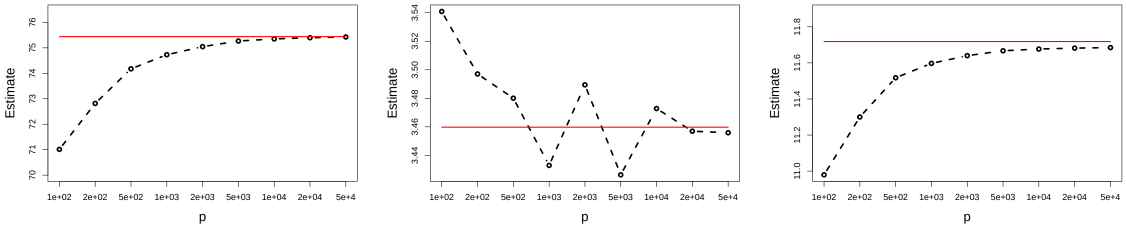


Figure 3.16: $\mathbb{E}\langle \text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{A}\tau)^\top \boldsymbol{\Pi} \rangle$ estimated by the quantiles and the closed-form conditional expectation (black dotted) against true expectation (red solid). Here $\delta = 0.3$ and A $\sim$ Exp(0.2)/10. Left: $\Pi \sim$ Exp(0.1). Middle: $\Pi \sim 10 \cdot$ Bernoulli(0.1). Right: $\Pi \sim \mathcal{N}(2, 25)$.

Substitute $u = q - x$,

$$\mathbb{E}[\Pi|\Pi + \tau Z = q] = q - \frac{\int_{\text{supp}(q-\Pi)} u \exp(-\frac{u^2}{2\tau^2}) p_\Pi(q-u) du}{\int_{\text{supp}(q-\Pi)} \exp(-\frac{u^2}{2\tau^2}) p_\Pi(q-u) du}.$$

Denote $s_q(u) := p_\Pi(q-u)\exp(-\frac{u^2}{2\tau^2}), S_q := \int_{-\infty}^{\infty} s_q(u)du$. Then $h_q(u) = s_q(u)/S_q$ is a normalized density of s_q. It can be viewed as a posterior density $h(u|q)$ with a Gaussian prior of u and $p_\Pi(q-u)$ as evidence. Then

$$\begin{aligned}\mathbb{E}[\Pi|\Pi + \tau Z = q] =& q - \frac{\int_{\text{supp}(q-\Pi)} u \cdot s_q(u) du}{\int_{\text{supp}(q-\Pi)} s_q(u) du} \\ =& q - \frac{\int_{\text{supp}(q-\Pi)} u \cdot h_q(u) du}{\int_{\text{supp}(q-\Pi)} h_q(u) du} = q - \mathbb{E}[U_q | U_q \in \text{supp}(q - \Pi)],\end{aligned}$$

where U_q is a univariate random variable with the density h_q.

Here we derive explicit formulae when the priors are Gaussian, exponential and Bernoulli. Two types of special generalization are worth mentioning: (1) when the support of Π is $(-\infty, \infty)$, the conditional expectation $\mathbb{E}[U_q|U_q \in \text{supp}(q - \Pi)]$ is indeed unconditional and the computation is simplified; (2) the cases of discrete priors can be easily derived in general besides the Bernoulli case.

Gaussian distribution When $\Pi = \mathcal{N}(\mu, \sigma^2)$, we have $\Pi + \tau Z = \mathcal{N}(\mu, \sigma^2 + \tau^2)$, and $h_q(u)$ is the density of $\mathcal{N}(\frac{(q-\mu)\tau^2}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$. Hence

$$\mathbb{E}[\Pi|\Pi + \tau Z = q] = q - \mathbb{E}[U_q] = q - \frac{(q-\mu)\tau^2}{\sigma^2+\tau^2} = \frac{q\sigma^2 + \mu\tau^2}{\sigma^2+\tau^2}.$$

Exponential distribution When $\Pi = \text{Exp}(c)$, we have $\Pi + \tau Z$ being an exponentially modified Gaussian (EMG) distribution. Then

$$s_q(u) = c\exp\left(-cq + \frac{c^2\tau^2}{2}\right)\left[\frac{1}{\sqrt{2\pi}\tau}\exp(-\frac{(u-c\tau^2)^2}{2\tau^2})\right],$$

and h_q is the density of $N(c\tau^2, \tau^2)$. And, denoting $\xi = \frac{q-c\tau^2}{\tau}$, we get

$$\mathbb{E}[\Pi|\Pi + \tau Z = q] = q - \mathbb{E}[U_q|U_q \in (-\infty, q)] = q - \mathbb{E}[U_q|U_q < q] = \tau\left(\xi + \frac{\phi(\xi)}{\Phi(\xi)}\right).$$

Discrete distribution We begin with $\Pi = k\cdot\text{Bernoulli}(\epsilon)$ and then generalize to any discrete priors. Writing the density as $(1-\epsilon)\delta(x) + \epsilon\delta(x-k)$,we have

$$\begin{aligned} h_q(u) &\propto [\epsilon\delta(q-u-k) + (1-\epsilon)\delta(q-u)]\exp(-\frac{u^2}{2\tau^2}) \\ &= [\epsilon\delta(u-(q-k)) + (1-\epsilon)\delta(u-q)]\exp(-\frac{u^2}{2\tau^2}). \end{aligned}$$

The last equality is true since the Dirac delta function is an even function. Hence

$$U_q = \begin{cases} q & \text{w.p. } (1-\epsilon)\exp(-\frac{q^2}{2\tau^2})/C \\ q-k & \text{w.p. } \epsilon\exp(-\frac{(q-k)^2}{2\tau^2})/C \end{cases}$$

where $C = (1-\epsilon)\exp(-\frac{q^2}{2\tau^2}) + \epsilon\exp(-\frac{(q-k)^2}{2\tau^2})$. We get

$$\begin{aligned} \mathbb{E}[\Pi|\Pi + \tau Z = q] =& q - \mathbb{E}[U_q|U_q \in \{q, q-k\}] = q - \mathbb{E}[U_q] \\ =& q - q\,\mathbb{P}[U_q = q] - (q-k)\,\mathbb{P}[U_q = q-k]. \end{aligned}$$

where both probabilities are given above.

Remark 3.7.10. It is easy to derive the conditional expectation for any discrete priors by writing the probability mass function as a sum of Dirac delta functions. In general, suppose the prior takes values in $a_1, ...a_n$ with probability $p_1, ...p_n$, then U_q takes values $(q-a_i)$ with probability $\mathbb{P}_i = p_i \exp(-\frac{(q-a_i)^2}{2\tau^2})/C$ where C is normalizing constant and the conditional expectation is $q - \sum_i (q-a_i)\,\mathbb{P}_i$.

Algorithm for state evolution term

Taking the quantile method and closed-form conditional expectation described in the previous sections, we present the following algorithm to compute the state evolution term efficiently.

Algorithm 3 Calculating $\mathbb{E}\langle[\text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{A}\tau) - \boldsymbol{\Pi}]^2\rangle$ given A, Π, τ

```
1. Derive quantiles $q_{\mathrm{A}}, q_{\Pi}$
2. Derive $D = \Pi + \tau Z$ by convolution and compute $q_D$
3. Compute $G = \text{prox}_J(q_D; q_{\mathrm{A}\tau}) \in \mathbb{R}^p$, (notice that $q_{\mathrm{A}\tau} = \tau q_{\mathrm{A}}$)
4. Compute $H = \mathbb{E}[\boldsymbol{\Pi} | \boldsymbol{\Pi} + \tau \boldsymbol{Z} = q_D] \in \mathbb{R}^p$
5. Return $\frac{1}{p}\left[G^\top G + q_\Pi^\top q_\Pi - 2G^\top H\right]$
```

We give some simulation results on different dimensions. Since all three priors show similar patterns, only the exponential prior case is plotted.

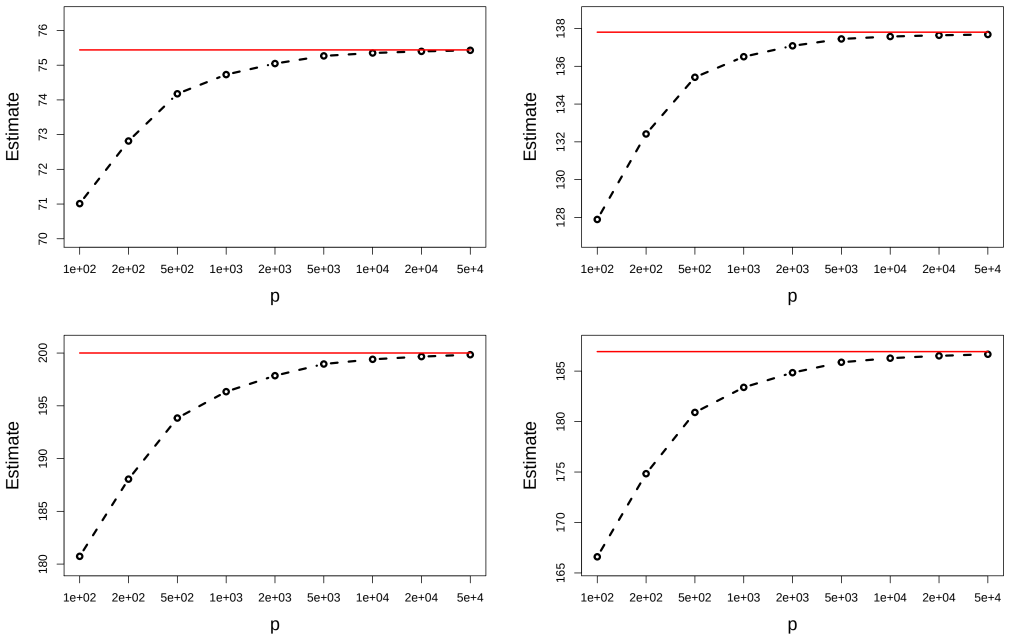


Figure 3.17: $\mathbb{E}\langle \text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{A}\tau)^\top \boldsymbol{\Pi}\rangle$ (top-left), $\mathbb{E}\langle[\text{prox}_J(\Pi + \tau Z; \alpha\tau)]^2\rangle$ (top-right), $\mathbb{E}[\Pi^2]$ (bottom-left) and $\mathbb{E}\langle[\text{prox}_J(\boldsymbol{\Pi} + \tau \boldsymbol{Z}; \boldsymbol{A}\tau) - \boldsymbol{\Pi}]^2\rangle$ (bottom-right) estimated by the quantiles (black dotted) against true expectation (red solid). Here $\sigma = 0$ (no noise), $\boldsymbol{X} \sim \mathcal{N}(0, 1/n), n/p = \delta = 0.3, \mathrm{A} \sim \text{Exp}(0.2)/10, \Pi \sim \text{Exp}(0.1)$.

3.7.9 Design of SLOPE penalty under fixed prior

This section studies the problem of minimizing FDP^∞ at fixed $\text{TPP}^\infty = u$ over all possible SLOPE penalties, when the prior Π is fixed. This problem has been investigated extensively in [HL19a] but not via the quadratic programming approach that we proposed in Section 3.4. We give the detailed procedure to find the SLOPE trade-off below.

1. Given Π, we try different τ from the small to the large, which defines $\pi := \Pi/\tau$

and $\pi^\star := \Pi^\star/\tau$. Denote the corresponding probability density function as $p_{\pi^\star}$.

2. Since $\mathrm{TPP}^\infty = u$, we require $\int_0^\infty [\Phi(t-\alpha)+\Phi(-t-\alpha)]p_{\pi^\star}(t)dt = u$ by the definition of the zero-threshold α. Note that $\alpha(\tau)$ is a unique scalar for a given π, or equivalently τ.

