Electronic Journal of Statistics
Vol. 16 (2022) 6512–6552
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2092

LASSO risk and phase transition under dependence[†][*]

Hanwen Huang

Department of Epidemiology and Biostatistics
University of Georgia, Athens, GA 30602 e-mail: huanghw@uga.edu

Abstract: We consider the problem of recovering a k-sparse signal $\beta_0 \in \mathbb{R}^p$ from noisy observations $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{w} \in \mathbb{R}^n$. One of the most popular approaches is the l_1-regularized least squares, also known as LASSO. We analyze the mean square error of LASSO in the case of random designs in which each row of $\mathbf{X}$ is drawn from distribution $N(0, \Sigma)$ with general Σ. We first derive the asymptotic risk of LASSO for $\mathbf{w} \neq 0$ in the limit of $n, p \to \infty$ with $n/p \to \delta \in [0, \infty)$. We then examine conditions on n, p, and k for LASSO to exactly reconstruct β_0 in the noiseless case $\mathbf{w} = 0$. A phase boundary $\delta_c = \delta(\epsilon)$ is precisely established in the phase space defined by $0 \leq \delta, \epsilon \leq 1$, where $\epsilon = k/p$. Above this boundary, LASSO perfectly recovers β_0 with high probability. Below this boundary, LASSO fails to recover β_0 with high probability. While the values of the non-zero elements of β_0 do not have any effect on the phase transition curve, our analysis shows that δ_c does depend on the signed pattern of the nonzero values of β_0 for general $\Sigma \neq \mathbf{I}_{p\times p}$. This is in sharp contrast to the previous phase transition results derived in i.i.d. case with $\Sigma = \mathbf{I}_{p\times p}$ where δ_c is completely determined by ϵ regardless of the distribution of β_0. Underlying our formalism is a recently developed efficient algorithm called approximate message passing (AMP) algorithm. We generalize the state evolution of AMP from i.i.d. case to general case with $\Sigma \neq \mathbf{I}_{p\times p}$. Extensive computational experiments confirm that our theoretical predictions are consistent with simulation results on moderate size system.

MSC2020 subject classifications: Primary 62F12, 62F12; secondary 62F12.
Keywords and phrases: Asymptotic risk, LASSO, mean square error, phase transition, state evolution.

Received December 2021.

Contents

[*] This research is supported in part by Division of Mathematical Sciences (National Science Foundation) Grant DMS-1916411.

[†] This is an original survey paper.

1. Introduction

1.1. LASSO phase transition

Consider the problem of recovering a sparse signal $\beta_0 \in \mathbb{R}^p$ from a undersampled collection of noisy measurements $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{w}$, where the matrix $\mathbf{X}$ is $n \times p$, the p-vector β_0 is k-sparse (i.e. it has at most k non-zero entries), and $\mathbf{w} \in \mathbb{R}^n$ is random noise. One of the most popular approaches for this problem is called LASSO which estimates β_0 by solving the following convex optimization problem

$$\hat{\beta}(\lambda) = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \right\}. \tag{1}$$

In the noiseless case $\mathbf{w} = 0$, exact reconstruction of β_0 through (1) is possible when $n \geq p$ or β_0 is sufficiently sparse for the case of $n < p$. Knowing the precise limits to such sparsity for the case of $n < p$ is important both for theory and practice.

In the noiseless case, the $\lambda = 0$ limit of (1) is identical to the solution of the following l_1 minimization problem

$$\begin{aligned} &\min \|\beta\|_1, \\ &\text{subject to } \mathbf{y} = \mathbf{X}\beta. \end{aligned} \tag{2}$$

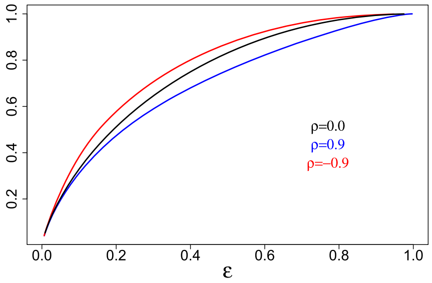

Fig 1. *Phase transition boundary in the plane (δ, ϵ) when the matrix $\mathbf{X}$ consisting of i.i.d. Gaussian rows $\mathbf{x}_i \sim N(0, \Sigma)$. Black curve: $\Sigma = \mathbf{I}$. Red curve: Σ is block-diagonal with AR(1) block structure Σ_s, i.e. $\Sigma_{s,ij} = \rho^{|i-j|}$ with block length $s = 2$ and $\rho = -0.9$. Blue curve: Σ is block-diagonal with AR(1) block structure, i.e. $\Sigma_{s,ij} = \rho^{|i-j|}$ with block length $s = 2$ and $\rho = 0.9$.*

The precise condition under which $\hat{\beta}(\lambda = 0)$ can successfully recover β_0 has been obtained through large system analysis by letting n, p, k tend to infinity with fixed rates n/p and k/p. Let $\epsilon = k/p$ and $\delta = n/p$ denote the sparsity and under-sampling fractions for sampling β_0 and $\mathbf{y}$ according to $\mathbf{y} = \mathbf{X}\beta_0$. Then $(\delta, \epsilon) \in [0, 1]$ defines a phase space which expresses different combinations of under-sampling δ and sparsity ϵ. When the elements of the matrix $\mathbf{X}$ are generated from i.i.d. Gaussian, the phase space can be divided into two phases: "success" and "failure" by a phase transition curve $\delta = \delta_c(\epsilon)$ which has been explicitly derived in the literature (see e.g. [14, 10, 19, 12]) as shown by the black curve in Figure 1. Above this curve, LASSO perfectly recovers the sparse signal β_0 with high probability, i.e. $\hat{\beta}(\lambda = 0) = \beta_0$. Below this curve, the reconstruction fails, i.e. $\hat{\beta}(\lambda = 0) \neq \beta_0$ also with high probability.

Our aim in this paper is to study the LASSO phase transition under arbitrary covariance dependence, i.e. $\mathbf{X}$ consists of i.i.d. Gaussian rows $\mathbf{x}_i \sim N(0, \Sigma)$ with general covariance matrix $\Sigma \succ 0$ and $\Sigma \neq \mathbf{I}_{p \times p}$. We present formulas that precisely characterize the LASSO sparsity/undersampling trade-off for arbitrary Σ. Our numerical results show that LASSO phase transition depends on the form of Σ. For example, the red and blue curves in Figure 1 correspond to the phase transition boundaries for block-diagonal covariance matrix Σ with AR(1) block structure Σ_s, i.e. $\Sigma_{s,ij} = \rho^{|i-j|}$ with block length $s = 2$ and $\rho = -0.9$ and $\rho = 0.9$ respectively. These results indicate that for a given sparsity fraction ϵ, the limits of allowable undersampling $\delta_c(\epsilon)$ of LASSO in the case when $\mathbf{X}$ has non-independent entries can be either higher or lower than the corresponding value in the case when $\mathbf{X}$ has i.i.d. entries. To the best of our knowledge, this is the first result to illustrate the LASSO phase transition for matrices $\mathbf{X}$ that have non-independent entries.

1.2. *Approximate message passing*

Our analysis is based on the asymptotic study of mean squared error (MSE) of the LASSO estimator, i.e. the quantity $\|\hat{\beta}(\lambda) - \beta_0\|^2/p$, in the large system limit $n, p \to \infty$ with $n/p = \delta \in [0, \infty)$ fixed. We derive the asymptotic MSE through the analysis of an efficient iterative algorithm first proposed by [12] called approximate message passing (AMP) algorithm. The AMP algorithms can be considered as quadratic approximations of loopy belief propagation algorithms on the dense factor graph corresponding to the LASSO model. A striking property of AMP algorithms is that their high-dimensional per-iteration behavior can be characterized by a one-dimensional recursion termed *state evolution*. The AMP's state evolution was first conjectured in [12] and subsequently proved rigorously in [5] for i.i.d. Gaussian matrices. This result was extended to i.i.d. non-Gaussian matrices in [4] under certain regularity conditions. [17] further extended the AMP's state evolution to independent but non-identical Gaussian matrices. But there remains the important question of how AMP behaves with non-independent matrices.

In this paper, we establish the AMP's state evolution for non-independent Gaussian matrices whose fixed points are consistent with the replica prediction derived in [18]. On the basis of this result, we first derive the MSE for AMP estimators using the fixed points of state evolution, then we obtain the MSE for LASSO by proving that, in the large system limits, the AMP algorithm converges to the LASSO optimum after enough iterations. Our analysis strategy is similar to the one used in [6] for i.i.d. Gaussian matrices. However, our main result cannot be seen as a straightforward extension of the ones in [6]. In particular, the proofs of some results for non-independent case are much more complicated than for i.i.d. case, and our proof techniques are hence of independent interest, see e.g. the proof of Lemma 1 for the concavity and strict increasing of ψ function defined in (27), the proof of Theorem 2 for deriving the phase transition curve, and the proof of Lemmas 4 and 5 for the structural property of LASSO under dependent designs.

Note that although this study is motivated by the phase transition problem shown in Figure 1 which is restricted to the case when $(\delta, \epsilon) \in [0, 1]$, the AMP and main results derived in Theorem 1 work fine for the entire range $\delta \in [0, \infty)$. The LASSO risk formulas derived in Theorem 1 apply to both noiseless and noisy cases with quite general i.i.d. random error. The phase transition results derived in Theorem 2 are only for the noiseless case. This result can also be generalized to the noisy case and we have some discussion about this in Section 6.

1.3. Related work

[24] derived expressions for the asymptotic mean square error of LASSO. Similar results were presented in [16, 18]. Unfortunately, these results were non-rigorous and were obtained through the famous replica method from statistical physics [22]. Some rigorous proofs were given in [3, 25, 6] to show that the replica symmetric prediction for LASSO is exact. However, all these rigorous proofs are limited to settings with i.i.d. Gaussian measurement matrices.

By now a large amount of empirical and theoretical studies have been conducted to understand the phase transitions of regularized reconstruction exhibited by different algorithms. In the noiseless case, the phase transition curve based on (2) was explored in [14] utilizing techniques of combinatorial geometry for entries of $\mathbf{X}$ being i.i.d. Gaussians. The AMP algorithm was proposed in [12] which produces the same phase transition curve. It has been proved in [6] that the limit of AMP estimate corresponds to the solution of LASSO in the asymptotic settings. Statistical physics methods were used to study l_q $(0 \leq q \leq 1)$ based reconstruction methods in [19]. [30] and [28] studied the phase transition for l_q penalized least square in the case of $0 \leq q < 1$ and $1 \leq q \leq 2$ respectively. [20] replaced the l_1 regularization with a probabilistic approach and studied its phase transition. [11] derived phase transition of AMP for a wide class of denoisers. In noisy case, [13] studied the noise sensitivity phase transition of LASSO through deriving the minimax formulation of the asymptotic MSE. [30, 28] studied the phase transition of l_q-regularized least squares using higher order analysis of regularization techniques. The phase transition in generalized linear models for i.i.d. matrices was characterized in [2]. [21] generalized AMP to complex approximate message passing methods and used it to study phase transitions for compressed sensing with complex vectors.

Most of the above results are for i.i.d. Gaussian matrices and some of them are for independent but non-identical Gaussian matrices. This paper performs the phase transition analysis of LASSO under dependent Gaussian matrices. We derive the basic relation between minimax MSE and the phase-transition boundary in the sparsity-undersampling plane. We adopt the message passing analysis whose state evolution allows to determine whether AMP recovers the signal correctly, by simply checking whether the MSE vanishes asymptotically or not. Most closely related to the current paper are results by [27] that derives the sharp thresholds for LASSO sparsity recovery in the case of random designs in which each row of $\mathbf{X}$ is drawn from a broad class of Gaussian ensembles $N(0, \Sigma)$. However, the major difference is that [27] only provides the necessary and sufficient conditions for the recovery of sparsity pattern, while we focus on the recovery of complete signal including both signed support and magnitude. Recently, based on Gordon's inequality, [9] derived the LASSO risk under non-standard Gaussian design for i.i.d. Gaussian random error, i.e. $w_i \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$. But they didn't study the phase transition problem and also we don't have Gaussian restriction here for random error $\mathbf{w}$.

2. LASSO risk

The Gaussian random design model for linear regression is defined as follows. We are given n i.i.d. pairs $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ with $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, and $\mathbf{x}_i \sim N(0, \Sigma)$ for some positive definite $p \times p$ covariance matrix $\Sigma \succ 0$. Further, y_i is a linear function of $\mathbf{x}_i$, plus noise

$$y_i = \mathbf{x}_i^T \beta_0 + w_i,$$

where $w_i \overset{i.i.d.}{\sim} p_w$ with mean 0 and variance σ_w^2, and $\beta_0 \in \mathbb{R}^p$ is a vector of parameters to be estimated. The special case $\Sigma = \mathbf{I}_{p \times p}$ is usually referred to as standard Gaussian design model. In matrix form, letting $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{w} = (w_1, \cdots, w_n)^T$, and denoting by $\mathbf{X}$ the matrix with rows $\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T$, we have

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{w}.$$

In this paper, our approach is based on the LASSO estimator

$$\hat{\beta} = \text{argmin}_\beta \mathcal{C}(\beta), \tag{3}$$

where

$$\mathcal{C}(\beta) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1.$$

We will consider sequences of instances of increasing sizes. The sequence of instances $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}$ parameterized by p is said to be a converging sequence if $\beta_0(p) \in \mathbb{R}^p, \mathbf{w}(p) \in \mathbb{R}^n, \Sigma(p) \in \mathbb{R}^{p\times p}, \mathbf{X}(p) \in \mathbb{R}^{n\times p}$ with $n = n(p)$ is such that $n/p \to \delta \in (0, \infty)$, and in addition the following conditions hold:

1. The empirical distribution of the entries of $\beta_0(p)$ converges weakly to a probability measure p_{β_0} on $\mathbb{R}$ with bounded second moment. Further $\sum_{i=1}^p \beta_{0,i}(p)^2/p \to E_{p_{\beta_0}}\{\beta_0^2\}$.
2. The empirical distribution of the entries of $\mathbf{w}(p)$ converges weakly to a probability measure p_w on $\mathbb{R}$ with $\sum_{i=1}^n w_i(p)^2/n \to \sigma_w^2 < \infty$.
3. For any $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_{\Sigma(p)}^2 = O(\|\mathbf{v}\|^2)$ and $\|\mathbf{v}\|_{\Sigma(p)^{-1}}^2 = O(\|\mathbf{v}\|^2)$, where $\|\mathbf{v}\|_\Sigma^2 = \mathbf{v}^T\Sigma\mathbf{v}$.
4. The rows of $\mathbf{X}(p)$ are drawn independently from distribution $N(0, \frac{1}{n}\Sigma(p))$.
5. The sequence of functions

$$\mathcal{E}^{(p)}(a, b) \equiv \frac{1}{p} E \min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\beta - \beta_0(p) - \sqrt{a}\Sigma(p)^{-1/2}\mathbf{z}\|_{\Sigma(p)}^2 + b\|\beta\|_1 \right\} \tag{4}$$

   admits a differentiable limit $\mathcal{E}(a, b)$ on $\mathbb{R}_+ \times \mathbb{R}_+$ with $\frac{\partial \mathcal{E}^{(p)}(a,b)}{\partial a} \to \frac{\partial \mathcal{E}(a,b)}{\partial a}$ and $\frac{\partial \mathcal{E}^{(p)}(a,b)}{\partial b} \to \frac{\partial \mathcal{E}(a,b)}{\partial b}$, where $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of $\beta_0(p)$.
6. For any $a_1, b_1, a_2, b_2 \in \mathbb{R}_+$ and any 2×2 positive definite matrix $\mathbf{S}$, the following limit exists and is finite

$$\lim_{p\to\infty} \frac{1}{p} \left\langle \hat{\beta}_1^{(p)}, \hat{\beta}_2^{(p)} \right\rangle,$$

   where $\langle\cdot, \cdot\rangle$ is the standard scalar product and

$$\begin{aligned} \hat{\beta}_1^{(p)} &= \text{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\beta - \beta_0(p) - \sqrt{a_1}\Sigma(p)^{-1/2}\mathbf{z}_1\|_{\Sigma(p)}^2 + b_1\|\beta\|_1 \right\}, \\ \hat{\beta}_2^{(p)} &= \text{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\beta - \beta_0(p) - \sqrt{a_2}\Sigma(p)^{-1/2}\mathbf{z}_2\|_{\Sigma(p)}^2 + b_2\|\beta\|_1 \right\}, \end{aligned}$$

   where $(\mathbf{z}_1, \mathbf{z}_2) \sim N(0, \mathbf{S} \otimes \mathbf{I}_{p\times p})$ and is independent of $\beta_0(p)$.

Conditions 1 and 2 have appeared in [6] which indicate that the entries of β_0 and $\mathbf{w}$ are drawn i.i.d. from certain distributions with bounded second order moment. Note that the entries of $\mathbf{w}$ are not necessarily normal. Denote $\lambda_{\min}(\Sigma(p))$ and $\lambda_{\max}(\Sigma(p))$ the smallest and largest eigenvalues of $\Sigma(p)$ respectively, then Condition 3 is equivalent to that $1/\lambda_{\min}(\Sigma(p)) = O(1)$ and $\lambda_{\max}(\Sigma(p)) = O(1)$. Condition 5 indicates that the covariance matrix should satisfy such conditions that the l_1 penalized quadratic loss function specified in (4) has a differentiable limit, i.e. the derivative over a, b and the limit of p are exchangeable. It is worth stressing that Conditions 5 and 6 are satisfied by a larger family of covariance matrices. For instance, based on law of large number, it can be proved that it holds for block-diagonal matrices Σ as long as the blocks have bounded length and the block's empirical distribution converges. This condition has also appeared in [18] and it ensures the existence of large dimensional limits of some functions such as (6), (10), and (12) that will be used in describing the main results of Theorems 1 and 2. It also allows us to exchange the order of operations such as taking limit and derivative over these functions. In Section 4.3, we will discuss the specific choice of covariance structure such that this condition can be satisfied. We insist on the fact that $\beta_0(p)$, $\mathbf{w}(p)$, $\Sigma(p)$, $\mathbf{X}(p)$ depend on p. However, we will drop this dependence most of the time to ease the reading.

In order to present our main result, for any $\theta > 0$ and $\Sigma \succ 0$, we need to introduce the soft-thresholding operation $\eta_\theta : \mathbb{R}^p \to \mathbb{R}^p$ which is defined as

$$\eta_\theta(\mathbf{v}) = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\beta - \mathbf{v}\|_\Sigma^2 + \theta \|\beta\|_1 \right\}. \tag{5}$$

Then for a converging sequence of instances, we can define the function

$$\psi(\tau^2, \theta) = \sigma_w^2 + \lim_{p \to \infty} \frac{1}{p\delta} E\left(\|\eta_\theta(\beta_0 + \tau \Sigma^{-1/2} \mathbf{z}) - \beta_0\|_\Sigma^2 \right), \tag{6}$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p \times p})$ is independent of β_0. Notice that the function ψ depends implicitly on the law p_{β_0}.

