Extending Approximate Message Passing (AMP) Algorithm to General Gaussian Covariates

Yisha Yao, Pierre Bellec, Cun-Hui Zhang

Abstract

Approximated Message Passing (AMP) refers to a class of efficient algorithms for statistical estimation in high-dimensional problems such as compressed sensing, low-rank matrix estimation, and sampling procedures. It has enjoyed several advantages, including nice empirical performances, faster than convex programming, and established asymptotic theory for independent Gaussian covariates under the regime $s \asymp n \asymp p$. However, the current AMP algorithm is largely restricted to independent covariates, and its asymptotic theory relies on some strong assumptions. In this paper, we modify the AMP iterations so that it applies to general Gaussian covariates. We also provide finite sample analysis for our modified version with explicit rates. Numerical simulations are supportive of our theory.

1 Introduction

Approximated message passing (AMP) algorithm is originally designed to solve the compressed sensing or sparse signal recovery problem [6], which stems from various engineering, bio-medical, and physical applications. The goal is to reconstruct a high-dimensional signal from some undersampled transformed measurements. Among existing methods for sparse signal recovery, convex optimization (*e.g.* linear programming, LP) has the best sparsity-undersampling tradeoff, *i.e.*, convex optimization requires the least sample size at the same sparsity level. However, convex optimization is computationally expensive. In [6], the authors proposed a new iterative thresholding algorithm i.e. AMP, which is much faster than the convex procedure while achieves a comparable sparsity-undersampling tradeoff under noiseless standard Gaussian design. Yet current AMP algorithm is limited to standard Gaussian design in the sense that its theory relies on the assumption of independent covariates. Moreover, implementing AMP algorithm requires the knowledge about the empirical distribution of the true signal, which is unrealistic in most applications. In this paper, we propose a revised AMP algorithm which applies to general Gaussian designs and does not require the empirical distribution of the true signal.

Consider the noise contaminated linear model (1.1).

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + \varepsilon_j, \qquad j = 1, 2, \cdots, n \tag{1.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the signal of interest and known to be sparse *e.g.* $\|\boldsymbol{\beta}\|_0 \leq s$, $\boldsymbol{x}_j$'s are the sensing vectors, and $\varepsilon_j \sim N(0, \sigma^2)$. We will carry on our discussion under the regime $s \asymp n \ll p$. We want to obtain a $\hat{\boldsymbol{\beta}}$ such that $\|\hat{\boldsymbol{\beta}}\|_0$ is small and $\boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}$ matches y_j closely. It is equivalent to solving a

constrained minimization problem.

$$\begin{cases} \textit{minimize } \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ \textit{subject to } \|\boldsymbol{\beta}\|_0 \leq s. \end{cases}$$

By convex relaxation [12], l_1 norm is used instead of l_0 norm because l_1 norm is convex and fits the problem into a framework of convex optimization.

$$\begin{cases} \textit{minimize } \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \\ \textit{subject to } \|\boldsymbol{\beta}\|_1 \leq t. \end{cases}$$

The solution to the above constrained minimization problem is well-known as LASSO estimator [13] [5]. There have been several algorithms developed for computing LASSO estimators, including conventional convex procedure, LAR [8], coordinate descent [9], *etc.* Among these algorithms, convex optimization has the best sparsity-undersampling tradeoff. However, as the dimension p or n increases, quadratic programming becomes computationally intractable. The AMP algorithm is calimed to achieve comparable performance as convex procedure and yet is much faster than convex procedure.

