# Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing

Zhiqi Bu* Jason Klusowski† Cynthia Rush‡ Weijie Su§

July 18, 2019

**Abstract**

SLOPE is a relatively new convex optimization procedure for high-dimensional linear regression via the sorted $\ell_1$ 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. In this paper, 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 $\ell_1$ 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.

## 1 Introduction

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tag{1.1}$$

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 [9]. This method reconstructs the signal by solving the minimization problem

$$\widehat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \sum_{i=1}^{p} \lambda_i |\boldsymbol{b}|_{(i)}, \tag{1.2}$$

*Department of Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: zbu@sas.upenn.edu

†Department of Statistics, Rutgers University, New Brunswick, NJ 08854, USA. Email: jason.klusowski@rutgers.edu

‡Department of Statistics, Columbia University, New York, NY 10027, USA. Email: cynthia.rush@columbia.edu

§Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: suw@wharton.upenn.edu This work was supported in part by NSF #1217023.

where $\|\cdot\|$ denotes the $\ell_2$ 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* $\ell_1$*-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 [38] and knockoffs [2]. Explicitly, on the estimation side, SLOPE achieves minimax estimation properties under certain random designs *without* requiring any knowledge of the sparsity degree of $\boldsymbol{\beta}$ [37, 7]. On the testing side, SLOPE controls the false discovery rate in the case of independent predictors [9, 11]. For completeness, we remark that [10, 39, 19] 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) [4, 16, 23, 31]. 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 [5]. 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 $t \geq 0$,

$$\boldsymbol{\beta}^{t+1} = \mathrm{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t), \tag{1.3a}$$

$$\boldsymbol{z}^{t+1} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{t+1} + \frac{\boldsymbol{z}^t}{n}\Big[\nabla \mathrm{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top \boldsymbol{z}^t + \boldsymbol{\beta}^t)\Big]. \tag{1.3b}$$

The non-increasing sequence $\boldsymbol{\theta}_t$ is proportional to $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and will be given explicitly in Section 2. Here, $\mathrm{prox}_{J_{\boldsymbol{\theta}}}$ is the proximal operator of the sorted $\ell_1$ norm, that is,

$$\mathrm{prox}_{J_{\boldsymbol{\theta}}}(\boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{b}} \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{b}\|^2 + J_{\boldsymbol{\theta}}(\boldsymbol{b}), \tag{1.4}$$

and $\nabla \mathrm{prox}_{J_\theta}$ denotes the divergence of the proximal operator (see an equivalent, but more explicit form, of this algorithm in Section 2 and further discussion of SLOPE and the prox operator in Section 5.1). Compared to the proximal gradient descent (ISTA) [13, 14, 29], 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 [16].

The *empirical* performance of AMP in solving SLOPE under i.i.d. Gaussian matrix $\boldsymbol{X}$ is illustrated in Figure 1 and Table 1, which suggest the superiority of AMP over ISTA and FISTA [6]—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, 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 [5], 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 [3, 36, 28, 35].

**Main contributions**. In this work, we prove that the AMP algorithm (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 [8] and applies this tool to derive the state evolution that describes the asymptotically exact behaviors of the AMP iterates $\boldsymbol{\beta}^t$ in (1.3). The next step, which is the core of our proof, is to relate the AMP estimates to the

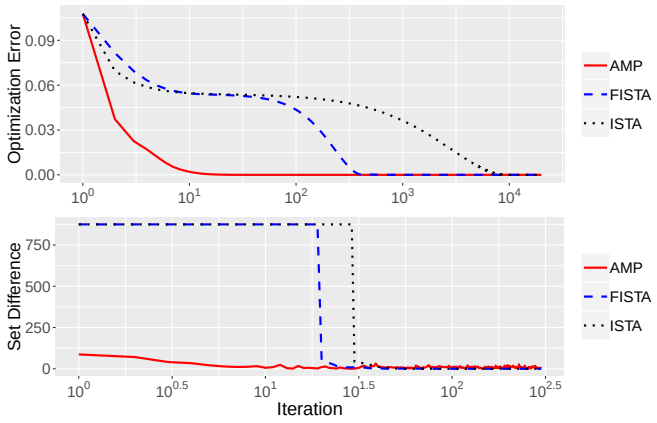


Figure 1: Optimization errors, $||\boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}}||^2/p$, and (symmetric) set difference of $\text{supp}(\boldsymbol{\beta}^t)$ and $\text{supp}(\widehat{\boldsymbol{\beta}})$.

| | | 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 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 1 and Table 1 Details: Design $X$ is $500 \times 1000$ 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).

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 [20]—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 we develop an AMP algorithm for finding the SLOPE estimator in (1.2). Specifically, it is through the threshold values $\boldsymbol{\theta}_t$ in the AMP algorithm in (1.3) that one can ensure the AMP estimates converge to the SLOPE estimator with parameter $\boldsymbol{\lambda}$, so in Section 2 we provide details for how one should calibrate the thresholds of the AMP iterations in (1.3) in order for the algorithm to solve SLOPE cost in (1.2). Then in Section 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 [8], like the one in (1.3). Finally, Sections 4-7 prove rigorously the theoretical results stated in Section 3 and we end with a discussion in Section 8.

## 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 (1.1) and parameter vector in (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. $\Lambda$, 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 $\Lambda$ 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:

$$\nabla \operatorname{prox}_f(\boldsymbol{v}) := \sum_{i=1}^{p} \frac{\partial}{\partial v_i} [\operatorname{prox}_f(\boldsymbol{v})]_i = \Big( \frac{\partial}{\partial v_1}, \frac{\partial}{\partial v_2}, \ldots, \frac{\partial}{\partial v_p} \Big) \cdot \operatorname{prox}_f(\boldsymbol{v}), \tag{2.1}$$

where [37, proof of Fact 3.4],

$$\frac{\partial [\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i}{\partial v_j} = \begin{cases} \frac{\operatorname{sign}([\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i) \cdot \operatorname{sign}([\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j)}{\#\{1 \le k \le p : |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_k| = |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j|\}}, & \text{if } |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_j| = |[\operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v})]_i|, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Hence the divergence takes the simplified form

$$\nabla \operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}) = \| \operatorname{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}) \|_0^*, \tag{2.3}$$

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 2.1.** *In AMP,* (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 5.1. Now we discuss the details of an AMP algorithm that can be used for finding the SLOPE estimator in (1.2).

## 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 [4, 8, 35, 21]. 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 (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 $\tau_t$ is defined by the state evolution. For this reason, the function used to update the estimate in (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 [4], and an analysis of the rate of this convergence was given in [35]. A 'separable' denoiser is one that applies the same (possibly non-linear) function to each element of its input. Recent work [8] 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 (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 $t \geq 0$,

$$\tau_{t+1}^2 = \sigma_w^2 + \lim_p \frac{1}{\delta p} \mathbb{E}\|\text{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{B} + \tau_t \boldsymbol{Z}) - \boldsymbol{B}\|^2. \tag{2.4}$$

Below we make rigorous the way that the recursion in (2.4) relates to the AMP iteration (1.3).

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

## 2.2 Analysis of the AMP State Evolution

As the state evolution (2.4) predicts the performance of the AMP algorithm (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 (2.4). Moreover, recall that through the sequence of thresholds $\boldsymbol{\theta}_t$, one can relate the AMP algorithm to the SLOPE estimator in (1.2) for a specific $\boldsymbol{\lambda}$, and the explicit form of this calibration, given in Section 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 (2.4). The definition of $\tau_{t+1}^2(p)$ is stated explicitly in (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 2.3. 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 (2.4), namely

$$\tau_{t+1}^2(p) = \sigma_w^2 + \frac{1}{\delta p} \mathbb{E}\|\text{prox}_{J_{\boldsymbol{\alpha}\tau_t(p)}}(\boldsymbol{\beta} + \tau_t(p)\boldsymbol{Z}) - \boldsymbol{\beta}\|^2, \tag{2.5}$$

where the difference between (2.5) and the state evolution (2.4) is via the large system limit in $p$. When we refer to the recursion in (2.5), we will always specify the $p$ dependence explicitly as $\tau_t(p)$. An analysis of the limiting properties (in $t$) of (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 (2.5) holds for (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.** *Any stationary point $\widehat{\boldsymbol{\beta}}$ (with corresponding $\widehat{\boldsymbol{z}}$) in the AMP algorithm* (1.3a)-(1.3b) *with $\boldsymbol{\theta}_* = \boldsymbol{\alpha}\tau_*$ is a minimizer of the SLOPE cost function in* (1.2) *with*

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

*Proof of Lemma 2.2.* Denote, $\omega := (\nabla \operatorname{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}}))/(\delta p)$. Now, by stationarity,

$$\widehat{\boldsymbol{\beta}} = \operatorname{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}}), \qquad \text{and} \qquad \widehat{\boldsymbol{z}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \frac{\widehat{\boldsymbol{z}}}{\delta p}(\nabla \operatorname{prox}_{J_{\boldsymbol{\theta}_*}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^\top \widehat{\boldsymbol{z}})). \tag{2.6}$$

From (2.6), notice that $\widehat{\boldsymbol{z}} = \frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}}{1-\omega}$. By Fact 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 5.1). Then, $\boldsymbol{X}^\top \widehat{\boldsymbol{z}} = \frac{\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{1-\omega} \in \partial J_{\boldsymbol{\theta}_*}(\widehat{\boldsymbol{\beta}})$, and therefore $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in \partial J_{\boldsymbol{\theta}_*(1-\omega)}(\widehat{\boldsymbol{\beta}})$ which is *exactly* the stationary condition of SLOPE with regularization parameter $\boldsymbol{\lambda} = (1-\omega)\boldsymbol{\theta}_*$, as desired. □

Now we present Theorem 1, which provides results about the $t$ asymptotics of the recursion in (2.5) and its proof is given in Appendix A. First, some notation must be introduced: let $\boldsymbol{A}_{\min}(\delta)$ be the set of solutions to

$$\delta = f(\boldsymbol{\alpha}), \quad \text{where} \quad f(\boldsymbol{\alpha}) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{\Big(1 - |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_i| \sum_{j \in I_i} \alpha_j\Big)\Big/[\boldsymbol{D}(\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z}))]_i\Big\}. \tag{2.7}$$

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 $\infty$ otherwise. Let $I_i = \{j : 1 \le j \le p \text{ and } |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_j| = |[\operatorname{prox}_{J_{\boldsymbol{\alpha}}}(\boldsymbol{Z})]_i|\}$. The expectation in (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.

**Theorem 1.** *For any $\boldsymbol{\alpha}$ strictly larger than at least one element in the set $\boldsymbol{A}_{\min}(\delta)$, the recursion in* (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 $F : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ as*

$$F\Big(\tau^2(p), \boldsymbol{\alpha}\tau(p)\Big) := \sigma_w^2 + \frac{1}{\delta p}\mathbb{E}\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau(p)}}(\boldsymbol{B} + \tau(p)\boldsymbol{Z}) - \boldsymbol{B}\|^2, \tag{2.8}$$

*where $\boldsymbol{B}$ is elementwise i.i.d. $B$ independent of $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$, so that $\tau_{t+1}^2(p) = F(\tau_t^2(p), \boldsymbol{\alpha}\tau_t(p))$. Then $|\frac{\partial F}{\partial \tau^2(p)}(\tau^2(p), \boldsymbol{\alpha}\tau(p))| < 1$ at $\tau(p) = \tau_*(p)$. Moreover, for $f(\boldsymbol{\alpha})$ defined in* (2.7)*, we show that $f(\boldsymbol{\alpha}) = \delta \lim_{\tau(p)\to\infty} dF/d\tau^2(p)$.*

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 (2.5), and equivalently, the calibration detailed in Section 2.3 below are well-defined.

Recall that it is actually the state evolution in (2.4) (and not that in (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 (2.5), can easily be used to give an analogous result for

(2.4). In particular analyzing (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 $\tau^2$. Indeed, to directly make use our proof for the finite-$p$ case given in Theorem 1, it is enough that

$$\frac{\partial}{\partial\tau^2}\lim_p \mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p)=\lim_p \frac{\partial}{\partial\tau^2}\mathbb{E}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p). \tag{2.9}$$

Note that we already have an argument (based on dominated convergence for fixed $p$, see (A.1) and Lemma A.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., [34, Theorem 7.17]) to justify the interchange in (2.9).