Condition 5 allows us to verify the existence of the limit in (6). Toward this end, we start from (4) and have

$$\mathcal{E}^{(p)}(\tau^2, \theta) = \frac{1}{p} E \left\{ \frac{1}{2} \|\hat{\beta} - \beta_0 - \tau \Sigma^{-1/2} \mathbf{z}\|_\Sigma^2 + \theta \|\hat{\beta}\|_1 \right\}, \tag{7}$$

where $\hat{\beta} = \eta_\theta(\beta_0 + \tau \Sigma^{-1/2} \mathbf{z})$. In order to take derivative over τ^2 and θ, we need to conduct integrals over $\mathbf{z} \in \mathbb{R}^p$. We first divide the p-dimensional space into regions such that $\hat{\beta}$ is differentiable in each region and continuous across the entire space (see Figure 7 for a simple 2-dimensional illustration). Then the derivative of $\mathcal{E}^{(p)}(\tau^2, \theta)$ involves the explicit derivative inside each region and integrals over the boundaries among different regions over $p-1$-dimensional measure. According to Stokes's theorem, as in Theorem 1 of [1], we conclude that the boundary effects are canceled and have no contribution due to the continuity of $\hat{\beta}$ (see detailed discussion in A.6). Further note that, according to the definition of $\hat{\beta}$, the derivative of the integrand in (7) over $\hat{\beta}$ is 0, therefore we only need to consider the explicit dependence of the integrand on τ^2 and θ in deriving the corresponding derivatives. We obtain

$$\frac{\partial \mathcal{E}^{(p)}(\tau^2,\theta)}{\partial \tau^2} = -\frac{1}{2p\tau} E\left\langle \hat{\beta} - \beta_0, \Sigma^{1/2}\mathbf{z} \right\rangle + \frac{1}{2}, \tag{8}$$

$$\frac{\partial \mathcal{E}^{(p)}(\tau^2,\theta)}{\partial \theta} = \frac{1}{p} E\|\hat{\beta}\|_1. \tag{9}$$

From Condition 5, all the limits of $\mathcal{E}^{(p)}(\tau^2,\theta)$, $\frac{\partial \mathcal{E}^{(p)}(\tau^2,\theta)}{\partial \tau^2}$, and $\frac{\partial \mathcal{E}^{(p)}(\tau^2,\theta)}{\partial \tau^2}$ exist, therefore, $\lim_{p\to\infty} \frac{1}{p\delta} E\left(\|\hat{\beta} - \beta_0\|_\Sigma^2\right)$ also exists, which is just the right hand side of (6). Taking $\beta_0 = 0$, we immediately obtain that the limit of the following equation (10) also exists.

We choose $\theta = \alpha\tau$, then we have the following result in order to establish a calibration mapping between α and λ.

Proposition 1. *Define function*

$$f(\alpha) \equiv \lim_{p\to\infty} \frac{1}{p\delta} E\left(\|\eta_\alpha(\Sigma^{-1/2}\mathbf{z})\|_\Sigma^2\right). \tag{10}$$

Then the equation $f(\alpha) = 1$ has a unique solution denoted by $\alpha_{\min}(\delta)$ when $\delta < 1$. Then for any $\delta \geq 1$ or $\delta < 1$ and $\alpha > \alpha_{\min}(\delta)$, the fixed point equation

$$\tau^2 = \psi(\tau^2, \alpha\tau) \tag{11}$$

admits a unique solution.

We then define a function $\alpha \to \lambda(\alpha)$ on $(\alpha_{\min}(\delta), \infty)$ by

$$\begin{aligned} &\lambda(\alpha) \\ = \quad &\alpha\tau_\star(\alpha)\left\{1 - \lim_{p\to\infty}\frac{1}{p\delta} E\left[\mathrm{div}\eta_{\alpha\tau_\star(\alpha)}(\beta_0 + \tau_\star(\alpha)\Sigma^{-1/2}\mathbf{z})\right]\right\}, \end{aligned} \tag{12}$$

where the divergence of the vector field is defined as $\mathrm{div}\eta_\theta(\mathbf{v}) = \sum_{j=1}^p \frac{\partial \eta_{\theta,j}(\mathbf{v})}{\partial v_j}$. This function defines a correspondence between α and λ. The existence of the limit of (12) can be obtained from the existence of the limit of $\frac{\partial \mathcal{E}^{(p)}(\tau^2,\theta)}{\partial \tau^2}$ in (8) following by integration by parts. In the following we will need to invert this function and define $\lambda \to \alpha(\lambda)$ on $(0,\infty)$ in such a way that

$$\alpha(\lambda) \in \{a \in (\alpha_{\min}, \infty) : \lambda(a) = \lambda\}. \tag{13}$$

The next result implies that the function $\lambda \to \alpha(\lambda)$ is well defined.

Proposition 2. *The function $\alpha \to \lambda(\alpha)$ is continuous on the interval $(\alpha_{\min}, \infty)$ and for any given λ there exist a unique α such that $\lambda(\alpha) = \lambda$.*

For two sequences (in n) of random variables $\mathbf{x}_n$ and $\mathbf{y}_n$, write $\mathbf{x}_n \overset{P}{\approx} \mathbf{y}_n$ when their difference convergences in probability to 0, i.e. $\mathbf{x}_n - \mathbf{y}_n \xrightarrow{P} 0$. For any $m \in \mathbb{N}_{>0}$, we say a function $\varphi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz if there exist a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m : |\varphi(\mathbf{x}, \mathbf{y})| \leq L(1 + \|\mathbf{x}\| + \|\mathbf{y}\|)\|\mathbf{x} - \mathbf{y}\|$. A sequence (in m) of pseudo-Lipschitz functions $\{\varphi_m\}_{m \in \mathbb{N}_{>0}}$ is called uniformly pseudo-Lipschitz if, denoting by L_m is the pseudo-Lipschitz constant, we have $L_m < \infty$ for each m and $\sup_{m \to \infty} L_m < \infty$. Note that the input and output dimensions of each φ_m can depend on m. We call any $L > \sup_{m \to \infty} L_m$ a pseudo-Lipschitz constant of the sequence. We can now state our main result.

Theorem 1. *Let $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}_{p \in \mathbb{N}}$ be a converging sequence of instances. Denote $\hat{\beta}(\lambda)$ the LASSO estimator for instance $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}$ with $\lambda > 0$ and $P\{\beta_0(p) \neq 0\} > 0$. For any sequence $\varphi_p : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, $p \geq 1$, of uniformly pseudo-Lipschitz functions, we have*

$$\varphi_p(\hat{\beta}(\lambda), \beta_0) \overset{P}{\approx} E\varphi_p(\eta_{\theta_\star}(\beta_0 + \tau_\star \Sigma^{-1/2}\mathbf{z}), \beta_0)$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p \times p})$ is independent of $\beta_0 \sim p_{\beta_0}$, $\tau_\star = \tau_\star(\alpha(\lambda))$, and $\theta_\star = \alpha(\lambda)\tau_\star(\alpha(\lambda))$.

Using function $\varphi_p(\mathbf{a}, \mathbf{b}) = \frac{1}{p}\|\mathbf{a} - \mathbf{b}\|^2$, we obtain LASSO MSE $\frac{1}{p}\|\hat{\beta}(\lambda) - \beta_0\|^2$ which can be used to evaluate competing optimization methods on large scale applications. Using Theorem 1, we get

$$\frac{1}{p}\|\hat{\beta}(\lambda) - \beta_0\|^2 \overset{P}{\approx} \frac{1}{p}E\|\eta_{\theta_\star}(\beta_0 + \tau_\star \Sigma^{-1/2}\mathbf{z}) - \beta_0\|^2, \tag{14}$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p \times p})$ is independent of $\beta_0 \sim p_{\beta_0}$, $\tau_\star = \tau_\star(\alpha(\lambda))$, and $\theta_\star = \alpha(\lambda)\tau_\star(\alpha(\lambda))$.

Therefore, for fixed λ, LASSO MSE explicitly depends on $\tau_\star^2$ which can be obtained by solving the fixed point equation $\tau_\star^2 = \psi(\tau_\star^2, \alpha\tau_\star)$ together with (12). Closer to the spirit of this paper, [18] non-rigorously derived the LASSO MSE under the same setting considered here using the replica method from statistical physics. The present paper is rigorous and putting on a firmer basis this line of research.

3. Phase transition of LASSO under dependence

Note that the LASSO risk results based on Theorem 1 work fine for entire $\sigma_w^2, \delta \in [0, \infty)$. To study phase transition, we only need to consider $\delta \in [0, 1]$ and evaluate the results in the noiseless setting $\sigma_w^2 = 0$ and understand the extend to which (3) accurately recovers β_0 under this setting. Consider a class of distributions $\mathcal{F}_\epsilon$ whose mass at zero is equal to $1 - \epsilon$, i.e.

$$\mathcal{F}_\epsilon \equiv \{p_{\beta_0} : p_{\beta_0}(\{0\}) = 1 - \epsilon\}.$$

When the matrix $\mathbf{X}$ has i.i.d. Gaussian elements, i.e. $\Sigma = \mathbf{I}_{p \times p}$, phase space $0 \leq \delta, \epsilon \leq 1$ can be divided into two components, or phases, separated by a curve $\delta_c = \delta(\epsilon)$, which does not depend on the actual distribution of p_{β_0} and can be explicitly computed. Above this curve, LASSO perfectly recovers the sparse signal β_0 with high probability. Below this curve, we have $\hat{\beta} \neq \beta_0$ with high probability.

For non-standard Gaussian design, i.e. $\Sigma \neq \mathbf{I}_{p\times p}$, we need to consider a more general class of distributions $\mathcal{F}_{\epsilon,\Delta}$ defined as

$$\mathcal{F}_{\epsilon,\Delta} \equiv \left\{ p_{\beta_0} : p_{\beta_0}(\{0\}) = 1-\epsilon \text{ and } \frac{|p_{\beta_0}(\{>0\}) - p_{\beta_0}(\{<0\})|}{|p_{\beta_0}(\{>0\}) + p_{\beta_0}(\{<0\})|} = \Delta \right\}.$$

Here we introduce an extra parameter $\Delta = \frac{|P(\beta_0>0)-P(\beta_0<0)|}{|P(\beta_0>0)+P(\beta_0<0)|}$ which represents the positive-negative asymmetry for the nonzero components of β_0. Clearly, $0 \leq \Delta \leq 1$, and if $\Delta = 0$, we have $P(\beta_0 > 0) = P(\beta_0 < 0)$, i.e. β_0 has positive and negative nonzero components with equal probability.

We denote by $[p] = \{1, \cdots, p\}$ the set of first p integers. For a subset $\mathbf{I} \subseteq [p]$, we let $|\mathbf{I}|$ denote its cardinality. For an $p \times p$ matrix Σ and set of indices $\mathbf{I} \subseteq [p]$, $\mathbf{J} \subseteq [p]$, we use $\Sigma_{\mathbf{IJ}}$ to denote the $|\mathbf{I}| \times |\mathbf{J}|$ sub-matrix formed by rows in $\mathbf{I}$ and columns in $\mathbf{J}$. Likewise, for a vector $\beta \in \mathbb{R}^p$, $\beta_{\mathbf{I}}$ is the restriction of β to indices in $\mathbf{I}$. The following Theorem shows that, under general covariance Σ, the phase transition curve exists and depends on the asymmetry parameter Δ.

Theorem 2. *Let $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}_{p\in\mathbb{N}}$ be a converging sequence of instances and $\mathbf{w}(p) = 0$. Assume $p_{\beta_0} \in \mathcal{F}_{\epsilon,\Delta}$. Then the phase space $0 \leq \delta, \epsilon \leq 1$ can be divided into two components separated by a curve $\delta_c = \delta(\epsilon)$. Above this curve, LASSO algorithm (3) perfectly recovers the sparse signal β_0 with high probability, i.e. $\frac{1}{p}\|\hat{\beta}(\lambda) - \beta_0\| \to 0$ after appropriately choosing the tuning parameter λ. Below this curve, we have $\hat{\beta} \neq \beta_0$ with high probability. For fixed ϵ, the δ_c is determined by*

$$\delta_c = \inf_{\alpha} M(\epsilon, \Delta, \alpha), \tag{15}$$

where

$$\begin{aligned} & M(\epsilon, \Delta, \alpha) \\ = \; & \lim_{p\to\infty} \frac{1}{p} E\{((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha sign(\hat{\beta}_{\mathcal{A}}))^T \Sigma_{\mathcal{A}\mathcal{A}}^{-1}((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha sign(\hat{\beta}_{\mathcal{A}}))\}, \end{aligned} \tag{16}$$

where the active set $\mathcal{A} = \mathcal{B} \cup \bar{\mathcal{B}}$ with $\mathcal{B} = \{j : \beta_{0,j} \neq 0\}$ and $\bar{\mathcal{B}}$ the active set of LASSO problem

$$\bar{\beta} = argmin_{\beta\in\mathbb{R}^{\bar{p}}} \left\{ \frac{1}{2}\|\bar{\mathbf{y}} - \bar{\mathbf{X}}\beta\|_2^2 + \alpha\|\beta\|_1 \right\} \tag{17}$$

with

$$\begin{aligned} \bar{\mathbf{X}} &= (\Sigma_{\mathcal{B}^c\mathcal{B}^c} - \Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\Sigma_{\mathcal{B}\mathcal{B}^c})^{1/2}, \\ \bar{\mathbf{y}} &= \bar{\mathbf{X}}^{-1}\left[(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}^c} - \Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\{(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}} - \alpha sign(\beta_{0,\mathcal{B}})\}\right], \end{aligned}$$

where $\bar{p} = |\mathcal{B}^c|$ and $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of p_{β_0}.

Note that all the nonzero components of β_0 contribute to function (16). Some zero components also have contribution if they are selected by the LASSO problem (17). Theorem 2 shows that the LASSO phase transition is independent of the actual distribution of p_{β_0} but depends on the positive-negative asymmetry of the nonzero components of β_0, i.e. depends on $\epsilon_+ = \#\{\beta_0 > 0\}/p$ and $\epsilon_- = \#\{\beta_0 < 0\}/p$ with $\epsilon = \epsilon_+ + \epsilon_-$ and $\Delta = (\epsilon_+ - \epsilon_-)/(\epsilon_+ + \epsilon_-)$.

The following two Corollaries provide the explicit phase transition curves for two special covariance matrices.

Corollary 1. *For $\Sigma = \mathbf{I}_{p\times p}$, the LASSO phase transition curve is determined by*

$$\begin{aligned} \delta &= \frac{2\phi(\alpha)}{\alpha + 2\phi(\alpha) - 2\alpha\Phi(-\alpha)}, \\ \epsilon &= \frac{2\phi(\alpha) - 2\alpha\Phi(-\alpha)}{\alpha + 2\phi(\alpha) - 2\alpha\Phi(-\alpha)}, \end{aligned} \tag{18}$$

This is equivalent to the result provided in [12] based on techniques of combinatorial geometry.

Corollary 2. *For block-diagonal matrix Σ with block $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, the LASSO phase transition curve is determined by*

$$\begin{aligned} \delta &= \epsilon^2 A(\alpha,\Delta) + \epsilon(1-\epsilon)B(\alpha) + (1-\epsilon)^2 C(\alpha), \\ \epsilon &= \frac{2C'(\alpha) - B'(\alpha) + \sqrt{B'(\alpha)^2 - 4\frac{\partial A(\alpha,\Delta)}{\partial\alpha}C'(\alpha)}}{2\{\frac{\partial A(\alpha,\Delta)}{\partial\alpha} - B'(\alpha) + C'(\alpha)\}}, \end{aligned} \tag{19}$$

where

$$A(\alpha,\Delta) = 1 + \frac{\alpha^2}{2}\left(\frac{1+\Delta^2}{1-\rho} + \frac{1-\Delta^2}{1+\rho}\right), \tag{20}$$

$$B(\alpha) = B_1(\alpha) + B_2(\alpha) + B_3(\alpha), \tag{21}$$

$$C(\alpha) = C_1(\alpha) + C_2(\alpha) + C_3(\alpha) + C_4(\alpha), \tag{22}$$

where

$$\begin{aligned} B_1(\alpha) &= E(\xi_1 - \alpha)^2 I(|\xi_2 - \rho\xi_1 + \rho\alpha| \le \alpha) \\ &\quad + E(\xi_2 - \alpha)^2 I(|\xi_1 - \rho\xi_2 + \rho\alpha| \le \alpha), \\ B_2(\alpha) &= E\frac{(\xi_1-\alpha)^2 + (\xi_2-\alpha)^2 - 2\rho(\xi_1-\alpha)(\xi_2-\alpha)}{1-\rho^2} \\ &\quad \{I(\xi_2 - \rho\xi_1 + \rho\alpha \ge \alpha) + I(\xi_1 - \rho\xi_2 + \rho\alpha \ge \alpha)\}, \\ B_3(\alpha) &= E\frac{(\xi_1-\alpha)^2 + (\xi_2+\alpha)^2 - 2\rho(\xi_1-\alpha)(\xi_2+\alpha)}{1-\rho^2} \\ &\quad I(\xi_2 - \rho\xi_1 + \rho\alpha \le -\alpha) \end{aligned}$$

$$
\begin{aligned}
& +E\frac{(\xi_1+\alpha)^2+(\xi_2-\alpha)^2-2\rho(\xi_1+\alpha)(\xi_2-\alpha)}{1-\rho^2} \\
& \quad I(\xi_1-\rho\xi_2+\rho\alpha\le-\alpha), \\
C_1(\alpha) &= E(\xi_1-\alpha)^2 I(|\xi_2-\rho\xi_1+\rho\alpha|\le\alpha)I(\xi_1\ge\alpha) \\
& +E(\xi_1+\alpha)^2 I(|\xi_2-\rho\xi_1+\rho\alpha|\le\alpha)I(\xi_1\le-\alpha), \\
C_2(\alpha) &= E(\xi_2-\alpha)^2 I(|\xi_1-\rho\xi_2+\rho\alpha|\le\alpha)I(\xi_2\ge\alpha) \\
& +E(\xi_2+\alpha)^2 I(|\xi_1-\rho\xi_2+\rho\alpha|\le\alpha)I(\xi_2\le-\alpha), \\
C_3(\alpha) &= E\frac{(\xi_1-\alpha)^2+(\xi_2-\alpha)^2-2\rho(\xi_1-\alpha)(\xi_2-\alpha)}{1-\rho^2} \\
& I(\xi_1-\rho\xi_2+\rho\alpha\ge\alpha)I(\xi_2-\rho\xi_1+\rho\alpha\ge\alpha) \\
& +E\frac{(\xi_1-\alpha)^2+(\xi_2+\alpha)^2-2\rho(\xi_1-\alpha)(\xi_2+\alpha)}{1-\rho^2} \\
& I(\xi_1-\rho\xi_2+\rho\alpha\ge\alpha)I(\xi_2-\rho\xi_1+\rho\alpha\le-\alpha), \\
C_4(\alpha) &= E\frac{(\xi_1+\alpha)^2+(\xi_2-\alpha)^2-2\rho(\xi_1+\alpha)(\xi_2-\alpha)}{1-\rho^2} \\
& I(\xi_1-\rho\xi_2+\rho\alpha\le-\alpha)I(\xi_2-\rho\xi_1+\rho\alpha\ge\alpha) \\
& +E\frac{(\xi_1+\alpha)^2+(\xi_2+\alpha)^2-2\rho(\xi_1+\alpha)(\xi_2+\alpha)}{1-\rho^2} \\
& I(\xi_1-\rho\xi_2+\rho\alpha\le-\alpha)I(\xi_2-\rho\xi_1+\rho\alpha\le-\alpha),
\end{aligned}
$$

where

$$
\begin{aligned}
\xi_1 &= \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}z_1+\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}z_2, \\
\xi_2 &= \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}z_1+\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}z_2,
\end{aligned}
$$

and $(z_1, z_2) \sim N(0, \mathbf{I}_{2\times 2})$.

For general Σ, it is difficult to derive closed form analytic result for δ_c due to the complicated expression of (16) and LASSO problem (17). We provide Monte Carlo based numerical solutions in following section.

4. Numerical illustration

In this section, we present some numerical studies to support our theoretical results in Section 2 and Section 3. Our studies are based on simulations on finite size systems of moderate dimensions. We compute asymptotic LASSO risks and compare them with simulation results in Section 4.1. In Section 4.2, we verify our theoretical prediction on LASSO phase transition through Monte Carlo simulations. In Section 4.3, we study the dependence of LASSO phase transition on the covariance structure Σ and positive negative asymmetrical parameter Δ under various settings.