AMP uses the iteration formula below with $\boldsymbol{\beta}^0 = 0$

$$\boldsymbol{\beta}^{t+1} = \eta(\boldsymbol{\beta}^t + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}^t; \alpha\boldsymbol{\tau}_t), \tag{1.2}$$

$$\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t + \frac{\|\boldsymbol{\beta}^t\|_0}{n}\boldsymbol{r}^{t-1}, \tag{1.3}$$

$$\tau_{t+1}^2 = \sigma^2 + \frac{p}{n}\mathbb{E}\Big\{[\eta_t(B + \alpha\tau_t Z) - B]^2\Big\}. \tag{1.4}$$

where η is the soft thresholded function

$$\eta(\mu;\tau) = \begin{cases} \mu - \tau & \text{if } \tau < \mu \\ 0 & \text{if } -\tau \leq \mu \leq \tau \\ \mu + \tau & \text{if } \mu < -\tau \end{cases}$$

$Z \sim N(0,1)$, B is the limit of the empirical distribution of $\boldsymbol{\beta}$ and α is a numerical tuning parameter. Such iteration has been shown to perform well under the regime $\boldsymbol{x}_j \sim N(0, I)$, and the empirical distribution of $\boldsymbol{\beta}$ is known.

1.1 Some intuition

First, we gain some intuition why the AMP formula would possibly work by studying the basic thresholding iteration

$$\begin{aligned} \boldsymbol{\beta}^{t+1} &= \eta_t(\boldsymbol{\beta}^t + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}^t), \\ \boldsymbol{r}^t &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t. \end{aligned} \tag{1.5}$$

Define the MSE of $\boldsymbol{\beta}^t + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}^t$ and the MSE of $\boldsymbol{\beta}^t$ by

$$\tau_t^2 \equiv \frac{\|\boldsymbol{\beta}^t + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}^t - \boldsymbol{\beta}\|_2^2}{p}$$
$$\hat{\tau}_t^2 \equiv \frac{\|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_2^2}{p}$$

Inside $\eta_t()$ is

$$\begin{aligned}\boldsymbol{\beta}^t + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}^t &= \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) + \boldsymbol{\beta}^t \\ &= \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon} + (I - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}^t - \boldsymbol{\beta}) + \boldsymbol{\beta}\end{aligned}$$

The first term above $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon}$ has entries iid. $N(0, \frac{\|\boldsymbol{\varepsilon}\|_2^2}{n^2})$. If $\boldsymbol{\beta}^t$ is independent of $\boldsymbol{X}$, the second term $(I - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}^t - \boldsymbol{\beta})$ is approximately Gaussian vector with entries iid. $N(0, \frac{p\hat{\tau}_t^2}{n})$. It can be shown that $\boldsymbol{X}^T\boldsymbol{\varepsilon}$ is asymptotically independent from $(I - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}^t - \boldsymbol{\beta})$. Therefore, the basic thresholded iteration (1.5) amounts to thresholding a Gaussian sequence with MSE approximate

$$\tau_t^2 = \sigma^2 + \frac{p}{n}\hat{\tau}_t^2$$

Meanwhile,

$$\begin{aligned}\hat{\tau}_{t+1}^2 &= \lim_{p\to\infty} \frac{\|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_2^2}{p} \\ &= \mathbb{E}\Big\{[\eta_t(B + \alpha\tau_t Z) - B]^2\Big\}.\end{aligned}$$

By the above two formula for τ_t^2 and $\hat{\tau}_{t+1}^2$, it follows

$$\tau_{t+1}^2 = \sigma^2 + \frac{p}{n}\mathbb{E}\Big\{[\eta_t(B + \alpha\tau_t Z) - B]^2\Big\}.$$

The above analysis is based on the fact that $\boldsymbol{\beta}^t$ is independent of $\boldsymbol{X}$. However, it is not true and the dependence can not be ignored. Extensive numerical simulation [11] has been carried out on the basic thresholded iteration formulae (1.5), showing that its MSE behavior is distinct from (1.4). While the AMP iterations (1.2) and (1.3) have the MSE dynamics matched to (1.4). The key for such phenomenon lies in the term $\frac{\|\boldsymbol{\beta}^t\|_0}{n}\boldsymbol{r}^{t-1}$. It makes the Gaussian sequence centered around the true $\boldsymbol{\beta}$, or in another word, it is the debias-correcting term. The authors of the AMP papers interpret it as the "Onsager" reaction term via the language of Statistical Physics. Yet there is also a statistical interpretation of this term, which is related to the material in [3] and [4].