3. By (3.7.28), we have the formula of $F_\alpha[\mathrm{A}_{\mathrm{eff}}, p_\pi]$ which we want to minimize over all $\mathrm{A}_{\mathrm{eff}}$ under the same constraints as in (3.4.6). The minimization can again be achieved by discretization and via quadratic programming in (3.4.9), though the forms of Q, d are different due to the more generalized form of the prior. I.e.

$$\begin{aligned} \min_{\mathbf{A}} \quad & \frac{1}{2}\mathbf{A}^\top\mathbf{Q}\mathbf{A} - \mathbf{A}^\top\mathbf{d} \\ \text{s.t.} \quad & \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \mathbf{A} \geq \begin{pmatrix} \alpha(\tau) \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned} \tag{3.7.31}$$

where

$$\mathbf{Q} = \mathrm{diag}\left(2(1-\epsilon)\phi(\boldsymbol{z}) + \epsilon\int_0^\infty \Big[\phi(\boldsymbol{z}-t)+\phi(-\boldsymbol{z}-t)\Big]p_{\pi^\star}(t)dt\right),$$
$$\mathbf{d} = 2(1-\epsilon)\boldsymbol{z}\phi(\boldsymbol{z}) + \epsilon\int_0^\infty \Big[(\boldsymbol{z}-t)\phi(\boldsymbol{z}-t)+(\boldsymbol{z}+t)\phi(\boldsymbol{z}+t)\Big]p_{\pi^\star}dt,$$

4. The smallest τ that is valid, i.e. $F_\alpha[\mathrm{A}^*_{\mathrm{eff}}, p_{\pi^\star}] \leq \delta$ with $\mathrm{A}^*_{\mathrm{eff}}$ being the optimal penalty from the above quadratic programming, can be shown to correspond

to the largest zero-threshold α by Fact 3.7.11.

5. The largest zero-threshold α gives the minimum FDP via (3.8):

$$\frac{2(1-\epsilon)\Phi(-\alpha(u))}{2(1-\epsilon)\Phi(-\alpha(u))+\epsilon u}.$$

Fact 3.7.11. *Fixing the prior $\Pi = \pi\tau$ and under the condition $\int_0^\infty [\Phi(t-\alpha)+\Phi(-t-\alpha)]p_{\pi^\star}(t)dt = u$, we have $\frac{d\alpha}{d\tau} < 0$.*

Proof of Fact 3.7.11.

$$\begin{aligned}
&\mathbb{E}[\Phi(\Pi/\tau-\alpha)+\Phi(-\Pi/\tau-\alpha)] = u\\
\Longrightarrow\quad &\mathbb{E}[-(\frac{\Pi}{\tau^2}+\frac{d\alpha}{d\tau})\phi(\Pi/\tau-\alpha)+(\frac{\Pi}{\tau^2}-\frac{d\alpha}{d\tau})\phi(-\Pi/\tau-\alpha)] = 0\\
\Longrightarrow\quad &\frac{d\alpha}{d\tau} = \frac{\mathbb{E}\left[-(\frac{\Pi}{\tau^2})[\phi(\Pi/\tau-\alpha)-\phi(-\Pi/\tau-\alpha)]\right]}{\mathbb{E}[\phi(\Pi/\tau-\alpha)+\phi(-\Pi/\tau-\alpha)]} < 0.
\end{aligned}$$

□

Chapter 4

Efficient Designs of SLOPE Penalty Sequences in Finite Dimension

This chapter is based on "Yiliang Zhang, and Zhiqi Bu. "Efficient designs of slope penalty sequences in finite dimension." In International Conference on Artificial Intelligence and Statistics, pp. 3277-3285. PMLR, 2021.".

4.1 Introduction

In sparse linear regression, we aim to find an accurate sparse estimator $\hat{\boldsymbol{\beta}}$ of the unknown truth $\boldsymbol{\beta}$ from

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}.$$

Here the response $\boldsymbol{y} \in \mathbb{R}^n$, the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, the true parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ and the noise $\boldsymbol{w} \in \mathbb{R}^n$. Specifically, in high dimension where $p > n$, ordinary linear regression fails to find a unique solution and L_1-related regularization is usually introduced to achieve sparse estimators, including the Lasso [Tib96b], elastic net [ZH05b], (sparse) group Lasso [YL06b], adaptive Lasso [Zou06b] and the recent SLOPE [Bog+15b]:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)} \tag{4.1.1}$$

Here $\sum_{i=1}^{p} \lambda_i |b|_{(i)}$ is the *sorted* ℓ_1 *norm* of $\boldsymbol{b}$ governed by the penalty vector $\boldsymbol{\lambda} \in \mathbb{R}^p$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, and $|b|_{(i)}$ is the ordered statistics of absolute values $|b_i|$ such that $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$. We pause here to remark that the Lasso is a sub-case of SLOPE when $\lambda_1 = \cdots = \lambda_p$, since there is no need to sort and hence sorted ℓ_1 norm is simply the ℓ_1 norm. Generally speaking, the sorting step in the norm allows SLOPE to work in a way similar to the taxation, assigning larger thresholds to larger fitted coefficients.

Many desirable properties have been proven for SLOPE. For example, SLOPE is a convex optimization that can be solved by existing gradient methods, such as the subgradient descent and the proximal gradient descent; SLOPE achieves minimax estimation properties without requiring knowledge of the sparsity degree of $\boldsymbol{\beta}$ [SC+16]; SLOPE controls the false discovery rate in the case of independent predictors. However, understanding the SLOPE problem is difficult. Questions such as what posterior distribution does SLOPE solution follow, can we characterize statistics (e.g. the false discovery rate and true positive rate) from SLOPE exactly, whether SLOPE has better estimation error than the Lasso, are not answered until recently [BLT18; Bu+20b; HL19b]. Still, the substantial difficulty imposed by the sorted penalty impedes the general application of SLOPE for two reasons. From the practical point of view, tuning a $\mathbb{R}^p$ penalty can be extremely costly for large p (e.g. in high dimensional regression or over-parameterized neural networks) and naive methods that work for the Lasso, such as the grid search, renders not pragmatic. From a theoretical perspective, the sorted norm is complicated in that the effect of thresholding of SLOPE is *non-separable* and *data-dependent*, unlike the Lasso, thus making the analysis much involved.

In this paper, we further exploit the advantage of the data-depending penalty in SLOPE and investigate, from the estimation error perspective, how to design the SLOPE penalty sequence to achieve better performance.

We give a computationally efficient framework to design the SLOPE penalty sequence $\boldsymbol{\lambda} \in \mathbb{R}^p$ which corresponds to an estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ that minimizes the estimation error. To be more specific, we derive the gradient of penalty for SLOPE under the Approximate Message Passing (AMP) regime [BM11b; BM11d; DMM10; DMM09b] and propose the *k-level* SLOPE for the general data matrcies. In words, k-level SLOPE is a sub-class of SLOPE, where the p elements in $\{\lambda_i\}$ have only k unique values. Under this definition, the general SLOPE is p-level SLOPE and the Lasso is indeed 1-level SLOPE. Additionally, k-level SLOPE is a sub-class of $(k+1)$-level SLOPE, and larger k leads to better performance but requires longer computation time. As a result, by choosing an appropriate k, we can establish a trade-off between speed and accuracy. We illustrate in various experiments that such a trade-off is of practical use as even a small k may improve the performance non-trivially.

4.1.1 Notations

We start by introducing the proximal operator of SLOPE,

$$\operatorname{prox}_{J_{\boldsymbol{\theta}}}(\boldsymbol{y}) := \underset{\boldsymbol{b}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{b}\|^2 + J_{\boldsymbol{\theta}}(\boldsymbol{b}), \tag{4.1.2}$$

where $J_{\boldsymbol{\theta}}(\boldsymbol{b}) := \sum_{i=1}^{p} \theta_i |b|_{(i)}$ and the proximal operator indeed solves (4.1.1) with an identity data matrix. This operator is the building block that is iteratively applied to derive the SLOPE estimator in the proximal gradient descent (ISTA [DDDM04]) and in FISTA [BT09]. We note that there is no closed form of $\operatorname{prox}_{J_{\boldsymbol{\theta}}}(\boldsymbol{x})$ but it can be efficiently computed as in [Bog+15b, Algorithm 3]. Next we denote the mean squared error ($\texttt{MSE}$) between two vectors in $\mathbb{R}^m$ as $\texttt{MSE}(\boldsymbol{u}, \boldsymbol{v}) := \|\boldsymbol{u}-\boldsymbol{v}\|^2$. Two performance measures that are investigated in this work are the prediction error, $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$, and the estimation error, $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$.

4.2 SLOPE penalty design under AMP regime

4.2.1 Computing the gradients with respect to the penalty

We introduce a special regime of the AMP for SLOPE [Bu+20b], within which the SLOPE estimator can be asymptotically exactly characterized. A similar regime is the case when Convex Gaussian Min-max Theorem (CGMT) [CMW20; TAH18; TOH15; TOH14] applies, which shares similar assumptions as those of AMP. We then derive the gradient of $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ with respect to the penalty $\boldsymbol{\lambda}$ and optimize our penalty design iteratively. Generally speaking, AMP is a class of gradient-based optimization algorithms that mainly work on independent Gaussian random data matrices, offering both a sequence of estimators that converges to the true minimizer and a distributional characterization of the latter (see [Bu+20b, Theorem 3] and [HL19b, Theorem 1]). Here we present the five assumptions of the SLOPE AMP [Bu+20b]:

- The data matrix $\boldsymbol{X}$ has independent and identically-distributed (i.i.d.) gaussian entries that have mean 0 and variance $1/n$.
- The signal $\boldsymbol{\beta}$ has elements that are i.i.d. and follow Π, with $\mathbb{E}\left(\Pi^2 \max\{0, \log \Pi\}\right) < \infty$.
- The noise $\boldsymbol{w}$ is elementwise i.i.d. and follows W, with $\sigma_w^2 := \mathbb{E}\left(W^2\right) < \infty$.
- The vector $\boldsymbol{\lambda}(p) = (\lambda_1, \ldots, \lambda_p)$ is elementwise i.i.d. and follows Λ, with $\mathbb{E}\left(\Lambda^2\right) < \infty$.
- The ratio n/p reaches a constant $\delta \in (0, \infty)$ in the large system limit, as n and $p \to \infty$.

Under these assumptions, Theorem 3 in [Bu+20b] provides an asymptotic characterization of $\hat{\boldsymbol{\beta}}$, which can be informally interpreted as

$$\hat{\boldsymbol{\beta}} \approx \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta} + \tau \boldsymbol{Z}) \tag{4.2.1}$$

in which $(\boldsymbol{\alpha}, \tau)$ are the unique solutions of two key equations, namely the calibration and the state evolution in the AMP (or CGMT) regime (see [Bu+20b; HL19b]):

$$\boldsymbol{\lambda} = \boldsymbol{\alpha}\tau \left(1 - \frac{1}{n}\mathbb{E}\left\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta} + \tau \boldsymbol{Z})\right\|_0^*\right) \tag{4.2.2}$$

$$\tau^2 = \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\left\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta} + \tau \boldsymbol{Z}) - \boldsymbol{\beta}\right\|^2 \tag{4.2.3}$$

Here we assume the noise $\boldsymbol{w}$ has variance σ_w^2, $\|\cdot\|_0^*$ is a modified ℓ_0 norm that counts the unique non-zero absolute values in a vector and $\boldsymbol{Z} \in \mathbb{R}^p$ is a vector in which each element is i.i.d. standard normal. We denote $\delta := \lim_p n/p$ as the aspect ratio or sampling ratio and $\epsilon := \lim_p |\{j : \beta_j \neq 0\}|/p$.