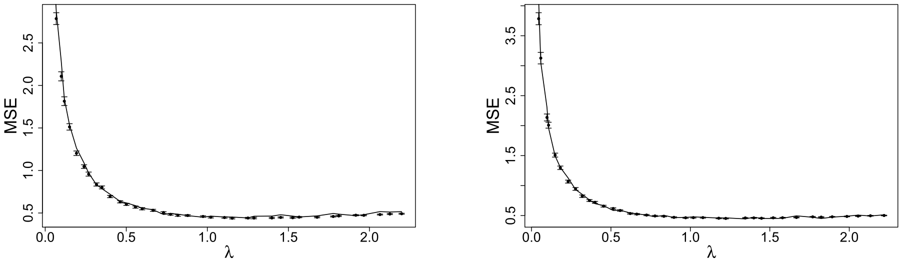

Fig 2. *LASSO MSE as a function of the regularization parameter* λ *compared to the asymptotic prediction. The solid curves represent theoretical prediction using (14) and the error bars are summaries over 100 simulated data with* $p = 400$. *Here the covariance matrix is block-diagonal with AR(1) block structure* $\Sigma_{s,ij} = \rho^{|i-j|}$ *with* $s = 2$ *and* $\rho = 0.5$. *The undersampling* $\delta = 1$ *and the sparsity* $\epsilon = 0.15$. *Left panel is for* $\Delta = 0$ *and right panel is for* $\Delta = 1$.

We consider block-diagonal covariance matrix with AR(1) block structure. For this choice, we can easily verify that Condition 5 is satisfied with limit

$$\lim_{p\to\infty} \mathcal{E}^{(p)}(a,b) = \frac{1}{s} E\left\{\frac{1}{2}\|\hat{\beta} - \beta_0 - \sqrt{a}\Sigma_s^{-\frac{1}{2}}\mathbf{z}\|^2_{\Sigma_s} + b\|\hat{\beta}\|_1\right\}, \tag{23}$$

where Σ_s is the block matrix with length s and $\hat{\beta} = \eta_b(\beta_0 + \sqrt{a}\Sigma_s^{-\frac{1}{2}}\mathbf{z})$ with covariance matrix Σ_s and $\mathbf{z} \sim N(0, \mathbf{I}_{s\times s})$. Similarly, we can verify that Condition 6 is also satisfied. We use $s = 2, 10, 20, 50$ in our numeric studies.

4.1. LASSO risk

We compute LASSO risk using (14) with $\tau_\star^2$ determined by solving the fixed point equation of $\tau^2 = \psi(\tau^2, \alpha(\lambda)\tau)$, where $\alpha(\lambda)$ is defined in (13). We use the bisection method to numerically solve the non-linear equation $f(\tau^2) = \tau^2 - \psi(\tau^2, \alpha\tau) = 0$.

For each setting, we generate 100 data sets with $p = 400$ consisting of design matrix $\mathbf{X} \sim N(0, \Sigma)$ and measurement vector $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{w}$ obtained from independent signal vector β_0 and independent noise vector $\mathbf{w}$. For each data set, we obtain the LASSO optimum estimator $\hat{\beta}(\lambda)$ using *glmnet*, an efficient package for fitting lasso or elastic-net regularization path for linear and generalized linear regression models. For each case, the dependence of MSE as a function of tuning parameter λ is plotted as shown in Figure 3. Here the random error $w_i \overset{i.i.d.}{\sim} N(0,1)$ and the magnitude for nonzero components of β_0 are sampled from uniform $[1, 2]$. The agreement is remarkably good already for p, n of a few hundred.

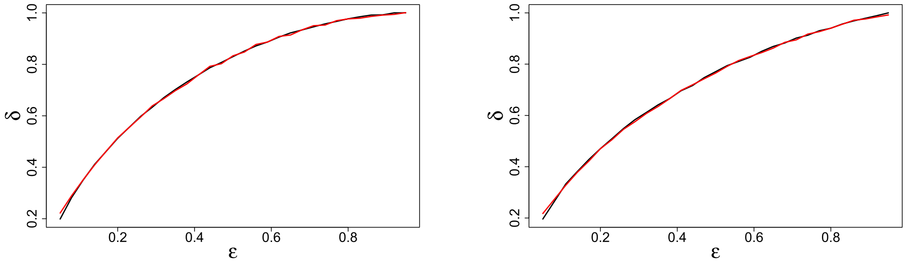

FIG 3. *Compare the theoretical phase transition curve with the one determined by simulation studies with $p = 500$. Here the covariance matrix is based on AR(1) model with $\rho = 0.5$. The black curves represent the theoretical estimations while the red curves represent the simulation results. Left panel is for $\Delta = 0$ and right panel is for $\Delta = 1$.*

4.2. Phase transition verification

For noiseless case, we compare the theoretical phase transition with the empirical one estimated by applying the following optimization algorithm to simulated data.

$$\begin{aligned}&\text{minimize}\|\beta\|_1,\\&\text{subject to } \mathbf{y} = \mathbf{X}\beta.\end{aligned}$$

Using the similar procedure as in [12], we first fix a grid of 31 ϵ values between 0.05 and 0.95. For each ϵ, we consider a series of δ values between $\max(0, \delta_c(\epsilon) - 0.2)$ and $\min(1, \delta_c(\epsilon) + 0.2)$, where $\delta_c(\epsilon)$ is the theoretically expected phase transition based on Theorem 2. We then have a grid of δ, ϵ values in parameter space $[0, 1]^2$. At each δ, ϵ, we generate $m = 100$ problem instances $(\mathbf{X}, \beta_0)$ with size $p = 500$. Then $\mathbf{y} = \mathbf{X}\beta_0$. For the ith problem instance, we obtain an output $\hat{\beta}_i$ by using the rq.lasso.fit function in package rqPen to the ith simulated data. We set the success indicator variable $S_i = 1$ if $\frac{\|\hat{\beta}_i - \beta_0\|}{\|\beta_0\|} \leq 10^{-4}$ and $S_i = 0$ otherwise. Then at each (δ, ϵ) combination, we have $S = \sum_{i=1}^m S_i$.

We analyze the simulated data-set to estimate the phase transition. At each fixed value of ϵ in our grid, we model the dependence of S on δ using logistic regression. We assume that S follows a binomial $B(\pi, 100)$ distribution with $\text{logit}(\pi) = a + b\delta$. We define the phase transition as the value of δ at which the success probability $\pi = 0.5$. In terms of the fitted parameters $\hat{a}, \hat{b}$, we have the estimated phase transition $\hat{\delta}(\epsilon) = -\hat{a}/\hat{b}$. Figure 3 shows that the agreement between the estimated phase transition curve based on the simulated finite-size systems and the analytical curve based on asymptotic theorem is remarkably good. We have tried different distributions for the random error $\mathbf{w}$ and nonzero components of β_0 and found that our phase transition results are dependent of those choices as illustrated by Theorem 2.

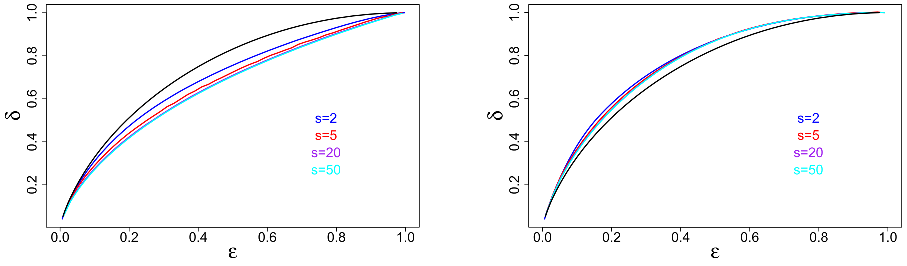

Fig 4. *The values of $\delta_c(\epsilon)$ as a function of ϵ for several different values of block length s with fixed $\Delta = 1$. Here the block covariance matrix is based on AR(1) model with $\Sigma_{s,ij} = \rho^{|i-j|}$. Left panel is for $\rho = 0.9$ and right panel is for $\rho = -0.9$.*

4.3. Phase transition under different dependent settings

In this section, we study the dependence of phase transition on the block length s, correlation coefficient ρ, and asymmetric coefficient Δ. Figure 4 shows the change of phase transition boundaries with the block length s for fixed $\Delta = 1$ and ρ. As s increases, the boundary moves further away from the i.i.d. boundary. For large s, in order to make a perfect recovery, less samples are needed under positive correlation $\rho = 0.9$ and more samples are needed under negative correlation $\rho = -0.9$. When s is big enough, e.g. $s = 20$, the boundaries only change slightly for further increasing of s.

Figure 5 shows the dependence of phase transition on ρ for fixed $s = 2$ and Δ. If the distribution of p_{β_0} is positive-negative symmetric, i.e. $\Delta = 0$, the boundaries are almost independent of ρ and very close to the Donoho-Tanner phase transition observed in [10] as illustrated by the left panel of Figure 5. If the distribution of the nonzero components of p_{β_0} is highly skewed, e.g. $\Delta = 1$, the phase transition curves fall below the Donoho-Tanner phase transition curve for $\rho > 0$ and above it for $\rho < 0$. As is clear from the right panel of Figure 5, for asymmetrically distributed signal β_0, the performance can be improved by increasing the correlation of covariance matrix Σ.

The phase transition curves for different Δ with fixed ρ are exhibited in Figure 6. For positive correlation, at the same sparsity level ϵ, the number of measurements δ that is required for successful recovery decreases as we increase Δ as shown by the left panel of Figure 6. For negative correlation, the conclusion is opposite as shown by the right panel of Figure 6.

5. Proof of the main results

We prove Theorem 1 using the limiting distribution of the approximate message passing (AMP) estimator. The AMP algorithm is a recently developed efficient iterative algorithm for solving the optimization problem (3). In order to define AMP algorithm, we need to use the soft-thresholding operation $\eta_\theta : \mathbb{R}^p \to \mathbb{R}^p$ defined in (5). For an arbitrary sequence of thresholds $\{\theta_t\}_{t\geq 0}$, the AMP

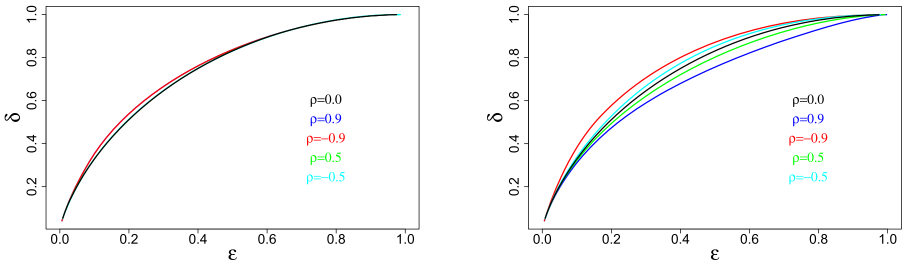

FIG 5. *The values of $\delta_c(\epsilon)$ as a function of ρ with fixed $s = 2$ and Δ. Left panel is for $\Delta = 0$ and right panel is for $\Delta = 1$.*

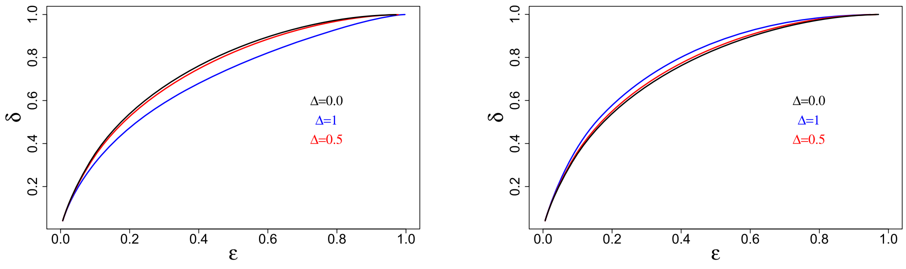

FIG 6. *The dependence of the phase transition curve on Δ for fixed $s = 2$ and ρ. Left panel is for $\rho = 0.9$ and right panel is for $\rho = -0.9$.*

constructs a sequence of estimates $\beta^t \in \mathbb{R}^p$, and residuals $\mathbf{z}^t \in \mathbb{R}^n$, according to the iteration

$$\begin{aligned} \beta^{t+1} &= \eta_{\theta_t}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}^t + \beta^t), \\ \mathbf{z}^t &= \mathbf{y} - \mathbf{X}\beta^t + \frac{1}{p\delta}\mathbf{z}^{t-1}\mathrm{div}\eta_{\theta_{t-1}}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}^{t-1} + \beta^{t-1}), \end{aligned} \tag{24}$$

where $\mathrm{div}\eta_\theta(\mathbf{v})$ is the divergence of the soft thresholding function. The algorithm (24) is mainly designed for theoretical analysis rather than practical use due to the fact that Σ is usually unknown. The following proposition shows the relation between the fixed-point solution of AMP algorithm (24) and the optimization solution of LASSO problem (3).

Proposition 3. *Any fixed point $\beta^t = \beta_\star, \mathbf{z}^t = \mathbf{z}_\star$ of the AMP iteration (24) with $\theta_t = \theta_\star$ is a minimizer of the LASSO cost function (3) with*

$$\lambda = \theta_\star \left\{ 1 - \frac{1}{p\delta} div\eta_{\theta_\star}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}_\star + \beta_\star) \right\}. \tag{25}$$

For a converging sequence of instances $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}$, the asymptotic behavior of the recursion (24) can be characterized as follows. Define the sequence $\{\tau_t^2\}_{t\geq 0}$ by setting $\tau_0^2 = \sigma_w^2 + \lim_{p\to\infty} E\{\|\beta_0\|_\Sigma^2\}/(p\delta)$ (for $\beta_0 \sim p_{\beta_0}$)

and letting, for all $t \geq 0$:

$$\tau_{t+1}^2 = \psi(\tau_t^2, \theta_t), \tag{26}$$

where the function $\psi(\cdot, \cdot)$ is defined in (6) which depends implicitly on the law p_{β_0}. The next proposition shows that the behavior of AMP can be tracked by the above one dimensional recursion which was often referred to as state evolution.

Proposition 4. *Let $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}_{p\in\mathbb{N}}$ be a converging sequence of instances and let sequence $\varphi_p : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, $p \geq 1$ be uniformly pseudo-Lipschitz functions. Then*

$$\varphi_p(\beta^{t+1}, \beta_0) \overset{P}{\approx} E\varphi_p(\eta_{\theta_t}(\beta_0 + \tau_t \Sigma^{-1/2}\mathbf{z}), \beta_0),$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of $\beta_0 \sim p_{\beta_0}$ and the sequence $\{\tau_t\}_{t\geq 0}$ is given by the recursion (26).

In order to establish the connection with LASSO, a specific policy has to be chosen for the thresholds $\{\theta_t\}_{t\geq 0}$. Throughout this paper we will take $\theta_t = \alpha\tau_t$ with α is fixed. The sequence $\{\tau_t\}_{t\geq 0}$ is given by the recursion

$$\tau_{t+1}^2 = \psi(\tau_t^2, \alpha\tau_t). \tag{27}$$

We prove Theorem 1 by proving the following result.

Theorem 3. *Assume the hypothesis of Theorem 1. Let $\hat{\beta}(\lambda; p)$ be the LASSO estimator for instance $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}$ and denote by $\{\beta^t(\alpha; p)\}_{t\geq 0}$ the sequence of estimators produced by AMP algorithm (24) with $\theta_t = \alpha(\lambda)\tau_t$, where $\alpha(\lambda)$ is the calibration mapping between α and λ defined in (13) and τ_t is updated by the recursion (27). Then*

$$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p} \|\beta^t(\alpha; p) - \hat{\beta}(\lambda; p)\|^2 = 0.$$

As mentioned by [6], Theorem 3 requires taking the limit of infinite dimensions $p \to \infty$ before the limit of an infinite number of $t \to \infty$.

6. Discussion

This paper focuses on the behavior of LASSO for learning the sparse coefficient vector in high-dimensional setting. We rigorously analyze the asymptotic behavior of LASSO for nonstandard Gaussian design models where the row of design matrix $\mathbf{X}$ are drawn independently from distribution $N(0, \Sigma)$. We first obtain the formula for the asymptotic mean square error (AMSE) characterized through a series of non-linear equations. Then we present an accurate characterization of the phase transition curve $\delta_c = \delta(\epsilon)$ for separating successful from unsuccessful reconstruction of β_0 by LASSO in the noiseless case $\mathbf{y} = \mathbf{X}\beta_0$. Our results show that the values of the non-zero elements of β_0 do not have any effect on the phase transition curve. However, for general Σ, the phase boundary δ_c

not only depends on the sparsity coefficient ϵ but also depends on the signed sparsity pattern of the nonzero components of β_0. This is in sharp contrast to the result for i.i.d. case where δ_c is completely determined by ϵ regardless of the distribution of β_0.

[30] shows that, in the noiseless setting, the l_q-regularized least squares exhibits the same phase transition for every $0 \leq q < 1$ and this phase transition is much better than that of LASSO. However, in the noisy setting, there is a major difference between the performance of l_p-regularized least squares with different values of q. For instance, $q = 0$ and $q = 1$ outperform the other values of q for very small and very large measurement noises. [28] further reveals some of the limitations and misleading features of the phase transition analysis. To overcome these limitations, they propose the small error analysis for l_q-regularized least squares to describe when phase transition analysis is reliable. [11] applied the AMP framework to a wider range of shrinkers including firm shrinkage and minimax shrinkage. Particularly, they show that the phase transition curve for AMP firm shrinkage and AMP minimax shrinkage are slightly better than that for LASSO.

An interesting future research direction is to generalize the results derived in [30, 28, 11] from the case of $\Sigma = \mathbf{I}_{p\times p}$ to the case of $\Sigma \neq \mathbf{I}_{p\times p}$. Our goal is to provide more accurate comparison for different regularizers in general setting for Σ. One of the major challenges in this direction is to establish the correspondence between regularized least square methods and specific AMP algorithms.

[23] introduces a class of generalized approximate message passing (GA MP) algorithms that cope with the case where the noisy measurement vector $\mathbf{y}$ can be non-linear function of the noiseless measurement $\mathbf{X}\beta_0$. [2] evaluate the asymptotic behavior of GLAM in standard Gaussian setting and locate the associated sharp phase transitions separating learnable and nonlearnable regions in phase space. Another interesting future direction is to generalize these GLM results from the case of i.i.d. design matrix to the case of general design matrix.

This work deals with the phase transition in noiseless case. For i.i.d. design matrix, [13] studied the phase transition behavior in the noisy case by introducing a quantity called noise sensitivity which is proportional to the mean-squared error of LASSO estimator. They found a boundary curve in the phase space $0 \leq \epsilon, \delta \leq 1$ such that the noise sensitivity is bounded above the curve and unbounded below the curve. This phase boundary is identical to the phase transition curve in the noiseless case for i.i.d. design. We plan to investigate if there is a similar phenomenon for LASSO phase transition with non-zero noise under non-i.i.d. design.