1.2 The AMP stationary solution is equivalent to a LASSO estimator

It has been perceived that the AMP iteration converges to a LASSO estimator [2]. Suppose $(\hat{\boldsymbol{\beta}}, \boldsymbol{r}, \tau)$ is a fixed point of the AMP iteration, then

$$\hat{\boldsymbol{\beta}} = \eta(\hat{\boldsymbol{\beta}} + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}; \alpha\tau), \tag{1.6}$$

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \frac{\|\hat{\boldsymbol{\beta}}\|_0}{n}\boldsymbol{r} \tag{1.7}$$

(1.6) implies there is a $\boldsymbol{v} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ such that $\hat{\boldsymbol{\beta}} + \alpha\tau\boldsymbol{v} = \hat{\boldsymbol{\beta}} + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r}$, which means $\boldsymbol{X}^T\boldsymbol{r} = n\alpha\tau\boldsymbol{v}$. (1.7) implies $\boldsymbol{r} = \frac{1}{1-\frac{\|\hat{\boldsymbol{\beta}}\|_0}{n}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. These two facts add up to

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \alpha\tau n(1 - \frac{\|\hat{\boldsymbol{\beta}}\|_0}{n})\boldsymbol{v}, \quad \boldsymbol{v} \in \partial\|\hat{\boldsymbol{\beta}}\|_1.$$

This is the KKT condition for a LASSO estimator with penalty

$$\lambda = \alpha\tau(1 - \frac{\|\hat{\boldsymbol{\beta}}\|_0}{n}).$$

Since the penalty level involves the sparsity of a pre-conceived estimator, the thresholding parameter α needs to be calibrated (according to λ) so that the AMP iterate converges in average to the LASSO estimator with a pre-specified penalty level λ. Let τ_* be the unique stationary solution for

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big[\eta_t(B + \tau Z; \alpha\tau) - B\Big]^2$$

with any fixed α and α needs to satisfy

$$\lambda(\alpha) = \alpha\tau_*\Big[1 - \frac{p}{n}\mathbb{P}\{\eta(B + \tau_* Z; \alpha\tau_*) \neq 0\}\Big]$$

Combining the two claims that the stationary solution $\hat{\boldsymbol{\beta}}$ is a LASSO estimator and the term $\hat{\boldsymbol{\beta}} + \frac{1}{n}\boldsymbol{X}^T\boldsymbol{r} = \hat{\boldsymbol{\beta}} + \frac{\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\|\hat{\boldsymbol{\beta}}\|_0}$ is centered around the true $\boldsymbol{\beta}$, it suggests a "debiased" LASSO estimator of the form (with $\Sigma = I$ under standard Gaussian design)

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{Lasso} + \frac{\Sigma^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{Lasso})}{n - \|\hat{\boldsymbol{\beta}}^{Lasso}\|_0} \tag{1.8}$$

which has appeared in [10]. Another paper also proposed the above estimator but with distinct derivation [4].

1.3 Limitations of the current AMP algorithm

All the analysis and derivation in the existing AMP literature is conducted under the standard Gaussian random design [1][7], which is unrealistic in many applications. It is mathematically challenge to extend the AMP theory to general Gaussian random design case, *i.e.* $\boldsymbol{x}_j$distributed as *i.i.d.* $N(0, \Sigma)$ for an arbitrary non-singular Σ. Besides, when deriving the AMP theory they made a quite strong assumption that the empirical distribution of the true $\boldsymbol{\beta}$ is known, which is almost impossible in most real-world problems. In the following section, we propose a modified AMP iteration formula that is applicable to general Gaussian random design and needs not any information on the empirical distribution of $\boldsymbol{\beta}$.