**Lemma 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 2.4.** *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 $\tau_1$, $\tau_2$ in a bounded set $D=\{\tau:0<r\le|\tau|\le R\}$,

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

This follows by the mean value theorem and (A.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 2.5.** *The boundedness of $\{\tau_t(p)\}$ is guaranteed by Proposition 2.6. In particular, since $\boldsymbol{\alpha}$ satisfies the assumption of Theorem 1, Proposition 2.6 guarantees $\boldsymbol{\lambda}$ is bounded and, consequently, so is $\tau$ (see the calibration in* (2.10) *below).*

### 2.3 Threshold Calibration

Motivated by Lemma 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

$$\boldsymbol{\lambda}(\boldsymbol{\alpha}):=\boldsymbol{\alpha}\tau_*(p)\Big(1-\frac{1}{n}\,\mathbb{E}\,\|\,\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau_*(p)}}(\boldsymbol{B}+\tau_*(p)\boldsymbol{Z})\|_0^*\Big), \tag{2.10}$$

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 (2.10) 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 2.6 we show that the calibration (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.

i.e. we can get $\alpha$ from $\lambda$

**Proposition 2.6.** *The function $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ defined in (2.10) is continuous on $\{\boldsymbol{\alpha} : f(\boldsymbol{\alpha}) < \delta\}$ for $f(\cdot)$ defined in (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 (2.10) exists. As $p \to \infty$, the function $\boldsymbol{\alpha} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha})$ becomes invertible (given $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ satisfying (2.10) exists uniquely). Furthermore, the inverse function is continuous non-decreasing for any $\boldsymbol{\lambda} > \boldsymbol{0}$.*

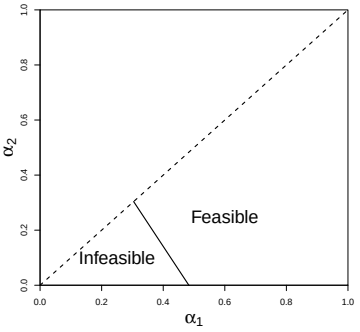

Figure 2: $\boldsymbol{A}_{\min}$ (black curve) when $p = 2$ and $\delta = 0.6$.

In [5, Proposition 1.4 (first introduced in [17]) and Corollary 1.7] this is proven rigorously for the analogous LASSO calibration and in Appendix A 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 (2.10), 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 $L : \mathbb{R} \to \mathbb{R}; \alpha \mapsto \operatorname{sign}(\boldsymbol{\lambda}(\alpha\boldsymbol{\ell}) - \boldsymbol{\lambda})$ **do**

2. Set $\alpha_1 = \alpha_2, \alpha_2 = 2\alpha_2$

**end while**

3. **return BISECTION** $(L(\alpha), \alpha_1, \alpha_2)$

Remark: $\operatorname{sign}(\boldsymbol{\lambda}(\cdot) - \boldsymbol{\lambda}) \in \mathbb{R}$ 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 (2.10) 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 $\Lambda$. It follows that the sequence $\{\boldsymbol{\alpha}(p)\}_{p\geq 0}$ defined for each $p$ by the finite-sample calibration (2.10) are i.i.d. from a distribution $A$, where $A$ satisfies $\mathbb{E}(A^2) < \infty$, and is defined via

$$\Lambda = A\tau_*\Big(1 - \lim_p \frac{1}{\delta p}\,\mathbb{E}\,||\operatorname{prox}_{J_{\boldsymbol{A}(p)\tau_*}}(\boldsymbol{B} + \tau_*\boldsymbol{Z})||_0^*\Big), \tag{2.11}$$

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

**Fact 2.7.** *The limits in (2.4) and (2.11) exist.*

This fact is proven in Appendix C. One idea used in the proof of Fact 2.7 is that the prox operator is *asymptotically* separable, a result shown by [20, 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 $\Lambda$, respectively, then there exists a limiting scalar function $h(\cdot) := h(\boldsymbol{v}(p); V, \Lambda)$ (determined by $V$ and $\Lambda$) of the proximal operator $\mathrm{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}(p))$. Further details are given in Lemma 3.3 in Section 3. Using $h(\cdot) := h(\cdot; B + \tau_* Z, A\tau_*)$, this argument implies that (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 (2.11) can be represented as

$$\mathbb{P}\left(B + \tau_* Z \in \left\{x \;\middle|\; h(x) \neq 0 \quad \text{and} \quad m\{z \mid |h(z)| = |h(x)|\} = 0\right\}\right). \tag{2.12}$$

In other words, the limit in (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).

# 3 Asymptotic Characterization of SLOPE

## 3.1 AMP Recovers the SLOPE Estimate

Here we show that the AMP algorithm converges in $\ell_2$ 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 [5, Lemma 3.1], however, major differences lie in the fact that the proximal operator used in the AMP updates (1.3a)-(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* (1.3a) *and the SLOPE estimate* (1.2)*,*

$$\operatorname*{plim}_{p\to\infty} \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2 = c_t, \quad \textit{where} \quad \lim_{t\to\infty} c_t = 0. \tag{3.1}$$

The proof of Theorem 2 can be found in Section 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 (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 (2.10), 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 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 (3.1). First, it implies that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|^2/p$ converges in probability to a constant, say $c_t$. Next, (3.1) says $c_t \to 0$ as $t \to \infty$.

### 3.2 Exact Asymptotic Characterization of the SLOPE Estimate

A consequence of Theorem 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 3.1. Uniformly pseudo-Lipschitz functions** *[8]: 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 $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,*

$$\|\phi(\boldsymbol{a}) - \phi(\boldsymbol{b})\| \leq L\Big(1 + (\|\boldsymbol{a}\|/\sqrt{d})^{k-1} + (\|\boldsymbol{b}\|/\sqrt{d})^{k-1}\Big)\Big(\|\boldsymbol{a} - \boldsymbol{b}\|/\sqrt{d}\Big). \tag{3.2}$$

*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 $\phi_p$, $L_p < \infty$ 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 $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$,*

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

*where $\tau_t$ is defined in* (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 $\operatorname{prox}_{J_{\boldsymbol{\alpha}(p)\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z})$ for large $p, t$.

The proof of Theorem 3 can be found in Section 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 [8], which can be used to demonstrate that the state evolution given in (2.4) characterizes the performance of the SLOPE AMP (1.3) via pseudo-Lipschitz loss functions.

We note that [20, Theorem 1] follows by Theorem 3 and their separability result [20, Proposition 1]. To see this, we use the following lemma that is a simple application of the Law of Large Numbers.

**Lemma 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*

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

*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 [20, 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

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

Then [20, Theorem 1] follows by [20, Proposition 1], restated below in Lemma 3.3, the Law of Large Numbers, and Theorem 1. Now we restate in Lemma 3.3, the result given in [20, Proposition 1], which says that $\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ becomes asymptotically separable as $p \to \infty$, for convenience.

**Lemma 3.3** (Proposition 1, [20]). *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 $\Lambda$, respectively, then there exists a limiting scalar function $h$ (determined by $V$ and $\Lambda$) such that as $p \to \infty$,*

$$\|\text{prox}_{J_{\boldsymbol{\lambda}(p)}}(\boldsymbol{v}(p)) - h(\boldsymbol{v}(p); V, \Lambda)\|^2/p \to 0, \tag{3.3}$$

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

Then [20, Theorem 1] follows from Theorem 3 by using the asymptotic separability of the prox operator. Namely, the result of Lemma 3.3 (using that $\boldsymbol{\alpha}(p)\tau_t$ has an empirical distribution that converges weakly to $A\tau_t$ for $A$ defined by (2.11)), along with Cauchy-Schwarz and the fact that $\psi$ is pseudo-Lipschitz, allow us to apply a dominated convergence argument (see Lemma B.2), from which it follows for some limiting scalar function $h^t$ as specified by Lemma 3.3,

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

Then the above allows us to apply Lemma 3.2 and the Law of Large Numbers to show

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

Finally we note that the result of [20, Theorem 1] follows since

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

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 $\tau_*$ and equivalently minimum error [20], 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 [12].

**Corollary 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 $\operatorname{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \lim_t \operatorname{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}) - \boldsymbol{\beta}\|^2]$. The desired result follows since $\lim_t \operatorname{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{\beta} + \tau_t \boldsymbol{Z}) - \boldsymbol{\beta}\|^2] = \delta(\tau_*^2 - \sigma_w^2)$. To see this, note that $\lim_t \delta(\tau_{t+1}^2 - \sigma_w^2) = \delta(\tau_*^2 - \sigma_w^2)$ and

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

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 3.2 are satisfied and follows similarly to that used to prove property **(P2)** stated in Section 4 and proved in Appendix B. □

# 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 (2.4) characterizes the performance of the SLOPE AMP algorithm (1.3b), given in Lemma 4.1 below. Specifically, Lemma 4.1 relates the state evolution (2.4) to the output of the AMP iteration (1.3b) for pseudo-Lipschitz loss functions. This result follows from [8, Theorem 14], which is a general result relating state evolutions to AMP algorithm with non-separable denoisers. In order to apply [8, Theorem 14], we need to demonstrate that our denoiser, i.e. the proximal operator $\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ defined in (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 (1.4), i.e. for a vector $\boldsymbol{v} \in \mathbb{R}^p$, define

$$\eta_p^t(\boldsymbol{v}) := \text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{v}). \tag{4.1}$$

**(P1)** For each $t$, denoisers $\eta_p^t(\cdot)$ defined in (4.1) are uniformly Lipschitz (i.e. uniformly pseudo-Lipschitz of order $k = 1$) per Definition 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 $2 \times 2$ covariance matrix, the following limits exist and are finite.

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

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

**Lemma 4.1.** *[8, Theorem 14] Under assumptions **(A1)** - **(A4)**, given that **(P1)** and **(P2)** are satisfied, for the AMP algorithm in* (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*

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

*where $\tau_t$ is defined in* (2.4).

We now show that Theorem 3 follows from Lemma 4.1 and Theorem 2.

*Proof of Theorem 3.* First, for any fixed $n$ and $t$, the following bound uses that $\psi_n$ 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

$$\lim_t \operatorname*{plim}_p \Big(\|\boldsymbol{\beta}^t\|/\sqrt{p}\Big)^{k-1}, \qquad \operatorname*{plim}_p \Big(\|\widehat{\boldsymbol{\beta}}\|/\sqrt{p}\Big)^{k-1}, \qquad \text{and} \qquad \operatorname*{plim}_p \Big(\|\boldsymbol{\beta}\|/\sqrt{p}\Big)^{k-1}, \tag{4.2}$$

are all finite, we have $\text{plim}_p \psi_p(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \lim_t \text{plim}_p \psi_p(\boldsymbol{\beta}^t, \boldsymbol{\beta})$. But by Theorem 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 (4.2) are finite. Since $k$ finite, that the third term in (4.2) is finite follows by property **(P2)**. Bounds for the first and second term are demonstrated in Lemma 7.1 found in Appendix 6.

□

# 5 Proof AMP Finds the SLOPE Solutions

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

$$\mathcal{C}(\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + J_{\boldsymbol{\lambda}}(\boldsymbol{b}), \tag{5.1}$$

where $J_{\boldsymbol{\lambda}}(\boldsymbol{b})$ is the sorted $\ell_1$-norm. The proof of Theorem 2 relies on a technical lemma, Lemma 5.5, stated in Section 5.2 below, that deals carefully with the fact that the SLOPE cost function given in (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 (2.10) 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 5.1) and LASSO's support set. This approach, taken in Lemma 5.5 below, allows us to deal with the fact that we are not guaranteed to have a strongly convex cost. Then Lemma 5.5 is used to prove Theorem 3.

Before we state Lemma 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 5.1.

## 5.1 Preliminaries on SLOPE

In general, we refer to the function $\mathcal{C}(\cdot)$ stated in (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 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 (1.4) is linked to the subgradient of the SLOPE norm in the following way.

**Fact 5.2.** *If* $\text{prox}_{J_{\boldsymbol{\lambda}}}(\boldsymbol{v}_1) = \boldsymbol{v}_2$, *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 $\nabla \text{prox}$, which equals the number of non-zero equivalence classes. For example, $\frac{\partial \text{prox}}{\partial \boldsymbol{v}}|_{\boldsymbol{v}=(1,0,-1,3)} = (\frac{1}{2}, 0, \frac{1}{2}, 1)$ 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 5.3.**

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

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,

$$\boldsymbol{A} = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/6 & 1/3 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix} \tag{5.2}$$

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 5.3 gives the usual subgradient of the $\ell_1$ norm.