Appendix A: Proofs

A.1. Proof of Proposition 3

Proof. First we need to prove that the fixed-point of iteration (24) is a solution of (3). Toward this end, the first equation of (24) implies that

$$\Sigma\{\beta_\star - (\Sigma^{-1}\mathbf{X}^T\mathbf{z}_\star + \beta_\star)\} + \theta_\star\partial\|\beta_\star\|_1 = 0.$$

Therefore

$$\mathbf{X}^T\mathbf{z}_\star = \theta_\star \partial \|\beta_\star\|_1.$$

The second equation of (24) implies that

$$(1-\omega_\star)\,\mathbf{z}_\star = \mathbf{y} - \mathbf{X}\beta_\star,$$

where

$$\omega_\star = \frac{1}{p\delta}\text{div}\eta_{\theta_\star}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}_\star + \beta_\star). \tag{28}$$

Therefore

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta_\star) = \theta_\star\,(1-\omega_\star)\,\partial\|\beta_\star\|_1,$$

which is the solution of (3) for appropriately choosing tuning parameter $\lambda = \theta_\star(1-\omega_\star)$. □

A.2. Proof of Proposition 4

Proof. Since the entries of $\mathbf{X}$ are not i.i.d. normal, we do transformation $\tilde{\mathbf{X}} = \mathbf{X}\Sigma^{-1/2}$ and consider a different problem from (3)

$$\hat{\tilde{\beta}} = \text{argmin}_{\tilde{\beta}}\tilde{\mathcal{C}}(\tilde{\beta}), \tag{29}$$

where

$$\tilde{\mathcal{C}}(\tilde{\beta}) = \frac{1}{2}\|\mathbf{y} - \tilde{\mathbf{X}}\tilde{\beta}\|^2 + \lambda\|\Sigma^{-1/2}\tilde{\beta}\|_1.$$

Here the design matrix $\tilde{\mathbf{X}}$ has i.i.d. normal entries but the penalty term is not component-wise. The AMP algorithm for solving $\tilde{\beta}$ in (29) constructs a sequence of estimates $\tilde{\beta}^t \in \mathbb{R}^p$, and residuals $\mathbf{z}^t \in \mathbb{R}^n$, according to the iteration

$$\begin{aligned} \tilde{\beta}^{t+1} &= \tilde{\eta}_{\theta_t}(\tilde{\mathbf{X}}^T\mathbf{z}^t + \tilde{\beta}^t), \\ \mathbf{z}^t &= \mathbf{y} - \tilde{\mathbf{X}}\tilde{\beta}^t + \frac{1}{p\delta}\mathbf{z}^{t-1}\text{div}\tilde{\eta}_{\theta_{t-1}}(\tilde{\mathbf{X}}^T\mathbf{z}^{t-1} + \tilde{\beta}^{t-1}), \end{aligned} \tag{30}$$

initialized with $\tilde{\beta}^0 = 0 \in \mathbb{R}^p$, where

$$\tilde{\eta}_\theta(\mathbf{v}) = \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\beta - \mathbf{v}\|^2 + \theta\|\Sigma^{-1/2}\beta\|_1\right\}. \tag{31}$$

Comparing (31) and (5), we have

$$\begin{aligned} \tilde{\eta}_\theta(\mathbf{v}) &= \text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\Sigma^{-1/2}\beta - \Sigma^{-1/2}\mathbf{v}\|_\Sigma^2 + \theta\|\Sigma^{-1/2}\beta\|_1\right\} \\ &= \Sigma^{1/2}\eta_\theta(\Sigma^{-1/2}\mathbf{v}). \end{aligned}$$

Substituting $\beta = \Sigma^{-1/2}\tilde{\beta}$ into (30), the AMP update for β^{t+1} is

$$\beta^{t+1} = \Sigma^{-1/2}\tilde{\beta}^{t+1} = \Sigma^{-1/2}\tilde{\eta}_{\theta_t}(\tilde{\mathbf{X}}^T\mathbf{z}^t + \tilde{\beta}^t)$$

$$
\begin{aligned}
&= \eta_{\theta_t}(\Sigma^{-1/2}(\tilde{\mathbf{X}}^T\mathbf{z}^t + \tilde{\beta}^t)) = \eta_{\theta_t}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}^t + \beta^t) \\
\mathbf{z}^t &= \mathbf{y} - \mathbf{X}\beta^t + \frac{1}{p\delta}\mathbf{z}^{t-1}\mathrm{div}\eta_{\theta_{t-1}}\left(\Sigma^{-1/2}(\tilde{\mathbf{X}}^T\mathbf{z}^{t-1} + \tilde{\beta}^{t-1})\right) \\
&= \mathbf{y} - \mathbf{X}\beta^t + \frac{1}{p\delta}\mathbf{z}^{t-1}\mathrm{div}\eta_{\theta_{t-1}}\left(\Sigma^{-1}\mathbf{X}^T\mathbf{z}^{t-1} + \beta^{t-1}\right)
\end{aligned}
$$

which is equal to the AMP (24) constructed for solving the original problem (3).

The asymptotic property of AMP algorithm (30) has been established in [7]. It can be verified that the assumptions (C1)-(C6) of Theorem 14 in [7] are satisfied for the AMP iteration problem (30) by using the Conditions 1-6 introduced in the definition of converging sequences. More specifically, assumption (C1) is trivial. Assumptions (C3) and (C4) can be implied by Conditions 1 and 2 respectively. Assumption (C2) is satisfied due to the fact that both $1/\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ are bounded. Assumptions (C5) and (C6) can be implied by Condition 6. Applying Theorem 14 in [7], for any sequence $\tilde{\varphi}_p : (\mathbb{R}^p)^2 \to \mathbb{R}$, $p \geq 1$, of uniformly pseudo-Lipschitz functions, we obtain

$$
\tilde{\varphi}_p\left(\tilde{\beta}^{t+1}, \tilde{\beta}_0\right) \overset{P}{\approx} E\tilde{\varphi}_p\left(\tilde{\eta}_{\theta_t}(\tilde{\beta}_0 + \tau_t\mathbf{z}), \tilde{\beta}_0\right), \tag{32}
$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of $\tilde{\beta}_0$ and τ_t is determined by the following state evolution recursion

$$
\begin{aligned}
\tau_0^2 &= \sigma_w^2 + \frac{1}{p\delta}E\|\tilde{\beta}_0\|^2, \\
\tau_{t+1}^2 &= \sigma_w^2 + \frac{1}{p\delta}E\left(\|\tilde{\eta}_{\theta_t}(\tilde{\beta}_0 + \tau_t\mathbf{z}) - \tilde{\beta}_0\|^2\right),
\end{aligned}
$$

where $\tilde{\beta}_0 = \Sigma^{1/2}\beta_0$.

Define sequence of functions: $\tilde{\varphi}_p(\mathbf{x}, \mathbf{y}) = \varphi_p\left(\Sigma^{-1/2}\mathbf{x}, \Sigma^{-1/2}\mathbf{y}\right)$ which is also uniformly pseudo-Lipschitz due to the fact that $\Sigma^{-1/2}$ is well-conditioned. Therefore, the distributional limit of $\beta^{t+1} = \Sigma^{-1/2}\tilde{\beta}^{t+1}$ can be described by

$$
\begin{aligned}
\varphi_p\left(\beta^{t+1}, \beta_0\right) &= \varphi_p\left(\Sigma^{-1/2}\tilde{\beta}^{t+1}, \Sigma^{-1/2}\tilde{\beta}_0\right) = \tilde{\varphi}_p\left(\tilde{\beta}^{t+1}, \tilde{\beta}_0\right) \\
&\overset{P}{\approx} E\tilde{\varphi}_p\left(\tilde{\eta}_{\theta_t}(\tilde{\beta}_0 + \tau_t\mathbf{z}), \tilde{\beta}_0\right) \\
&= E\varphi_p\left(\Sigma^{-1/2}\tilde{\eta}_{\theta_t}(\tilde{\beta}_0 + \tau_t\mathbf{z}), \Sigma^{-1/2}\tilde{\beta}_0\right) \\
&= E\varphi_p\left(\eta_{\theta_t}(\Sigma^{-1/2}(\tilde{\beta}_0 + \tau_t\mathbf{z})), \beta_0\right) \\
&= E\varphi_p\left(\eta_{\theta_t}(\beta_0 + \tau_t\Sigma^{-1/2}\mathbf{z}), \beta_0\right). \quad \square
\end{aligned}
$$

A.3. Proof of Proposition 1

Proof. In order to prove Proposition 1, we need the following Lemma.

Lemma 1. *For any fixed $\alpha > 0$, the function $\psi(\tau^2, \alpha\tau)$ is strictly increasing and concave with respect to τ^2.*

We first prove that $f(\alpha) = 1$ has a unique solution when $\delta < 1$. From the definition (5), we get

$$\begin{aligned} \eta_\alpha(\Sigma^{-1/2}\mathbf{z}) &= \hat{\beta} \\ &= \operatorname{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\beta - \Sigma^{-1/2}\mathbf{z}\|_\Sigma^2 + \alpha\|\beta\|_1 \right\}, \end{aligned} \tag{33}$$

which is equivalent to the solution of LASSO problem with $\mathbf{X} = \Sigma^{1/2}$, $\mathbf{y} = \mathbf{z}$, and $\lambda = \alpha$. It can be easily verified that $f(\alpha) = \frac{1}{p\delta}E\|\hat{\mathbf{y}}\|^2 = \frac{1}{p\delta}E\|\mathbf{X}\hat{\beta}\|^2$ with $f(0) = 1/\delta$ and $f(\infty) = 0$. Thus in order to find unique solution of $f(\alpha) = 1$, it is enough if we can prove that $f(\alpha)$ is a strictly decreasing function. Denote $\mathcal{A} = \{j : \hat{\beta}_j \neq 0\}$ the active set of LASSO solution $\hat{\beta}$. From (33), we obtain

$$\Sigma(\hat{\beta} - \Sigma^{-1/2}\mathbf{z}) + \alpha\partial\|\hat{\beta}\|_1 = 0,$$

which implies

$$\Sigma_{\mathcal{A}\mathcal{A}}(\hat{\beta}_\mathcal{A} - (\Sigma^{-1/2}\mathbf{z})_\mathcal{A}) - \Sigma_{\mathcal{A}\mathcal{A}^c}(\Sigma^{-1/2}\mathbf{z})_{\mathcal{A}^c} + \alpha\text{sign}(\hat{\beta}_\mathcal{A}) = 0.$$

Taking derivative over α on both side, we obtain

$$\Sigma_{\mathcal{A}\mathcal{A}}\frac{\partial\hat{\beta}_\mathcal{A}}{\partial\alpha} + \text{sign}(\hat{\beta}_\mathcal{A}) + h(\alpha, \mathbf{z}) = 0,$$

where $h(\alpha, \mathbf{z})$ is the contribution from the changing of active set $\mathcal{A}$ with α. Since $\|\eta_\alpha(\Sigma^{-1/2}\mathbf{z})\|_\Sigma^2$ is continuous across the entire space $\mathbf{z} \in \mathbb{R}^p$, according to the discussion before (48) in Section A.6, the term $h(\alpha, \mathbf{z})$ disappears after taking expectation over $\mathbf{z}$. Therefore

$$\frac{df(\alpha)}{d\alpha} = \frac{2}{p\delta}E\left(\hat{\beta}_\mathcal{A}^T\Sigma_{\mathcal{A}\mathcal{A}}\frac{\partial\hat{\beta}_\mathcal{A}}{\partial\alpha}\right) = -\frac{2}{p\delta}E\left(\hat{\beta}_\mathcal{A}^T\text{sign}(\hat{\beta}_\mathcal{A})\right) < 0,$$

and we prove that $f(\alpha)$ is a decreasing function from $1/\delta$ to 0 as α increasing from 0 to ∞. Hence $f(\alpha) = 1$ has a unique solution denoted by $\alpha_{\min}$.

Next we prove that for fixed $\alpha > \alpha_{\min}$, the solution of equation (11) exists. According to the definition (6), we have

$$\lim_{\tau^2\to\infty} E\left(\|\eta_{\alpha\tau}(\beta_0 + \tau\Sigma^{-1/2}\mathbf{z}) - \beta_0\|_\Sigma^2\right) \to E\left(\|\eta_\alpha(\Sigma^{-1/2}\mathbf{z})\|_\Sigma^2\right)\tau^2,$$

which implies

$$\lim_{\tau^2\to\infty}\frac{\psi(\tau^2, \alpha\tau)}{\tau^2} = f(\alpha)$$

based on the definition (10). From Lemma 1, we have that $\psi(\tau^2, \alpha\tau)$ is strictly increasing and concave function. Further, we have $\psi(\tau^2, \alpha\tau)|_{\tau^2=0} = \sigma_w^2 > 0$.

Therefore, in order for the fixed point equation $\tau^2 = \psi(\tau^2, \alpha\tau)$ to have solutions, it is enough to show that $f(\alpha) < 1$ for $\alpha > \alpha_{\min}(\delta)$. This can be obtained from the fact that $f(\alpha)$ is decreasing and $f(\alpha_{\min}) = 1$. Thus we conclude that $\psi(\tau^2, \alpha\tau) < \tau^2$ as $\tau^2 \to \infty$ and prove that the solution of (11) exists and is unique. □

A.4. Proof of Proposition 2

Proof. Consider a system of equations

$$\begin{aligned} \tau^2 &= \psi(\tau^2, \theta), && (34) \\ \theta &= 1 - \frac{1}{\delta} E\left\langle \eta_\theta(\beta_0 + \tau\Sigma^{-1/2}\mathbf{z}), \mathbf{z} \right\rangle. && (35) \end{aligned}$$

According to Theorem 1 in [9], for $\sigma_w^2 > 0$, equations (34) and (35) have a unique solution denoted by $\tau^\star$, $\theta^\star$. Therefore, for any give λ, let $\alpha = \lambda/(\theta^\star\tau^\star)$, then α satisfies equation (12) and is also unique. □

A.5. Proof of Theorem 3

Proof. The proof of Theorem 3 is based on a series of Lemmas. The first Lemma implies that, asymptotically for large p, the AMP estimates converge.

Lemma 2. *The estimates $\{\beta^t\}_{t\geq 0}$ and residuals $\{\mathbf{z}^t\}_{t\geq 0}$ of AMP (24) almost surely satisfy*

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\beta^t - \beta^{t-1}\|^2 = 0, \lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\mathbf{z}^t - \mathbf{z}^{t-1}\|^2 = 0.$$

Denote $\sigma_{\min}(\mathbf{X})$ and $\sigma_{\max}(\mathbf{X})$ the maximum and minimum non-zero singular value of $\mathbf{X}$. Then the second Lemma implies that with high probability, $\sigma_{\min}(\mathbf{X})$ is lower bounded and $\sigma_{\max}(\mathbf{X})$ is upper bounded.

Lemma 3. *For every $t \geq 0$, there exists $c_5 > 0$ such that*

$$\mathbb{P}\left(c_5^{-1} \leq \sigma_{\min}(\mathbf{X}) \leq \sigma_{\max}(\mathbf{X}) \leq c_5\right) > 1 - 2\exp(-t^2/2).$$

According to the first equation of (24), denote the subgradient $\mathbf{v}^t \in \partial\|\beta^t\|_1$ such that

$$\Sigma\{\beta^t - (\Sigma^{-1}\mathbf{X}^T\mathbf{z}^{t-1} + \beta^{t-1})\} + \theta_{t-1}\mathbf{v}^t = 0. \tag{36}$$

Then the next Lemma implies that with high probability, the subgradient $\mathbf{v}^t$ cannot have too many coordinates with magnitude close to 1.

Lemma 4. *For large enough t, there exists $c, C, c_2 > 0$ such that*

$$\mathbb{P}\left(\frac{\left|j \in [p] : |v_j^t| \geq 1 - c_2\right|}{n} \geq 1 - \omega^\star/2\right) \leq C\exp(-cn),$$

where $\omega^\star$ is defined in (28) and

$$\omega^\star = \frac{1}{n} E\left(\left\| \eta_{\theta_\star}\left(\beta_0 + \tau_\star \Sigma^{-1/2}\mathbf{z}\right)\right\|_0 \right).$$

Define the minimum singular value of $\mathbf{X}$ over a set $S \subset [p]$ by

$$\kappa_-(\mathbf{X}, S) = \inf\left\{ \|\mathbf{X}\mathbf{w}\|_2 : \ \mathrm{supp}(\mathbf{w}) \subset S, \|\mathbf{w}\|_2 = 1 \right\},$$

and the s sparse singular value by

$$\kappa_-(\mathbf{X}, s) = \min_{|S| \le s} \kappa_-(\mathbf{X}, S).$$

Then the next Lemma implies that $\kappa_-(\mathbf{X}, s)$ is lower bounded with high probability.

Lemma 5. *For every $c_4 \ge 0$, there exists $C, c > 0$ such that*

$$\mathbb{P}\left(\kappa_-(\mathbf{X}, n(1 - \omega^\star/4)) \le c_4\right) \le Ce^{-cn}.$$

We are now ready to prove Theorem 3. The remainder of the argument takes place on the high-probability event determined by Lemmas 3, 4, and 5.

Let $\mathbf{r} = \hat{\beta} - \beta^t$ denote the distance between the LASSO optimum and the AMP estimate at t-th iteration, then

$$\begin{aligned} 0 &\ge \frac{\mathcal{C}(\beta^t + \mathbf{r}) - \mathcal{C}(\beta^t)}{p} \\ &= \frac{1}{2p}\|\mathbf{y} - \mathbf{X}(\beta^t + \mathbf{r})\|^2 + \frac{\lambda}{p}\|\beta^t + \mathbf{r}\|_1 - \frac{1}{2p}\|\mathbf{y} - \mathbf{X}\beta^t\|^2 - \frac{\lambda}{p}\|\beta^t\|_1 \\ &= \frac{1}{2p}\|\mathbf{X}\mathbf{r}\|^2 - \frac{\mathbf{r}^T\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta^t)}{p} + \frac{\lambda}{p}(\|\beta^t + \mathbf{r}\|_1 - \|\beta^t\|_1). \end{aligned}$$

Then by using equation (24) we have

$$0 \ge \underbrace{\frac{1}{2p}\|\mathbf{X}\mathbf{r}\|^2}_{\mathrm{I}} + \underbrace{\frac{1}{p}\langle \mathbf{r}, \mathrm{sg}\mathcal{C}(\beta^t)\rangle}_{\mathrm{II}} + \underbrace{\frac{\lambda}{p}(\|\beta^t + \mathbf{r}\|_1 - \|\beta^t\|_1 - \mathbf{r}^T\mathbf{v}^t)}_{\mathrm{III}}. \tag{37}$$

where the sub-gradient $\mathrm{sg}\mathcal{C}(\beta^t) = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta^t) + \lambda\mathbf{v}^t$ and $\mathbf{v}^t$ is defined in (36).