2 Generalized AMP iteration

We propose the iteration formula (2.1)-(2.3), which performs well in general Gaussian design and circumvents the requirement of the empirical distribution of $\boldsymbol{\beta}$. Some heuristic theories will be established to give some insights on the proposed iteration.

$$\begin{aligned}
\boldsymbol{\beta}^0 = 0,\ \boldsymbol{r}^0 = \boldsymbol{y},\ s^0 = 0 \\
\boldsymbol{\beta}^{t+1} &= \eta(\boldsymbol{\beta}^t + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\boldsymbol{r}^t; \alpha\boldsymbol{\tau}^t), \qquad (2.1)\\
\boldsymbol{r}^t &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t + \frac{s^t}{n}\boldsymbol{r}^{t-1}, \qquad (2.2)\\
(\tau_j^t)^2 &= \Sigma_{jj}^{-1} \cdot \frac{\|\boldsymbol{r}^t\|^2}{n^2}. \qquad (2.3)
\end{aligned}$$

In our iteration formula, α could depend on (n, p, s), *e.g.*, $\alpha = c_0\sqrt{\log p}$. This is different from the original AMP iterations, where α can be solved from a fixed point equation involving the empirical distribution of the true $\boldsymbol{\beta}$.

2.1 Intuition

Follow the same reasoning in section 1.1, we first gain some intuition starting from the basic iteration formula.

$$\begin{aligned}
\boldsymbol{\beta}^{t+1} &= \eta_t(\boldsymbol{\beta}^t + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\boldsymbol{r}^t),\\
\boldsymbol{r}^t &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t.
\end{aligned}$$

Inside $\eta_t()$ is

$$\begin{aligned}
\boldsymbol{\beta}^t + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\boldsymbol{r}^t &= \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t) + \boldsymbol{\beta}^t\\
&= \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon} + (I - \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}^t - \boldsymbol{\beta}) + \boldsymbol{\beta}
\end{aligned}$$

Again, the first and second terms above are asymptotically Gaussian vector with mean zero provided $\boldsymbol{\beta}^t \perp\!\!\!\perp \boldsymbol{X}$. However, this assumption obviously does not hold. And consequently, this Gaussian vector actually has nonzero mean. The extra term $\frac{s^t}{n}\boldsymbol{r}^{t-1}$ in (2.2) corrects the bias and centers the Gaussian vector. Next we will show why $\frac{s^t}{n}\boldsymbol{r}^{t-1}$ corrects the bias through heuristics. We shall also derive (2.3), the approximately variance of $\boldsymbol{\beta}^t + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\boldsymbol{r}^t$, by heuristics.

2.2 Bias correction via Stein's method

We start our analysis from the easiest case, when the iteration (2.1)-(2.3) has reached its stationary solution. For the moment, let us assume there exists a unique stationary solution without rigorous proof. Suppose $\hat{\boldsymbol{\beta}}$ is the stationary solution, $\hat{\mathcal{S}} = support(\hat{\boldsymbol{\beta}})$, $\hat{s} = |\hat{\mathcal{S}}|$, and $\hat{\boldsymbol{r}}$ be the stationary residue.

By stationary condition, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \eta(\hat{\boldsymbol{\beta}} + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^T\hat{\boldsymbol{r}}; \alpha\hat{\boldsymbol{\tau}}),\\
\hat{\boldsymbol{r}} &= \frac{n}{n-\hat{s}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),\\
\hat{\tau}_j^2 &= \frac{\Sigma_{jj}^{-1}}{(n-\hat{s})^2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2
\end{aligned}$$

And the above iterations can be written as

$$\hat{\boldsymbol{\beta}} = \eta(\hat{\boldsymbol{\beta}} + \frac{\Sigma^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}}; \alpha\hat{\boldsymbol{\tau}}) \tag{2.4}$$

which implies

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}^c} = \boldsymbol{0}\\
&\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}} = \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}} + \frac{\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}} - \alpha\cdot\frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|}{n-\hat{s}}\cdot\boldsymbol{v}_{\hat{\mathcal{S}}} \;\Rightarrow\\
&\frac{\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}} = \alpha\cdot\frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|}{n-\hat{s}}\cdot\boldsymbol{v}_{\hat{\mathcal{S}}} \qquad (2.5)\\
&\left|\frac{\Sigma_{\hat{\mathcal{S}}^c,\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}}\right| \le \alpha\cdot\hat{\boldsymbol{\tau}}_{\hat{\mathcal{S}}^c}
\end{aligned}$$

where $\boldsymbol{v}_{\hat{\mathcal{S}}} = \begin{bmatrix} \vdots \\ \sqrt{\Sigma_{jj}^{-1}}\cdot sign(\hat{\beta}_j) \\ \vdots \end{bmatrix}$ for all the $j \in \hat{\mathcal{S}}$.