Using (4.2.1), we observe that to minimize $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ is equivalent to finding desirable $(\boldsymbol{\alpha}, \tau)$. We now introduce some properties that are useful in deriving the desirable $\boldsymbol{\lambda}$, which uniquely defines $(\boldsymbol{\alpha}, \tau)$. By [Bu+20b, Proposition 2.3], the calibration (4.2.2) describes a bijective, monotone and parallel mapping Λ[1] between $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ [Bu+20b, Proposition 2.3], which allows us to work with $\boldsymbol{\alpha}$ easily instead of $\boldsymbol{\lambda}$. By [Bu+20b, Theorem 1], the state evolution (4.2.3) can be solved via a fixed point recursion, which converges to the unique solution $\tau(\boldsymbol{\alpha})$ monotonically under any initial condition.

[1] For a given $\boldsymbol{\alpha}$, we can use (4.2.3) to obtain a unique $\tau(\boldsymbol{\alpha})$ and leverage (4.2.2) to obtain a corresponding penalty vector $\boldsymbol{\lambda}(\boldsymbol{\alpha})$.

Under AMP region, our strategy is to design $\boldsymbol{\lambda} \in \mathbb{R}^p$ in SLOPE that, by quoting [Bu+20b, Corollary 3.2], minimizes:

$$\operatorname{plim} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p = \delta(\tau^2 - \sigma_w^2)$$

where plim is the probability limit. Hence minimizing $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ is in fact equivalent to minimizing τ, which depends on $\boldsymbol{\alpha}$ and leads to differentiating (4.2.3) against each of α_i for $i \in [p]$. In what follows, we view the scalar τ as a function of the penalty $\boldsymbol{\alpha}$ given the prior. Next, we use the gradient information to descend (with the projection elaborated in Algorithm 4) till convergence. Once the minimizer $\boldsymbol{\alpha}$ is obtained, we leverage the calibration (4.2.2) to map to the corresponding $\boldsymbol{\lambda}(\boldsymbol{\alpha})$.

In what follows, we shorthand $\operatorname{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})$ by using $\boldsymbol{\eta}(\boldsymbol{a}; \boldsymbol{b})$. In particular $\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})$ is denoted by $\boldsymbol{\eta}$ and we let η_j represent its j-th element. We define a set $I_j := \{k : |\eta_k| = |\eta_j|\}$, which will be used in characterization of gradients. We also define an inverse mapping for ranking of indices: $\sigma : \{1, \ldots, p\} \to \{1, \ldots, p\}$ such that $\sigma(i) = j$ representing $|\boldsymbol{\eta}|_{(i)} = |\eta_j|$. Consider a toy example $\boldsymbol{\eta} = (-2, -4, 3, 1)$, then the ranking of magnitudes is $(3, 1, 2, 4)$ whose inverse gives: $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$ and $\sigma(4) = 4$. This mapping is useful in assigning the penalties to coefficients in $\hat{\boldsymbol{\beta}}$ due to the sorting procedure.

We state the following theorem to give a concrete form of gradients $\partial\tau/\partial\alpha_i$, which is used in the projected gradient descent (PGD) in Algorithm 5.

Theorem 7. *The gradients satisfy*

$$\frac{\partial \tau}{\partial \alpha_i} = \mathbb{E}\frac{1}{|I_{\sigma(i)}|D(\boldsymbol{\alpha}, \tau)} \sum_{j \in I_{\sigma(i)}} (\eta_j - \beta_j) \operatorname{sign}(\eta_j)\tau \tag{4.2.4}$$

where $D(\boldsymbol{\alpha}, \tau)$ is a negative constant that is independent of index i.

Here the expectation is taken with respect to $\boldsymbol{Z}$ in $\boldsymbol{\eta}$, which in turn also affects $I_{\sigma(i)}$. The detailed form of $D(\boldsymbol{\alpha}, \tau)$ in the denominator and the proof of Theorem 7 can be found in Appendix 4.6.2, where we also claim that $D(\boldsymbol{\alpha}, \tau)$ is always negative. In practice, we can either set the step size s_t as constant or simply set $D = 1$ to save computation time. We remark that, using a constant step size and $\mathbb{E}\left(\sum_{j \in I_{\sigma(i)}} (\eta_j - \beta_j) \operatorname{sign}(\eta_j)\tau\right)/|I_{\sigma(i)}|$ as the gradient is equivalent to using a time-dependent $s_t = s \cdot D(\boldsymbol{\alpha}^t, \tau_t)$ and the actual gradient $\frac{\partial \tau}{\partial \alpha_i}$.

4.2.2 Projection onto non-negative decreasing vectors

We notice that $\boldsymbol{\alpha}$ (and $\boldsymbol{\lambda}$) must be non-negative and decreasing, hence the vanilla gradient descent is unsuitable for this constrained optimization problem of $\boldsymbol{\alpha}$. We design a projected gradient descent (PGD) in the following. To do so, we first give Algorithm 4 to compute the projection and establish the correctness of the algorithm in Theorem 8.

Let $\mathcal{S}$ denote the set of non-negative and decreasing vectors in $\mathbb{R}^p$ (i.e. $\boldsymbol{\lambda} \in \mathcal{S} \Rightarrow \lambda_i \geq \lambda_{i+1} \geq 0, \forall i$). Define the projection on to $\mathcal{S}$ as

$$\Pi_{\mathcal{S}}(\boldsymbol{\gamma}) = \operatorname{argmin}_{\boldsymbol{\gamma}' \in \mathcal{S}} \frac{1}{2} \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2^2. \tag{4.2.5}$$

Algorithm 4 `ProjectOnS` ($\Pi_{\mathcal{S}}$)

```
Input:   Arbitrary sequence $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$
for $i = 1, \cdots, p$ do
    ▷ Identify the shortest sub-sequence $\{\gamma_j, \ldots, \gamma_i\}$ whose average is smaller than its left neighbor (with $\gamma_0 = \infty$):
        $$\frac{1}{i-j+1} \sum_{k=j}^{i} \gamma_k \leq \gamma_{j-1}$$
    ▷ Assign the average value to such sub-sequence for $(\gamma_j, \ldots, \gamma_i)$:
        $$\gamma_j, \ldots, \gamma_i \leftarrow \frac{1}{i-j+1} \sum_{k=j}^{i} \gamma_k$$
Output:   $\max\{\boldsymbol{\gamma}, 0\}$        ▷ Element-wise truncation
```

We show that Algorithm 4 indeed finds the minimizer of (4.2.5) and provide the proof in Appendix 4.6.3.

Theorem 8. *Given an arbitrary $\boldsymbol{\gamma} \in \mathbb{R}^p$ as input, Algorithm 4 outputs the projection of $\boldsymbol{\gamma}$ on $\mathcal{S}$, that is, $\Pi_{\mathcal{S}}(\boldsymbol{\gamma})$.*

On high level, the proof consists of two parts. In the first part we provide a detailed characterization of $\Pi_{\mathcal{S}}(\gamma)$ by partitioning the index sequence $\{1, \ldots, p\}$ into a number of carefully selected sub-sequences. We prove that within each sub-sequence, $\Pi_{\mathcal{S}}(\boldsymbol{\gamma})$ takes the same value at each index, and such value is exactly the average of the sub-sequence $\boldsymbol{\gamma}$'s values at these indices. In the second part, we prove that Algorithm 1 indeed finds such sub-sequences and thus operates in a way that matches the goal of the projection $\Pi_{\mathcal{S}}(\gamma)$. The final truncation of the averaged sequence at 0 is a trivial method to guarantee the non-negativity.

4.2.3 Projected Gradient Descents

Now that we have the gradients in Theorem 7 and the projection in Algorithm 4, the projected gradient descent is straight-forward. In each iteration we first conduct gradient descent using Theorem 7 and transform the sequence from $\boldsymbol{\alpha}$ regime to $\boldsymbol{\lambda}$ regime. The transformation Λ is defined in Section 2.1 using the calibration and the state evolution in AMP. Then we project the gradient onto the constrained space $\mathcal{S}$ and transform it back to $\boldsymbol{\alpha}$ regime. The procedure is summarized in Algorithm 5.

Algorithm 5 `Projected Gradient Descent` (PGD)

```
  Input: initial $\boldsymbol{\alpha}^0$, step size $\{s_t\}$
  for $t = 1, \cdots, T$ do
      ▷ Gradient descent on $\boldsymbol{\alpha}$ and transform to $\boldsymbol{\lambda}$ regime
      $\boldsymbol{\gamma}^{t+1} = \Lambda(\boldsymbol{\alpha}^t - s_t \nabla_{\boldsymbol{\alpha}}(\tau(\boldsymbol{\alpha}^t)))$
      ▷ Project onto $\mathcal{S}$
      $\lambda^{t+1} = \texttt{ProjectOnS}(\gamma^{t+1})$
      ▷ Transform back to $\boldsymbol{\alpha}$ regime
      $\boldsymbol{\alpha}^{t+1} = \Lambda^{-1}(\lambda^{t+1})$
  Output: $\boldsymbol{\alpha}^{T+1}$
```

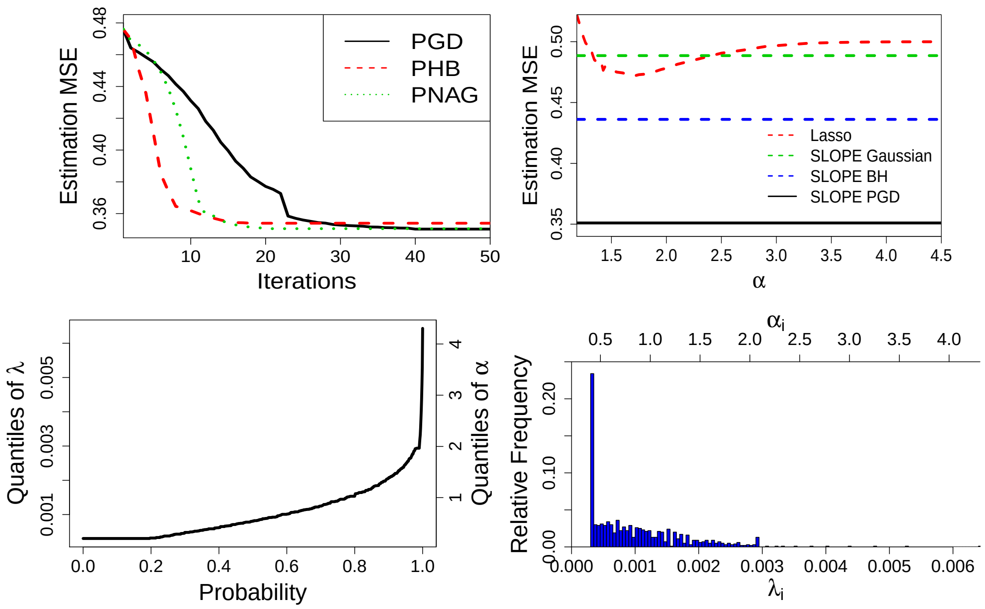


Figure 4.1: $\boldsymbol{X}$ i.i.d. $\mathcal{N}(0,1/n), n = 300, p = 1000, \Pi$ is Bernoulli(ϵ), $\delta = n/p = 0.3, \epsilon = 0.5, \sigma_w = 0$. Top-left: A sample run of PGD that finds the minimizing $\boldsymbol{\alpha}$, *not* the minimizing $\hat{\beta}$ for SLOPE in [Bu+20b]. Additionally, we plot two gradient descent methods with 0.9 momentum: the projected heavy ball (PHB) and the projected Nesterov accelerated gradient. All methods use the fine-tuned Lasso penalty as their starting points. Top-right: Red dashed line is Lasso MSE path (each point corresponds to one Lasso penalty, as λ varies from 0 to large values); other lines are different SLOPE MSE for a single SLOPE penalty. BH here stands for "Benjamini and Hochberg". Bottom-left: Best $\boldsymbol{\alpha}$ (right y-axis) and best $\boldsymbol{\lambda}$ (left y-axis) sequences found by PGD. Bottom-right: Histogram of best $\boldsymbol{\lambda}$ (bottom) and $\boldsymbol{\alpha}$ (upper) sequences found by PGD.