Using the subgradient definition in Fact 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 (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 $\lambda_1$ and $2/3$ in $\lambda_2$; the second entry of $\boldsymbol{v}_I$ is a weighted average with $1/6$ its weight in $\lambda_1$, $1/3$ in $\lambda_2$, and $1/2$ in $\lambda_2$, 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 $\lambda_1$'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 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 $\ell_1$ 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 5.3, we refer the reader to [32, 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 [32, Exercise 8.31]) the subgradient takes the form of a convex hull of the partial derivatives of the maximizing elements:

$$\partial J_{\boldsymbol{\lambda}}(\boldsymbol{s}) \in \text{conv}\{\nabla_{\boldsymbol{s}}(\boldsymbol{\lambda}^\top f_i(\boldsymbol{s})) : i \in A(\boldsymbol{s})\} \equiv \text{conv}\{f_i^{-1}(\boldsymbol{\lambda}) : i \in A(\boldsymbol{s})\}, \tag{5.3}$$

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.

## 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 5.5, let us introduce a very important definition:

**Definition 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*

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

For example, if $\boldsymbol{v} = (1, 1, -1, 0, 2, -1)$, then $\text{supp}^\star(\boldsymbol{v}) = \{\{1, 2, 3, 6\}, \{5\}\}$. Now we state and prove Lemma 5.5.

**Lemma 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 $s_t(c_2) := \{I \subset [p] : |\boldsymbol{\nu}_I^t| \succeq [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I(1 - c_2)\}$ 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 5.5. Once we have proved Lemma 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 5.5 are satisfied. Then the result of Theorem 2 is immediate. We show the five assumptions are met in Sections 7.1 - 7.5. Now we prove the Lemma.

*Proof of Lemma 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 $\epsilon$ such that $\xi_i(\epsilon) \to 0$ as $\epsilon \to 0$ (we omit the dependence of $\xi_i$ on $\epsilon$). 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. [5, 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,

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

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 (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

$$\left[\langle \hat{\boldsymbol{\lambda}}_S, |\boldsymbol{\beta}^t_S + \boldsymbol{r}_S| - |\boldsymbol{\beta}^t_S| \rangle - \langle \boldsymbol{\nu}_S, \boldsymbol{r}_S \rangle\right] + \left[\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle\right] + \frac{\|\boldsymbol{X}\boldsymbol{r}\|^2}{2p} \leq \frac{\|sg(\mathcal{C}, \boldsymbol{\beta}^t)\|\|\boldsymbol{r}\|}{p} \leq c_1\epsilon \quad (5.4)$$

We now show all three terms on the left side of (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 (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 $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$ 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 $\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$ follows since $\boldsymbol{\nu}$ is a valid subgradient of $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ (see Fact 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 5.3, since $|\boldsymbol{\nu}_I| \in [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I$, we know that $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I(1-c_2)$ 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 $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I(1-c_2)$ since $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$. In the other direction, if the non-zero equivalence class $I$ is such that $|\boldsymbol{\nu}_I| \succeq [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I(1-c_2)$ then there exists a vector $\widetilde{\boldsymbol{\nu}}_I \in [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I$ such that $|\boldsymbol{\nu}_I| \geq \widetilde{\boldsymbol{\nu}}_I(1-c_2)$ elementwise. However since $\widetilde{\boldsymbol{\nu}}_I \in [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I$, this implies that $|\boldsymbol{\nu}_I| \geq (1-c_2)\hat{\boldsymbol{\lambda}}_I$ is also true since $\hat{\boldsymbol{\lambda}}_I \in [\mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))]_I$ 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 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 $\hat{\boldsymbol{\lambda}} \in \mathcal{P}(\hat{\Pi}^{-1}_{\boldsymbol{\beta}^t}(\boldsymbol{\lambda}))$).

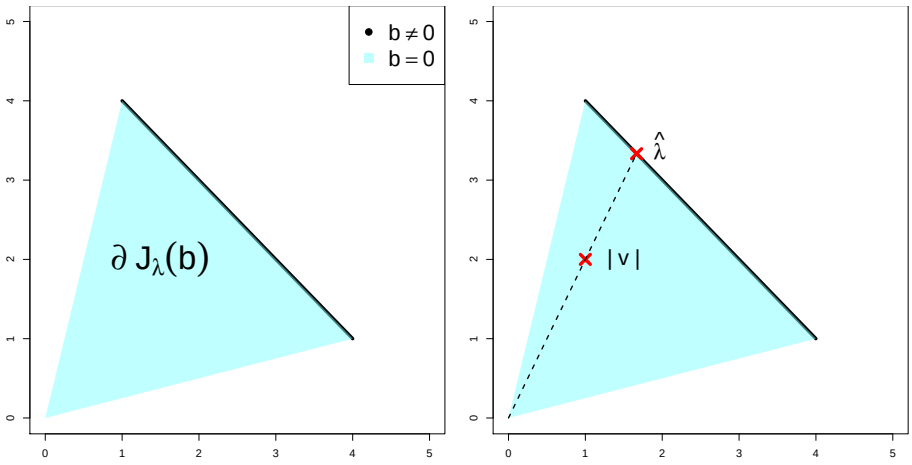


Figure 3: The blue area contained by the black line segment is the set of subgradients; Red crosses are examples of $\boldsymbol{\nu}_I$ and $\hat{\boldsymbol{\lambda}}_I$ correspondingly when $\boldsymbol{b}_I = \mathbf{0}$.

Now we would like to show that the first term in (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 $(\beta_i^t + r_i)\left[\operatorname{sign}(\beta_i^t + r_i) - \operatorname{sign}(\beta_i^t)\right]$ 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 (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 (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,

$$
\begin{aligned}
\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}} \rangle &\leq \xi_1(\epsilon), &(5.5)\\
\|\boldsymbol{X}\boldsymbol{r}\|^2 &\leq p\xi_1(\epsilon). &(5.6)
\end{aligned}
$$

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

Decompose $\boldsymbol{r}$ as $\boldsymbol{r} = \boldsymbol{r}^\perp + \boldsymbol{r}^\parallel$, with $\boldsymbol{r}^\parallel \in \ker(\boldsymbol{X})$ and $\boldsymbol{r}^\perp \in \ker^\perp(\boldsymbol{X})$ so that $\boldsymbol{X}\boldsymbol{r} = \boldsymbol{X}\boldsymbol{r}^\perp$. We will now use (5.5) and (5.6) to obtain bounds for $\|\boldsymbol{r}^\perp\|^2$ and $\|\boldsymbol{r}^\parallel\|^2$. First notice that by (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 $\|\boldsymbol{r}^\parallel\|^2$. To bound $\|\boldsymbol{r}^\parallel\|^2$, 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 $\|\boldsymbol{r}^\parallel_{\bar{Q}}\|^2$ and show a high probability lower bound for $\sigma^2_{min}(\boldsymbol{X}_Q)$.

In (5.5), decompose $\boldsymbol{r}_{\bar{S}} = \boldsymbol{r}_{\bar{S}}^{\perp} + \boldsymbol{r}_{\bar{S}}^{\|}$ and observe that by Cauchy Schwarz inequality and the bound just obtained,

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}^{\perp}| \rangle \leq \frac{1}{p} \|\hat{\boldsymbol{\lambda}}_{\bar{S}}\| \|\boldsymbol{r}_{\bar{S}}^{\perp}\| \leq \frac{1}{p} \|\hat{\boldsymbol{\lambda}}\| \|\boldsymbol{r}^{\perp}\| \leq \frac{1}{\sqrt{p}} \|\hat{\boldsymbol{\lambda}}\| \sqrt{c_5 \xi_1(\epsilon)}. \tag{5.7}$$

Then we use the fact that

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

to get from (5.5) and (5.7) that

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}^{\|}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}^{\|} \rangle \leq \xi_2(\epsilon). \tag{5.8}$$

Next we would like to show

$$\langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}, |\boldsymbol{r}_{\bar{S}(c_2)}^{\|}| \rangle - \langle \boldsymbol{\nu}_{\bar{S}(c_2)}, \boldsymbol{r}_{\bar{S}(c_2)}^{\|} \rangle (1 - c_2)^{-1} \geq 0. \tag{5.9}$$

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,

$$\begin{aligned}
\langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)}, |\boldsymbol{r}_{\bar{S}(c_2)}^{\|}| \rangle &\overset{(a)}{\leq} \frac{1}{c_2} \langle \boldsymbol{\lambda}_{\bar{S}(c_2)}, |\boldsymbol{r}_{\bar{S}(c_2)}^{\|}| \rangle - \frac{1}{c_2} \langle \boldsymbol{\nu}_{\bar{S}(c_2)}, \boldsymbol{r}_{\bar{S}(c_2)}^{\|} \rangle = \frac{1}{c_2} \langle \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)} - \boldsymbol{\nu}_{\bar{S}(c_2)} \operatorname{sign}(\boldsymbol{r}_{\bar{S}(c_2)}^{\|}), |\boldsymbol{r}_{\bar{S}(c_2)}^{\|}| \rangle \\
&\overset{(b)}{\leq} \frac{1}{c_2} \langle \hat{\boldsymbol{\lambda}}_{\bar{S}} - \boldsymbol{\nu}_{\bar{S}} \operatorname{sign}(\boldsymbol{r}_{\bar{S}}^{\|}), |\boldsymbol{r}_{\bar{S}}^{\|}| \rangle = \frac{1}{c_2} \langle \hat{\boldsymbol{\lambda}}_{\bar{S}}, |\boldsymbol{r}_{\bar{S}}^{\|}| \rangle - \frac{1}{c_2} \langle \boldsymbol{\nu}_{\bar{S}}, \boldsymbol{r}_{\bar{S}}^{\|} \rangle \overset{(c)}{\leq} c_2^{-1} \xi_2(\epsilon).
\end{aligned} \tag{5.10}$$

In particular, step $(a)$ follows by (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 (5.8). We now use the bound in (5.10) to bound components of $\boldsymbol{r}^{\|}$.

In order to bound $\|\boldsymbol{r}^{\|}\|^2$, we would like to exploit a relationship between the $\ell_1$ and $\ell_2$ 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 $\ell$, it should be smaller than the average of all the elements in the previous group $\ell - 1$.

Then,

$$\begin{aligned}
\|\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2 \overset{(a)}{=} \sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_\ell}^{\|}\|^2 \overset{(b)}{\leq} \sum_{\ell=2}^{K} |S_\ell| \frac{\|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1^2}{|S_{\ell-1}|^2} \overset{(c)}{\leq} \frac{4}{pc_3} \sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1^2 \leq \frac{4}{pc_3} \Big[ \sum_{\ell=2}^{K} \|\boldsymbol{r}_{S_{\ell-1}}^{\|}\|_1 \Big]^2 \\
\overset{(d)}{\leq} \frac{4}{pc_3} \|\boldsymbol{r}_{\bar{S}(c_2)}^{\|}\|_1^2 \overset{(e)}{\leq} \frac{4 \xi_2(\epsilon)^2 p}{c_2^2 c_3 (\min \hat{\boldsymbol{\lambda}}_{\bar{S}(c_2)})^2} =: p \xi_3(\epsilon).
\end{aligned} \tag{5.11}$$

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,

$$c_4^2\|\boldsymbol{r}_{S_+}^{\|}\|^2 \leq \sigma_{min}^2(\boldsymbol{X}_{S_+})\|\boldsymbol{r}_{S_+}^{\|}\|^2 \leq \|\boldsymbol{X}_{S_+}\boldsymbol{r}_{S_+}^{\|}\|^2 \overset{(a)}{=} \|\boldsymbol{X}_{\bar{S}_+}\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2 \overset{(b)}{\leq} 2c_5\|\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2. \tag{5.12}$$

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 $\|\boldsymbol{X}_{\bar{S}_+}\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2 \leq \sigma_{\max}^2(\boldsymbol{X})\|\boldsymbol{r}_{\bar{S}_+}^{\|}\|^2$. 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 (5.11) and (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. □

# 6 Expansion of the AMP State Evolution Ideas

In this section, we develop ideas and notation specifically for the SLOPE AMP algorithm given in (1.3). Most are adapted from the work in [8] 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 5.5. Throughout this section, we use the $\{\eta_p^t\}_{p\in\mathbb{N}_{>0}}$ notation introduced in Section 4 and defined in (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 $\text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$ defined in (1.4), i.e. $\eta_p^t(\boldsymbol{v}) := \text{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\boldsymbol{v})$ 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 $t \geq 0$,

$$\boldsymbol{h}^{t+1} = \boldsymbol{\beta} - (\boldsymbol{X}^\top\boldsymbol{z}^t + \boldsymbol{\beta}^t) \quad \text{and} \quad \boldsymbol{m}^t = \boldsymbol{w} - \boldsymbol{z}^t. \tag{6.1}$$

We next introduce a generalization to the state evolution given in (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

$$\Sigma_{0,t+1} = \sigma_w^2 + \lim_p \frac{1}{\delta p}\mathbb{E}\{-\boldsymbol{B}^\top[\eta_p^t(\boldsymbol{B} + \tau_t\boldsymbol{Z}_t) - \boldsymbol{B}]\}, \tag{6.2}$$

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

$$\Sigma_{s+1,t+1} = \sigma_w^2 + \lim_p \frac{1}{\delta p}\mathbb{E}\Big\{[\eta_p^s(\boldsymbol{B} + \tau_s\boldsymbol{Z}_s) - \boldsymbol{B}]^\top[\eta_p^t(\boldsymbol{B} + \tau_t\boldsymbol{Z}_t) - \boldsymbol{B}]\Big\}, \tag{6.3}$$

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 (2.4).