From (2.5), we obtain

$$\begin{aligned}
\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}} &= \Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{y} - \alpha\cdot\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|\cdot\boldsymbol{v}_{\hat{\mathcal{S}}}\\
\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}} &= (\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{y} - \alpha\cdot\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|\cdot(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\boldsymbol{v}_{\hat{\mathcal{S}}} \qquad (2.6)
\end{aligned}$$

Next we analyze the term inside $\eta(\cdot)$. Let $\Sigma^{-1}_{\cdot j}$ be the jth column of Σ^{-1}, Σ^{-1}_{jj} be the jth diagonal element of Σ^{-1}. Let

$$\boldsymbol{z}_j = \frac{\boldsymbol{X}\Sigma^{-1}_{\cdot j}}{\Sigma^{-1}_{jj}},$$

and

$$\boldsymbol{X}\boldsymbol{Q}_j = \boldsymbol{X} - \boldsymbol{z}_j \boldsymbol{e}_j^T.$$

It is easy to check that $\boldsymbol{z}_j \sim N(0, \frac{1}{\Sigma^{-1}_{jj}} I)$. Besides, it can be shown that $\boldsymbol{z}_j$ and $\boldsymbol{X}\boldsymbol{Q}_j$ are independent. Intuitively, $\boldsymbol{z}_j$ is the component in the column space of $\boldsymbol{X}$ that can only be explained by β_j. Taking jth coordinate as an instance, we expand the term below inside $\eta(\cdot)$.

$$\begin{aligned}
\hat{\beta}_j + \frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - \hat{s}} - \beta_j &= \hat{\beta}_j - \beta_j + \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \hat{\beta}_j - \beta_j + \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{\varepsilon} - \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \hat{\beta}_j - \beta_j + \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{\varepsilon} - \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T(\boldsymbol{X}\boldsymbol{Q}_j + \boldsymbol{z}_j\boldsymbol{e}_j^\top)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \hat{\beta}_j - \beta_j + \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{\varepsilon} - \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\|\boldsymbol{z}_j\|^2(\hat{\beta}_j - \beta_j)
\end{aligned} \tag{2.7}$$

It is straightforward that the second term above has mean zero, and the forth term has mean

$$\mathbb{E}\Big[\frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\|\boldsymbol{z}_j\|^2(\hat{\beta}_j - \beta_j)\Big] = \frac{n}{n-\hat{s}}(\hat{\beta}_j - \beta_j).$$

To calculate the mean of the second term in (2.7), Let

$$f(\boldsymbol{z}_j) = \boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{2.8}$$

By Stein's Lemma,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{z}_j^T\boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right] &= \mathbb{E}\left[\boldsymbol{z}_j^T f(\boldsymbol{z}_j)\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right] = \frac{1}{\Sigma^{-1}_{jj}}\mathbb{E}\left[div f(\boldsymbol{z}_j)\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right] \\
&= \frac{1}{\Sigma^{-1}_{jj}}\mathbb{E}\left[trace\left\{\frac{\partial f(\boldsymbol{z}_j)}{\partial \boldsymbol{z}_j^T}\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right] \\
&= \frac{1}{\Sigma^{-1}_{jj}}\mathbb{E}\left[trace\left\{[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right]
\end{aligned} \tag{2.9}$$

It would be tedious to compute $\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}$ directly from (2.6). We will detour by differentiating both

sides of (2.5) and then solving a differential equation.