Let's first take a look at the second term of (37). Substituting (24) and $\mathbf{v}^t$ from (36), we obtain

$$\begin{aligned} \mathrm{sg}\mathcal{C}(\beta^t) &= \mathbf{X}^T(\omega_t\mathbf{z}^{t-1} - \mathbf{z}^t) - \frac{\lambda}{\theta_{t-1}}\{\Sigma(\beta^t - \beta^{t-1}) - \mathbf{X}^T\mathbf{z}^{t-1}\} \\ &= \frac{\lambda - \theta_{t-1}(1 - \omega_t)}{\theta_{t-1}}\mathbf{X}^T\mathbf{z}^{t-1} - \mathbf{X}^T(\mathbf{z}^t - \mathbf{z}^{t-1}) - \frac{\lambda}{\theta_{t-1}}\Sigma(\beta^t - \beta^{t-1}), \end{aligned}$$

where $\omega_t = \mathrm{div}\eta_{\theta_{t-1}}(\Sigma^{-1}\mathbf{X}^T\mathbf{z}^{t-1} + \beta^{t-1})/p/\delta$. Hence

$$\begin{aligned} \frac{1}{\sqrt{p}}\|\mathrm{sg}\mathcal{C}(\beta^t)\| \quad &\leq \quad \frac{|\lambda - \theta_{t-1}(1-\omega_t)|}{\theta_{t-1}}\sigma_{\max}(\mathbf{X})\frac{\|\mathbf{z}^{t-1}\|}{\sqrt{p}} + \sigma_{\max}(\mathbf{X})\frac{\|\mathbf{z}^t - \mathbf{z}^{t-1}\|}{\sqrt{p}} \\ &\quad + \frac{\lambda}{\theta_{t-1}}\sigma_{\max}(\Sigma)\frac{\|\beta^t - \beta^{t-1}\|}{\sqrt{p}}. \end{aligned}$$

By Lemmas 2, 3 and the fact that $\lambda_{\max}(\Sigma)$ is bounded as $p \to \infty$, we deduce that the last two terms converge to 0 as $p \to \infty$ and then $t \to \infty$. For the first term, using state evolution, we obtain $\frac{\|\mathbf{z}^{t-1}\|}{\sqrt{p}} = O(1)$. Finally, using the calibration relation (25), we get

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{|\lambda - \theta_{t-1}(1-\omega_t)|}{\theta_{t-1}} \overset{a.s.}{=} \frac{1}{\theta_\star}|\lambda - \theta_\star(1-\omega_\star)| = 0.$$

Therefore $\frac{1}{\sqrt{p}}\|\mathrm{sg}\mathcal{C}(\beta^t)\| \to 0$ almost surely. Since $\frac{\|\hat{\beta}\|}{\sqrt{p}} = O(1)$ and $\frac{\|\beta^t\|}{\sqrt{p}} = O(1)$, we get that $\frac{\|\mathbf{r}\|}{\sqrt{p}} = O(1)$ and hence the second term of (37) $\langle \mathbf{r}, \mathrm{sg}\mathcal{C}(\beta^t)\rangle \to 0$ almost surely. From (37), we have

$$\frac{1}{2p}\|\mathbf{X}\mathbf{r}\|^2 + \frac{\lambda}{p}(\|\beta^t + \mathbf{r}\|_1 - \|\beta^t\|_1 - \mathbf{r}^T\mathbf{v}^t) \leq c_1\varepsilon.$$

Both the first and third terms on the right-hand side of (37) are non-negative. The first one is trivial. Denote $S \equiv \{j \in \mathbb{N} : \beta_j^t \neq 0\}$ the support of β^t. The third one is non-negative since

$$\begin{aligned} & \|\beta^t + \mathbf{r}\|_1 - \|\beta^t\|_1 - \mathbf{r}^T\mathbf{v}^t \\ = \quad & \|\beta_S^t + \mathbf{r}_S\|_1 - \|\beta_S^t\|_1 - \mathbf{r}_S^T\mathrm{sign}(\beta_S^t) + \|\mathbf{r}_{\bar{S}}\|_1 - \mathbf{r}_{\bar{S}}^T\mathbf{v}_{\bar{S}}^t \\ = \quad & (\beta_S^t + \mathbf{r}_S)\{\mathrm{sign}(\beta_S^t + \mathbf{r}_S) - \mathrm{sign}(\beta_S^t)\} + \|\mathbf{r}_{\bar{S}}\|_1 - \mathbf{r}_{\bar{S}}^T\mathbf{v}_{\bar{S}}^t \geq 0. \end{aligned}$$

Since $(\beta_S^t + \mathbf{r}_S)\{\mathrm{sign}(\beta_S^t + \mathbf{r}_S) - \mathrm{sign}(\beta_S^t)\} \geq 0$ and $\|\mathbf{v}_{\bar{S}}^t\|_1 \leq 1$, we have

$$\begin{aligned} \frac{\|\mathbf{X}\mathbf{r}\|^2}{p} \quad &\leq \quad \xi_1(\varepsilon), && (38) \\ \|\mathbf{r}_{\bar{S}}\|_1 - \mathbf{r}_{\bar{S}}^T\mathbf{v}_{\bar{S}}^t \quad &\leq \quad p\xi_1(\varepsilon), && (39) \end{aligned}$$

where $\xi_1(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Consider $\mathbf{r} = \mathbf{r}^\perp + \mathbf{r}^\|$ with $\mathbf{r}^\| \in \ker(\mathbf{X})$ and $\mathbf{r}^\perp \perp \ker(\mathbf{X})$. It follows from (38) and Lemma 3 that

$$\|\mathbf{r}^\perp\|^2 \leq pc_5\xi_1(\varepsilon). \tag{40}$$

We need to prove an analogous bound for $\mathbf{r}^\|$. Note that $\|\mathbf{r}_{\bar{S}}^\perp\|_1 \leq \sqrt{p}\|\mathbf{r}_{\bar{S}}^\perp\|_2 \leq \sqrt{p}\|\mathbf{r}^\perp\|_2 \leq p\sqrt{\xi_1(\varepsilon)}$, from (39), we get

$$\|\mathbf{r}_{\bar{S}}^\|\|_1 - (\mathbf{r}_{\bar{S}}^\|)^T\mathbf{v}_{\bar{S}}^{t\|} \leq p\xi_2(\varepsilon). \tag{41}$$

Define $S(c_2) \equiv \{j \in \mathbb{N} : |v_j^t| \geq 1 - c_2\}$, then $\bar{S}(c_2) \subseteq \bar{S}$. We have

$$\|\mathbf{r}^{\|}_{\bar{S}}\|_1 - (\mathbf{r}^{\|}_{\bar{S}})^T \mathbf{v}^{t\|}_{\bar{S}} \geq \|\mathbf{r}^{\|}_{\bar{S}(c_2)}\|_1 - |\mathbf{r}^{\|}_{\bar{S}(c_2)}|^T |\mathbf{v}^{t\|}_{\bar{S}(c_2)}| \geq c_2 \|\mathbf{r}^{\|}_{\bar{S}(c_2)}\|_1.$$

Therefore using (41), we have

$$\|\mathbf{r}^{\|}_{\bar{S}(c_2)}\|_1 \leq c_2^{-1} p \xi_2(\varepsilon). \tag{42}$$

Denote $c_3 = \delta\omega^\star/4$. Then from Lemma 4, we have $|S(c_2)| \leq n - 2pc_3$. Thus if $|\bar{S}(c_2)| \leq pc_3/2$, one obtains $p \leq n - 3pc_3/2$. In this case, $\ker(\mathbf{X}) = \{0\}$ and the proof is concluded. Let us now consider the case $|\bar{S}(c_2)| \geq pc_3/2$. Then partition $\bar{S}(c_2) = \cup_{l=1}^K S_l$, where $pc_3/2 \leq |S_l| \leq pc_3$, and for each $i \in S_l$, $j \in S_{l+1}$, $|r_i^{\|}| \geq |r_j^{\|}|$. Also define $\bar{S}_+ \equiv \cup_{l=2}^K S_l \subseteq \bar{S}(c_2)$. Since, for any $i \in S_l$, $|r_i^{\|}| \leq \|\mathbf{r}^{\|}_{S_{l-1}}\|_1 / |S_{l-1}|$, we have

$$\begin{aligned}
\|\mathbf{r}^{\|}_{\bar{S}_+}\|_2^2 &= \sum_{l=2}^K \|\mathbf{r}^{\|}_{S_l}\|_2^2 \leq \sum_{l=2}^K |S_l| \left(\frac{\|\mathbf{r}^{\|}_{S_{l-1}}\|_1}{|S_{l-1}|} \right)^2 \\
&\leq \frac{4}{pc_3} \sum_{l=2}^K \|\mathbf{r}^{\|}_{S_{l-1}}\|_1^2 \leq \frac{4}{pc_3} \left(\sum_{l=2}^K \|\mathbf{r}^{\|}_{S_{l-1}}\|_1 \right)^2 \\
&\leq \frac{4}{pc_3} \|\mathbf{r}^{\|}_{\bar{S}(c_2)}\|_1^2 \leq \frac{4\xi_2(\varepsilon)^2}{c_2^2 c_3} p \equiv p\xi_3(\varepsilon).
\end{aligned} \tag{43}$$

To conclude the proof, it is sufficient to prove an analogous bound for $\|\mathbf{r}^{\|}_{S_+}\|_2^2$ with $S_+ = S(c_2) \cup S_1$. Since $|S_1| \leq pc_3$ and $|S(c_2)| \leq n - 2pc_3$, we have $|S_+| \leq n - pc_3$ and by Lemma 5 that $\sigma_{\min}(\mathbf{X}_{S_+}) \geq c_4$. Since $0 = \mathbf{X}\mathbf{r}^{\|} = \mathbf{X}_{S_+}\mathbf{r}^{\|}_{S_+} + \mathbf{X}_{\bar{S}_+}\mathbf{r}^{\|}_{\bar{S}_+}$, we have

$$c_4^2 \|\mathbf{r}^{\|}_{S_+}\|_2^2 \leq \|\mathbf{X}_{\bar{S}_+}\mathbf{r}^{\|}_{\bar{S}_+}\|_2^2 = \|\mathbf{X}_{S_+}\mathbf{r}^{\|}_{S_+}\|^2 \leq c_5 \|\mathbf{r}^{\|}_{\bar{S}_+}\|_2^2 \leq c_5 p \xi_3(\varepsilon). \tag{44}$$

Combining (40), (43), and (44), we conclude the proof. □

A.6. Proof of Theorem 2

Proof. Since there is no measurement noise, i.e. $\sigma_w^2 = 0$, we have $\psi(\tau^2, \alpha\tau)|_{\tau^2=0} = 0$. Thus in order for the fixed point equation $\tau^2 = \psi(\tau^2, \alpha\tau)$ to have unique solution $\tau_\star^2 = 0$, we need to have $\inf_\alpha \frac{d\psi(\tau^2,\alpha\tau)}{d\tau^2}|_{\tau^2=0} \leq 1$ due to the fact that $\psi(\tau^2, \alpha\tau)$ is a increasing and concave function of τ^2 for fixed α. Since $\psi(\tau^2, \alpha\tau)$ decreases with δ, the critical value δ_c is defined as

$$\delta_c = \inf \left\{ \delta : \inf_\alpha \frac{d\psi(\tau^2, \alpha\tau)}{d\tau^2}|_{\tau^2=0} \leq 1 \right\}. \tag{45}$$

Then for any $\delta > \delta_c$, we have unique solution $\tau_\star^2 = 0$; for any $\delta < \delta_c$, we also have solution $\tau_\star^2 > 0$. According to Theorem 1, we can consider the following solution

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\beta - \beta_0 - \tau \Sigma^{-1/2} \mathbf{z}\|_\Sigma^2 + \alpha\tau \|\beta\|_1 \right\},$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{p\times p})$ is independent of p_{β_0}. Define $\mathcal{A} = \{j : \hat{\beta}_j \neq 0\}$, then we have $\hat{\beta}_{\mathcal{A}^c} = 0$ and

$$\{\Sigma(\hat{\beta} - \beta_0) - \tau\Sigma^{1/2}\mathbf{z}\}_{\mathcal{A}} + \alpha\tau \text{sign}(\hat{\beta}_{\mathcal{A}}) = 0,$$

which implies

$$\Sigma_{\mathcal{A}\mathcal{A}}(\hat{\beta}_{\mathcal{A}} - \beta_{0,\mathcal{A}}) = \tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\tau\text{sign}(\hat{\beta}_{\mathcal{A}}) + \Sigma_{\mathcal{A}\mathcal{A}^c}\beta_{0,\mathcal{A}^c},$$

and

$$\begin{aligned} & \hat{\beta}_{\mathcal{A}} - \beta_{0,\mathcal{A}} \\ = \; & \Sigma_{\mathcal{A}\mathcal{A}}^{-1}\{\tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\tau\text{sign}(\hat{\beta}_{\mathcal{A}}) + \Sigma_{\mathcal{A}\mathcal{A}^c}\beta_{0,\mathcal{A}^c}\}. \end{aligned} \tag{46}$$

Substituting into the definition, we get

$$\begin{aligned} & E\{\|\hat{\beta} - \beta_0\|_\Sigma^2\} \\ = \; & E\{(\hat{\beta}_{\mathcal{A}} - \beta_{0,\mathcal{A}})^T \Sigma_{\mathcal{A}\mathcal{A}} (\hat{\beta}_{\mathcal{A}} - \beta_{0,\mathcal{A}}) \\ & - 2(\hat{\beta}_{\mathcal{A}} - \beta_{0,\mathcal{A}})^T \Sigma_{\mathcal{A}\mathcal{A}^c}\beta_{0,\mathcal{A}^c} + \beta_{0,\mathcal{A}^c}^T \Sigma_{\mathcal{A}^c\mathcal{A}^c}\beta_{0,\mathcal{A}^c}\} \\ = \; & E\{\tau^2((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}}))^T \Sigma_{\mathcal{A}\mathcal{A}}^{-1}((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}}))\} \\ & + E\{\beta_{0,\mathcal{A}^c}^T(\Sigma_{\mathcal{A}^c\mathcal{A}^c} - \Sigma_{\mathcal{A}^c\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\Sigma_{\mathcal{A}\mathcal{A}^c})\beta_{0,\mathcal{A}^c}\}. \end{aligned} \tag{47}$$

To perform the integrals over $\mathbf{z} \in \mathbb{R}^p$, we divide the p-dimensional space into regions such that the active set of $\hat{\beta}(\mathbf{z})$ keeps the same in each region and changes by one variable between two neighboring regions that share a common boundary hyperplane. In each region, the sign of $\beta(\mathbf{z})$ also keeps the same. A illustration of this space separation is shown in Figure 7 for a simple two dimensional example. Let S_i and S_j denote two neighboring regions that share a common hyperplane F_{ij} determined by equation $g_{ij}(\mathbf{z},\tau) = 0$ with $g_{ij}(\mathbf{z},\tau) > 0$ in S_i and $g_{ij}(\mathbf{z},\tau) < 0$ in S_j. Denote $f_i(\mathbf{z},\tau)$ the function form of $\|\hat{\beta}(\mathbf{z},\tau) - \beta_0\|_\Sigma^2$ in region S_i. Then $f_i(\mathbf{z},\tau)$ is differentiable over τ^2 inside S_i and the derivative of $Ef_i(\mathbf{z},\tau)I(\mathbf{z} \in S_i)$ over τ^2 involves integrals over face F_{ij} with respect to $d-1$ dimensional measure $\sigma_{F_{ij}}(\cdot)$. An application of Stokes's theorem, as in Theorem 1 of [1], establishes differentiability of this integral which is given by $\sigma_{F_{ij}}(f_i(\mathbf{z},\tau)\frac{\partial g_{ij}(\mathbf{z},\tau)}{\partial\tau^2})$. Similarly, we can obtain the boundary contribution of F_{ij} to the derivative of $Ef_j(\mathbf{z},\tau)I(\mathbf{z} \in S_j)$ over τ^2 which is given by $-\sigma_{F_{ij}}(f_j(\mathbf{z},\tau)\frac{\partial g_{ij}(\mathbf{z},\tau)}{\partial\tau^2})$. Since $\hat{\beta}(\mathbf{z},\tau)$ is continuous across F_{ij}, we have $f_i(\mathbf{z},\tau) = f_j(\mathbf{z},\tau)$ on F_{ij} and thus the contributions of the boundary effects due to F_{ij} cancel each other between the derivative of $Ef_i(\mathbf{z},\tau)I(\mathbf{z}\in S_i)$ over τ^2 and the derivative of $Ef_j(\mathbf{z},\tau)I(\mathbf{z}\in S_j)$ over τ^2. Therefore, in taking derivative over τ^2 for (47), the boundary effects are canceled and one gets

$$\begin{aligned} & \frac{d\psi(\tau^2,\alpha\tau)}{d\tau^2} \\ = & \lim_{p\to\infty}\frac{1}{p\delta}E\{((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}}-\alpha\text{sign}(\hat\beta_{\mathcal{A}}))^T\Sigma^{-1}_{\mathcal{A}\mathcal{A}}((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}}-\alpha\text{sign}(\hat\beta_{\mathcal{A}}))\}\end{aligned} \tag{48}$$

which only depends on the sign of the non-zero components of $\hat\beta$.

We need to consider situations as $\tau^2\to 0$. Since $\hat\beta\to\beta_0$, we have $\hat\beta=\beta_0+o_P(1)$ as $\tau^2\to 0$. Let $\mathcal{B}=\{j:\beta_{0,j}\neq 0\}$, clearly $\mathcal{B}\subseteq\mathcal{A}$ and $\mathcal{A}^c\subseteq\mathcal{B}^c$ as $\tau^2\to 0$. For $\mathcal{B}$ part, from (46), we obtain

$$\{\Sigma(\hat\beta-\beta_0)-\tau\Sigma^{1/2}\mathbf{z}\}_{\mathcal{B}}+\alpha\tau\text{sign}(\hat\beta_{\mathcal{B}})=0,$$

which implies

$$\begin{aligned} & \Sigma_{\mathcal{B}\mathcal{B}}(\hat\beta_{\mathcal{B}}-\beta_{0,\mathcal{B}}) \\ = & \tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}}-\alpha\tau\text{sign}(\hat\beta_{\mathcal{B}})-\Sigma_{\mathcal{B}\mathcal{B}^c}\hat\beta_{\mathcal{B}^c},\end{aligned}$$

and thus

$$\begin{aligned} & \hat\beta_{\mathcal{B}}-\beta_{0,\mathcal{B}} \\ = & \Sigma^{-1}_{\mathcal{B}\mathcal{B}}\{\tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}}-\alpha\tau\text{sign}(\hat\beta_{\mathcal{B}})-\Sigma_{\mathcal{B}\mathcal{B}^c}\hat\beta_{\mathcal{B}^c}\}.\end{aligned} \tag{49}$$

For $\mathcal{B}^c$ part, we have

$$\{\Sigma(\hat\beta-\beta_0)-\tau\Sigma^{1/2}\mathbf{z}\}_{\mathcal{B}^c}+\alpha\tau\partial\|\hat\beta_{\mathcal{B}^c}\|_1=0$$

which implies

$$\begin{aligned} & \Sigma_{\mathcal{B}^c\mathcal{B}}(\hat\beta_{\mathcal{B}}-\beta_{0,\mathcal{B}}) \\ = & \tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}^c}-\alpha\tau\partial\|\hat\beta_{\mathcal{B}^c}\|_1-\Sigma_{\mathcal{B}^c\mathcal{B}^c}\hat\beta_{\mathcal{B}^c}.\end{aligned}$$

Using (49), we have

$$\begin{aligned} & \Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}}\{\tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}}-\alpha\tau\text{sign}(\hat\beta_{\mathcal{B}})-\Sigma_{\mathcal{B}\mathcal{B}^c}\hat\beta_{\mathcal{B}^c}\} \\ = & \tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}^c}-\alpha\tau\partial\|\hat\beta_{\mathcal{B}^c}\|_1-\Sigma_{\mathcal{B}^c\mathcal{B}^c}\hat\beta_{\mathcal{B}^c}.\end{aligned}$$

The final equation for $\hat\beta_{\mathcal{B}^c}$ is

$$\begin{aligned} & (\Sigma_{\mathcal{B}^c\mathcal{B}^c}-\Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}^c})\hat\beta_{\mathcal{B}^c}-\tau(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}^c} \\ & +\tau\Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma^{-1}_{\mathcal{B}\mathcal{B}}\{(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}}-\alpha\text{sign}(\hat\beta_{\mathcal{B}})\}+\alpha\tau\partial\|\hat\beta_{\mathcal{B}^c}\|_1=0.\end{aligned}$$

Therefore $\hat\beta_{\mathcal{B}^c}$ is equivalent to the solution of the following LASSO problem

$$\hat\beta_{\mathcal{B}^c}=\text{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1\right\}$$

with

$$\begin{aligned}\mathbf{X} &= (\Sigma_{\mathcal{B}^c\mathcal{B}^c} - \Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\Sigma_{\mathcal{B}\mathcal{B}^c})^{1/2},\\ \mathbf{y} &= \tau\mathbf{X}^{-1}\left[(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}^c} - \Sigma_{\mathcal{B}^c\mathcal{B}}\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\{(\Sigma^{1/2}\mathbf{z})_{\mathcal{B}} - \alpha\text{sign}(\beta_{0,\mathcal{B}})\}\right],\end{aligned}$$

and $\lambda = \alpha\tau$. Since (48) only involves the sign of $\hat{\beta}$, without loss of generality, we can take $\tau = 1$. Therefore $\hat{\beta}_{\mathcal{B}^c}$ is independent of the actual distribution of β_0 but depends on ϵ and Δ. Denote $\bar{\mathcal{B}} = \{j : j \in \mathcal{B}^c \text{ and } \hat{\beta}_j \neq 0\}$, then we have $\mathcal{A} = \mathcal{B} \cup \bar{\mathcal{B}}$. Define function

$$M(\epsilon, \Delta, \alpha) = \lim_{p\to\infty}\frac{1}{p}E\{((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}}))^T\Sigma_{\mathcal{A}\mathcal{A}}^{-1}((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}}))\},$$

which exists according to Condition 5. Substituting (48) into (45), we obtain

$$\delta_c = \inf_{\alpha} M(\epsilon, \Delta, \alpha).$$

A.7. Proof of Lemma 1

Proof. For fixed α, in order to prove that $\psi(\tau^2, \alpha\tau)$ is an increasing and concave function of τ^2, we need to show that $\frac{d\psi(\tau^2,\alpha\tau)}{d\tau^2} > 0$ and $\frac{d^2\psi(\tau^2,\alpha\tau)}{(d\tau^2)^2} < 0$. Since $\Sigma_{\mathcal{A}\mathcal{A}}^{-1}$ is positive definite, from (48), we get $\frac{d\psi(\tau^2,\alpha\tau)}{d\tau^2} > 0$ and prove that $\psi(\tau^2, \alpha\tau)$ is an increasing function of τ^2.