Using the above covariances, we have the following result that characterizes the asymptotic empirical distributions of the difference vectors defined in (F.1) and generalizes Lemma (4.1). This result follows by [8, Theorem 1].

**Lemma 6.1.** *[8, 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 $k$,*

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

*for $(\boldsymbol{Z}_0, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t)$ defined in (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 $\mathbb{E}[\boldsymbol{Z}_s'] = \mathbb{E}[\boldsymbol{Z}_r'] = \boldsymbol{0}$, $\mathbb{E}\{([\boldsymbol{Z}_s']_i)^2\} = \mathbb{E}\{([\boldsymbol{Z}_r']_i)^2\} = 1$ 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 6.1 to explicitly state asymptotic characterizations of AMP quantities that will be useful in our analysis.

**Lemma 6.2.** *Under the condition of Theorem 3, for $\boldsymbol{z}^t$ and $\boldsymbol{\beta}^{t+1}$ defined in (1.3) and the generalized state evolution sequence defined in (6.3),*

$$\operatorname*{plim}_n \Big( \frac{1}{n}\|\boldsymbol{z}^t - \boldsymbol{z}^{t-1}\|^2 - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) \Big) = 0, \tag{6.4}$$

$$\operatorname*{plim}_p \Big( \frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 - (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) \Big) = 0. \tag{6.5}$$

*Proof.* The major tools in proving (6.4)-(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 (F.1) and finally making an appeal to the Law of Large Numbers. We prove (6.5) and (6.4) follows similarly.

By (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 6.1 for the uniformly pseudo-Lipschitz function

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

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 6.1,

$$\operatorname*{plim}_p \Big( \frac{1}{\delta p}\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 - \frac{1}{\delta p}\mathbb{E}\|\eta_p^t(\boldsymbol{\beta} - \tau_t \boldsymbol{Z}_t) - \eta_p^{t-1}(\boldsymbol{\beta} - \tau_{t-1}\boldsymbol{Z}_{t-1})\|^2 \Big) = 0. \tag{6.6}$$

Now to prove result (6.4), we note that by Lemma 3.2,

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

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 3.2 are met follows like that used in Appendix B in the proof of Proposition **(P2)** introduced in Section 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 $\ell_2$ norm of the residuals in AMP algorithm (1.3b) following from Lemma 4.1.

**Lemma 6.3.** *For $\boldsymbol{z}^t$ defined in* (1.3b) *and $\tau_t^2$ given in* (2.4)*,*

$$\operatorname*{plim}_n \left(\|\boldsymbol{z}^t\|^2/n - \tau_t^2\right) = 0. \tag{6.7}$$

*Proof.* This follows from Lemma 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, $\operatorname{plim}_n \|\boldsymbol{z}^t\|^2/n = \operatorname{plim}_n \mathbb{E}_{\boldsymbol{Z}}[\|\boldsymbol{w}+\sqrt{\tau_t^2-\sigma_w^2}\boldsymbol{Z}\|^2]/n$ 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 $\operatorname{plim}_n \|\boldsymbol{w}\|^2/n = \sigma_w^2$ by the Law of Large Numbers,

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

□

# 7 Verification of Main Technical Lemma Conditions

We now verify that the Lemma 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 7.4 and Section 7.5 below.

## 7.1 Condition (4)

This follows by standard limit theorems about the singular values of Wishart matrices (see Appendix G, Theorem H.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, $\mathcal{C}(\boldsymbol{\beta}^t) \geq \mathcal{C}(\hat{\boldsymbol{\beta}}) = \mathcal{C}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}) \geq \mathcal{C}_{\boldsymbol{\beta}^t}(\hat{\boldsymbol{\beta}})$.

## 7.3 Condition (1)

Condition (1) follows, for large enough $p$, from Lemma 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 7.1.** *For any parameter vector $\boldsymbol{\lambda} \in \mathbb{R}^p$ defining a SLOPE cost as in* (1.2)*, let $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\lambda})$, then for $t \geq 0$,*

$$\operatorname*{plim}_p \frac{1}{p}\|\boldsymbol{\beta}^t\|^2 = \operatorname*{plim}_p \frac{1}{p}\mathbb{E}_{\boldsymbol{Z}}[\|\eta_p^t(\boldsymbol{\beta}+\tau_t\boldsymbol{Z})\|^2] \leq 2\sigma_{\boldsymbol{\beta}}^2 + 2\tau_t^2, \tag{7.1}$$

*for $\eta_p^t(\cdot)$ defined in* (4.1) *with $\sigma_{\boldsymbol{\beta}}^2 := \mathbb{E}[B^2] < \infty$ and $\sigma_{\boldsymbol{\beta}}^2 + \tau_*^2 < \infty$ and*

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|^2 \leq \mathsf{C}, \tag{7.2}$$

*where* $\mathsf{C} := \mathsf{C}(\delta, \sigma_{\boldsymbol{\beta}}^2, \sigma_w^2, \mathsf{B}_{max}, \mathsf{B}_{min}, \lambda_{min})$ *is a positive constant depending on* $\delta, \sigma_{\boldsymbol{\beta}}^2, \sigma_w^2$*, 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 D. □

### 7.4 Condition (2)

Condition (2) follows from Lemma 7.2 stated below, for $\epsilon$ arbitrarily small when $t$ is large enough.

**Lemma 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* (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 [5, Lemma 3.3], though, the subgradient for the SLOPE cost function (studied extensively in Section 5.1) is quite different than that of the LASSO cost and our analysis requires handling this carefully. Before we prove Lemma 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 $\ell_2$ norm as the algorithm runs. This result is crucial to the proof of Lemma 7.2.

**Lemma 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 7.3.* This result uses Lemma 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 (6.3). Then Lemma E.1 (which is stated and proved in Appendix E) shows that $\lim_t (\tau_t^2 - 2\Sigma_{t,t-1} + \tau_{t-1}^2) = 0$. □

*Proof of Lemma 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 5.1. Moreover, by AMP (1.3b), $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t = \boldsymbol{z}^t - \omega^t \boldsymbol{z}^{t-1}$ with $\omega^t := \frac{1}{\delta p}[\nabla \eta^{t-1}(\boldsymbol{\beta}^{t-1} + \boldsymbol{X}^\top \boldsymbol{z}^{t-1})]$. Therefore we can write,

$$\begin{aligned} &\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) = \boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{z}^t - \omega^t \boldsymbol{z}^{t-1}) = \boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) - (1-\omega^t)\boldsymbol{X}^\top \boldsymbol{z}^{t-1} \\ &= (\boldsymbol{\nu}^t - \mu_t \boldsymbol{X}^\top \boldsymbol{z}^{t-1}) - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + (\mu_t - (1-\omega^t))\boldsymbol{X}^\top \boldsymbol{z}^{t-1}, \end{aligned} \tag{7.3}$$

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 (1.3a) and by Fact 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

$$\boldsymbol{\nu}^t = \mu_t(\boldsymbol{X}^\top \boldsymbol{z}^{t-1} + \boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t) \in \partial J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t), \tag{7.4}$$

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 (7.3),

$$\boldsymbol{\nu}^t - \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) = \mu_t(\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^t) - \boldsymbol{X}^\top(\boldsymbol{z}^t - \boldsymbol{z}^{t-1}) + (\mu_t - (1-\omega^t))\boldsymbol{X}^\top \boldsymbol{z}^{t-1}. \tag{7.5}$$

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

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

Using Lemma 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 (2.11) is finite, the first two terms on the right side of the above $\to 0$. Finally, for the third term, Lemma 6.3 gives $\lim_t \operatorname{plim}_p \|z^t\|/\sqrt{p} = \tau_*$, and together with the calibration formula (2.11), that $\sigma_{\max}(\boldsymbol{X})$ is almost surely bounded as $p \to \infty$, and the definition of $\omega$ in the proof of Lemma 2.2, we find $\lim_t \lim_p(\mu_t - (1-\omega^t)) = 0$, and thus the third term $\to 0$. 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. □

## 7.5 Condition (3)

We take $\boldsymbol{\nu}^t$ to be the subgradient defined in (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 7.4, stated below, whose proof is the main challenge. We state the proposition and then we identify two auxiliary lemmas, Lemma 7.5 and 7.6, that will be used to ultimately prove Proposition 7.4.

**Proposition 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 7.4 will use two auxiliary lemmas, Lemma 7.5 and 7.6, stated below.

**Lemma 7.5.** *Let the set $s_t$ be measurable on the $\sigma$-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 7.5 is given in Appendix F. The key difference in SLOPE case (Lemma 7.5) and LASSO case (cf. [5, 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 7.5. □

**Lemma 7.6.** *[5, 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*

$$\frac{1}{p}|S_{t_2}(\gamma) \setminus S_{t_1}(\gamma)| < \xi\,, \tag{7.6}$$

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

*Proof.* For LASSO, this result was given in [5, 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 [5] can appeal to AMP analysis in [4], for SLOPE, we appeal to similar results given in [8] (e.g. Lemma 6.1). □

*Proof of Proposition 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. (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 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 $c_2 < 0.4$.

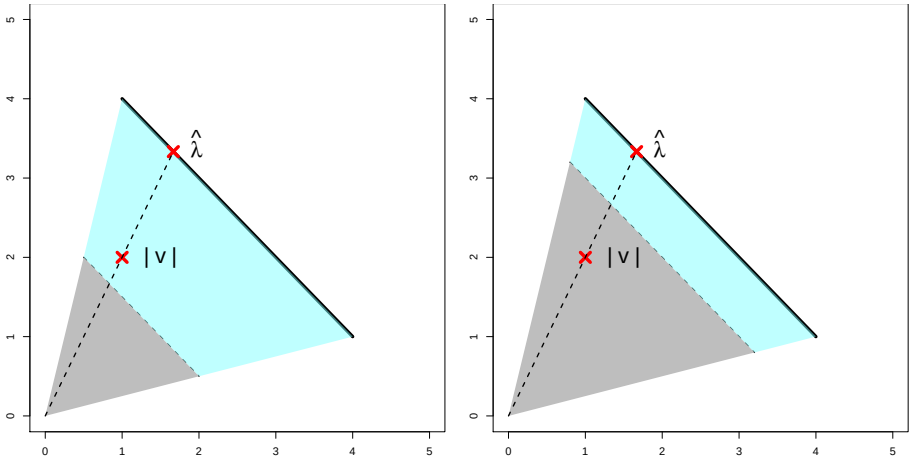


Figure 4: Left: $c_2 = 0.5$; Right: $c_2 = 0.2$; Blue area is $\{\boldsymbol{\nu} \in \partial J_{\boldsymbol{\lambda}}(0,0) : |\boldsymbol{\nu}| \succeq (1-c_2)\mathcal{P}(\lambda_1, \lambda_2)\}$ and grey area is complement of blue area in $\partial J_{\boldsymbol{\lambda}}(0,0)$.

We know from the proof of Lemma 7.2 Eq. (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 $I$,

$$\begin{aligned} |s_t(c_2)| &= \sum_I \mathbb{I}\{|\boldsymbol{\nu}_I^t| \succeq \mathcal{P}([\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda})]_I)(1-c_2)\} \\ &= \sum_I \mathbb{I}\Big\{|\boldsymbol{\beta}^t - [\boldsymbol{X}^\top \boldsymbol{z}^{t-1}] - \boldsymbol{\beta}^{t-1}|_I \succeq \mathcal{P}([\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\theta}^{t-1})]_I)(1-c_2)\Big\}. \end{aligned} \tag{7.7}$$

As detailed in the proof of Lemma 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 $\mathcal{P}(\hat{\Pi}_{\boldsymbol{\beta}^t}^{-1}(\boldsymbol{\lambda}))$ follows since $\boldsymbol{\nu}$ is a valid subgradient of $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^t)$ (see Fact 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 (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 (F.1), $\boldsymbol{\beta}^t - [\boldsymbol{X}^\top \boldsymbol{z}^{t-1}] - \boldsymbol{\beta}^{t-1} = \boldsymbol{h}^t + \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t) - \boldsymbol{\beta}$ and $\boldsymbol{\beta}^t = \eta^{t-1}(\boldsymbol{\beta} - \boldsymbol{h}^t)$ and therefore by (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 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 $\tau_{t-1}$ given by the state evolution (2.4). Therefore a careful argument using continuous approximations to indicators gives,

$$\begin{aligned}&\operatorname*{plim}_p \frac{1}{p}\sum_I \mathbb{I}\Big\{|\boldsymbol{h}^t + \eta^{t-1}(\boldsymbol{\beta}-\boldsymbol{h}^t)-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big\}\\ &= \lim_p \mathop{\mathbb{E}}_{\boldsymbol{Z}_{t-1}}\Big\{\frac{1}{p}\sum_I \mathbb{I}\Big\{|\tau_{t-1}\boldsymbol{Z}_{t-1}+\eta^{t-1}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big\}\Big\},\end{aligned} \tag{7.8}$$

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 $\eta^{t-1}$ that is non-flat, to which we apply the Portmanteau Theorem (cf. [20], Lemma 1(b) in [4] and Lemma F.3(b) in [5]).