$$
\begin{aligned}
\frac{\partial LHS}{\partial \boldsymbol{z}_j^T} &= \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\frac{\partial \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{z}_j^T} \\
&= \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\frac{\partial\Big((\boldsymbol{X}\boldsymbol{Q}_j)^\top + \boldsymbol{e}_j\boldsymbol{z}_j^\top\Big)(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{z}_j^\top} \\
&= \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}(\boldsymbol{X}\boldsymbol{Q}_j)^\top\frac{\partial(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{z}_j^\top} + \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\boldsymbol{e}_j\frac{\partial \boldsymbol{z}_j^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{z}_j^\top} \\
&= \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\left[(\beta_j-\hat{\beta}_j)(\boldsymbol{X}\boldsymbol{Q}_j)^\top - (\boldsymbol{X}\boldsymbol{Q}_j)^\top\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{z}_j^\top} + \boldsymbol{e}_j(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top + (\beta_j-\hat{\beta}_j)\boldsymbol{e}_j\boldsymbol{z}_j^\top - \boldsymbol{e}_j\boldsymbol{z}_j^\top\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{z}_j^\top}\right] \\
&= (\beta_j-\hat{\beta}_j)\Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\boldsymbol{X}^\top - \Sigma^{-1}_{\hat{\mathcal{S}},\cdot}\boldsymbol{X}^\top\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{z}_j^\top} + \Sigma^{-1}_{\hat{\mathcal{S}},j}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top
\end{aligned}
$$

Meanwhile, we differentiate the RHS of (2.5).

$$
\begin{aligned}
\frac{\partial RHS}{\partial \boldsymbol{z}_j^T} &= \alpha\boldsymbol{v}_{\hat{\mathcal{S}}}\frac{\partial\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|}{\partial \boldsymbol{z}_j^T} \\
&= \alpha\boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^\top\frac{\partial(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{z}_j^\top} \\
&= (\beta_j-\hat{\beta}_j)\cdot\alpha\boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^T - \alpha\cdot\boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^T\boldsymbol{X}\frac{\partial\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{z}_j^T},
\end{aligned}
\tag{2.10}
$$

where $\hat{\boldsymbol{u}} = \frac{\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|} = \frac{\hat{\boldsymbol{r}}}{\|\hat{\boldsymbol{r}}\|} \in \mathbb{R}^n$ is a unit vector.

Remark

In the above derivation, we are treating $\hat{s}$ as a local constant with respect to $\boldsymbol{z}_j$. It is verified rigorously in Appendix. Here we just give some intuition.

$$
\hat{\beta}_j + \frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}} \approx \hat{\beta}_j + \frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n},
$$

and the right hand side of " $\approx$ " is Lipschitz continuous with respect to $\boldsymbol{z}_j$. While $\hat{\tau}_j$ is also Lipschitz continuous with respect to $\boldsymbol{z}_j$. If the "thresholding condition" holds strictly, *i.e.*

$$
\begin{aligned}
j \in \hat{\mathcal{S}} &\iff \left|\frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}\right| > \alpha\hat{\tau}_j \\
j \in \hat{\mathcal{S}}^c &\iff \left|\frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}\right| < \alpha\hat{\tau}_j
\end{aligned}
$$

the stationary support $\hat{\mathcal{S}}$ and stationary sparsity $\hat{s}$ remains constant in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$.

Setting $\frac{\partial LHS}{\partial \boldsymbol{z}_j^T} = \frac{\partial RHS}{\partial \boldsymbol{z}_j^T}$, we get

$$\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}} - \alpha \boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} = \frac{n-\hat{s}}{n}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{e}_j\hat{\boldsymbol{r}}^T - (\hat{\beta}_j - \beta_j)\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T + (\hat{\beta}_j - \beta_j)\cdot\alpha\boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^T \quad (2.11)$$

Massaging (2.5) results in

$$\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{u}} = \alpha\boldsymbol{v}_{\hat{\mathcal{S}}},$$

so (2.11) can be equivalently written as

$$\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} = \frac{n-\hat{s}}{n}\Sigma_{\hat{\mathcal{S}},j}^{-1}\hat{\boldsymbol{r}}^T - (\hat{\beta}_j - \beta_j)\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top + (\hat{\beta}_j - \beta_j)\cdot\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T \quad (2.12)$$