We need to take further derivative over τ^2 to obtain $\frac{d^2\psi(\tau^2,\alpha\tau)}{(d\tau^2)^2}$. Toward this end, consider the LASSO problem with

$$\mathbf{X} = \Sigma^{1/2}, \mathbf{y} = \tau\mathbf{z} + \Sigma^{1/2}\beta_0,$$

and $\lambda = \alpha\tau$. Following the discussion in deriving (48), we can divide the p-dimensional space $\mathbf{z} \in \mathbb{R}^p$ into regions such that the active set and the sign of each variable are fixed in each region. Denote by $\mathcal{A}_i$ and $\mathcal{A}_j$ the active sets in two neighboring regions S_i and S_j respectively. Further denote by F_{ij} the boundary hyperplane between S_i and S_j. Assume that $|\mathcal{A}_i| = k$, $|\mathcal{A}_j| = k - 1$, and denote $\mathbf{x}_k$ the active variable that drops when moving from S_i to S_j. Therefore, $\mathcal{A}_j \subset \mathcal{A}_j$ and $\mathcal{A}_i \setminus \mathcal{A}_j = \mathbf{x}_k$. Then, from (46), we obtain that the solution of $\hat{\beta}$ inside S_i is differentiable over τ^2 and can be written as $\hat{\beta}_{\mathcal{S}_i} = \Sigma_{\mathcal{A}_i\mathcal{A}_i}^{-1}\left\{(\Sigma^{1/2}\mathbf{y})_{\mathcal{A}_i} - \alpha\tau\text{sign}(\hat{\beta}_{\mathcal{S}_i})\right\}$. Assume that the k-th component of $\hat{\beta}_{\mathcal{S}_i}$, i.e. $\hat{\beta}_{\mathcal{S}_i}[k] > 0$ in S_i and $\hat{\beta}_{\mathcal{S}_i}[k] = 0$ in S_j, then the boundary hyperplane F_{ij} is determined by equation

$$g_{ij}(\mathbf{z}, \tau) = \mathbf{e}_{(k)}^T\hat{\beta}_{\mathcal{S}_i} = \mathbf{e}_{(k)}^T\Sigma_{\mathcal{A}_i\mathcal{A}_i}^{-1}\left\{(\Sigma^{1/2}\mathbf{y})_{\mathcal{A}_i} - \alpha\tau\text{sign}(\hat{\beta}_{\mathcal{S}_i})\right\} = 0, \tag{50}$$

where $\mathbf{e}_{(k)}$ represents the k-th coordinate vector for $\hat{\beta}_{\mathcal{A}_i}$. Denote by $\bar{S}_i$ and $\bar{S}_j$ the other two neighboring regions that have the same active sets but opposite sign of variables comparing to S_i to S_j, i.e. $\text{sign}(\hat{\beta}_{\bar{\mathcal{S}}_i}) = -\text{sign}(\hat{\beta}_{\mathcal{S}_i})$ and $\text{sign}(\hat{\beta}_{\bar{\mathcal{S}}_j}) = -\text{sign}(\hat{\beta}_{\mathcal{S}_j})$. Then their boundary hyperplane $\bar{F}_{ij}$ is determined by equation

$$\bar{g}_{ij}(\mathbf{z},\tau) = \mathbf{e}_{(k)}^T\hat{\beta}_{\bar{\mathcal{S}}_i} = \mathbf{e}_{(k)}^T\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}\left\{(\Sigma^{1/2}\mathbf{y})_{\mathcal{A}_i} - \alpha\tau\text{sign}(\hat{\beta}_{\bar{\mathcal{S}}_i})\right\} = 0,$$

Denote $f_i(\mathbf{z},\tau)$ the integrand inside the expectation on the right hand side of (48) in region S_i, i.e.

$$f_i(\mathbf{z},\tau) = ((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i}))^T\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}((\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i})),$$

which does not depend on τ^2 explicitly, thus the dependence of the expectation on τ^2 only comes from the boundary effects. From (47), the continuity of $\|\hat{\beta}(\mathbf{z},\tau) - \beta_0\|^2_\Sigma$ leads to

$$\begin{aligned} & \tau^2[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i})]^T\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i})] \\ & +\beta^T_{0,\mathcal{A}_i^c}(\Sigma_{\mathcal{A}_i^c\mathcal{A}_i^c} - \Sigma_{\mathcal{A}_i^c\mathcal{A}_i}\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}\Sigma_{\mathcal{A}_i\mathcal{A}_i^c})\beta_{0,\mathcal{A}_i^c} \\ = \; & \tau^2[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_j} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_j})]^T\Sigma^{-1}_{\mathcal{A}_j\mathcal{A}_j}[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_j} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_j})] \\ & +\beta^T_{0,\mathcal{A}_j^c}(\Sigma_{\mathcal{A}_j^c\mathcal{A}_j^c} - \Sigma_{\mathcal{A}_j^c\mathcal{A}_j}\Sigma^{-1}_{\mathcal{A}_j\mathcal{A}_j}\Sigma_{\mathcal{A}_j\mathcal{A}_j^c})\beta_{0,\mathcal{A}_j^c}. \end{aligned}$$

Therefore, the difference of the integrand function on (48) caused by the change of active set from region S_i to region S_j can be written as

$$\begin{aligned} \Delta_{ij} &= f_i(\mathbf{z},\tau) - f_j(\mathbf{z},\tau) \\ &= [(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i})]^T\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_i} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_i})] \\ &\quad -[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_j} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_j})]^T\Sigma^{-1}_{\mathcal{A}_j\mathcal{A}_j}[(\Sigma^{1/2}\mathbf{z})_{\mathcal{A}_j} - \alpha\text{sign}(\hat{\beta}_{\mathcal{A}_j})] \\ &= \{\beta^T_{0,\mathcal{A}_j^c}(\Sigma_{\mathcal{A}_j^c\mathcal{A}_j^c} - \Sigma_{\mathcal{A}_j^c\mathcal{A}_j}\Sigma^{-1}_{\mathcal{A}_j\mathcal{A}_j}\Sigma_{\mathcal{A}_j\mathcal{A}_j^c})\beta_{0,\mathcal{A}_j^c} \\ &\quad -\beta^T_{0,\mathcal{A}_i^c}(\Sigma_{\mathcal{A}_i^c\mathcal{A}_i^c} - \Sigma_{\mathcal{A}_i^c\mathcal{A}_i}\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}\Sigma_{\mathcal{A}_i\mathcal{A}_i^c})\beta_{0,\mathcal{A}_i^c}\}/\tau^2, \end{aligned} \tag{51}$$

which only depends on the active sets $\mathcal{A}_i$ and $\mathcal{A}_j$. Therefore, we also have $\bar{\Delta}_{ij} = \Delta_{ij}$, where $\bar{\Delta}_{ij}$ represents the difference of the integrand function caused by the change of active set from region $\bar{S}_i$ to region $\bar{S}_j$.

According to Stokes's theorem shown in Theorem 1 of [1], the contribution of boundary F_{ij} to the derivative of integral $Ef_i(\mathbf{z},\tau)I(\mathbf{z}\in S_i)+Ef_j(\mathbf{z},\tau)I(\mathbf{z}\in S_j)$ over τ^2 is given by $\sigma_{F_{ij}}(\Delta_{ij}\frac{\partial g_{ij}(\mathbf{z},\tau)}{\partial\tau^2})$. Similarly, we derive that the contribution of boundary $\bar{F}_{ij}$ to the derivative of integral $E\bar{f}_i(\mathbf{z},\tau)I(\mathbf{z}\in\bar{S}_i)+E\bar{f}_j(\mathbf{z},\tau)I(\mathbf{z}\in\bar{S}_j)$ over τ^2 is given by $-\sigma_{\bar{F}_{ij}}(\Delta_{ij}\frac{\partial\bar{g}_{ij}(\mathbf{z},\tau)}{\partial\tau^2})$. Define

$$\begin{aligned} \mathbf{a}_k &= \Sigma^{1/2}_{,\mathcal{A}_i}\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}\mathbf{e}_{(k)}, \\ b_k &= \mathbf{e}^T_{(k)}\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}\text{sign}(\hat{\beta}_{\mathcal{S}_i}), \\ c_k &= \mathbf{e}^T_{(k)}\Sigma^{-1}_{\mathcal{A}_i\mathcal{A}_i}(\Sigma\beta_0)_{\mathcal{A}_i}. \end{aligned}$$

Then from (50), we have $g_{ij}(\mathbf{z},\tau) = \tau\mathbf{a}_k^T\mathbf{z} - \alpha\tau b_k + c_k$. Therefore, $\frac{\partial g_{ij}(\mathbf{z},\tau)}{\partial\tau^2} = \frac{1}{2\tau}(\mathbf{a}_k^T\mathbf{z} - \alpha b_k)$. We obtain the boundary contributions of F_{ij} and $\bar{F}_{ij}$ as

$$\sigma_{F_{ij}}\left\{\frac{\Delta_{ij}c_k}{2\tau^3\|\mathbf{a}_k\|}\left[\phi\left(\frac{c_k}{\tau\|\mathbf{a}_k\|}+\frac{\alpha b_k}{\|\mathbf{a}_k\|}\right)-\phi\left(\frac{c_k}{\tau\|\mathbf{a}_k\|}-\frac{\alpha b_k}{\|\mathbf{a}_k\|}\right)\right]\right\}.$$

From (51), since $\mathcal{A}_j \subset \mathcal{A}_i$, we get $\mathcal{A}_j^c \supset \mathcal{A}_i^c$ and hence $\Delta_{ij} \geq 0$. Then we conclude that the boundary contribution is less than or equal to zero since $x(\phi(x+c)-\phi(x-c)) \leq 0$ for any x and $c \geq 0$. This complete the proof of the concavity of function $\psi(\tau^2, \alpha\tau)$. □

A.8. Proof of Lemma 2

Proof. We begin with the convergence of the state evolution (27) iteration described by the following lemma which can be immediately proved using the concavity of $\psi(\tau^2, \alpha\tau)$ over τ^2.

Lemma 6. *For any $\alpha \geq \alpha_{\min}$. The iteration (27) converges to the unique solution of the fixed-point equation $\tau_\star^2 = \psi(\tau_\star^2, \alpha\tau_\star)$, i.e. $\tau_t^2 \to \tau_\star^2$ as $t \to \infty$.*

Next we need to generalize state evolution to compute large system limits for functions of β^t, β^s, with $t \neq s$. To this purpose, we define the covariances $\{\tau_{s,t}\}_{s,t\geq 0}$ recursively by

$$\begin{aligned}\tau_{s+1,t+1} \quad = \quad & \sigma_w^2 + \lim_{p\to\infty}\frac{1}{p\delta}E\Big\{[\eta_{\theta_s}(\beta_0+\Sigma^{-1/2}\mathbf{z}_s)-\beta_0]^T\Sigma \\ & \qquad [\eta_{\theta_t}(\beta_0+\Sigma^{-1/2}\mathbf{z}_t)-\beta_0]\Big\},\end{aligned} \tag{52}$$

where $(\mathbf{z}_s, \mathbf{z}_t)$ jointly Gaussian, independent from $\beta_0 \sim p_{\beta_0}$ with mean 0 and covariance given by $E(\mathbf{z}_s\mathbf{z}_s^T) = \tau_{s,s}\mathbf{I}_{p\times p} = \tau_s^2\mathbf{I}_{p\times p}$, $E(\mathbf{z}_t\mathbf{z}_t^T) = \tau_{t,t}\mathbf{I}_{p\times p} = \tau_t^2\mathbf{I}_{p\times p}$, and $E(\mathbf{z}_s\mathbf{z}_t^T) = \tau_{s,t}\mathbf{I}_{p\times p}$. The boundary condition is fixed by letting $\tau_{0,0} = \sigma_w^2 + E\{\|\beta\|_\Sigma^2\}/\delta$ and $\tau_{0,1} = \sigma_w^2 + \lim_{p\to\infty} E\{[\beta_0 - \eta_{\theta_0}(\beta_0 + \Sigma^{-1/2}\mathbf{z}_0)]^T\Sigma\beta_0\}/p/\delta$. With this definition, we have the following generalization of Proposition 1.

Lemma 7. *Let $\{\beta_0(p), \mathbf{w}(p), \Sigma(p), \mathbf{X}(p)\}_{p\in\mathbb{N}}$ be a converging sequence of instances and let sequence $\varphi_p : (\mathbb{R}^p)^3 \to \mathbb{R},\ p \geq 1$ be uniformly pseudo-Lipschitz functions. Then for all $s \geq 0$ and $t \geq 0$, we get*

$$\varphi_p(\beta^{s+1}, \beta^{t+1}, \beta_0) \overset{P}{\sim} \varphi_p(\eta_{\theta_s}(\beta_0 + \Sigma^{-1/2}\mathbf{z}_s), \eta_{\theta_t}(\beta_0 + \Sigma^{-1/2}\mathbf{z}_t), \beta_0),$$

where $(\mathbf{z}_s, \mathbf{z}_t)$ jointly Gaussian, independent from $\beta_0 \sim p_{\beta_0}$ with mean 0 and covariance given by $E(\mathbf{z}_s\mathbf{z}_s^T) = \tau_s^2\mathbf{I}_{p\times p}$, $E(\mathbf{z}_t\mathbf{z}_t^T) = \tau_t^2\mathbf{I}_{p\times p}$, and $E(\mathbf{z}_s\mathbf{z}_t^T) = \tau_{s,t}\mathbf{I}_{p\times p}$. The recursion $\tau_{s,,t}$ for all $s, t \geq 0$ is determined by (6) and (52).

Proof of Lemma 2. Define sequence of $\{y_t\}_{t\geq 0}$ as

$$y_t = \lim_{p\to\infty}\frac{1}{p\delta}E\left\|\eta_{\theta_t}(\beta_0 + \Sigma^{-1/2}\mathbf{z}_t) - \eta_{\theta_{t-1}}(\beta_0 + \Sigma^{-1/2}\mathbf{z}_{t-1})\right\|_\Sigma^2.$$

From (52), we have

$$y_t = \tau_t^2 + \tau_{t-1}^2 - 2\tau_{t,t-1}. \tag{53}$$

Take $\theta_t = \alpha\tau_t$ with α is fixed, then according to Lemma 6, we have $\tau_t^2 \to \tau_\star^2$ and $\theta_t \to \theta_\star = \alpha\tau_\star$ as $t \to \infty$. We will show that $y_t \to 0$ which in turn yields $\tau_{t,t-1} \to \tau_\star^2$ based on (53). For large enough t, we have the representation as follows in terms of the two independent random vectors $\mathbf{z}, \mathbf{w} \sim N(0, \mathbf{I}_{p\times p})$:

$$\begin{aligned} y_t \quad = \quad & \lim_{p\to\infty} \frac{1}{p\delta} E \left\| \eta_{\theta_\star}\left(\beta_0 + \sqrt{\tau_\star^2 - \frac{y_{t-1}}{4}}\Sigma^{-1/2}\mathbf{z} + \sqrt{\frac{y_{t-1}}{4}}\Sigma^{-1/2}\mathbf{w}\right)\right. \\ & \left. -\eta_{\theta_\star}\left(\beta_0 + \sqrt{\tau_\star^2 - \frac{y_{t-1}}{4}}\Sigma^{-1/2}\mathbf{z} - \sqrt{\frac{y_{t-1}}{4}}\Sigma^{-1/2}\mathbf{w}\right)\right\|_\Sigma^2 . \end{aligned}$$

Consider y_t as a function of y_{t-1} denoted by $y_t = R(y_{t-1})$. A straightforward calculation yields

$$\begin{aligned} R'(y_{t-1}) \quad = \quad & \lim_{p\to\infty} \frac{1}{p\delta} E\left(Tr\left[\Sigma^{-1}\left\{\nabla\eta_{\theta_\star}\left(\beta_0 + \Sigma^{-1/2}\mathbf{z}_t\right)\right\}^T\right.\right. \\ & \left.\left. \Sigma\nabla\eta_{\theta_\star}\left(\beta_0 + \Sigma^{-1/2}\mathbf{z}_{t-1}\right)\right]\right), \end{aligned}$$

where

$$\mathbf{z}_t = \sqrt{\tau_\star^2 - \frac{y_{t-1}}{4}}\mathbf{z} + \sqrt{\frac{y_{t-1}}{4}}\mathbf{w}, \mathbf{z}_{t-1} = \sqrt{\tau_\star^2 - \frac{y_{t-1}}{4}}\mathbf{z} - \sqrt{\frac{y_{t-1}}{4}}\mathbf{w},$$

and ∇ denotes the vector differential operator. For $y_{t-1} = 0$, we have $\mathbf{z}_t = \mathbf{z}_{t-1}$ and

$$R'(0) = \lim_{p\to\infty} \frac{1}{p\delta} E\left(Tr\left[\Sigma^{-1}\left\{\nabla\hat{\eta}\right\}^T \Sigma\nabla\hat{\eta}\right]\right), \tag{54}$$

where $\hat{\eta} = \eta_{\alpha\tau_\star}\left(\beta_0 + \tau_\star\Sigma^{-1/2}\mathbf{z}\right)$. Denote $\mathcal{A} = \{j : \hat{\eta}_j \neq 0\}$. From the definition (5), we get

$$\Sigma(\hat{\eta} - (\beta_0 + \Sigma^{-1/2}\mathbf{z})) + \theta_\star\partial\|\hat{\eta}\|_1 = 0,$$

which implies that

$$\Sigma_{\mathcal{A}\mathcal{A}}(\hat{\eta}_\mathcal{A} - (\beta_0 + \Sigma^{-1/2}\mathbf{z})_\mathcal{A}) - \Sigma_{\mathcal{A}\mathcal{A}^c}(\beta_0 + \Sigma^{-1/2}\mathbf{z})_{\mathcal{A}^c} + \theta_\star \text{sign}(\hat{\eta}_\mathcal{A})) = 0.$$

Taking derivatives, we obtain

$$\Sigma_{\mathcal{A}\mathcal{A}}(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}} = \Sigma_{\mathcal{A}\mathcal{A}}\text{and}\Sigma_{\mathcal{A}\mathcal{A}}(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}^c} = \Sigma_{\mathcal{A}\mathcal{A}^c}.$$