We highlight that Algorithm 5 is only one form of PGD. In fact, with a concrete form of the gradients, we can use any off-the-shelf first-order optimizer to find $\boldsymbol{\alpha}$ iteratively. Some examples include projected versions of stochastic gradient descent, Heavy Ball method [Pol64], Nesterov accelerated gradient descent [Nes83], Adagrad [DHS11], AdaDelta [Zei12] and Adam [KB14]. We include some of these optimizers in Figure 4.1.

To understand the convergence behavior of PGD, we need to study the convexity of the domain $\mathcal{S}$ and the objective function τ. Clearly $\mathcal{S}$ is convex by simply applying the definition. Unfortunately, $\tau(\boldsymbol{\alpha})$ for SLOPE is in general non-convex: even in the Lasso AMP regime, it is shown that $\tau(\alpha)$ is only a quasi-convex function of α [MMB+18b, Theorem 3.3]. We note that some non-convex problems may enjoy desirable properties such as having unique global minimum or local minima do not exist. As for SLOPE, the analysis on quasi-convexity of $\tau(\alpha)$ has not been well established but in practice, we do not observe any local minimum.

Remarkably, the gradient information that we use distinguishes our work from [HL19b]. We pause a bit and compare our approach with theirs, as they work under very similar assumptions as our AMP regime (in fact, both AMP and CGMT regimes agree asymptotically). Instead of optimizing directly on τ, they propose to optimize the proximal operator η in the functional space [HL19b, Proposition 3]: for a fixed candidate τ, they use the finite approximation with 2048 grids to solve a functional optimization, whose minimum is $\mathcal{L}(\tau)$. Next, they check the feasibility of the candidate τ by whether $\mathcal{L}(\tau) \leq \delta(\tau^2 - \sigma_w^2)$. Lastly, a binary search is conducted to find the optimal τ (smallest feasible τ) and the optimal design can be derived from the corresponding η. In summary, this approach took a detour by using a zeroth-order optimization algorithm, as the authors did not search over $\boldsymbol{\lambda}$ (or $\boldsymbol{\alpha}$) directly. Our first-order algorithm overcomes the seemingly unwieldy computation burden, especially in the high dimension when p is very large.

4.2.4 Transforming from α to λ

Once we find the desirable $\boldsymbol{\alpha}$ with Algorithm 5, the calibration (4.2.2) allows us to convert $\boldsymbol{\alpha}$ to $\boldsymbol{\lambda}$ in the original SLOPE problem. We demonstrate in Figure 4.1 and Figure 4.2 that SLOPE can outperform the best-tuned Lasso significantly. In Figure 1, SLOPE reduces $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ from 0.473 by Lasso to 0.350 by SLOPE, a 26% improvement in the estimation error. In fact, we observe from Figure 4.2 that SLOPE is even comparable to Minimum Mean Squared Error (MMSE; proposed by [BM11b]) estimator, which produces the lowest MSE possible. We emphasize that our result does not contradict [WWM19] which states that, under some conditions, the Lasso is the optimal SLOPE. We note that the condition in [WWM19, Theorem 2] does not hold for large systems: the premise of Lasso being optimal is that the Lasso achieves exact recovery, which requires $n \sim p \log p$ (see [Wai09]). Therefore, in our setting where $n/p \to \delta$, the Lasso is incapable of achieving the exact recovery nor outperforming general SLOPE.

4.3 k-level SLOPE

In this section we propose a method, described in Algorithm 6, that works on the general linear model. I.e. our method works on arbitrary data $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{w}$ and does not require $n/p \to \delta$ when $n, p \to \infty$.

In contrast to the AMP regime, we directly search on $\boldsymbol{\lambda}$ without implicitly using $\boldsymbol{\alpha}$ and we do not try to use the gradient information. To avoid searching in the high dimension of $\boldsymbol{\lambda}$ space, we propose to restrict that the penalty $\boldsymbol{\lambda}$ only contains k different non-negative values, which is denoted by $(\lambda_1, \cdots, \lambda_k; S_1, \cdots, S_{k-1})$. Here λ_i denotes the penalty magnitude and S_i represents the splitting index in $[p]$, where the penalty magnitudes change, i.e, $S_i - S_{i-1}$ entries in $\boldsymbol{\lambda}$ take the value λ_i. We note that λ_i is decreasing in i while S_i is increasing, guaranteeing that $\boldsymbol{\lambda}$ satisfies the assumption of SLOPE. As an example in $\mathbb{R}^5$, $\boldsymbol{\lambda} = (7, 5, 1; 2, 3) = (7, 7, 5, 1, 1)$. We name this restricted SLOPE problem as the *k-level SLOPE* and design the $(2k-1)$ degree of freedom penalty $\boldsymbol{\lambda}$, so as to only search in the reduced dimension $k \ll p$.

Notice that the original SLOPE is the p-level SLOPE and the Lasso is the 1-level SLOPE. We note that k-level SLOPE is always a sub-case of $(k+1)$-level SLOPE. Therefore intuitively, by allowing k to take values other than 1 and p, we can trade off the difficulty of designing the penalty and the accuracy gain by employing more penalty levels. We demonstrate that empirically, the trade-off is surprisingly encouraging: even 2 or 3 levels of penalty is sufficient to exploit the benefit of SLOPE.

4.3.1 Practical penalty design for k-level SLOPE

We emphasize that in the general regime beyond AMP and CGMT, we cannot access the gradient information nor the functional optimization in [HL19b] for two reasons: the true $\boldsymbol{\beta}$ distribution is not known in real data and the data matrix $\boldsymbol{X}$ is general (not i.i.d. with a specific variance). To design the k-level SLOPE penalty in the real-world datasets, we propose the Coordinate Descent (CD, Algorithm 6)[2] and compare to the PGD in Algorithm 5 under the AMP and CGMT regimes in Figure 4.2.

Algorithm 6 `Coordinate Descent` (CD)

```
Input: initial $\boldsymbol{\lambda}, \texttt{MSE}_{old} = \infty$, level $k$
while $\texttt{MSE} < \texttt{MSE}_{old}$ do
    Set $\texttt{MSE}_{old} = \texttt{MSE}$
    for $i \in \{1, \ldots, k\}$ do
        ▷ Search on magnitudes $\lambda_i$
        $\lambda_i = \underset{\lambda_i \in (\lambda_{i+1}, \lambda_{i-1})}{\text{argmin}} \ \texttt{MSE}(\lambda_1, \cdots; S_1, \cdots)$
    for $i \in \{1, \ldots, k-1\}$ do
        ▷ Search on splits $S_i$
        $S_i = \underset{S_i \in (S_{i-1}, S_{i+1})}{\text{argmin}} \ \texttt{MSE}(\lambda_1, \cdots; S_1, \cdots)$
Output: $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k; S_1, \cdots, S_{k-1})$
```

[2] We slightly abuse the notation of `MSE` to mean either the estimation error (only available in synthetic data) or the prediction error.

We highlight some details of Algorithm 6 that make it efficient and practical. First of all, Algorithm 6 directly works on $\boldsymbol{\lambda}$ instead of $\boldsymbol{\alpha}$ (the calibration is generally unavailable). Second, the projection is not needed as in Algorithm 5 since $\boldsymbol{\lambda}$ is decreasing and non-negative by our definition of the search domain. Third, Algorithm 6 is flexible in the following sense: (1) we can choose any order of coordinates to successively minimize the error, e.g. by $\lambda_1, S_1, \lambda_2, S_2, \cdots$; (2) we can use any zeroth-order search method such as the grid search or the binary search for the magnitudes and splits.

4.4 Experiments

In this section, we justify the effectiveness of k-level SLOPE on various synthetic and real datasets, on linear and logistic regression tasks. For the sake of implementation consistency, we adopt R package `SLOPE` to run both Lasso and SLOPE in experiments. Empirically, we remark that PGD and k-level CD are both significantly fast in all experimental settings, taking only a few minutes to converge even for $p = 1000$.

4.4.1 Synthetic datasets

Independent case

In this experiment we investigate the performance of k-level SLOPE in the AMP regime: data matrix $\boldsymbol{X}$ is i.i.d. $\mathcal{N}(0, 1/n), n = 300$. The signal distribution Π is

Gaussian-Bernoulli with probability 0.5 being standard normal and 0 otherwise. We work on a high-dimensional setting where $\delta = n/p = 0.3$ and hence $\boldsymbol{X} \in \mathbb{R}^{300\times 1000}$. The noise σ_w is 0. We observe from Figure 4.2 that employing more levels of penalty is beneficial and fast, suggesting that even a small k may be sufficient to reduce the errors significantly.

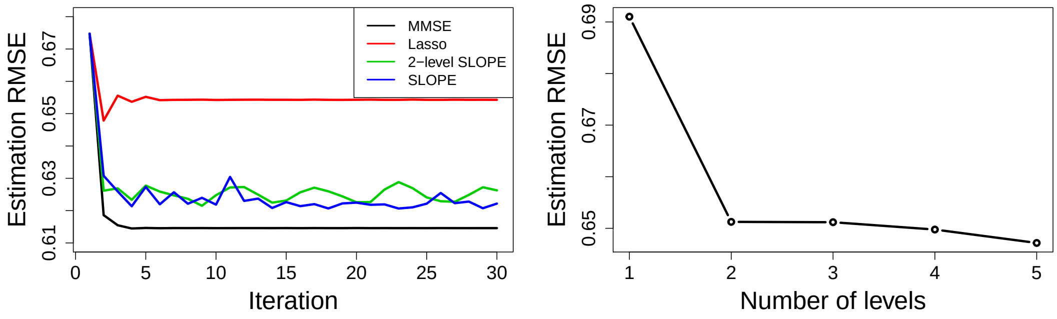


Figure 4.2: Left: the result of a single run for MMSE AMP, Lasso AMP, 2-level SLOPE AMP (by CD) and the p-level SLOPE AMP (by PGD). Right: averaged result over 10 independent runs of different k-level SLOPE by CD.

Dependent case

Different from the AMP regime in which each entry in the design matrix is i.i.d. gaussian, we study the performance of 2-level SLOPE in a synthetic dataset with features strongly correlated with each other. We include three other methods: Lasso and SLOPE with two other designs: Benjamini Hochberg design and the MR design proposed in [BLT18]. The data $\boldsymbol{X}$ is generated from an ARMA(1,1) model:

$$X_t = \varepsilon_t + 0.8X_{t-1} + 0.8\varepsilon_{t-1} \tag{4.4.1}$$

where X_t denote the t-th feature and ε_t follows i.i.d. $\mathcal{N}(0,1)$. We set Π, the asymptotic distribution of $\boldsymbol{\beta}$, to be i.i.d. Gaussian-Binomial: $\boldsymbol{\beta}_i \sim \boldsymbol{BZ}$ with $\boldsymbol{B} \sim B(5, 0.3)$ and $\boldsymbol{Z}$ being standard normal. In terms of the dimension, we study two cases (1) $n = 20, p = 50$; (2) $n = 200, p = 500$. 10-fold cross-validation $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ are calculated for both the Lasso and SLOPE. We highlight that, different than the previous section, we investigate the prediction error instead of the estimation error here. Curves for $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ with different iterations in both cases are shown in Figure 4.3. In the first case, using grid search, the optimal prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ given by Lasso is 0.128 while the optimal prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ given by 2-level SLOPE (using Algorithm 6) is 0.083. Prediction errors of SLOPE with other two penalty sequences are also under 0.1, but worse than that of 2-level SLOPE. We observe a 35% improvement on prediction error when using 2-level SLOPE for this case of small sample size and dimension, compared with Lasso. In the second case, the optimal prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ given by Lasso and other two SLOPEs are no smaller than 0.2, while that of 2-level SLOPE is 0.186, giving a 7.5% reduction in the prediction error.