By Sherman-Morrison formula,

$$\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]^{-1} = (\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1} - \frac{(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}}{1 + \hat{\boldsymbol{u}}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\hat{\boldsymbol{u}}},$$

and $\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]$ is invertible if and only if $1 + \hat{\boldsymbol{u}}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\hat{\boldsymbol{u}} \neq 0$. If a "general position" condition is imposed on $\boldsymbol{X}$, $\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}$ is invertible with probability one. Let

$$\boldsymbol{P} \triangleq \boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T, \quad (2.13)$$

which is an oblique projection matrix with $\boldsymbol{u}^T\boldsymbol{P}\boldsymbol{u} \geq 0$ for $\forall \boldsymbol{u}$. Hence $\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]$ is invertible. For simplicity, we denote

$$\boldsymbol{M} = (\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}(\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}})^{-1}.$$

Hence, we obtain

$$\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} = \left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]^{-1}\left[\frac{n-\hat{s}}{n}\Sigma_{\hat{\mathcal{S}},j}^{-1}\hat{\boldsymbol{r}}^T - (\hat{\beta}_j - \beta_j)\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top + (\hat{\beta}_j - \beta_j)\cdot\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^T\right] \quad (2.14)$$

Case 1: $j \notin \hat{\mathcal{S}}$, then $[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}} = \boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}$.

$$\begin{aligned}\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} &= \boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\left[\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \hat{\boldsymbol{u}}\hat{\boldsymbol{u}}^\top)\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\right]^{-1}\left[\frac{n-\hat{s}}{n}\Sigma_{\hat{\mathcal{S}},j}^{-1}\hat{\boldsymbol{r}}^T - (\hat{\beta}_j - \beta_j)\Sigma_{\hat{\mathcal{S}},\cdot}^{-1}\boldsymbol{X}^\top + (\hat{\beta}_j - \beta_j)\cdot\alpha\boldsymbol{v}_{\hat{\mathcal{S}}}\hat{\boldsymbol{u}}^T\right] \\ &\asymp -\hat{s}\cdot(\hat{\beta}_j - \beta_j).\end{aligned}$$

The above bound is derived under the assumption that $\Sigma^{-1}_{j,j} \leq A$ for some constant $A > 0$. Further assume that $L \leq \lambda_{min}(\Sigma^{-1}_{\mathcal{S}',\mathcal{S}'}) \leq \lambda_{max}(\Sigma^{-1}_{\mathcal{S}',\mathcal{S}'}) \leq U$ for absolute constants U, L, and for all $\mathcal{S}' \subset [p]$ with $|\mathcal{S}'| = O(s)$. If $\|\hat{\boldsymbol{r}}\|^2 = O(s \log p)$, $n \gg \log p$, $\hat{s} = O(s)$, and $s \asymp n$

$$\begin{aligned} trace\{term1\} &= \frac{n-\hat{s}}{n}\|\hat{\boldsymbol{r}}\| \cdot \frac{\hat{\boldsymbol{u}}^T \boldsymbol{P}\boldsymbol{u}_0}{1+\hat{\boldsymbol{u}}^T\boldsymbol{P}\hat{\boldsymbol{u}}} \\ &\leq \frac{n-\hat{s}}{n}\|\hat{\boldsymbol{r}}\| \cdot \|\boldsymbol{u}_0\| \\ &\leq C_3 \frac{(n-s)s}{n}\sqrt{\frac{\log p}{n}} \\ &= o(s), \end{aligned}$$

Case 2: $j \in \hat{\mathcal{S}}$, then $[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}} = \boldsymbol{X}_{\cdot,\hat{\mathcal{S}}} - \boldsymbol{z}_j \boldsymbol{e}^T_{j\hat{\mathcal{S}}}$.