Substituting into (54), we obtain

$$\begin{aligned} & R'(0) \\ = \quad & \lim_{p\to\infty} \frac{1}{p\delta} E\left(Tr\left[\left\{(\Sigma^{-1})_{\mathcal{A}\mathcal{A}}\left[(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}}\right]^T + (\Sigma^{-1})_{\mathcal{A}\mathcal{A}^c}\left[(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}^c}\right]^T\right\}\Sigma_{\mathcal{A}\mathcal{A}}\right.\right. \\ & \left.\left. + \left\{(\Sigma^{-1})_{\mathcal{A}^c\mathcal{A}}\left[(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}}\right]^T + (\Sigma^{-1})_{\mathcal{A}^c\mathcal{A}^c}\left[(\nabla\hat{\eta})_{\mathcal{A}\mathcal{A}^c}\right]^T\right\}\Sigma_{\mathcal{A}\mathcal{A}^c}\right]\right) \end{aligned}$$

$$= \lim_{p\to\infty}\frac{1}{p\delta}E\Big(Tr\Big[[(\nabla\hat\eta)_{\mathcal{A}\mathcal{A}}]^T\left\{\Sigma_{\mathcal{A}\mathcal{A}}(\Sigma^{-1})_{\mathcal{A}\mathcal{A}}+\Sigma_{\mathcal{A}\mathcal{A}^c}(\Sigma^{-1})_{\mathcal{A}^c\mathcal{A}}\right\}$$
$$+\ [(\nabla\hat\eta)_{\mathcal{A}\mathcal{A}^c}]^T\left\{\Sigma_{\mathcal{A}\mathcal{A}}(\Sigma^{-1})_{\mathcal{A}\mathcal{A}^c}+\Sigma_{\mathcal{A}\mathcal{A}^c}(\Sigma^{-1})_{\mathcal{A}^c\mathcal{A}^c}\right\}\Big]\Big)$$
$$= \lim_{p\to\infty}\frac{1}{p\delta}E\{\mathrm{div}(\hat\eta)\} = \lim_{p\to\infty}\frac{1}{p\delta}E\left\{\sum_{j=1}^{p}I(\hat\eta_j\neq 0)\right\}\leq 1,$$

for any $\alpha\geq\alpha_{\min}$ according to Propositions (1) and (2). By the argument in [6], the covariance of z_t and z_{t-1} is $\tau_\star^2 - y_{t-1}/2$ decreasing with y_{t-1} which implies that $R'(y_{t-1})$ is a decreasing function. Moreover $R(0)=0$. Therefore $R(y)$ is concave with $R'(0)\leq 1$ and $R(0)=0$. For any $y_0>0$, the iteration procedure $y_t = R(y_{t-1})$ leads to a convergent result with $y_t \xrightarrow{t\to\infty} 0$. Therefore,

$$\lim_{p\to\infty}\frac{1}{p}\left\|\beta^{t+1}-\beta^t\right\|^2$$
$$= \lim_{p\to\infty}\frac{1}{p}E\left\|\eta_{\theta_t}(\beta_0+\Sigma^{-1/2}\mathbf{z}_t)-\eta_{\theta_{t-1}}(\beta_0+\Sigma^{-1/2}\mathbf{z}_{t-1})\right\|^2$$

which vanishes as $t\to\infty$. The statement of $\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{p}\|\mathbf{z}^t-\mathbf{z}^{t-1}\|^2=0$ can be proved similarly. □

A.9. Proof of Lemma 7

Proof. Applying Corollary 2 of [7] to the AMP iteration (30), for any sequence $\tilde\varphi_p : (\mathbb{R}^p)^3\to\mathbb{R},\ p\geq 1$, of uniformly pseudo-Lipschitz functions, we obtain

$$\tilde\varphi_p\left(\tilde\beta^{t+1},\tilde\beta^{s+1},\tilde\beta_0\right)\overset{P}{\approx}\tilde\varphi_p\left(\tilde\eta_{\theta_t}(\tilde\beta_0+\mathbf{z}_t),\tilde\eta_{\theta_t}(\tilde\beta_0+\mathbf{z}_s),\tilde\beta_0\right), \tag{55}$$

where $(\mathbf{z}_s,\mathbf{z}_t)$ jointly Gaussian, independent from $\beta_0\sim p_{\beta_0}$ with mean 0 and covariance given by $E(\mathbf{z}_s\mathbf{z}_s^T)=\tau_s^2\mathbf{I}_{p\times p}$, $E(\mathbf{z}_t\mathbf{z}_t^T)=\tau_t^2\mathbf{I}_{p\times p}$, and $E(\mathbf{z}_s\mathbf{z}_t^T)=\tau_{s,t}\mathbf{I}_{p\times p}$. The recursion $\tau_{s,,t}$ for all $s,t\geq 0$ is determined by

$$\begin{aligned}
\tau_{t+1}^2 &= \sigma_w^2+\lim_{p\to\infty}\frac{1}{p\delta}E\left(\|\tilde\eta_{\theta_t}(\tilde\beta_0+\tau_t\mathbf{z})-\tilde\beta_0\|^2\right),\\
\tau_{s+1}^2 &= \sigma_w^2+\lim_{p\to\infty}\frac{1}{p\delta}E\left(\|\tilde\eta_{\theta_s}(\tilde\beta_0+\tau_s\mathbf{z})-\tilde\beta_0\|^2\right),\\
\tau_{t+1,s+1} &= \sigma_w^2+\lim_{p\to\infty}\frac{1}{p\delta}E\left(\tilde\eta_{\theta_t}(\tilde\beta_0+\tau_t\mathbf{z})-\tilde\beta_0\right)\left(\tilde\eta_{\theta_s}(\tilde\beta_0+\tau_t\mathbf{z})-\tilde\beta_0\right).
\end{aligned}$$

Then define sequence of functions: $\tilde\varphi_p(\mathbf{x},\mathbf{y},\mathbf{z})=\varphi_p\left(\Sigma^{-1/2}\mathbf{x},\Sigma^{-1/2}\mathbf{y},\Sigma^{-1/2}\mathbf{z}\right)$ which is also uniformly pseudo-Lipschitz. We then obtain the distributional limit for $\beta^{t+1}=\Sigma^{-1/2}\tilde\beta^{t+1}$ and $\beta^{s+1}=\Sigma^{-1/2}\tilde\beta^{s+1}$ using (55). □

A.10. Proof of Lemma 3

Proof. The matrix $\mathbf{X} = \tilde{\mathbf{X}}\Sigma^{1/2}$, where $\tilde{\mathbf{X}}$ has entries distributed i.i.d. $N(0, 1/n)$. Thus, one has ([26], Corollary 5.35)

$$\mathbb{P}\left(\sqrt{\delta} - 1 - t \le \sigma_{\min}(\tilde{\mathbf{X}}) \le \sigma_{\max}(\tilde{\mathbf{X}}) \le \sqrt{\delta} + 1 + t\right) \ge 1 - 2\exp(-t^2/2).$$

From the fact that

$$\sigma_{\min}(\mathbf{X}) \ge \sigma_{\min}(\tilde{\mathbf{X}})\sigma_{\min}(\Sigma^{1/2}), \text{and} \sigma_{\max}(\mathbf{X}) \le \sigma_{\max}(\tilde{\mathbf{X}})\sigma_{\max}(\Sigma^{1/2}),$$

We conclude that, for every $t \ge 0$, there exists $c_5 > 0$ such that

$$\mathbb{P}\left(c_5^{-1} \le \sigma_{\min}(\mathbf{X}) \le \sigma_{\max}(\mathbf{X}) \le c_5\right) > 1 - 2\exp(-t^2/2).$$

A.11. Proof of Lemma 4

Proof. Define $S(c_2) = \{j \in [p] : |v_j^t| \ge 1 - c_2\}$, we have almost surely

$$\begin{aligned} \frac{|S(c_2)|}{p} &= \frac{1}{p}\sum_{i=1}^p \mathbb{I}\left\{\frac{1}{\theta_{t-1}}|\mathbf{X}^T\mathbf{z}^{t-1} + \Sigma(\beta^{t-1} - \beta^t)|_i \ge 1 - c_2\right\} \\ &\to \frac{1}{p}\sum_{i=1}^p \mathbb{I}\left\{\frac{1}{\theta_{t-1}}|\Sigma\{\beta_0 + \tau_{t-1}\Sigma^{-1/2}\mathbf{z} - \eta_{\theta_{t-1}}(\beta_0 + \tau_{t-1}\Sigma^{-1/2}\mathbf{z})\}|_i \right. \\ &\qquad\qquad \left. \ge 1 - c_2\right\}. \end{aligned}$$

Let us write $\bar{\Sigma} = \Sigma/\lambda_{\min}, \bar{\tau}_{t-1} = \tau_{t-1}/\lambda_{\min}^{1/2}, \bar{\theta}_{t-1} = \theta_{t-1}/\lambda_{\min}$, so that

$$\begin{aligned} \hat{\beta} &= \eta_{\theta_{t-1}}(\beta_0 + \tau_{t-1}\Sigma^{-1/2}\mathbf{z}) \\ &= \text{argmin}_{\beta \in \mathbb{R}^p}\left\{\frac{1}{2}\|\Sigma^{1/2}(\beta - \beta_0) - \tau_{t-1}\mathbf{z}\|_2^2 + \theta_{t-1}\|\beta\|_1\right\} \\ &= \text{argmin}_{\beta \in \mathbb{R}^p}\left\{\frac{1}{2}\|\bar{\Sigma}^{1/2}(\beta - \beta_0) - \bar{\tau}_{t-1}\mathbf{z}\|_2^2 + \bar{\theta}_{t-1}\|\beta\|_1\right\}. \end{aligned} \tag{56}$$

The KKT conditions of this optimization problem are

$$\bar{\Sigma}^{1/2}(\bar{\tau}_{t-1}\mathbf{z} + \bar{\Sigma}^{1/2}(\beta_0 - \hat{\beta})) \in \bar{\theta}_{t-1}\partial\|\hat{\beta}\|_1.$$

Define $\hat{\mathbf{y}} = \hat{\beta} + \bar{\Sigma}^{1/2}(\bar{\tau}_{t-1}\mathbf{z} + \bar{\Sigma}^{1/2}(\beta_0 - \hat{\beta}))$, we have

$$\hat{\beta} = \eta_{soft}(\hat{\mathbf{y}}; \bar{\theta}_{t-1}),$$

where $\eta_{soft}(x;\alpha) = \text{sign}(x)(|x| - \alpha)_+$ and applies coordinates-wise. Define $\mathbf{f}(\bar{\tau}_{t-1}\mathbf{z}) = (\mathbf{I}_{p\times p} - \bar{\Sigma}^{-1})\bar{\Sigma}^{1/2}(\beta_0 - \hat{\beta})$, then $\hat{\mathbf{y}}$ can be written as

$$\hat{\mathbf{y}} = \beta_0 + \bar{\Sigma}^{1/2}(\bar{\tau}_{t-1}\mathbf{z} + (\mathbf{I}_{p\times p} - \bar{\Sigma}^{-1})\bar{\Sigma}^{1/2}(\beta_0 - \hat{\beta}))$$

$$= \quad \beta_0 + \bar{\Sigma}^{1/2}(\bar{\tau}_{t-1}\mathbf{z} + \mathbf{f}(\bar{\tau}_{t-1}\mathbf{z})).$$

Denote σ_j the j-th row of $\bar{\Sigma}^{1/2}$ and $\sigma_j^T\mathbf{z} = x$, then $x \sim N(0, \|\sigma_j\|_2^2)$. Let $P_j^{\perp}$ be the projection operator onto the orthogonal complement of the span of σ_j. Then

$$\begin{aligned} \hat{y}_j &= \beta_0 + \bar{\tau}_{t-1}\sigma_j^T\mathbf{z} + \sigma_j^T\mathbf{f}(\bar{\tau}_{t-1}(\sigma_j^T\mathbf{z})\sigma_j/\|\sigma_j\|_2^2 + \bar{\tau}_{t-1}P_j^{\perp}\mathbf{z}) \\ &= \beta_0 + \bar{\tau}_{t-1}x + \sigma_j^T\mathbf{f}(\bar{\tau}_{t-1}x\sigma_j/\|\sigma_j\|_2^2 + \bar{\tau}_{t-1}P_j^{\perp}\mathbf{z}) \equiv h(x). \end{aligned} \tag{57}$$

By (56), $\bar{\Sigma}^{1/2}(\beta_0 - \hat{\beta})$ is 1-Lipschitz in $\bar{\tau}_{t-1}\mathbf{z}$. Thus, $\mathbf{f}(\bar{\tau}_{t-1}\mathbf{z})$ is $(1-\kappa_{cond}^{-1})$-Lipschitz in $\bar{\tau}_{t-1}\mathbf{z}$ and $\bar{\tau}_{t-1}(1-\kappa_{cond}^{-1})/\|\sigma_j\|_2$-Lipschitz in x, where $\kappa_{cond} = \lambda_{\max}/\lambda_{\min}$. For any $x_1, x_2 \in \mathbb{R}$, we have

$$\begin{aligned} & |h(x_1) - h(x_2)| \\ \geq\ & \bar{\tau}_{t-1}|x_1 - x_2| - \left|\sigma_j^T\left\{\mathbf{f}\left(\frac{\bar{\tau}_{t-1}x_1\sigma_j}{\|\sigma_j\|_2^2} + \bar{\tau}_{t-1}P_j^{\perp}\mathbf{z}\right)\right.\right. \\ & \left.\left. \qquad - \mathbf{f}\left(\frac{\bar{\tau}_{t-1}x_2\sigma_j}{\|\sigma_j\|_2^2} + \bar{\tau}_{t-1}P_j^{\perp}\mathbf{z}\right)\right\}\right| \\ \geq\ & \bar{\tau}_{t-1}|x_1 - x_2| - \bar{\tau}_{t-1}(1-\kappa_{cond}^{-1})|x_1 - x_2| = \bar{\tau}_{t-1}\kappa_{cond}^{-1}|x_1 - x_2|. \end{aligned} \tag{58}$$

According to (36), we have

$$\mathbf{v}^t = \frac{1}{\bar{\theta}_{t-1}}(\hat{\mathbf{y}} - \eta_{soft}(\hat{\mathbf{y}}; \bar{\theta}_{t-1})). \tag{59}$$

By the definition of $S(c_2)$, one obtains

$$S(c_2) = \{j \in [p] : |\hat{y}_j| \geq \bar{\theta}_{t-1}(1 - c_2)\}.$$

Therefore

$$\frac{|S(c_2)|}{n} = \frac{\{j \in [p] : |\hat{y}_j| > \bar{\theta}_{t-1}\}}{n} + \frac{\{j \in [p] : 1 - |\hat{y}_j|/\bar{\theta}_{t-1} \in [0, c_2]\}}{n}. \tag{60}$$

Consider the function

$$g(\hat{\mathbf{y}}, c_2) = \frac{1}{n}\sum_{j=1}^{p} g_1(\hat{y}_j, c_2),$$

where $g_1(\hat{y}, c_2) = \min\left\{1, \left(\frac{|\hat{y}|}{\bar{\theta}_{t-1}c_2} - \frac{1}{c_2} + 2\right)_+\right\}$. Since

$$\begin{aligned} |g(\hat{\mathbf{y}}_1, c_2) - g(\hat{\mathbf{y}}_2, c_2)| &\leq \frac{1}{n}\sum_{j=1}^{p}\{|g_1(\hat{y}_{1,j}, c_2) - g_1(\hat{y}_{2,j}, c_2)|\} \\ &\leq \frac{1}{n}\sum_{j=1}^{p}\frac{1}{\bar{\theta}_{t-1}c_2}|\hat{y}_{1,j} - \hat{y}_{2,j}| \end{aligned}$$

$$\leq \frac{\sqrt{p}}{n\theta_{t-1}c_2}\|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|_2,$$

the function $g(\hat{\mathbf{y}}, c_2)$ is $\frac{\sqrt{p}}{n\bar{\theta}_{t-1}c_2}$-Lipschitz in $\hat{\mathbf{y}}$. For all $\hat{\mathbf{y}}$, by definition we have $\frac{|S(c_2)|}{n} \leq g(\hat{\mathbf{y}}, c_2) \leq \frac{|S(2c_2)|}{n}$. Moreover, by (58) and (60), one obtains

$$\begin{aligned}
\mathbb{E}(g(\hat{\mathbf{y}}, c_2)) &\leq \mathbb{E}\left(\frac{\|\hat{\beta}\|_0}{n}\right) + \mathbb{E}\left(\frac{\{j \in [p] : 1 - |\hat{y}_j|/\bar{\theta}_{t-1} \in [0, 2c_2]\}}{n}\right) \\
&\leq 1 - \omega^\star + \sup_a \mathbb{E}_x \left(\mathbb{I}\left(a \leq \frac{h(x)}{\bar{\theta}_{t-1}} \leq a + 4c_2\right)\right) \\
&\leq 1 - \omega^\star + \sup_a \mathbb{E}_x \left(\mathbb{I}\left(\frac{a\kappa_{cond}}{\bar{\tau}_{t-1}} \leq \frac{x}{\bar{\theta}_{t-1}} \leq \frac{(a + 4c_2)\kappa_{cond}}{\bar{\tau}_{t-1}}\right)\right) \\
&\leq 1 - \omega^\star + \frac{4c_2\kappa_{cond}\bar{\theta}_{t-1}}{\sqrt{2\pi}\bar{\tau}_{t-1}}.
\end{aligned}$$

From (57), $\hat{\mathbf{y}}$ is $2\kappa_{cond}^{1/2}\bar{\tau}_{t-1}$-Lipschitz in $\mathbf{z}$. Therefore, $g(\hat{\mathbf{y}}, c_2)$ is $\frac{2\sqrt{p}\kappa_{cond}^{1/2}\bar{\tau}_{t-1}}{n\bar{\theta}_{t-1}c_2}$-Lipschitz in $\mathbf{z}$. By Gaussian concentration of Lipschitz functions

$$\begin{aligned}
&\mathbb{P}\left(\frac{|S(c_2)|}{n} \geq 1 - \omega^\star + \frac{4c_2\kappa_{cond}\bar{\theta}_{t-1}}{\sqrt{2\pi}\bar{\tau}_{t-1}} + \epsilon\right) \\
\leq\ &\mathbb{P}\left(g(\hat{\mathbf{y}}, c_2) \geq 1 - \omega^\star + \frac{4c_2\kappa_{cond}\bar{\theta}_{t-1}}{\sqrt{2\pi}\bar{\tau}_{t-1}} + \epsilon\right) \\
\leq\ &\mathbb{P}(g(\hat{\mathbf{y}}, c_2) \geq \mathbb{E}(g(\hat{\mathbf{y}}, c_2)) + \epsilon) \\
\leq\ &\exp\left(-\frac{n\delta\bar{\theta}_{t-1}^2 c_2^2}{8\kappa_{cond}\bar{\tau}_{t-1}^2}\epsilon^2\right).
\end{aligned}$$

Absorbing constants appropriately, we conclude there exists $C, c_1 > 0$ such that

$$\mathbb{P}\left(\frac{|S(c_2)|}{n} \geq 1 - \omega^\star/2\right) \leq C\exp\left(-nc_1\right).$$

A.12. Proof of Lemma 5

Proof. Let $k = [n(1 - \omega^\star/4)]$ and note that $k < p$. Because for $k > p$, we have $\kappa_-(\mathbf{X}, n(1 - \omega^\star/4)) = \kappa_-(\mathbf{X}, p)$ and thus $\mathbb{P}(\kappa_-(\mathbf{X}, n(1 - \omega^\star/4)) \geq c_4) \geq 1 - C\exp(-cn)$.