4.4.2 Real datasets for linear and logistic regression

To further demonstrate the utility of k-level SLOPE in practice, we apply the model to real datasets, where $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ is intractable, and focus on the prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$. In this experiment, again we compare the performance of 2-level SLOPE in a linear regression setting with three other methods we studied in Section 4.1. The dataset we adopt is atherosclerosis cardiovascular disease (ASCVD), which records medical information of 236 patients and their corresponding ASCVD risk score (outcome variable). We select 1000 features out of 4216 features, which have the largest correlation with the outcome variable. We conduct 20-fold cross-validation and calculate the cross-validation prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$. Using grid search, the optimal prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ given by Lasso is 0.528. Interestingly, prediction $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ given by other two SLOPEs are worse than that of Lasso while that given by 2-level SLOPE (using Algorithm 6) is 0.489. This result clearly demonstrates the outperformance of k-level SLOPE compared to Lasso and SLOPE using other penalty sequences. A curve for $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ with different iterations is shown in Figure 4.3.

We further extend the idea of k-level SLOPE in logistic regression and justify the results on Alzheimer's Disease Neuroimaging Initiative (ADNI) gene dataset. The dataset contains over 19000 genomic features of 649 patients, along with a binary disease status (normal or ill). We select the first 300 patients in the original dataset and 500 features out of the total features, which has the largest correlation with the outcome variable. We conduct 10-fold cross-validation and calculate the cross-validation prediction accuracy. Using grid search, the optimal prediction accuracy given by Lasso is 0.62. The optimal prediction accuracy given by 2-level SLOPE (using Algorithm 6) is 0.66.

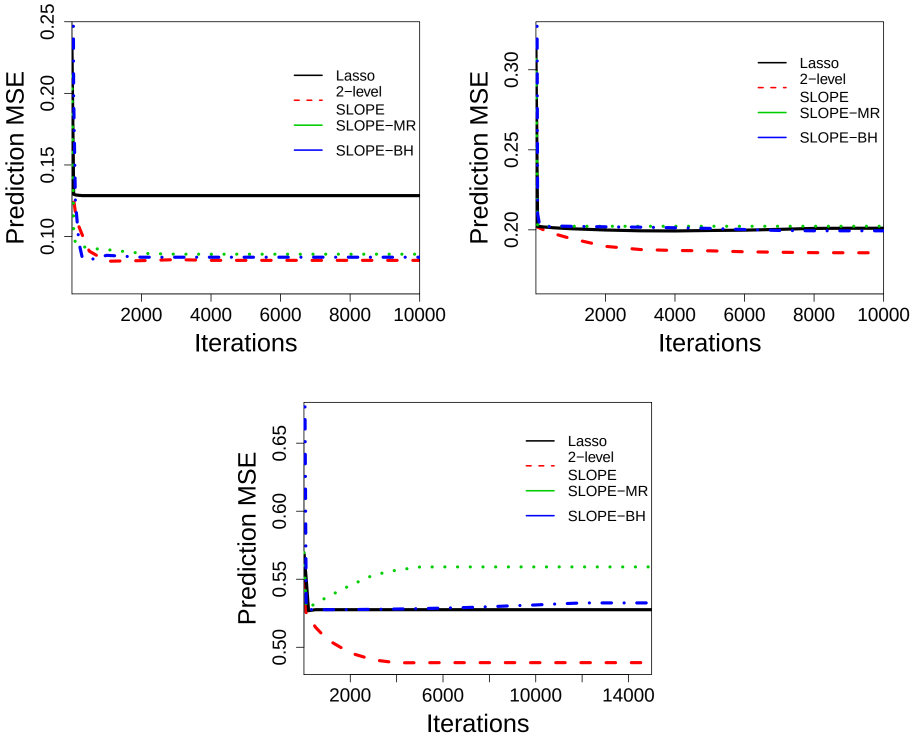


Figure 4.3: $\texttt{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ in linear regression cases. SLOPE-MR: SLOPE using penalty sequence suggested in [BLT18]; SLOPE-BH: SLOPE using Benjamini–Hochberg penalty sequence in R function 'SLOPE'. Top-left: Synthetic data with $\boldsymbol{X}$ i.i.d. drawn from ARMA(1,1) model (4.4.1), $n = 20$, $p = 50$. Top-right: $\boldsymbol{X}$ i.i.d. drawn from (4.4.1) with $n = 200$, $p = 500$. Bottom: the results of ASCVD dataset.

4.5 Discussion

In this work, we propose a framework to flexibly and efficiently design the SLOPE penalty sequence. Under the AMP setting, our first-order PGD approach is capable of finding the effective penalty sequence with reasonable computation budget. The key is to use the gradient with respect to the penalty instead of using zeroth-order search as previous works have proposed. In the practical world beyond the AMP setting, via various experiments, we illustrate that the proposed k-level SLOPE with penalty sequence determined by Algorithm 6 can provide decent results. Although Algorithm 6 loses the access to the first-order information when compared to Algorithm 5, the universal ability to search good penalty is desirable for practical use, as we can view the algorithm as a dimension-reduction trick. In many cases even 2-level SLOPE, the simplest k-level SLOPE (other than Lasso), can outperform the Lasso in accuracy as well as the (p-level) SLOPE in computation speed. Additionally, our framework indeed generalizes to other high-dimensional penalty designs. Some direct extensions include group SLOPE and weighted Lasso.

Much room is left for future study. From a theoretical perspective, the quasi-convexity of $\tau(\alpha)$ in AMP setting is still not well studied. The asymptotic $\texttt{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ (i.e. Equation (7) in [MMB+18b]) is shown to be quasi-convex in Lasso case. However, no such theoretical property has been shown for SLOPE. If the quasi-convexity indeed holds true for SLOPE AMP, then we can guarantee that the minimizing $\boldsymbol{\lambda}$ by PGD is indeed the global minimizer and thus claim our design is optimal.

It would also be interesting to develop PGD (based on AMP regime) for k-level SLOPE, i.e. using gradient descent to find the optimal magnitudes and splits. One could then derive a theoretical trade-off curve between the minimum τ and each k, similarly to Figure 2 bottom subplot. This would suggest a proper choice of k for our k-level SLOPE.

From a practical perspective, we anticipate that k-level SLOPE can also be explored in various applications that already employ the Lasso, such as the matrix completion, the compressed sensing and the neural network regularization.

4.6 Appendix

4.6.1 Introduction to MMSE AMP

We firstly introduce the procedure for general AMP procedure.

$$\begin{aligned} s^{(t+1)} &= X^\top \boldsymbol{Z}^{(t)} + \boldsymbol{\beta}^{(t)} \\ \boldsymbol{\beta}^{(t+1)} &= \eta^{(t+1)}(s^{(t+1)}) \\ \boldsymbol{Z}^{(t+1)} &= y - X\boldsymbol{\beta}^{(t+1)} + \frac{1}{n}\boldsymbol{Z}^{(t)}[\nabla\eta^{(t)}(s^{(t)})] \end{aligned} \tag{4.6.1}$$

Different η functions give different AMP, e.g. the soft-thresholding η gives the Lasso AMP; the SLOPE proximal operator η gives the SLOPE AMP.

The MMSE AMP adopts the following denoiser $\eta^{(t)}$ [BM11b]

$$\eta_i^{(t)}(s) = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} + \tau_t \boldsymbol{z} = s_i] \quad i = 1, \ldots, p$$

with $\boldsymbol{z} \sim \mathcal{N}(0,1)$. In above, using the state evolution [Bu+20b], τ_t^2 can be calculated iteratively as:

$$\tau_t^2 = \sigma_w^2 + \frac{1}{\delta}\mathbb{E}[(\eta^{(t-1)}(\boldsymbol{\beta} + \tau_{t-1}\boldsymbol{z}) - \boldsymbol{\beta})^2]$$

Assume that the measurement matrix X has i.i.d. $\mathcal{N}(0, 1/n)$ entries. In many scenarios, the denoiser $\eta^{(t)}$ might be hard to calculate. Here we provide a derivation about calculating $\eta^{(t)}$ in the Bernoulli-Gaussian case: we assume that true signal $\boldsymbol{\beta} \overset{i.i.d.}{\sim} \boldsymbol{B}$ where $\boldsymbol{B}$ is a Bernoulli-Gaussian distribution, i.e. $\boldsymbol{\beta}_i = 0$ with probability $e \in [0,1]$, otherwise $\boldsymbol{\beta}_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{B}}^2)$.

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \neq 0, \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i]\mathbb{P}(\boldsymbol{\beta} \neq 0|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i) \qquad (4.6.2)$$

It's straightforward to see that, with f denoting the corresponding probability density function,

$$\mathbb{P}(\boldsymbol{\beta} \neq 0|\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i) = \frac{f(\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2))(1-e)}{f(\boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2))(1-e) + f(\tau_t\boldsymbol{z} = s_i)e} \qquad (4.6.3)$$

Meanwhile. we have

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \neq 0, \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2), \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i]$$

since $\boldsymbol{\beta} + \tau_t\boldsymbol{z} \sim \mathcal{N}(0, \sigma_B^2 + \tau_t^2)$, conditional expectation on joint normal distribution yields

$$\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_B^2), \boldsymbol{\beta} + \tau_t\boldsymbol{z} = s_i] = \frac{\sigma_B^2}{\sigma_B^2 + \tau_t^2}s_i \qquad (4.6.4)$$

(4.6.3) and (4.6.4) give a simple way to calculate the denoiser using (4.6.2).

4.6.2 Analysis of Gradient in PGD for α

Proof of Theorem 7. Minimizing the estimation error is equivalent to minimizing τ. Since the AMP algorithms are working on the finite dimension, we analyze the finite-size approximation of the state evolution [Bu+20b, Equation (2.5)]:

$$\tau^2 = \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\left\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z}) - \boldsymbol{\beta}\right\|^2$$

In finite dimensions, the expectation is taken with respect to $\boldsymbol{Z}$. Differentiating both sides of the state evolution with respect to α_i and denoting $\tau' = \frac{\partial\tau}{\partial\alpha_i}$ gives:

$$\begin{aligned}2\tau\tau' &= \frac{\partial}{\partial\alpha_i}\left(\sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z}) - \boldsymbol{\beta}\|^2\right)\\ &= \frac{1}{n}\frac{\partial}{\partial\alpha_i}\sum_{j=1}^{p}\mathbb{E}\left([\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})]_j - \boldsymbol{\beta}_j\right)^2 \qquad (4.6.5)\end{aligned}$$

Recall η_j represents the j-th element of $\boldsymbol{\eta} := \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{\beta}+\tau\boldsymbol{Z})$. By chain rule

$$2\tau\tau' = \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\frac{\partial\eta_j}{\partial\alpha_i} = \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\left[\sum_{k=1}^{p}\frac{d\eta_j}{da_k}\frac{\partial a_k}{\partial\alpha_i} + \frac{d\eta_j}{db_k}\frac{\partial b_k}{\partial\alpha_i}\right]$$

where we define $a_k := \beta_k + \tau Z_k, b_k := \alpha_k\tau$. To calculate the derivatives, we pause to discuss forms of general derivatives of $\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})$. Define

$$\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) := \operatorname{diag}\left[\frac{\partial}{\partial a_1},\frac{\partial}{\partial a_2},\ldots,\frac{\partial}{\partial a_p}\right]\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) \qquad (4.6.6)$$

$$\partial_2\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}) := \operatorname{diag}\left[\frac{\partial}{\partial b_1},\frac{\partial}{\partial b_2},\ldots,\frac{\partial}{\partial b_p}\right]\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b}). \qquad (4.6.7)$$