Now, using (7.8), we can simplify:

$$\operatorname*{plim}_p \frac{1}{p}|s_t(c_2)| = \lim_p \frac{1}{p}\sum_I \mathop{\mathbb{P}}_{\boldsymbol{Z}_{t-1}}\Big(|\tau_{t-1}\boldsymbol{Z}_{t-1}-\eta^{t-1}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big), \tag{7.9}$$

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 $\eta^{t-1}$, dropping the input.

$$\begin{aligned}&\mathbb{P}\Big(|\tau_{t-1}\boldsymbol{Z}_{t-1}+\eta^{t-1}-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big)\\ &= \mathbb{P}\Big(|\tau_{t-1}\boldsymbol{Z}_{t-1}+\eta^{t-1}-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2), \eta_1^{t-1}=\mathbf{0}\Big)\\ &\quad + \mathbb{P}\Big(|\tau_{t-1}\boldsymbol{Z}_{t-1}+\eta^{t-1}-\boldsymbol{\beta}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2), \eta_I^{t-1}\neq\mathbf{0}\Big)\\ &\overset{(a)}{=} \mathbb{P}\Big(\hat{\boldsymbol{\theta}}_I^{t-1} \geq |\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big) + \mathbb{P}\Big(\hat{\boldsymbol{\theta}}_I^{t-1} \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big)\,\mathbb{P}(\eta_I^{t-1}\neq\mathbf{0}).\\ &= \mathbb{P}\Big(\hat{\boldsymbol{\theta}}_I^{t-1} \geq |\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1}|_I \geq \hat{\boldsymbol{\theta}}_I^{t-1}(1-c_2)\Big) + \mathbb{P}(\eta_I^{t-1}\neq 0).\end{aligned} \tag{7.10}$$

In the above, step $(a)$ follows when $\eta_I^{t-1} = [\operatorname{prox}_{J_{\boldsymbol{\theta}^{t-1}}}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})]_I = \mathbf{0}$, 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 5.2 and Fact 5.3, we know that $|\eta^{t-1}(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})-(\boldsymbol{\beta}-\tau_{t-1}\boldsymbol{Z}_{t-1})|_I \in \mathcal{P}([\hat{\Pi}^{-1}_{\eta^{t-1}}(\boldsymbol{\theta}^{t-1})]_I)$.

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 $\mathcal{P}([\hat{\Pi}^{-1}_{\eta^{t-1}}(\boldsymbol{\theta}^{t-1})]_I)$ 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

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

and that $\eta^{t-1}$ is asymptotically separable by Lemma 3.3. Define $h^{t-1}(x) = h(x; B+\tau_{t-1}Z, \Theta^{t-1})$ with $\Theta^{t-1}$ being the distribution to which the empirical distribution of $\boldsymbol{\theta}^{t-1}$ converges, and also define

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

similarly to (2.12), where $m$ is the Lebesgue measure. Then,

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

where the last equality holds by Lemma 3.2.

Then (2.10) gives this term is smaller than $\delta$ for large $t$. Hence, by (7.9) and (7.10),

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

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 $\delta$, 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 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 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 7.6 and let $a_2 = a_2(c, t_{\min})$. Then, by Lemma 7.5 and the fact that $a_2(c,t)$ is non-increasing in $t$,

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

In addition, by Lemma 7.6, $|S_t(c_2) \setminus S_{t_{\min}}(c_2)| \leq p a_1/2$. 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 $t$. $\square$

# 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 [8], 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 $\ell_2$ 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 [20] 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 $n$.

## References


[1] Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific, 2008.

[2] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[3] M. Bayati, M. A. Erdogdu, and A. Montanari. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pages 944–952, 2013.

[4] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inf. Theory*, 57(2):764–785, 2011.

[5] M. Bayati and A. Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[7] P. C. Bellec, G. Lecué, and A. B. Tsybakov. SLOPE meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[8] R. Berthier, A. Montanari, and P.-M. Nguyen. State evolution for approximate message passing with non-separable functions. *arXiv preprint arXiv:1708.03950*, 2017.

[9] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103, 2015.

[10] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[11] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan. Group SLOPE—adaptive selection of groups of predictors. *Journal of the American Statistical Association*, pages 1–15, 2018.

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

[13] A. Chambolle, R. A. De Vore, N.-Y. Lee, and B. J. Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998.

[14] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[15] D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[16] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[17] D. L. Donoho, A. Maleki, and A. Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[18] J. L. Doob. *Stochastic processes*, volume 101. New York Wiley, 1953.

[19] M. Figueiredo and R. Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Artificial Intelligence and Statistics*, pages 930–938, 2016.

[20] H. Hu and Y. M. Lu. Asymptotics and optimal designs of SLOPE for sparse linear regression. *arXiv preprint arXiv:1903.11582*, 2019.

[21] A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[22] B. S. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 41(2):334–351, 1977.

[23] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mech. Theory Exp.*, (8), 2012.

[24] M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[25] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195(2):491–523, 2005.

[26] F. J. MacWilliams and N. J. A. Sloane. *The theory of error-correcting codes*, volume 16. Elsevier, 1977.

[27] A. Montanari. Graphical models concepts in compressed sensing. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 394–438. Cambridge University Press, 2012.

[28] A. Mousavi, A. Maleki, R. G. Baraniuk, et al. Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148, 2018.

[29] N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[30] L. D. Pitt. Positively correlated normal variables are associated. *The Annals of Probability*, pages 496–499, 1982.

[31] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 2168–2172, 2011.

[32] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[33] H. L. Royden. *Real analysis*. Krishna Prakashan Media, 1968.

[34] W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[35] C. Rush and R. Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Trans. on Inf. Theory*, 64(11):7264–7286, 2018.

[36] W. Su, M. Bogdan, and E. Candès. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150, 2017.

[37] W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[38] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[39] X. Zeng and M. A. Figueiredo. Decreasing weighted sorted $\ell_1$ regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, 2014.

# A State Evolution Analysis

We first prove Theorem 1 and then provide a proof of Proposition 2.6.

## A.1 Proving Theorem 1

*Proof of Theorem 1.* To begin with, we prove that $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau)$ defined in (2.8) is concave with respect to $\tau^2$. The proof follows along the same lines as the proof of [5, Proposition 1.3], however, whereas the proof of [5, Proposition 1.3] proceeds by explicitly expressing the first derivative of the corresponding function $\mathsf{F}$, 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,

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

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

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

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 $\mathsf{I}$ indicate the collection of unique equivalence classes.

**Lemma A.1.**

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

Lemma A.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 A.1 can be used to produce the dominated convergence result needed in step $(a)$ of (A.1). First note,

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

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

Now we want to further simplify (A.1). For each $1 \le i \le p$, 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,

$$\frac{\partial}{\partial \tau^2}[\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i = \boldsymbol{J}_{f\circ g}(\tau^2) \overset{(a)}{=} \boldsymbol{J}_f(g(\tau^2))\boldsymbol{J}_g(\tau^2) = \Big[\nabla_{\boldsymbol{a}} f(g(\tau^2)), \nabla_{\boldsymbol{b}} f(g(\tau^2))\Big]\Big[\frac{\boldsymbol{Z}}{2\tau}, \frac{\boldsymbol{\alpha}}{2\tau}\Big]^\top, \tag{A.3}$$

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 $\eta(\boldsymbol{a}, \boldsymbol{b}) := \mathrm{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})$ and consider the partial derivatives of $\eta$ with respect to its first and second arguments. Denote

$$\partial_1\eta(\boldsymbol{a}, \boldsymbol{b}) := \mathrm{diag}\Big[\frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}, \ldots, \frac{\partial}{\partial a_p}\Big]\eta(\boldsymbol{a}, \boldsymbol{b}), \quad \text{and} \quad \partial_2\eta(\boldsymbol{a}, \boldsymbol{b}) := \mathrm{diag}\Big[\frac{\partial}{\partial b_1}, \frac{\partial}{\partial b_2}, \ldots, \frac{\partial}{\partial b_p}\Big]\eta(\boldsymbol{a}, \boldsymbol{b}). \tag{A.4}$$

Recall that the derivatives computed in $\partial_1\eta(\boldsymbol{a}, \boldsymbol{b})$ are defined in (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):

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

we have

$$\frac{d}{da_j}f(\boldsymbol{a}, \boldsymbol{b}) = \frac{d}{da_j}[\eta(\boldsymbol{a}, \boldsymbol{b})]_i = \mathbb{I}\{|[\eta(\boldsymbol{a}, \boldsymbol{b})]_i| = |[\eta(\boldsymbol{a}, \boldsymbol{b})]_j|\}\operatorname{sign}([\eta(\boldsymbol{a}, \boldsymbol{b})]_i[\eta(\boldsymbol{a}, \boldsymbol{b})]_j)[\partial_1\eta(\boldsymbol{a}, \boldsymbol{b})]_i, \tag{A.5}$$

and similarly,

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

Now plugging the above into (A.3), we have

$$\begin{aligned}&\frac{\partial}{\partial \tau^2}[\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau\boldsymbol{Z})]_i \\ &= \frac{1}{2\tau}\big[\partial_1\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)\big]_i \operatorname{sign}\big([\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)]_i\big)\sum_{j\in I_i}\big(\operatorname{sign}([\eta(\boldsymbol{B} + \tau\boldsymbol{Z}, \boldsymbol{\alpha}\tau)]_j)Z_j - \alpha_j\big)\end{aligned} \tag{A.6}$$

In what follows, we drop the explicit statement of the $\eta(\cdot, \cdot)$ input to save space, writing $\eta_i$ 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 (A.6) in (A.1),

$$\begin{aligned}\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) &= \frac{1}{\delta p\tau}\sum_{i=1}^p\sum_{j\in I_i}\mathbb{E}\Big\{(\eta_i - B_i)\,[\partial_1\eta]_i \operatorname{sign}(\eta_i)(\operatorname{sign}(\eta_j)Z_j - \alpha_j)\Big\} \\ &= \frac{1}{\delta p}\sum_{i=1}^p\sum_{j\in I_i}\mathbb{E}\Big\{([\partial_1\eta]_i)^2 + (\eta_i - B_i)[\partial_1^2\eta]_i\Big\} - \frac{1}{\delta p\tau}\sum_{i=1}^p\sum_{j\in I_i}\mathbb{E}\Big\{(\eta_i - B_i)\,[\partial_1\eta]_i \operatorname{sign}(\eta_i)\alpha_j\Big\}.\end{aligned} \tag{A.7}$$

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,

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

where the last step uses the definition of $\frac{d}{da_j}[\eta(\boldsymbol{a},\boldsymbol{b})]_i$ given in (A.5) and the fact that $\frac{d}{da_j}[\partial_1\eta(\boldsymbol{a},\boldsymbol{b})]_i = \operatorname{sign}(\eta_i)\operatorname{sign}(\eta_j)[\partial_1^2\eta(\boldsymbol{a},\boldsymbol{b})]_i$.

Therefore, simplifying (A.7), we have shown

$$(\delta p\tau)\times\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)=\sum_{i=1}^{p}\mathbb{E}\Big\{\tau|I_i|\Big([\partial_1\eta]_i^2+(\eta_i-B_i)[\partial_1^2\eta]_i\Big)-[\partial_1\eta]_i\operatorname{sign}(\eta_i)(\eta_i-B_i)\sum_{j\in I_i}\alpha_j\Big\}. \tag{A.8}$$

We now have the tools to prove Lemma A.1.