Because $\kappa_-(\mathbf{X}, S') \geq \kappa_-(\mathbf{X}, S)$ when $S' \subset S$, we have that $\kappa_-(\mathbf{X}, n(1 - \omega^\star/4)) = \min_{|S|=k} \kappa_-(\mathbf{X}, S)$. By a union bound, for any $t > 0$

$$\mathbb{P}(\kappa_-(\mathbf{X}, n(1 - \omega^\star/4)) \leq t) \leq \sum_{|S|=k} \mathbb{P}(\kappa(\mathbf{X}_S) \leq t). \tag{61}$$

The matrix $\mathbf{X}_S = \tilde{\mathbf{X}}_S \Sigma_{S,S}^{1/2}$ where $\tilde{\mathbf{X}}_S$ has entries distribution i.i.d. $N(0, 1/n)$. Thus, one has

$$\kappa_-(\mathbf{X}_S) \geq \kappa_-(\tilde{\mathbf{X}}_S)\kappa_-(\Sigma_{S,S}^{1/2}) \geq \kappa_-(\tilde{\mathbf{X}}_S)\kappa_{\min}^{1/2}$$

Invoking the fact that $\tilde{\mathbf{X}}_S$ has the same distribution for all $|S| = k$, expression (61) implies

$$\mathbb{P}(\kappa_-(\mathbf{X}, n(1-\omega^\star/4)) \leq t) \leq \left(\begin{array}{c} p \\ k \end{array}\right) \mathbb{P}(\kappa_-(\tilde{\mathbf{X}}_S) \leq t/\kappa_{\min}^{1/2}).$$

Let $f_{\min}(k, n, \lambda)$ denote the probability density function for the smallest eigenvalue $\kappa_-(\tilde{\mathbf{X}}_S)$. By Prop. 5.2, pp.553 [15], $f_{\min}(k, n, \lambda)$ satisfies

$$\begin{aligned} f_{\min}(k, n; \lambda) &\leq g_{\min}(k, n; \lambda) \\ &\equiv \frac{\Gamma((n+1)/2)}{\Gamma(k/2)\Gamma((n-k+1)/2)\Gamma((n-k+2)/2)} \\ &\quad \left(\frac{\pi}{2n\lambda}\right)^{1/2} \left(\frac{n\lambda}{2}\right)^{(n-k)/2} \exp(-n\lambda/2). \end{aligned}$$

It can be verified that the quantity $g_{\min}(k, n; \lambda)$ is strictly increasing in λ on $[0, (n-k-1)/n)$. Lemma 2.9 of [8] states that as $n, k \to \infty$ with $k/n \to \rho \in (0, 1]$,

$$g_{\min}(k, n; \lambda) \to p_{\min}(n, \lambda) \exp(n\psi_{\min}(\lambda, \rho)),$$

where $p_{\min}(n, \lambda)$ is a polynomial in n, λ, and $\psi_{\min}(\lambda, \rho) = H(\rho) + \frac{1}{2}[(1-\rho)\log\lambda + 1 - \rho + \rho\log\rho - \lambda]$, where $H(\rho) = \rho\log(1/\rho) + (1-\rho)\log(1/(1-\rho))$. Therefore, for $t/\kappa_{\min}^{1/2} \leq 1 - \rho$, we have

$$\begin{aligned} \mathbb{P}(\kappa_-(\tilde{\mathbf{X}}_S) \leq t/\kappa_{\min}^{1/2}) &= \int_0^{t/\kappa_{\min}^{1/2}} f_{\min}(k, n; \lambda) d\lambda \\ &\leq \int_0^{t/\kappa_{\min}^{1/2}} g_{\min}(k, n; \lambda) d\lambda \\ &\leq t/\kappa_{\min}^{1/2} g_{\min}(k, n; t/\kappa_{\min}^{1/2}) \\ &= C(n, t/\kappa_{\min}^{1/2}) \exp(n\psi(\rho, t/\kappa_{\min}^{1/2})), \end{aligned}$$

where $C(a, b)$ is a polynomial in a, b. To simplify $\left(\begin{array}{c} p \\ k \end{array}\right)$, we apply the second of Binet's log gamma formulas [29] and obtain

$$\frac{1}{n}\log\left(\begin{array}{c} p \\ k \end{array}\right) \to \rho\log\frac{1}{\rho\delta} + (\frac{1}{\delta} - \rho)\log\frac{1}{1-\rho\delta} = H(\rho\delta)/\delta.$$

We conclude that

$$\mathbb{P}(\kappa_-(\mathbf{X}, n(1-\zeta^\star/4)) \leq t) \leq C(n, t/\kappa_{\min}^{1/2}) \exp(n(H(\rho\delta)/\delta + \psi(\rho, t/\kappa_{\min}^{1/2}))).$$

Note that $H(\rho) \leq 1/2$ for $\rho \in (0,1)$. Thus, there exists $c > 0$ such that

$$H(\rho\delta)/\delta + \psi(\rho, t/\kappa_{\min}^{1/2}) \leq -c$$

for all $\log(t/\kappa_{\min}^{1/2}) \leq -1 - \frac{8(1/\delta+1)\log 2}{\omega^\star} - \frac{8c}{\omega^\star}$. Because $C(n, t/\kappa_{\min}^{1/2})e^{-cn}$ is upper bounded by a constant C, we conclude there exists $C, c > 0$ such that

$$\mathbb{P}(\kappa_-(\mathbf{X}, n(1-\omega^\star/4)) \leq t) \leq Ce^{-cn}.$$

A.13. Proof of Corollary 1

Proof. For $\Sigma = \mathbf{I}_{p\times p}$, (16) can be simplified as

$$\begin{aligned} M(\epsilon,\Delta,\alpha) &= \epsilon_+ E(z-\alpha)^2 + \epsilon_- E(z+\alpha)^2 + (1-\epsilon)E[(z-\alpha)^2 I(z \geq \alpha) \\ &\quad + (z+\alpha)^2 I(z \leq -\alpha)] \\ &= \epsilon(1+\alpha^2) + 2(1-\epsilon)[(1+\alpha^2)(1-\Phi(\alpha)) - \alpha\phi(\alpha)], \end{aligned}$$

where the first term comes from the non-zero components of β_0 and the second term comes from the zero components of β_0. To determine $\delta_c = \inf_\alpha M(\epsilon,\Delta,\alpha)$, we can solve $\frac{\partial M(\epsilon,\Delta,\alpha)}{\partial\alpha} = 0$ and thus obtain the phase transition curve as shown in (18). □

A.14. Proof of Corollary 2

Proof. For block-diagonal matrix with block $\Sigma_s = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, (16) can be simplified as

$$\begin{aligned} & M(\epsilon,\Delta,\alpha) \\ = \; & \frac{1}{2}E\{((\Sigma_s^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat\beta_{\mathcal{A}}))^T \Sigma_{s\mathcal{A}\mathcal{A}}^{-1}((\Sigma_s^{1/2}\mathbf{z})_{\mathcal{A}} - \alpha\text{sign}(\hat\beta_{\mathcal{A}}))\}, \end{aligned} \quad (62)$$

where $\mathbf{z} \sim N(0, \mathbf{I}_{2\times 2})$. Note that $\Sigma_s^{1/2} = \begin{pmatrix} \rho_1 & \rho_2 \\ \rho_2 & \rho_1 \end{pmatrix}$, where $\rho_1 = \frac{\sqrt{1+\rho}}{2}$ and $\rho_2 = \frac{\sqrt{1-\rho}}{2}$.

There are three scenarios. In the first scenario, both components of β_0 are non-zero, which means that $\mathcal{B} = \{1,2\}$ and $\mathcal{B}^c = \varnothing$. Its contribution to (16) can be written as

$$\begin{aligned} M_1(\epsilon,\Delta,\alpha) &= \epsilon_+^2(1 + \frac{\alpha^2}{1+\rho}) + \epsilon_-^2(1 + \frac{\alpha^2}{1+\rho}) + 2\epsilon_+\epsilon_-(1 + \frac{\alpha^2}{1-\rho}) \\ &= \epsilon^2 A(\alpha,\Delta), \end{aligned}$$

where $A(\alpha,\Delta)$ is defined in (20). In the second scenario, only one component of β_0 are non-zero, i.e. $\mathcal{B} = \{1\}$ or $\mathcal{B} = \{2\}$. In the situation where $\mathcal{B} = \{1\}$ and

$\beta_{0,1} > 0$, we need to consider the one-dimensional LASSO problem specified by (17) with $\bar{x} = \sqrt{1-\rho^2}$ and $\bar{y} = \bar{x}^{-1}(\xi_2 - \rho\xi_1 + \rho\alpha)$ whose solution is

$$\begin{cases} \text{positive} & if \quad \xi_2 - \rho\xi_1 + \rho\alpha \geq \alpha \\ 0 & if \quad |\xi_2 - \rho\xi_1 + \rho\alpha| < \alpha \\ \text{negative} & if \quad \xi_2 - \rho\xi_1 + \rho\alpha \leq -\alpha \end{cases} .$$

Plugging this result into (17), we obtain its contribution to (62) to be

$$\begin{aligned} & \epsilon_+(1-\epsilon)\{E(\xi_1-\alpha)^2 I(|\xi_2 - \rho\xi_1 + \rho\alpha| < \alpha) \\ + \quad & E\frac{(\xi_1-\alpha)^2 + (\xi_2-\alpha)^2 - 2\rho(\xi_1-\alpha)(\xi_2-\alpha)}{1-\rho^2} I(\xi_2 - \rho\xi_1 + \rho\alpha \geq \alpha) \\ + \quad & E\frac{(\xi_1-\alpha)^2 + (\xi_2+\alpha)^2 - 2\rho(\xi_1-\alpha)^2(\xi_2+\alpha)^2}{1-\rho^2} I(\xi_2 - \rho\xi_1 + \rho\alpha \leq -\alpha)\}. \end{aligned}$$

The other situations in this scenario can be considered in a similar way. The total contribution of the second scenario to (62) is

$$M_2(\epsilon, \Delta, \alpha) = \epsilon(1-\epsilon)B(\alpha),$$

where $B(\alpha)$ is defined in (21).

In the third scenario, both components of β_0 are zero, i.e. $\mathcal{B} = \varnothing$ and $\mathcal{B}^c = \{1, 2\}$. According to (17), we need to consider the following two dimensional LASSO problem

$$\bar{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^2} \left\{ \frac{1}{2} \|\mathbf{z} - \Sigma_s^{1/2}\beta\|_2^2 + \alpha\|\beta\|_1 \right\}.$$

There exists subgradients $\partial\|\beta_1\|_1$ and $\partial\|\beta_2\|_1$ such that

$$\begin{aligned} \bar{\beta}_1 + \rho\bar{\beta}_2 &= \xi_1 - \alpha\partial\|\bar{\beta}_1\|_1, \\ \rho\bar{\beta}_1 + \bar{\beta}_2 &= \xi_2 - \alpha\partial\|\bar{\beta}_2\|_1. \end{aligned} \tag{63}$$

By dividing the two dimensional space into nine regions (as illustrated by Figure 7), we obtain the following solution for $\bar{\beta}$

$$\begin{cases} \bar{\beta}_1 = \bar{\beta}_2 = 0 & if & \|\xi_1\| < \alpha \ \& \ \|\xi_2\| < \alpha \\ \bar{\beta}_1 > 0, \ \bar{\beta}_2 = 0 & if & \|\xi_1\| \geq \alpha \ \& \ |\xi_2 - \rho\xi_1 + \rho\alpha| < \alpha \\ \bar{\beta}_1 < 0, \ \bar{\beta}_2 = 0 & if & \|\xi_1\| \leq -\alpha \ \& \ |\xi_2 - \rho\xi_1 - \rho\alpha| < \alpha \\ \bar{\beta}_1 = 0, \ \bar{\beta}_2 > 0 & if & \|\xi_2\| \geq \alpha \ \& \ |\xi_1 - \rho\xi_2 + \rho\alpha| < \alpha \\ \bar{\beta}_1 = 0, \ \bar{\beta}_2 < 0 & if & \|\xi_2\| \leq -\alpha \ \& \ |\xi_1 - \rho\xi_2 - \rho\alpha| < \alpha \\ \bar{\beta}_1 > 0, \ \bar{\beta}_2 > 0 & if & \xi_1 - \rho\xi_2 + \rho\alpha \geq \alpha \ \& \ \xi_2 - \rho\xi_1 + \rho\alpha \geq \alpha \\ \bar{\beta}_1 > 0, \ \bar{\beta}_2 < 0 & if & \xi_1 - \rho\xi_2 - \rho\alpha \geq \alpha \ \& \ \xi_2 - \rho\xi_1 + \rho\alpha \leq -\alpha \\ \bar{\beta}_1 < 0, \ \bar{\beta}_2 > 0 & if & \xi_1 - \rho\xi_2 + \rho\alpha \leq -\alpha \ \& \ \xi_2 - \rho\xi_1 - \rho\alpha \geq \alpha \\ \bar{\beta}_1 < 0, \ \bar{\beta}_2 < 0 & if & \xi_1 - \rho\xi_2 - \rho\alpha \leq -\alpha \ \& \ \xi_2 - \rho\xi_1 - \rho\alpha \leq -\alpha \end{cases} . \tag{64}$$

Substituting into (17), the total contribution of the third scenario to (62) can be written as

$$M_3(\epsilon, \Delta, \alpha) = (1-\epsilon)^2 C(\alpha),$$

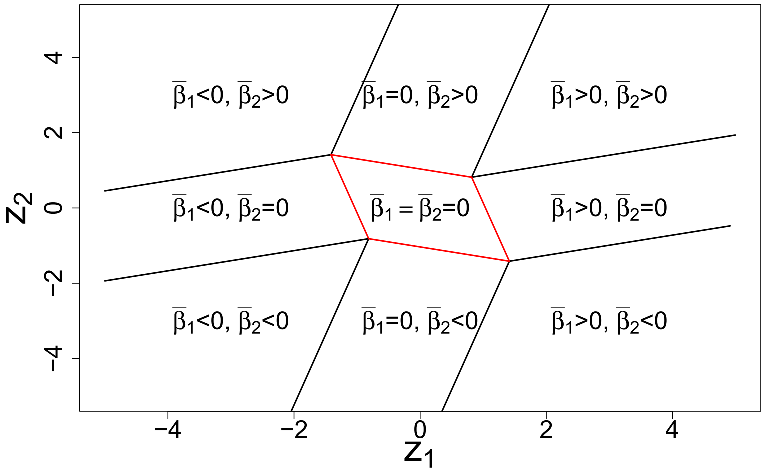
$\bar\beta_1<0, \bar\beta_2>0$
$\bar\beta_1=0, \bar\beta_2>0$
$\bar\beta_1>0, \bar\beta_2>0$
$\bar\beta_1<0, \bar\beta_2=0$
$\bar\beta_1 = \bar\beta_2=0$
$\bar\beta_1>0, \bar\beta_2=0$
$\bar\beta_1<0, \bar\beta_2<0$
$\bar\beta_1=0, \bar\beta_2<0$
$\bar\beta_1>0, \bar\beta_2<0$
z_2
4
2
0
−2
−4
−4
−2
0
2
4
z_1

Fig 7. *Illustration of the solution (64) for equation (63) in two dimensional space. Here $\rho = 0.5$ and $\alpha = 1$.*

where $C(\alpha)$ is defined in (22). Therefore

$$\begin{aligned} M(\epsilon,\Delta,\alpha) &= M_1(\epsilon,\Delta,\alpha)+M_2(\epsilon,\Delta,\alpha)+M_3(\epsilon,\Delta,\alpha) \\ &= \epsilon^2 A(\alpha,\Delta)+\epsilon(1-\epsilon)B(\alpha)+(1-\epsilon)^2C(\alpha). \end{aligned} \qquad (65)$$

To get δ_c, we need to solve the equation $\frac{\partial M(\epsilon,\Delta,\alpha)}{\partial\alpha} = 0$ for ϵ which is given by

$$\epsilon = \frac{2C'(\alpha) - B'(\alpha) + \sqrt{B'(\alpha)^2 - 4\frac{\partial A(\alpha,\Delta)}{\partial\alpha}C'(\alpha)}}{2\{\frac{\partial A(\alpha,\Delta)}{\partial\alpha} - B'(\alpha) + C'(\alpha)\}}.$$

Substituting into (65), we conclude that the transition curve is determined by (19). □

Acknowledgments

The author thanks the editor, associate editor, and two referees for many helpful comments and suggestions which led to a much improved presentation.

References

[1] Baddeley, A. (1977). Integrals on a moving manifold and geometrical probability. *Advances in Applied Probability 9*(3), 588–603.

[2] Barbier, J., F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences 116*(12), 5451–5460.

[3] Barbier, J. and N. Macris (2019, Aug). The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields 174*(3), 1133–1185.

[4] Bayati, M., M. Lelarge, and A. Montanari (2015, 04). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab. 25*(2), 753–822.

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

[6] Bayati, M. and A. Montanari (2012). The lasso risk for gaussian matrices. *IEEE Trans. Information Theory 58*(4), 1997–2017.

[7] Berthier, R., A. Montanari, and P.-M. Nguyen (2019, 01). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA 00*, 1–47.

[8] Blanchard, J. D., C. Cartis, and J. Tanner (2011). Compressed sensing: How sharp is the restricted isometry property? *SIAM Review 53*(1), 105–125.

[9] Celentano, M., A. Montanari, and Y. Wei (2020). The lasso with general gaussian designs with applications to hypothesis testing.

[10] Donoho, D. and J. Tanner (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 367*(1906), 4273–4293.

[11] Donoho, D. L., I. Johnstone, and A. Montanari (2013, June). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inf. Theor. 59*(6), 3396–3433.

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

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

[14] Donoho, D. L. and J. Tanner (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences 102*(27), 9446–9451.

[15] Edelman, A. (1988). Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications 9*(4), 543–560.

[16] Guo, D., D. Baron, and S. Shamai (2009, Sep.). A single-letter characterization of optimal noisy compressed sensing. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 52–59.

[17] Javanmard, A. and A. Montanari (2013). 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.

[18] Javanmard, A. and A. Montanari (2014, Oct). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymp-

totic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

[19] Kabashima, Y., T. Wadayama, and T. Tanaka (2009). A typical reconstruction limit of compressed sensing based on Lp-norm minimization. *Journal of Statistical Mechanics Theory and Experiment*, L09003.

[20] Krzakala, F., M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová (2012, May). Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X 2*, 021005.

[21] Maleki, A., L. Anitori, Z. Yang, and R. G. Baraniuk (2013, July). Asymptotic analysis of complex lasso via complex approximate message passing (camp). *IEEE Transactions on Information Theory 59*(7), 4290–4308.

[22] Mezard, M. and A. Montanari (2009). *Information, Physics, and Computation.* New York, NY, USA: Oxford University Press, Inc.

[23] Rangan, S. (2011, July). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2168–2172.

[24] Rangan, S., V. Goyal, and A. K. Fletcher (2009). Asymptotic analysis of map estimation via the replica method and compressed sensing. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, pp. 1545–1553. Curran Associates, Inc.

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

[26] Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

[27] Wainwright, M. J. (2009, May). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

[28] Weng, H., A. Maleki, and L. Zheng (2018, 12). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Ann. Statist. 46*(6A), 3099–3129.

[29] Whittaker, E. T. and G. N. Watson (1996). *A Course of Modern Analysis* (4 ed.). Cambridge Mathematical Library. Cambridge University Press.

[30] Zheng, L., A. Maleki, H. Weng, X. Wang, and T. Long (2017, Nov). Does ℓ_p -minimization outperform ℓ_1 -minimization? *IEEE Transactions on Information Theory 63*(11), 6896–6935.