According to [SC+16, Proof of Fact 3.4] and [Bu+20b, Proof of Theorem 1], we have

$$[\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = \frac{1}{\#\{1\le k\le p : |[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_k| = |[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j|\}}$$

and that

$$\frac{d}{da_k}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j = \mathbb{I}\{|\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_j = |\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})|_k\}\operatorname{sign}(\boldsymbol{\eta}_j\boldsymbol{\eta}_k)[\partial_1\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j$$

for the derivative regardng the first variable. Recall that the permutation $\sigma : \{1,\dots,p\} \to \{1,\dots,p\}$ is the inverse mapping for ranking of indices such that $|\boldsymbol{\eta}|_{(i)} = |[\boldsymbol{\eta}]_{\sigma(i)}|$. Similarly, according to [Bu+20b, Proof of Theorem 1]:

$$\begin{aligned}\frac{d}{db_k}[\eta(\boldsymbol{a},\boldsymbol{b})]_j &= -\operatorname{sign}([\eta(\boldsymbol{a},\boldsymbol{b})]_{\sigma(k)})\frac{d}{da_{\sigma(k)}}[\boldsymbol{\eta}(\boldsymbol{a},\boldsymbol{b})]_j \\ &= \mathbb{I}\Big\{|\eta(\boldsymbol{a},\boldsymbol{b})|_j = |\eta(\boldsymbol{a},\boldsymbol{b})|_{\sigma(k)}\Big\}\operatorname{sign}\Big(\eta_j\Big)\Big[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})\Big]_j. \qquad (4.6.8)\end{aligned}$$

In addition to I_j defined in Section 2, we let $K_j := \{k : |\eta_{\sigma(k)}| = |\eta_j|\}$, which is the set of ranking indices whose corresponding entries share the same absolute value with η_j. This notion will be used to replace the indicator term $\mathbb{I}\Big\{|\eta(\boldsymbol{a},\boldsymbol{b})|_j = |\eta(\boldsymbol{a},\boldsymbol{b})|_{\sigma(k)}\Big\}$ above. We can rewrite (4.6.6) as

$$\begin{aligned}2\tau\tau' &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\left[\sum_{k\in I_j}\frac{d\eta_j}{da_k}\frac{\partial a_k}{\partial\alpha_i} + \sum_{k\in K_j}\frac{d\eta_j}{db_k}\frac{\partial b_k}{\partial\alpha_i}\right] \\ &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\left[\frac{1}{|I_j|}\sum_{k\in I_j}\operatorname{sign}(\eta_k)\frac{\partial a_k}{\partial\alpha_i} - \frac{1}{|K_j|}\sum_{k\in K_j}\frac{\partial b_k}{\partial\alpha_i}\right] \\ &= \frac{2}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\Bigg[\frac{1}{|I_j|}\sum_{k\in I_j}\operatorname{sign}(\eta_k)Z_k\tau' \\ &\quad - \frac{1}{|K_j|}\sum_{k\in K_j}(\alpha_k\tau' + \mathbb{I}\{k=i\}\tau)\Bigg] \qquad (4.6.9)\end{aligned}$$

Merging the terms containing the derivative τ' on one side gives

$$\begin{aligned}&\frac{1}{n}\sum_{j\in I_{\sigma(i)}}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)/|K_j| \\ &= \frac{1}{n}\sum_{j=1}^{p}\mathbb{E}(\eta_j-\beta_j)\operatorname{sign}(\eta_j)\left[\frac{1}{|I_j|}\sum_{k\in I_j}\operatorname{sign}(\eta_k)Z_k\tau' - \frac{1}{|K_j|}\sum_{k\in K_j}\alpha_k\tau'\right] - \tau\tau'\end{aligned}$$

Notice that $|I_j| = |K_j|$ due to σ being a permutation, we can simplify above as

$$\frac{\partial \tau}{\partial \alpha_i} = \mathbb{E}\frac{1}{|I_{\sigma(i)}|D(\boldsymbol{\alpha}, \tau)} \sum_{j \in I_{\sigma(i)}} (\eta_j - \beta_j) \operatorname{sign}(\eta_j)\tau \tag{4.6.10}$$

where $D(\boldsymbol{\alpha}, \tau)$ in the denominator is

$$D(\boldsymbol{\alpha}, \tau) = -n\tau + \sum_{j=1}^{p} \mathbb{E}\frac{1}{|I_j|}(\eta_j - \beta_j) \operatorname{sign}(\eta_j) \sum_{k \in I_j} (\operatorname{sign}(\eta_k) Z_k - \alpha_{\sigma^{-1}(k)})$$

We next show that $D(\boldsymbol{\alpha}, \tau)$ is always negative. Firstly observe from (4.2.3) that

$$\tau^2 > \frac{1}{n}\sum_{j=1}^{p} \mathbb{E}(\eta_j - \beta_j)^2 \tag{4.6.11}$$

Now for the set I_i with a fixed index i,

$$\sum_{j \in I_i} (\eta_j - \beta_j)^2 \geq \frac{1}{|I_i|}(\sum_{j \in I_i} |\eta_j - \beta_j|)^2 \tag{4.6.12}$$

$$\geq \frac{1}{|I_i|}(\sum_{j \in I_i} (\eta_j - \beta_j) \operatorname{sign}(\eta_j))^2 \tag{4.6.13}$$

$$= \frac{1}{|I_i|}\sum_{j \in I_i} (\eta_j - \beta_j) \operatorname{sign}(\eta_j) \sum_{k \in I_i} \tau Z_k \operatorname{sign}(\eta_k) - \alpha_{\sigma^{-1}(k)}\tau \tag{4.6.14}$$

$$\geq \frac{\tau}{|I_i|}\sum_{j \in I_i} (\eta_j - \beta_j) \operatorname{sign}(\eta_j) \sum_{k \in I_j} Z_k \operatorname{sign}(\eta_k) - \alpha_{\sigma^{-1}(k)} \tag{4.6.15}$$

This in turn implies that

$$\begin{aligned}\sum_{j=1}^{p} (\eta_j - \beta_j)^2 &= \sum_{j=1}^{p} \frac{1}{|I_j|}\sum_{k \in I_j} (\eta_k - \beta_k)^2 \\ &\geq \sum_{j=1}^{p} \frac{\tau}{|I_j|}(\eta_j - \beta_j) \operatorname{sign}(\eta_j) \sum_{k \in I_j} Z_k \operatorname{sign}(\eta_k) - \alpha_{\sigma^{-1}(k)} \end{aligned}\tag{4.6.16}$$

Combining with (4.6.11) yields $D < 0$.

□

4.6.3 Analysis of Projection in PGD for α

Characterization of projection on $\mathcal{S}$

We firstly prove that Algorithm 4 indeed finds the projection. To do so we firstly provide a detailed characterization of the projection, then prove that the output of Algorithm 4 matches the form of projection. We start by defining *blocks* and *segmentation blocks*, upon which our proof highly relies. Suppose $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$, *blocks* are subsequences defined as $B(\boldsymbol{\gamma}, u) := \{\gamma_u, \ldots, \gamma_{u+L(\boldsymbol{\gamma},u)-1}\}$ where length $L(\boldsymbol{\gamma}, u)$ is defined as

$$L(\boldsymbol{\gamma}, u) = \begin{cases} L^* & \text{if } L^* \neq \emptyset \\ \\ p & \text{otherwise} \end{cases} \tag{4.6.17}$$

where

$$L^* \triangleq \min \left\{ 1 \leq L \leq p - u \middle| \forall 0 \leq k \leq p - u - L, \frac{1}{k+1} \sum_{i=0}^{k} \gamma_{u+L+i} < \frac{1}{L} \sum_{i=0}^{L-1} \gamma_{u+i} \right\}$$

Roughly speaking, $L(\boldsymbol{\gamma}, u)$ is the minimum value of a finite set (truncated at p when the set is empty). For each element L in this set, the average value in sequence $\{\gamma_u, \ldots, \gamma_{u+L-1}\}$ is always larger than that of arbitrary sequence $\{\gamma_{u+L}, \ldots, \gamma_{u+L+k}\}$ whose left start is γ_{u+L}. With such definition of blocks, we can now segment $\boldsymbol{\gamma}$ into $q \leq p$ blocks:

$$\begin{aligned} \boldsymbol{\gamma} &= \{B(\boldsymbol{\gamma}, 1), B(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, 1) + 1), B(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, L(\boldsymbol{\gamma}, 1) + 1) + L(\boldsymbol{\gamma}, 1) + 1), \ldots\} \\ &\triangleq \{B_1, \ldots, B_q\} \end{aligned}$$

We call $B_1, \dots, B_q$ *segmentation blocks* for vector $\boldsymbol{\gamma}$. It's straightforward to see that $B_k = B(\boldsymbol{\gamma}, L_k)$ where L_k satisfies $L_1 = L(\boldsymbol{\gamma}, 1)$ and

$$L_k = L(\boldsymbol{\gamma}, \sum_{i=1}^{k-1} L_i + 1)$$

Our result shows that for input vector γ, its projection vector $\Pi_{\mathcal{S}}(\boldsymbol{\gamma})$ takes identical values inside each of the segmentation blocks. Before formally stating the theorem, We first highlight the following fact that will be frequently used in the proof of the theorem.

Fact 4.6.1. *For two sequences of length p: $\{a_i\}$ and $\{b_i\}$, if $\sum a_i = \sum b_i$, then function $g(C) := \sum (b_i - a_i + C)^2$ is monotonically increasing with respect to $|C|$.*

Proof. Notice that

$$\sum (b_i - a_i + C)^2 = \sum (b_i - a_i)^2 + \sum 2C(b_i - a_i) + pC^2 = pC^2 + \sum (b_i - a_i)^2$$

Hence $g(C)$ is is monotonically increasing with respect to $|C|$. □

Theorem 9. *Let B denote the segmentation block that contains γ_i, then*

$$(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \max \left\{ \frac{1}{|B|} \sum_{\gamma_j \in B} \gamma_j, 0 \right\}$$

Proof. The proof consists of two steps. In the first step, we prove that for each segmentation block B, the projection of each coordinates share the same value. That is, $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \mathcal{C}(B)$ as long as $\gamma_i \in B$. In the second step, we show that this constant is the mean of the block truncated at 0: $\mathcal{C}(B) = \max \left\{ \frac{1}{|B|} \sum_{\gamma_j \in B} \gamma_j, 0 \right\}$.

Step 1 Without loss of generality, we consider $B = B(\boldsymbol{\gamma}, u)$. We know from definition of blocks that $\forall 1 \leq l \leq L-1$, $\exists k_l$ s.t. $\frac{1}{k_l}\sum_{i=1}^{k_l}\gamma_{u+l-1+i} \geq \frac{1}{l}\sum_{i=1}^{l}\gamma_{u+i-1}$. We use induction to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l}$, $\forall 1 \leq l \leq L(\gamma, u) - 1$. For $l = 1$, assume $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}$. Consider two cases: (i) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > \gamma_u$. (ii) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u \leq \gamma_u$. We now show that both cases lead to contradiction and hence do not hold. In case (i), we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} \max\{\gamma_u, (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}\} & \text{if } i = u \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise} \end{cases}$$

then obviously,

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_u - \gamma_u\right| < |(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u - \gamma_u|$$

which leads to that $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$. This contradicts to the definition of projection. In case (ii), from definition of blocks we have that $\exists k_0 \geq 1$ s.t. $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i} \geq \gamma_u$. Consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u & \text{if } i \in \{u+1, \ldots, u+k_0\} \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise} \end{cases}$$

Notice that $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i} \geq \gamma_u \geq (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+1}$, we have for $i \in \{u+1, \ldots, u+k_0\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}\right|$$

According to Fact 4.6.1, we define substitution for $i \in \{u+1, \ldots, u+k_0\}$: $b_i = \frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+i}$, $a_i = \gamma_{u+i}$, $b_i + C_1 = (\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ and $b_i + C_2 = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i$. Then since

$|C_1 < C_2|$, we have $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which contradicts to the definition of projection.