*Proof of Lemma A.1.* First,

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

As in the work above, we denote the proximal operator using a function $\eta:\mathbb{R}^{2p}\to\mathbb{R}^p$ as $\eta(\boldsymbol{a},\boldsymbol{b}):=\mathrm{prox}_{J_{\boldsymbol{b}}}(\boldsymbol{a})$. Now from (A.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,

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

Therefore,

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

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

$$\begin{aligned}\Big|\frac{\partial}{\partial\tau^2}\frac{1}{p}\|\mathrm{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B}+\tau\boldsymbol{Z})-\boldsymbol{B}\|^2\Big|&\le\frac{1}{p}\Big|\sum_{i=1}^{p}\sum_{j\in I_i}(Z_i-\operatorname{sign}(\eta_i)\alpha_i)\,[\partial_1\eta]_i\operatorname{sign}(\eta_i)(\operatorname{sign}(\eta_j)Z_j-\alpha_j)\Big|\\&=\frac{1}{p}\Big|\sum_{i=1}^{p}\sum_{j\in I_i}(\operatorname{sign}(\eta_i)Z_i-\alpha_i)(\operatorname{sign}(\eta_j)Z_j-\alpha_j)\,[\partial_1\eta]_i\Big|.\end{aligned} \tag{A.9}$$

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

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

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

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

Now plugging back into (A.9),

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

□

Now considering (A.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 $\tau$. 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 (A.8) becomes

$$(\delta p\tau)\times\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)=\sum_{i=1}^{p}\Big[\mathbb{E}\{\tau([\partial_1\eta]_i)^2+\tau(\eta_i-B_i)[\partial_1^2\eta]_i-\alpha_i\,\mathrm{sign}(\eta_i)(\eta_i-B_i)[\partial_1\eta]_i\}\Big]. \quad \text{(A.10)}$$

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 $\eta$ 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

$$[\partial_1^2\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i=\delta(B_i+\tau Z_i-\alpha_i\tau)-\delta(B_i+\tau Z_i+\alpha_i\tau) \quad \text{(A.11)}$$

and

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{Z},\boldsymbol{B}}\Big\{([\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i-B_i)[\partial_1^2\eta(\boldsymbol{B}+\tau\boldsymbol{Z},\boldsymbol{\alpha}\tau)]_i\Big\} \\ &=-\mathop{\mathbb{E}}_{\boldsymbol{B}}\mathop{\mathbb{E}}_{\boldsymbol{Z}|\boldsymbol{B}}\big\{B_i\big[\delta(B_i+\tau Z_i-\alpha_i\tau)-\delta(B_i+\tau Z_i+\alpha_i\tau)\big]\big\} \\ &=-\frac{1}{\tau}\mathop{\mathbb{E}}_{\boldsymbol{B}}\big\{B_i\big[\phi(\alpha_i-\frac{1}{\tau}B_i)-\phi(-\alpha_i-\frac{1}{\tau}B_i)\big]\big\}.\end{aligned} \quad \text{(A.12)}$$

Therefore, denoting $\odot$ as elementwise multiplication of vectors, by (A.10) and (A.12),

$$\begin{aligned}&(\delta p\tau)\times\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau) \\ =&\tau\mathbb{E}||\partial_1\eta||^2-\mathop{\mathbb{E}}_{\boldsymbol{B}}\big\{\boldsymbol{B}^\top\big[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})\big]\big\}-\mathbb{E}\big\{\big[\boldsymbol{\alpha}\odot\mathrm{sign}(\eta)\odot(\eta-\boldsymbol{B})\big]^\top\partial_1\eta\big\}.\end{aligned} \quad \text{(A.13)}$$

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 $\frac{\partial}{\partial\tau}[\frac{\partial\mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)]\le 0$ 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,

$$\begin{aligned}(\delta p)\times\frac{\partial}{\partial\tau}\Big[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)\Big] = \frac{\partial}{\partial\tau}\mathbb{E}||\partial_1\eta||^2 - \frac{\partial}{\partial\tau}\frac{1}{\tau}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\}\\ -\frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}\{\big[\boldsymbol{\alpha}\odot\operatorname{sign}(\eta)\odot(\eta-\boldsymbol{B})\big]^\top\partial_1\eta\}.\end{aligned} \tag{A.14}$$

To show that (A.14) is $\leq 0$, 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 (A.14) can be simplified to the following:

$$\frac{\partial}{\partial\tau}\mathbb{E}||\partial_1\eta||^2 = -\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B}))-\phi(\boldsymbol{\alpha}+\frac{1}{\tau}\boldsymbol{B}))]\}. \tag{A.15}$$

Doing so requires smart uses of the chain rule, a dominated convergence argument, the partials in (A.6), and special care for the 'kink' points as discussed above. Similarly, using (A.12), one can easily show for the third term on the right side of (A.14),

$$\frac{\partial}{\partial\tau}\frac{1}{\tau}\mathbb{E}\{\big[\boldsymbol{\alpha}\odot\operatorname{sign}(\eta)\odot(\eta-\boldsymbol{B})\big]^\top\partial_1\eta\} \geq \frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{[\boldsymbol{\alpha}\odot\boldsymbol{B}^2]^\top[\phi(\boldsymbol{\alpha}+\frac{1}{\tau}\boldsymbol{B})+\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\Big\}. \tag{A.16}$$

Finally, using $\phi'(u) = -u\phi(u)$ and a dominated convergence argument, the second term on the right side of (A.14) equals

$$\begin{aligned}&-\frac{\partial}{\partial\tau}\frac{1}{\tau}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\}\\ &\quad=\frac{1}{\tau^2}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\}\\ &\qquad-\frac{1}{\tau^3}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{(\boldsymbol{B}^2)^\top[(\frac{1}{\tau}\boldsymbol{B}-\boldsymbol{\alpha})\odot\phi(\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})-(\boldsymbol{\alpha}+\frac{1}{\tau}\boldsymbol{B})\odot\phi(-\boldsymbol{\alpha}-\frac{1}{\tau}\boldsymbol{B})]\}.\end{aligned} \tag{A.17}$$

Now we plug (A.15),(A.16), and (A.17) back into (A.14) to show that $\frac{\partial}{\partial\tau}[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)] \leq 0$.

$$\begin{aligned}&(\delta p)\times\frac{\partial}{\partial\tau}\Big[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)\Big]\\ &\leq -\frac{1}{\tau^2}\mathbb{E}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau))-\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau))]\}+\frac{1}{\tau^2}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{\boldsymbol{B}^\top[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-\phi(-\boldsymbol{\alpha}-\boldsymbol{B}/\tau)]\}\\ &\quad-\frac{1}{\tau^3}\mathop{\mathbb{E}}_{\boldsymbol{B}}\{(\boldsymbol{B}^2)^\top[(\boldsymbol{B}/\tau-\boldsymbol{\alpha})\odot\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)\odot\phi(-\boldsymbol{\alpha}-\boldsymbol{B}/\tau)]\}\\ &\quad-\frac{1}{\tau^3}\mathbb{E}_{\boldsymbol{B}}\Big\{[\boldsymbol{\alpha}\odot\boldsymbol{B}^2]^\top[\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)+\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)]\Big\}\\ &=-\frac{1}{\tau^4}\mathbb{E}_{\boldsymbol{B}}\{[\boldsymbol{B}^3]^\top[\phi(\boldsymbol{\alpha}-\boldsymbol{B}/\tau)-\phi(\boldsymbol{\alpha}+\boldsymbol{B}/\tau)]\}.\end{aligned} \tag{A.18}$$

We justify non-positivity of (A.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 (A.18), implies $\frac{\partial}{\partial\tau}[\frac{\partial \mathsf{F}}{\partial\tau^2}(\tau^2,\boldsymbol{\alpha}\tau)] \leq 0$ and therefore, we have shown that $\mathsf{F}(\tau^2,\boldsymbol{\alpha}\tau)$ defined in (2.8), is concave with respect to $\tau^2$.

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 (A.8). In particular,

$$\delta \frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\Big\{ |I_i| \Big( [\partial_1 \eta]_i^2 + (\eta_i - B_i)[\partial_1^2 \eta]_i \Big) - \frac{1}{\tau}[\partial_1 \eta]_i \operatorname{sign}(\eta_i)(\eta_i - B_i) \sum_{j \in I_i} \alpha_j \Big\}, \quad \text{(A.19)}$$

Then taking $\tau \to \infty$ in the above, it is easy to see that $f(\boldsymbol{\alpha})$ is equivalent to setting $\boldsymbol{B} = \boldsymbol{0}$ 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 (A.11)-(A.12), means that this term will go to zero as $\tau \to \infty$. Therefore, using $\operatorname{sign}(\eta(\boldsymbol{Z}, \boldsymbol{\alpha})) \odot \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = |\eta(\boldsymbol{Z}, \boldsymbol{\alpha})|$,

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

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 $\infty$ otherwise. Using that $\partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \frac{1}{\boldsymbol{D}(\eta(\boldsymbol{Z}, \boldsymbol{\alpha}))}$,

$$f(\boldsymbol{\alpha}) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{E}\left\{ \Big( 1 - |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i| \sum_{j : |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_j| = |[\eta(\boldsymbol{Z}, \boldsymbol{\alpha})]_i|} \alpha_j \Big) \frac{1}{[D(\eta(\boldsymbol{Z}, \boldsymbol{\alpha}))]_i} \right\} \quad \text{(A.20)}$$

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

Now considering (A.20), let $\boldsymbol{\alpha} \to \infty$ and note that since $|\boldsymbol{Z}| < \boldsymbol{\alpha}$ almost surely as $\boldsymbol{\alpha} \to \infty$, it follows that $\eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \partial_1 \eta(\boldsymbol{Z}, \boldsymbol{\alpha}) = \boldsymbol{0}$. 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 $\mathsf{F}$ is now obvious: since $\mathsf{F}$ is concave (implying $\frac{\partial \mathsf{F}}{\partial \tau^2}(\tau^2, \boldsymbol{\alpha}\tau)$ is non-increasing) and strictly increasing for $\tau^2$ large enough, it is increasing everywhere. Moreover, the monotonicity of $\mathsf{F}$ implies the monotonicity of the sequence $\{\tau_t^2(p)\}_{t \geq 0}$.

Finally we show that there exists a unique $\tau_*$ 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 (A.20), we know $f(\boldsymbol{0}) = \mathbb{E}\|\partial_1 \eta(\tau \boldsymbol{Z}, \boldsymbol{0})\|^2/p = \mathbb{E}\|\boldsymbol{1}\|^2/p = 1$. 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 $\tau^2$ and $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) < \tau^2$ for large enough $\tau^2$. Therefore, there is at least one $\tau_*$ such that $\mathsf{F}(\tau_*^2, \boldsymbol{\alpha}\tau_*) = \tau_*^2$ since $\mathsf{F}$ is continuous in $\tau$. It follows from the concavity of $\mathsf{F}$ 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 $\tau^2$. 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 (2.8), namely, $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) = \sigma_w^2 + \mathbb{E}\|\operatorname{prox}_{J_{\boldsymbol{\alpha}\tau}}(\boldsymbol{B} + \tau \boldsymbol{Z}) - \boldsymbol{B}\|^2/(\delta p)$, clearly $\mathsf{F}(0, \boldsymbol{0}) = \sigma_w^2 \geq 0$ 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 $\tau^2$. 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 $\tau^2$ grows) for any $\boldsymbol{\alpha} \succeq \boldsymbol{A}_{\min}$. Also, $G(\tau^2)$ is concave and therefore for $\tau^2$ large enough we will have $G(\tau^2) < 0$, in which case $\mathsf{F}(\tau^2, \boldsymbol{\alpha}\tau) < \tau^2$.

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

## A.2 Proving Proposition 2.6

*Proof of Proposition 2.6.* This proof is a generalized result of [5, Proposition 1.4] (originally proved in [17]) and [5, Corollary 1.7]. Here we fixed $p$ and denote $\tau(p)$ as $\tau$.

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 $\tau^2$ 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 $\tau$,

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

Recall $\mathsf{F}(0, \boldsymbol{0}) = \sigma_w^2$ 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 $\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)$,

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

We claim that $\ell_* < 0$. Using the definition of the vector $\boldsymbol{D}$ and the set $\boldsymbol{A}_{\min}(\delta)$ in (2.7),

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

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

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

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

$$\lim_{\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = \ell_* \cdot \lim_{\boldsymbol{\alpha} \downarrow \boldsymbol{A}_{\min}(\delta)} \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha}) = -\infty\,. \tag{A.21}$$

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

$$\lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\lambda}(\boldsymbol{\alpha}) = 1 \cdot \lim_{\boldsymbol{\alpha}\to\infty} \boldsymbol{\alpha}\tau_*(\boldsymbol{\alpha}) = \infty\,. \tag{A.22}$$

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 (A.21) and (A.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 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 $\tau_*$ 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.

□

# B Verifying Properties (P1) and (P2)

In this appendix we demonstrate that the properties **(P1)** and **(P2)** given in Section 4 and relating to the denoiser $\eta_p^t(\cdot)$ defined in (4.1) are true.

*Verifying Properties **(P1)** and **(P2)**.* Property **(P1)** follows since $\eta_p^t(\cdot) = \operatorname{prox}_{J_{\boldsymbol{\alpha}\tau_t}}(\cdot)$, as it is easy to show that proximal operators are Lipschitz continuous with Lipschitz constant one. Namely

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

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 $2 \times 2$ covariance matrix, the following limits exist and are finite.

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

We first note that the first limit in (B.1) exists by Assumption **(A2)** and the strong law of large numbers. We focus on the other two limits. These results follow by [20, Proposition 1] given in Lemma 3.3 and the following lemma, which is a classic result in probability theory.

**Lemma B.1** (Doob's $L^1$ maximal inequality, [18] 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 (B.1) exist and are finite, we state one more result that will be helpful in the proof. This result uses Lemma B.1 along with a Dominated Convergence argument to study expectations taken with respect to $(\boldsymbol{Z}, \boldsymbol{Z}')$ like those in (B.1).

**Lemma B.2.** *Consider a function $\psi_p : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that for iterations $s, t \geq 0$,*

$$\frac{1}{p}\Big|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}')) - \psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big| \to 0, \quad \text{as} \quad p \to \infty, \tag{B.2}$$

*where $h^s, h^t$ are the unspecified functions of Lemma 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 $p$,*

$$\frac{1}{p}\Big|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}')) - \psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big| \leq L\Big(1 + \frac{\|\boldsymbol{\beta}\|^2}{p} + \frac{\|\boldsymbol{Z}\|^2}{p} + \frac{\|\boldsymbol{Z}'\|^2}{p}\Big). \tag{B.3}$$

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

$$\frac{1}{p}\Big|\mathop{\mathbb{E}}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}))\Big\} - \mathop{\mathbb{E}}_{\boldsymbol{Z},\boldsymbol{Z}'}\Big\{\psi_p(\boldsymbol{\beta}, h^s(\boldsymbol{\beta}+\boldsymbol{Z}), h^t(\boldsymbol{\beta}+\boldsymbol{Z}'))\Big\}\Big| \to 0. \tag{B.4}$$

*Proof.* We begin by showing that $\mathbb{E}_{\boldsymbol{Z},\boldsymbol{Z}'}\{\sup_{p\geq 1}\frac{1}{p}|\psi_p(\boldsymbol{\beta}, \eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}), \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}'))|\} < \infty$. Using (B.3), it is clear that this expectation is finite almost surely if

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

where we have made the dependence of the vectors on the dimension $p$ explicit. But Lemma B.1 immediately implies the above since $\mathbb{E}[B^2 \max\{0, \log B\}] < \infty$ by assumption **(A2)**.

Now by dominated convergence we have,

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

Then the above implies the desired result (B.4) from assumption (B.2). □

First consider the second limit in (B.1). By Cauchy-Schwarz, (3.3) of Lemma 3.3 implies that $|\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - \boldsymbol{\beta}^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})|/p \to 0$, as $p \to \infty$. This follows because

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

Then the right side of the above $\to 0$ with growing $p$ because $\|\boldsymbol{\beta}\|/\sqrt{p}$ limits to a constant as justified above (this is the limit in (B.1)), and the other term $\to 0$ by (3.3) of Lemma 3.3. This means that assumption (B.2) of Lemma B.2 is satisfied. Assumption (B.3) of Lemma B.2 is also satisfied since both $\eta_p^t$ and $h^t$ are Lipschitz(1), by Cauchy-Schwarz inequality. Therefore Lemma B.2 implies $|\,\mathbb{E}_{\boldsymbol{Z}}\{\boldsymbol{\beta}^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z})\} - \mathbb{E}_{\boldsymbol{Z}}\{\boldsymbol{\beta}^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})\}|/p \to 0$, 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 $\mathbb{E}[h^t(B+Z)^2] < \mathbb{E}[(B+Z)^2] \le \mathbb{E}[B^2] + \mathbb{E}[Z^2] = \sigma_{\boldsymbol{\beta}}^2 + \Sigma_{11} < \infty$.

Finally consider the third limit in (B.1). Similarly to the work in studying the second limit in (B.1), we will appeal to Lemma B.2. First we will show that

$$|\eta_p^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top \eta_p^t(\boldsymbol{\beta}+\boldsymbol{Z}) - h^s(\boldsymbol{\beta}+\boldsymbol{Z}')^\top h^t(\boldsymbol{\beta}+\boldsymbol{Z})|/p \to 0, \quad \text{as} \quad p \to \infty, \tag{B.5}$$

meaning that assumption (B.2) of Lemma B.2 is satisfied. Then, again, assumption (B.3) of Lemma B.2 is satisfied since both $\eta_p^t(\cdot)$ and $h^t(\cdot)$ are Lipschitz(1), using Cauchy-Schwarz.

Now we want to prove (B.5). By repeated applications of Cauchy-Schwarz it is not hard to show,

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

Now, (B.5) follows since the right side of the above goes to 0 as $p$ grows. This follows since, by (3.3) of Lemma 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 $\mathbb{E}[(h^s(B+Z'))^2] \le \mathbb{E}[(B+Z')^2] \le \sigma_{\boldsymbol{\beta}}^2 + \Sigma_{22} < \infty$ and $\mathbb{E}[(h^t(B+Z))^2] \le \mathbb{E}[(B+Z)^2] \le \sigma_{\boldsymbol{\beta}}^2 + \Sigma_{11} < \infty$. This proves (B.5) and therefore we can apply Lemma B.2.

Then Lemma B.2 implies,

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

But now, using the above, we find that

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

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

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

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

□

# C Proof of Fact 2.7

*Proof.* The fact follows from the asymptotic separability of the proximal operator [20, Proposition 1] (restated in Lemma 3.3) and the dominated convergence theorem [33] allowing for interchange of limit and expectation. We sketch the proof of the existence of the limit in (2.4) (and the result for the limit in (2.11) follows similarly). By Lemma 3.3, the weak convergence of $\boldsymbol{\alpha}(p)$ to $A$, and the Weak Law of Large Numbers, one can argue that

$$\lim_p \|\mathrm{prox}_{J_{\boldsymbol{\alpha}(p)\tau_*}}(\boldsymbol{B}+\tau_*\boldsymbol{Z})-\boldsymbol{B}\|^2/(\delta p) = \mathbb{E}\{(h(B+\tau_*Z)-B)^2\}/\delta, \tag{C.1}$$

where $h(\cdot) := h(\cdot; B+\tau_*Z, A\tau_*)$ is the unspecified, separable function of Lemma 3.3. This is consistent with [Lemma 29, [20]]. The limit in (2.4) exists if $\mathbb{E}\{(h(B+\tau_*Z)-B)^2\}/\delta<\infty$ and

$$\begin{aligned}&\mathbb{E}\{(h(B+\tau_*Z)-B)^2\} \le 2\mathbb{E}\{h(B+\tau_*Z)^2+B^2\} \le 2\mathbb{E}\{(B+\tau_*Z)^2+B^2\}\\ &\le 2\mathbb{E}\{2B^2+2\tau_*^2Z^2+B^2\} = 6\mathbb{E}\{B^2\}+4\tau_*^2<\infty.\end{aligned}$$

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

# D Proof of Lemma 7.1

*Proof.* First, the proof of (7.1) follows from Theorem 4.1. To see this, note that by (1.3a), we have $\boldsymbol{\beta}^{t+1} = \mathrm{prox}_{J_{\boldsymbol{\theta}_t}}(\boldsymbol{X}^\top\boldsymbol{z}^t+\boldsymbol{\beta}^t) = \eta_p^t(\boldsymbol{X}^\top\boldsymbol{z}^t+\boldsymbol{\beta}^t)$, and therefore we apply Theorem 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

$$\operatorname*{plim}_p \|\boldsymbol{\beta}^t\|^2/p \overset{p}{=} \operatorname*{plim}_p \mathbb{E}_{\boldsymbol{Z}}[\|\eta_p^t(\boldsymbol{\beta}+\tau_t\boldsymbol{Z})\|^2]/p, \tag{D.1}$$

for $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbb{I}_p)$. By the Lipschitz property of $\eta_p^t$ (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 (D.1), we find $\operatorname{plim}_p \|\boldsymbol{\beta}^t\|^2/p \stackrel{p}{=} 2 \operatorname{plim}_p \|\boldsymbol{\beta}\|^2/p + 2\tau_t^2 = 2\sigma_{\boldsymbol{\beta}}^2 + 2\tau_t^2$, where the final inequality follows by Assumption **(A2)**.

Now consider the $\widehat{\boldsymbol{\beta}}$ result in (7.2). First, note that by definition $\mathcal{C}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{C}(\mathbf{0})$ where the cost function $\mathcal{C}(\cdot)$ is defined in (1.2). Using that

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

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, [8, Lemma F.2]). Therefore,

$$\operatorname*{plim}_p \mathcal{C}(\widehat{\boldsymbol{\beta}})/p \leq \operatorname*{plim}_p \sigma_{\max}^2(\boldsymbol{X})\|\boldsymbol{\beta}\|^2/p + \operatorname*{plim}_p \|\boldsymbol{w}\|^2/p \leq \mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2 + \sigma_w^2, \tag{D.3}$$

where we've defined $\mathsf{B}_{max}$ 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

$$\|\widehat{\boldsymbol{\beta}}^{\|}\|_2^2 \leq \nu_1 \|\widehat{\boldsymbol{\beta}}^{\|}\|_1^2/p. \tag{D.4}$$

Then we have the following bound

$$\|\widehat{\boldsymbol{\beta}}\|^2 = \|\widehat{\boldsymbol{\beta}}^{\|}\|^2 + \|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \stackrel{(a)}{\leq} \nu_1\|\widehat{\boldsymbol{\beta}}^{\|}\|_1^2/p + \|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \stackrel{(b)}{\leq} 2\nu_1\|\widehat{\boldsymbol{\beta}}\|_1^2/p + (2\nu_1+1)\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2, \tag{D.5}$$

where step $(a)$ holds by (D.4) 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 (D.5). 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,

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

Therefore, using (D.2) and (D.3), we have

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}^{\perp}\|^2 \leq \operatorname*{plim}_p \frac{\frac{2}{p}\mathcal{C}(\mathbf{0})}{\hat{\sigma}_{min}^2(\boldsymbol{X})} \leq \frac{2(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2 + \sigma_w^2)}{\mathsf{B}_{min}}. \tag{D.6}$$

where we've defined $\mathsf{B}_{min}$ 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 (D.5). Recall the definition of the sort-ed $\ell_1$ 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 (D.2) and (D.3), we see

$$\operatorname*{plim}_p \frac{1}{p}\|\widehat{\boldsymbol{\beta}}\|_1 \le \operatorname*{plim}_p \frac{1}{\lambda_{min}}\Big(\frac{1}{p}\mathcal{C}(\mathbf{0})\Big) \le \frac{1}{\lambda_{min}}(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2+\sigma_w^2). \tag{D.7}$$

By (D.7), along with the upper bound in (D.5), we have

$$\operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}\|^2}{p} \le 2\nu_1 \operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}\|_1^2}{p^2}+(2\nu_1+1)\operatorname*{plim}_p \frac{\|\widehat{\boldsymbol{\beta}}^\perp\|^2}{p} \le \Big[\frac{2\nu_1(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2+\sigma_w^2)}{\lambda_{min}}\Big]^2+\frac{2(2\nu_1+1)(\mathsf{B}_{max}\sigma_{\boldsymbol{\beta}}^2+\sigma_w^2)}{\mathsf{B}_{min}}.$$

□

# E Proof of Lemma 7.3

The proof of Lemma 7.3 relies on the following result, Lemma E.1, about the exponential rate of the convergence of the state evolution sequence defined in (6.3). We state and prove Lemma E.1, and Lemma 7.3 is proved afterward.

**Lemma E.1.** *Assume $\boldsymbol{\alpha} > \boldsymbol{A}_{\min}(\delta)$ and let $\{\Sigma_{s,t}\}_{s,t\ge 0}$ be defined by the recursion (6.3) with initial condition (6.2). Then there exists constants $B_1, r_1 > 0$ such that for all $t \ge 0$, letting $\tau_* := \lim_t \tau_t$,*

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

*Proof.* Throughout the proof, we use the $\{\eta_p^t\}_{p\in\mathbb{N}_{>0}}$ notation introduced in Section 4 and defined in (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 (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 (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 $\mathbb{E}[\boldsymbol{Z}_s]=\mathbb{E}[\boldsymbol{Z}_r]=\mathbf{0}$, 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 (2.4) and by Theorem 1 we know that $\{E_{t,t}\}_{t\ge0}$ is monotone and converges to $\tau_*^2$ as $t\to\infty$. To prove exponential convergence of $\{E_{t-1,t}\}_{t\ge0}$ as claimed in the lemma statement, we construct a discrete dynamical system below.