Now assume the statement holds for $1 \leq l \leq l_0 - 1$, that is $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = \cdots = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0-1}$, we want to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$. Since the projection is on $\mathcal{S}$, by definition we know $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u$ can never be smaller than $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$. We now assume $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$ and consider two cases: (i) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. (ii) $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u \leq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. To complete the proof, it suffices for us to show that neither of the cases can hold without contradictions. In case (i), we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} \max\{\frac{1}{l_0}\sum_{j=0}^{l_0-1}\gamma_{u+j}, (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}\} \\ \qquad\qquad \text{if } i \in \{u, \ldots, u+l_0-1\} \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i \qquad\qquad\qquad \text{otherwise} \end{cases}$$

then obviously for $i \in \{u, \ldots, u+l_0-1\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}\right|$$

According to Fact 4.6.1, using the same substitution as that in analysis of $l = 1$, we have that $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which makes contradiction to the definition of projection. In case (ii), from definition of blocks we have that $\exists k_0 \geq 1$ s.t. $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+l_0-1+i} \geq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i}$. Now we consider

$$(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \begin{cases} (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u & \text{if } i \in \{u+l_0, \ldots, u+l_0-1+k_0\} \\ (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i & \text{otherwise} \end{cases}$$

Notice that $\frac{1}{k_0}\sum_{i=1}^{k_0}\gamma_{u+l_0-1+i} \geq \frac{1}{l_0}\sum_{i=0}^{l_0-1}\gamma_{u+i} \geq (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$, we have for $i \in \{u+l_0, \ldots, u+l_0-1+k_0\}$, $\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i\right|$ is a constant independent of i and that

$$\left|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=0}^{k_0-1}\gamma_{u+l_0+i}\right| < \left|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i - \frac{1}{k_0}\sum_{i=0}^{k_0-1}\gamma_{u+l_0+i}\right|$$

Again according to Fact 4.6.1, we have $\frac{1}{2}\|(\widetilde{\Pi}_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2 < \frac{1}{2}\|(\Pi_{\mathcal{S}}(\boldsymbol{\gamma})) - \boldsymbol{\gamma}\|_2^2$, which contradicts to the definition of projection. This implies that it can never happen that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u > (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+l_0}$, which completes the induction. We have proved that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_u = \cdots = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_{u+L(\boldsymbol{\gamma},u)-1} \triangleq \mathcal{C}(B(u))$ for each segmentation block $B(u)$ of vector $\boldsymbol{\gamma}$.

Step 2 Now we already know that inside each segmentation block, the projection of each coordinate is a constant $\mathcal{C}(B)$, we now optimize the sequence $\{\mathcal{C}(B_i)\}_{i=1}^{q}$. According to Fact 4.6.1, inside each B_i, the optimal constant (i.e. constant gives smallest ℓ_2 error $\text{argmin}_{C\geq 0}\frac{1}{2}\sum_{\gamma_j\in B_i}(\gamma_j - C)^2$) is : $\max\left\{\frac{1}{|B_i|}\sum_{\gamma_j\in B_i}\gamma_j, 0\right\}$. Meanwhile, it's feasible to set

$$(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = \max\left\{\frac{1}{|B|}\sum_{\gamma_j\in B}\gamma_j, 0\right\}$$

since we have that $\max\left\{\frac{1}{|B_i|}\sum_{\gamma_j\in B_i}\gamma_j, 0\right\} \geq \max\left\{\frac{1}{|B_{i+1}|}\sum_{\gamma_j\in B_{i+1}}\gamma_j, 0\right\}$ by definition of blocks. This wraps up the proof.

□

Proof of Theorem 8

We next prove the validity of Algorithm 4.

Proof. Suppose $\boldsymbol{\gamma}$ has segmentation blocks $B_1, \ldots, B_q$, we firstly prove that $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ for $i \leq |B_1|$. We let $\gamma_j(t)$ denote the value of γ_j at the moment i was assigned from t to $t+1$ in Algorithm 4 (i.e. the time when first t iterations are finished). We also let $\gamma_j(0)$ denote the initial value of γ_j in the input. Then clearly $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_j = \max\{\gamma_j(p), 0\}$. During the value-averaging step, the algorithm is constantly transporting values from elements with larger index to those with smaller. Hence it's straightforward to see that

$$\sum_{j=1}^{J} \gamma_j(t) \geq \sum_{j=1}^{J} \gamma_j(t-1) \tag{4.6.18}$$

for arbitrary $J, t \in \{1, \ldots, p\}$. First assume $\gamma_1(p) = \cdots = \gamma_{\widetilde{L}_1}(p) > \gamma_{\widetilde{L}_1+1}(p)$. Since Algorithm 4 only involves averaging values among subsequences, we have that $\sum_{j=1}^{p} \gamma_j(p) = \sum_{j=1}^{p} \gamma_j$. Moreover since $\gamma_{\widetilde{L}_1}(p) > \gamma_{\widetilde{L}_1+1}(p)$, there's no value-averaging steps between any one of the first $\widetilde{L}_1$ elements and one of the rest elements. This implies

$$\sum_{j=1}^{\widetilde{L}_1} \gamma_j(p) = \sum_{j=1}^{\widetilde{L}_1} \gamma_j \tag{4.6.19}$$

By definition of blocks, we know that $\exists k$ such that $\frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i} \geq \frac{1}{\widetilde{L}_1}\sum_{i=1}^{\widetilde{L}_1} \gamma_i = \gamma_1(p)$. By (4.6.18) we have that

$$\frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i} \leq \frac{1}{k}\sum_{i=1}^{k} \gamma_{\widetilde{L}_1+i}(p) \leq \gamma_{\widetilde{L}_1+1}(p)$$

Together with above, this implies that $\gamma_1(p) \leq \gamma_{\widetilde{L}_1+1}(p)$, which contradicts to the assumption. Hence we have that $\widetilde{L}_1 \geq L_1$.

On the other hand, if $\widetilde{L}_1 > L_1$, then at the moment i is assigned to be $\widetilde{L}_1 + 1$ in the algorithm (i.e. the time when first $\widetilde{L}_1$ iterations are finished), we must have that

$$\frac{\sum_{j=1}^{\widetilde{L}_1}\gamma_j(\widetilde{L}_1-1)}{\widetilde{L}_1} \geq \frac{\sum_{j=1}^{L_1}\gamma_j(\widetilde{L}_1-1)}{L_1}$$

This implies that

$$\frac{\sum_{j=L_1+1}^{\widetilde{L}_1}\gamma_j(\widetilde{L}_1-1)}{\widetilde{L}_1-L_1} \geq \frac{\sum_{j=1}^{L_1}\gamma_j(\widetilde{L}_1-1)}{L_1} \tag{4.6.20}$$

By (4.6.18) we have

$$\frac{\sum_{j=1}^{L_1}\gamma_j}{L_1} \leq \frac{\sum_{j=1}^{L_1}\gamma_j(\widetilde{L}_1-1)}{L_1} \tag{4.6.21}$$

Meanwhile at $t = \widetilde{L}_1 - 1$, the sum of first L_1 terms is the same as that in $\boldsymbol{\gamma}$. This implies

$$\begin{aligned}\sum_{j=L_1+1}^{\widetilde{L}_1}\gamma_j(\widetilde{L}_1-1) &= \sum_{j=1}^{L_1}\gamma_j + \sum_{j=L_1+1}^{\widetilde{L}_1}\gamma_j - \sum_{j=1}^{L_1}\gamma_j(\widetilde{L}_1-1) \\ &\leq \sum_{j=L_1+1}^{\widetilde{L}_1}\gamma_j\end{aligned} \tag{4.6.22}$$

where the last inequality is given by (4.6.18). Combining (4.6.20), (4.6.21) and ((4.6.22)) yields

$$\frac{\sum_{j=L_1+1}^{\widetilde{L}_1}\gamma_j}{\widetilde{L}_1-L_1} \geq \frac{\sum_{j=1}^{L_1}\gamma_j}{L_1}$$

This contradicts to definition of L_1 in (4.6.17). Hence we have that $\widetilde{L}_1 = L_1$. This means $\gamma_1(p) = \cdots = \gamma_{L_1}(p) > \gamma_{L_1+1}(p)$. Recall that $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_j = \max\{\gamma_j(p), 0\}$, this together with (4.6.19) yields

$$\begin{aligned}(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_1 = \max\left\{\frac{1}{|B_1|}\sum_{j=1}^{L_1}\gamma_j, 0\right\} &= (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_1 \\ = \cdots = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{L_1} &> (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{L_1+1}\end{aligned}$$

Now we have prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ for $i \leq |B_1|$ and that there is no interaction between element in B_1 and that outside B_1. This implies that the existence of B_1 does *not* affect the rest of output values $(\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_{i>|B_1|}$. Hence we can ignore B_1 and repeat exactly the same procedure to prove that $(\Pi_{\mathcal{S}}(\boldsymbol{\gamma}))_i = (\Lambda_{\mathcal{S}}(\boldsymbol{\gamma}))_i$ when $|B_1| + 1 \leq i \leq |B_2|$ and that there is no interactions between element in B_2 and that outside B_2. Iteratively we can prove $\Pi_{\mathcal{S}}(\boldsymbol{\gamma}) = \Lambda_{\mathcal{S}}(\boldsymbol{\gamma})$

□

Bibliography

[Abr+06] Felix Abramovich et al. “Adapting to unknown sparsity by controlling the false discovery rate”. In: *The Annals of Statistics* 34.2 (2006), pp. 584–653.

[BC+15] Rina Foygel Barber, Emmanuel J Candès, et al. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[BEM13] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari. “Estimating lasso risk and noise level”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 944–952.

[BLT18] Pierre C Bellec, Guillaume Lecue, and Alexandre B Tsybakov. “SLOPE meets lasso: improved oracle bounds and optimality”. In: *The Annals of Statistics* 46.6B (2018), pp. 3603–3642.

[BM11a] Mohsen Bayati and Andrea Montanari. “The dynamics of message passing on dense graphs, with applications to compressed sensing”. In: *IEEE Trans. on Inf. Theory* 57.2 (2011), pp. 764–785.

[BM11b] Mohsen Bayati and Andrea Montanari. “The dynamics of message passing on dense graphs, with applications to compressed sensing”. In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 764–785.

[BM11c] Mohsen Bayati and Andrea Montanari. “The LASSO risk for Gaussian matrices”. In: *IEEE Transactions on Information Theory* 58.4 (2011), pp. 1997–2017.

[BM11d] Mohsen Bayati and Andrea Montanari. “The LASSO risk for Gaussian matrices”. In: *IEEE Transactions on Information Theory* 58.4 (2011), pp. 1997–2017.

[BMN20] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. “State evolution for approximate message passing with non-separable functions”. In: *Information and Inference: A Journal of the IMA* 9.1 (2020), pp. 33–79.

[Bog+13a] Małgorzata Bogdan et al. “Statistical estimation and testing via the sorted L1 norm”. In: *arXiv preprint arXiv:1310.1969* (2013).

[Bog+13b] Małgorzata Bogdan et al. “Supplementary materials for Statistical Estimation and Testing via the Sorted l1 Norm.” In: Available at *https://statweb.stanford.edu/~candes/publications/downloads/SortedL1_SM.pdf* (2013).