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

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

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 [5, Lemma 5.8] is strictly positive definite. Using the definition of the $\Sigma$ recursion in (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

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

$$G_2(\boldsymbol{y}_t) \equiv \sigma_w^2+\lim_p \mathbb{E}\{\|\eta_p(\boldsymbol{B}+\sqrt{y_{t,2}}\boldsymbol{Z}_t;\boldsymbol{\alpha}\sqrt{y_{t,2}})-\boldsymbol{B}\|^2\}/(\delta p), \tag{E.3}$$

$$G_3(\boldsymbol{y}_t) \equiv \lim_p \mathbb{E}\{\|\eta_p(\boldsymbol{B}+\sqrt{y_{t,2}}\boldsymbol{Z}_t;\boldsymbol{\alpha}\sqrt{y_{t,2}})-\eta_p(\boldsymbol{B}+\sqrt{y_{t,1}}\boldsymbol{Z}_{t-1};\boldsymbol{\alpha}\sqrt{y_{t,1}})\|^2\}/(\delta p), \tag{E.4}$$

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. (6.2) is

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

It follows that

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

Using the above, it is easy to show $y_{1,3} < y_{1,1} + y_{1,2}$. This follows since $\mathbb{E}_{\boldsymbol{B}}\{\boldsymbol{B}^\top(\boldsymbol{B} - \mathbb{E}_{\boldsymbol{Z}_0}\{\eta_p^0(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0)\})\}$ is asymptotically separable using Lemma 3.3 and because the function $x \mapsto x - \mathbb{E}_Z h^0(x + \tau_0 Z)$ is monotone increasing. It follows that $\lim_p \mathbb{E}_{\boldsymbol{B}}\{\boldsymbol{B}^\top(\boldsymbol{B} - \mathbb{E}_{\boldsymbol{Z}_0}\{\eta_p^0(\boldsymbol{B} + \tau_0 \boldsymbol{Z}_0)\})\}/(\delta p) > 0$.

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.

$$\begin{aligned} &y_{t+1,1} + y_{t+1,2} - y_{t+1,3} \\ &= 2\sigma_w^2 + \lim_p 2\mathbb{E}\{[\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \boldsymbol{B}]^\top[\eta_p(\boldsymbol{B} + \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\sqrt{y_{t,1}}) - \boldsymbol{B}]\}/(\delta p). \end{aligned} \tag{E.5}$$

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 [30], 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, $\mathbb{E}\{[\eta_p(\boldsymbol{B} + \sqrt{y_{t,2}}\boldsymbol{Z}_t; \boldsymbol{\alpha}\sqrt{y_{t,2}}) - \boldsymbol{B}]^\top[\eta_p(\boldsymbol{B} + \sqrt{y_{t,1}}\boldsymbol{Z}_{t-1}; \boldsymbol{\alpha}\sqrt{y_{t,1}}) - \boldsymbol{B}]\} \geq 0$, 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 $y_{t,3} \leq (y_{t,1} + y_{t,2})$ 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 (2.8). 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

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

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 $\boldsymbol{Z}$, $\boldsymbol{W} \sim N(0, \mathbb{I})$:

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

where

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

Obviously $G_*(0) = 0$. A simple Taylor expansion about the first argument around $\boldsymbol{B}$ yields (recall higher derivatives of $\eta$ are 0 almost everywhere)

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

Using the above, we study $G_*'(x)$. First, we can exchange the limit and differentiation because $f_p(x) := x\, \mathbb{E}\left\{\|\partial_1 \eta_p(\boldsymbol{B}; \boldsymbol{\alpha}\tau_*)]\|^2\right\}/(\delta p)$ converges uniformly to $f(x) := \lim_p x\, \mathbb{E}\left\{\|\partial_1 \eta_p(\boldsymbol{B}; \boldsymbol{\alpha}\tau_*)]\|^2\right\}/(\delta p)$. 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

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

By uniform convergence we have,

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

Hence $G_*'(0) < 1$, using (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 $\sigma(J) = \max\{\mathsf{F}'(\tau_*^2), G_*'(0)\}$. Since $G_*'(0) < 1$ proved above and $\mathsf{F}(\tau_*^2) < 1$ by Theorem 1, the claim follows. □

*Proof of Lemma 7.3.* We show that Lemma 7.3 follows by Lemmas E.1 and 6.2. By Lemma 6.2,

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

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 E.1 since $\tau_t^2 = \Sigma_{t,t}$ and $\tau_{t-1}^2 = \Sigma_{t-1,t-1}$ both converge to $\tau_*^2$ as does $\Sigma_{t,t-1}$.

□

# F 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 [5, Section 5.3], with the key difference being the concept of equivalence classes as described in Section 5.1.

We now introduce a more general recursion than the AMP algorithm in (1.3a)-(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 $\boldsymbol{z}^0 = \boldsymbol{y}$.

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

Note that these definitions of $\boldsymbol{h}^t$ and $\boldsymbol{m}^t$ match those used in Section 6.

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

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

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

Define the $\sigma$-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 [4, 8], says that the conditional distribution of the random matrix $\boldsymbol{X}$ given $\mathfrak{S}_t$ is

$$\boldsymbol{X}|_{\mathfrak{S}_t} \stackrel{d}{=} \boldsymbol{E}_t + \boldsymbol{P}_{\boldsymbol{M}_t}^{\perp} \tilde{\boldsymbol{X}} \boldsymbol{P}_{\boldsymbol{Q}_t}^{\perp}, \tag{F.3}$$

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 (F.3), we use the notation $\boldsymbol{P}_{\boldsymbol{M}_t}^{\perp} = \mathbb{I} - \boldsymbol{P}_{\boldsymbol{M}_t}$ and$\boldsymbol{P}_{\boldsymbol{Q}_t}^{\perp} = \mathbb{I} - \boldsymbol{P}_{\boldsymbol{Q}_t}$ 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 (F.3) can be found in [4, 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 7.5 is implied by the following statement.

**Lemma F.1.** *Let $s$ be a set of maximal atoms in $[p]$ such that $|s| \le p(\delta-\gamma)$, 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 $\gamma$) with*

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

*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 7.5.* The proof is adapted from [5, Section 5.3.1]. First note that by Borel-Cantelli, it is sufficient to show that, for $s$ measurable on $\mathfrak{S}_t$ and $|s|\le p(\delta-c)$ there exist $a_1=a_1(c)>0$ and $a_2=a_2(c,t)>0$, such that

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

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,

$$\begin{aligned}
&\mathbb{P}\{ \min_{|s'|\le a_1 p} \min_{\boldsymbol{v}(s')} \|\boldsymbol{X}\boldsymbol{v}\| < a_2 \big| \mathfrak{S}_t\} \le \sum_{|s'|\le a_1 p} \mathbb{P}\{ \min_{\boldsymbol{v}(s')} \|\boldsymbol{X}\boldsymbol{v}\| < a_2 \big| \mathfrak{S}_t\} \\
&\le \Big[ \sum_{k=1}^{a_1 p} \binom{p}{k} \Big] \max_{|s'|\le p a_1} \mathbb{P}\{ \min_{\boldsymbol{v}(s')} \|\boldsymbol{X}\boldsymbol{v}\| < a_2 \big| \mathfrak{S}_t\} \le e^{ph(a_1)} \max_{|s'|\le a_1 p} \mathbb{P}\{ \min_{\boldsymbol{v}(s')} \|\boldsymbol{X}\boldsymbol{v}\| < a_2 \big| \mathfrak{S}_t\}\,,
\end{aligned} \tag{F.4}$$

where $h(a) = -a\log a - (1-a)\log(1-a)$ is the binary entropy function (cf. [26, Chapter 10, Corollary 9]). Therefore, using iterated expectation and (F.4),

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

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 $\alpha_1$ and $\alpha_2$ are defined by Lemma F.1. Then,

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

Finally, using (cf. [5, Lemma 5.1]),

$$\boldsymbol{X}\boldsymbol{v}|_{\mathfrak{S}} \stackrel{d}{=} \boldsymbol{Y}(\boldsymbol{Q}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}^*\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v} + \boldsymbol{M}(\boldsymbol{M}^*\boldsymbol{M})^{-1}\boldsymbol{X}^*\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\,. \tag{F.5}$$

to estimate $\boldsymbol{X}\boldsymbol{v}$ and applying Lemma F.1, 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 F.1, using a proof that is similar to that of [5, 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 [5] 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 $\text{supp}^*$ equals $s$. In the LASSO case where $\text{supp}^* \equiv \text{supp}$ 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 $1 \le j \le p$ 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 F.2.** *[Adapted from [5, 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 $\boldsymbol{Ev} = \boldsymbol{Y}(\boldsymbol{Q}^\top\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top\boldsymbol{P_Q v} + \boldsymbol{M}(\boldsymbol{M}^\top\boldsymbol{M})^{-1}\boldsymbol{A}^\top\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}$ and consider the event*

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

*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:

$$\widetilde{\varepsilon}_1 = \Big\{|(\boldsymbol{Ev} - \boldsymbol{P_M}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v})^\top(\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v})| \ \le \ \Big(1 - \frac{\gamma}{2\delta}\Big)^{1/2}\|\boldsymbol{Ev} - \boldsymbol{P_M}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|\,\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|\Big\}, \quad \text{(F.6)}$$

where the event $\widetilde{\varepsilon}_1$ is meant to hold for all $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$ and $\text{supp}^*(\boldsymbol{v}) \subseteq s$. 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 $\text{im}(\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{P}_s)$ with dimension at most $p(\delta - \gamma)$. Also, $\boldsymbol{Ev} - \boldsymbol{P_M}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}$ 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 (F.1) and (F.2), having dimension at most $2t$. Applying Proposition G.1 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

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

holds with the desired probability, proving the claim. Conditional on event (F.6), one can show

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

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 $\varepsilon_1$ occurs, giving the desired probability of $\varepsilon_1$ as well. □

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

**Lemma F.3.** *[Adapted from [5, 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*

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

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

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

$$\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq \sigma_{\min}(\tilde{\boldsymbol{X}}|_V)\,\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|\,, \tag{F.7}$$

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 [25] 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 F.4.** *[5, Lemma 5.6] There exists a constant $c = c(t) > 0$ such that, defining the event,*

$$\mathcal{E}_3 := \{\|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\| \geq c(t)\|\boldsymbol{P}_{\boldsymbol{Q}}\boldsymbol{v}\|\,, \|\boldsymbol{E}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \leq c(t)^{-1}\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|, \ \ \textit{for all} \ \ \boldsymbol{v} \in \mathbb{R}^p\}\,, \tag{F.8}$$

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

Finally, we can now prove Lemma F.1 with the ingredients given in Lemmas F.2-F.4. We restate the proof from [5, Lemma 5.3] with minor changes.

*Proof of Lemma F.1.* We start with Lemma F.4 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

$$\mathcal{E} := \Big\{ \min_{\|\boldsymbol{v}\|=1,\, \operatorname{supp}^*(\boldsymbol{v}) \subseteq s} \|\boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \leq \alpha_2(t) \Big\}\,. \tag{F.9}$$

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 F.2, we get

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

and the thesis follows from large deviation bounds on the norm $\|\tilde{\boldsymbol{X}}\|_2$ (see [24]) 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 F.2 and F.3, we can assume events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. Therefore $\|\boldsymbol{E}\boldsymbol{v} + \boldsymbol{P}_{\boldsymbol{M}}^{\perp}\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq (\frac{\gamma}{4\delta})^{1/2}\|\tilde{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\| \geq (\frac{\gamma}{4\delta})^{1/2}c_1(\gamma)\|\boldsymbol{P}_{\boldsymbol{Q}}^{\perp}\boldsymbol{v}\|$, proving our thesis. □

# G 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 [5] that we repeat here for the reader.

The following proposition is used in the proof of Lemma F.2. The proof is identical to that of [5, Proposition E.1] and it follows from a standard concentration of measure argument in [24]. For this reason, we don't repeat it here.

**Proposition G.1.** *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 $\boldsymbol{P}_\lambda[:, j] = \boldsymbol{0}$. 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 [22] relating to the equivalence of $\ell^2$ and $\ell^1$ norms on random vector spaces (cf. also [5, Theorem F.1]).

**Theorem G.1.** *[22] For any positive number $\upsilon$ there exist a universal constant $c_\upsilon$ 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 [1] (cf. also [5, Theorem F.2]).

**Theorem G.2.** *[1] 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,*

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