[Bog+15a] Małgorzata Bogdan et al. “SLOPE—Adaptive variable selection via convex optimization”. In: *The Annals of Applied Statistics* 9.3 (2015), p. 1103.

[Bog+15b] Małgorzata Bogdan et al. “SLOPE—adaptive variable selection via convex optimization”. In: *The annals of applied statistics* 9.3 (2015), p. 1103.

[BR08] Howard D Bondell and Brian J Reich. “Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR”. In: *Biometrics* 64.1 (2008), pp. 115–123.

[Brz+18] Damian Brzyski et al. “Group SLOPE—Adaptive selection of groups of predictors”. In: *Journal of the American Statistical Association* (2018), pp. 1–15.

[Brz+19] Damian Brzyski et al. “Group SLOPE—Adaptive selection of groups of predictors”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 419–433.

[BS13] J Frederic Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[BS98] J Frederic Bonnans and Alexander Shapiro. “Optimization problems with perturbations: A guided tour”. In: *SIAM review* 40.2 (1998), pp. 228–264.

[BT09] Amir Beck and Marc Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

[Bu+20a] Zhiqi Bu et al. “Algorithmic analysis and statistical estimation of SLOPE via approximate message passing”. In: *IEEE Transactions on Information Theory* 67.1 (2020), pp. 506–537.

[Bu+20b] Zhiqi Bu et al. “Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing”. In: *IEEE Transactions on Information Theory* 67.1 (2020), pp. 506–537.

[BY08] ZD Bai and YQ Yin. “Limit of the smallest eigenvalue of a large dimensional sample covariance matrix”. In: *Advances In Statistics*. World Scientific, 2008, pp. 108–127.

[Cha+98] Antonin Chambolle et al. “Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage”. In: *IEEE Transactions on Image Processing* 7.3 (1998), pp. 319–335.

[CM19] Michael Celentano and Andrea Montanari. “Fundamental Barriers to High-Dimensional Regression with Convex Penalties”. In: *arXiv preprint arXiv:1903.10603* (2019).

[CMW20] Michael Celentano, Andrea Montanari, and Yuting Wei. “The Lasso with general Gaussian designs with applications to hypothesis testing”. In: *arXiv preprint arXiv:2007.13716* (2020).

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint”. In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 57.11 (2004), pp. 1413–1457.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization.” In: *Journal of machine learning research* 12.7 (2011).

[Dik67] II Dikin. “Iterative solution of problems of linear and quadratic programming”. In: *Doklady Akademii Nauk*. Vol. 174. 4. Russian Academy of Sciences. 1967, pp. 747–748.

[DMM09a] David L Donoho, Arian Maleki, and Andrea Montanari. “Message-passing algorithms for compressed sensing”. In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919.

[DMM09b] David L Donoho, Arian Maleki, and Andrea Montanari. “Message-passing algorithms for compressed sensing”. In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919.

[DMM10] David L Donoho, Arian Maleki, and Andrea Montanari. "Message passing algorithms for compressed sensing: I. motivation and construction". In: *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*. IEEE. 2010, pp. 1–5.

[DMM11] David L Donoho, Arian Maleki, and Andrea Montanari. "The noise-sensitivity phase transition in compressed sensing". In: *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6920–6941.

[Don05] David L. Donoho. "Neighborly polytopes and sparse solutions of underdetermined linear equations". In: (2005).

[Don06] David L Donoho. "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension". In: *Discrete & Computational Geometry* 35.4 (2006), pp. 617–652.

[Doo53] Joseph Leo Doob. *Stochastic processes*. Vol. 101. New York Wiley, 1953.

[DT09a] David Donoho and Jared Tanner. "Counting faces of randomly projected polytopes when the projection radically lowers dimension". In: *Journal of the American Mathematical Society* 22.1 (2009), pp. 1–53.

[DT09b] David Donoho and Jared Tanner. "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing". In: *Philosophical Transactions of* *the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906 (2009), pp. 4273–4293.

[Fer+14] Hans Joachim Ferreau et al. “qpOASES: A parametric active-set algorithm for quadratic programming”. In: *Mathematical Programming Computation* 6.4 (2014), pp. 327–363.

[FN16] Mario Figueiredo and Robert Nowak. “Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects”. In: *Artificial Intelligence and Statistics*. 2016, pp. 930–938.

[FW56] Marguerite Frank and Philip Wolfe. “An algorithm for quadratic programming”. In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110.

[GHT13] Max Grazier GSell, Trevor Hastie, and Robert Tibshirani. “False variable selection rates in regression”. In: *arXiv preprint arXiv:1302.2303* (2013).

[GI83] Donald Goldfarb and Ashok Idnani. “A numerically stable dual method for solving strictly convex quadratic programs”. In: *Mathematical programming* 27.1 (1983), pp. 1–33.

[HL19a] Hong Hu and Yue M Lu. “Asymptotics and optimal designs of SLOPE for sparse linear regression”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019, pp. 375–379.

[HL19b] Hong Hu and Yue M Lu. “Asymptotics and optimal designs of SLOPE for sparse linear regression”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019, pp. 375–379.

[JM13] Adel Javanmard and Andrea Montanari. “State evolution for general approximate message passing algorithms, with applications to spatial coupling”. In: *Information and Inference: A Journal of the IMA* 2.2 (2013), pp. 115–144.

[JP12] Richard Johnsonbaugh and William E Pfaffenberger. *Foundations of mathematical analysis*. Courier Corporation, 2012.

[Kas77] Boris Sergeevich Kashin. “Diameters of some finite-dimensional sets and classes of smooth functions”. In: *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 41.2 (1977), pp. 334–351.

[KB14] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[KB20] Michał Kos and Małgorzata Bogdan. “On the asymptotic properties of SLOPE”. In: *Sankhya A* 82.2 (2020), pp. 499–532.

[KF75] Andrey Nikolaevich Kolmogorov and Sergey Vasil’evich Fomin. *Introductory real analysis*. Courier Corporation, 1975.

[Krz+12] Florent Krzakala et al. “Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices”. In: *J. Stat. Mech. Theory Exp.* 8 (2012).

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. 89. American Mathematical Soc., 2001.

[Lit+05] Alexander E Litvak et al. “Smallest singular value of random matrices and geometry of random polytopes”. In: *Advances in Mathematics* 195.2 (2005), pp. 491–523.

[MMB+18a] Ali Mousavi, Arian Maleki, Richard G Baraniuk, et al. “Consistent parameter estimation for LASSO and approximate message passing”. In: *The Annals of Statistics* 46.1 (2018), pp. 119–148.

[MMB+18b] Ali Mousavi, Arian Maleki, Richard G Baraniuk, et al. “Consistent parameter estimation for LASSO and approximate message passing”. In: *The Annals of Statistics* 46.1 (2018), pp. 119–148.

[MS77] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes*. Vol. 16. Elsevier, 1977.

[MY88] Katta G Murty and Feng-Tien Yu. *Linear complementarity, linear and nonlinear programming*. Vol. 3. Citeseer, 1988.

[Nes83] Yurii Nesterov. “A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2)”. In: *Doklady an ussr*. Vol. 269. 1983, pp. 543–547.

[PB14] Neal Parikh and Stephen Boyd. “Proximal algorithms”. In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[Pit82] Loren D Pitt. “Positively correlated normal variables are associated”. In: *The Annals of Probability* (1982), pp. 496–499.

[Pol64] Boris T Polyak. “Some methods of speeding up the convergence of iteration methods”. In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17.

[Ran11] Sundeep Rangan. “Generalized approximate message passing for estimation with random linear mixing”. In: *Proc. IEEE Int. Symp. Inf. Theory*. 2011, pp. 2168–2172.

[Roy68] Halsey Lawrence Royden. *Real analysis*. Krishna Prakashan Media, 1968.

[Rud+64] Walter Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1964.

[Rud+76] Walter Rudin et al. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York, 1976.

[RV18] Cynthia Rush and Ramji Venkataramanan. "Finite sample analysis of approximate message passing algorithms". In: *IEEE Trans. on Inf. Theory* 64.11 (2018), pp. 7264–7286.

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009.

[SBC17] Weijie Su, Małgorzata Bogdan, and Emmanuel J Candès. "False discoveries occur early on the Lasso path". In: *The Annals of Statistics* 45.5 (2017), pp. 2133–2150.

[SC+16] Weijie Su, Emmanuel Candes, et al. "SLOPE is adaptive to unknown sparsity and asymptotically minimax". In: *The Annals of Statistics* 44.3 (2016), pp. 1038–1068.

[SC16] Weijie Su and Emmanuel J Candès. "SLOPE is adaptive to unknown sparsity and asymptotically minimax". In: *The Annals of Statistics* 44.3 (2016), pp. 1038–1068.

[SCC19] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. "The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square". In: *Probability Theory and Related Fields* 175.1-2 (2019), pp. 487–558.

[Sha92] Alexander Shapiro. “Perturbation analysis of optimization problems in Banach spaces”. In: *Numerical Functional Analysis and Optimization* 13.1-2 (1992), pp. 97–116.

[SNW12] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. MIT Press, 2012.

[Su18] Weijie J Su. “When is the first spurious variable selected by sequential regression procedures?” In: *Biometrika* 105.3 (2018), pp. 517–527.

[SZ49] Herbert E Salzer and Ruth Zucker. “Table of the zeros and weight factors of the first fifteen Laguerre polynomials”. In: *Bulletin of the American Mathematical Society* 55.10 (1949), pp. 1004–1012.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. “Precise Error Analysis of Regularized M-Estimators in High Dimensions”. In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5592–5628.

[Tib96a] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1 (1996), pp. 267–288.

[Tib96b] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[TOH14] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. “The Gaussian min-max theorem in the presence of convexity”. In: *arXiv preprint arXiv:1408.4837* (2014).

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. “Regularized linear regression: A precise analysis of the estimation error”. In: *Proceedings of Machine Learning Research* 40 (2015), pp. 1683–1709.

[Wai09] Martin J Wainwright. “Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)”. In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

[Wan+20] Hua Wang et al. “The complete Lasso tradeoff diagram”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[WMZ18] Haolei Weng, Arian Maleki, and Le Zheng. “Overcoming the limitations of phase transition by higher order analysis of regularization techniques”. In: *Annals of Statistics* 46.6A (2018), pp. 3099–3129.

[WWM17] Shuaiwen Wang, Haolei Weng, and Arian Maleki. “Which bridge estimator is optimal for variable selection?” In: *arXiv preprint arXiv:1705.08617* (2017).

[WWM19] Shuaiwen Wang, Haolei Weng, and Arian Maleki. "Does SLOPE outperform bridge regression?" In: *arXiv preprint arXiv:1909.09345* (2019).

[WYS20] Hua Wang, Yachong Yang, and Weijie J Su. "The price of competition: Effect size heterogeneity matters in high dimensions". In: *arXiv preprint arXiv:2007.00566* (2020).

[YL06a] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[YL06b] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[ZB21] Yiliang Zhang and Zhiqi Bu. "Efficient designs of SLOPE penalty sequences in finite dimension". In: *The 24th International Conference on Artificial Intelligence and Statistics* (2021).

[Zei12] Matthew D Zeiler. "Adadelta: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701* (2012).

[ZF14] Xiangrong Zeng and Mario AT Figueiredo. "Decreasing Weighted Sorted ℓ_1 Regularization". In: *IEEE Signal Processing Letters* 21.10 (2014), pp. 1240–1244.

[ZH05a] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[ZH05b] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[Zou06a] Hui Zou. “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101.476 (2006), pp. 1418–1429.

[Zou06b] Hui Zou. “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101.476 (2006), pp. 1418–1429.