$$\begin{aligned} [\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} &= \boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} - \boldsymbol{z}_j \boldsymbol{e}^T_{j\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} \\ trace\{[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\} &= trace\{\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\} - \boldsymbol{e}^T_{j\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\boldsymbol{z}_j = trace\{\boldsymbol{X}_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\} - \frac{\partial \hat{\beta}_j}{\partial \boldsymbol{z}_j^T}\boldsymbol{z}_j \\ &\asymp -\hat{s}\cdot(\hat{\beta}_j - \beta_j). \end{aligned}$$

Now it follows that

$$\mathbb{E}\left[\boldsymbol{z}_j^T\boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] = \frac{1}{\Sigma^{-1}_{jj}}\mathbb{E}\left[trace\left\{[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\hat{\mathcal{S}}}\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T}\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \asymp \frac{-(\hat{\beta}_j-\beta_j)\hat{s}}{\Sigma^{-1}_{jj}}, \quad (2.15)$$

and hence,

$$\mathbb{E}\left[\hat{\beta}_j + \frac{\Sigma^{-1}_{j\cdot}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}} - \beta_j\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] = -\frac{\Sigma^{-1}_{jj}}{n-\hat{s}}\mathbb{E}\left[\boldsymbol{z}_j^T\boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] - \frac{\hat{s}}{n-\hat{s}}(\hat{\beta}_j-\beta_j) \asymp 0$$

Note in the above differentiation procedure, we treat

$$\frac{\partial \hat{s}}{\partial \boldsymbol{z}_j} = 0$$

It is shown in Appendix that in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$, $\hat{\mathcal{S}}$, $\hat{s}$ remain constant, and hence $\frac{\partial \boldsymbol{v}_{\hat{\mathcal{S}}}}{\partial \boldsymbol{z}_j^T} = \boldsymbol{0}$.

So far we have shown that when the iteration has reached stationary, the term inside $\eta(\cdot)$ is a random variable centered at true β_j. However, it becomes much more challenging to analyze the intermediate case. It is intractable to find an explicit formulae for $f(\boldsymbol{z}_j) = \boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t - \boldsymbol{\beta})$ since $\boldsymbol{\beta}^t$ is a propagation of all the previous iterations.

2.3 Predicting the thresholding value via second order Stein method

It is natural to set $(\hat{\tau}_j)^2$ equal to the variance of $\hat{\beta}_j + \frac{\Sigma_{j\cdot}^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-\hat{s}}$. Let $g(\boldsymbol{z}_j) = \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, we need to compute $Var[\boldsymbol{z}_j^T g(\boldsymbol{z}_j)]$. By Theorem 1.1 in [3], we obtain

$$
\begin{aligned}
& Var\left[\frac{\Sigma_{jj}^{-1}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{\varepsilon} - \frac{\Sigma_{jj}^{-1}}{n-\hat{s}}\boldsymbol{z}_j^T\boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \\
&= \frac{\Sigma_{jj}^{-1}}{(n-\hat{s})^2}Var\left[\sqrt{\Sigma_{jj}^{-1}}\boldsymbol{z}_j^T g(\boldsymbol{z}_j)\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \\
&= \frac{\Sigma_{jj}^{-1}}{(n-\hat{s})^2}\mathbb{E}\left[\|\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{Q}_j(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2 + trace\left\{\frac{1}{\Sigma_{jj}^{-1}}[\frac{\partial g}{\partial \boldsymbol{z}_j^\top}]^2\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \\
&= \frac{\Sigma_{jj}^{-1}}{(n-\hat{s})^2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + (\hat{\beta}_j - \beta_j)\boldsymbol{z}_j\|^2 + \frac{(\hat{\beta}_j-\beta_j)^2}{(n-\hat{s})^2}\hat{s} \\
&\approx \frac{\Sigma_{jj}^{-1}}{(n-\hat{s})^2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2
\end{aligned}
\tag{2.16}
$$

provided $\hat{\beta}_j - \beta_j$ is sufficiently small.

2.4 Approximately centered mean in each iteration

In section 2.2, we showed that the term inside the soft-thresholding function in (2.7) is approximately centered at the true β_j for $1 \le j \le p$ after the iterations reach stationary. Here we will show that the same conclusion holds for each iteration, *i.e.*, the term inside the soft-thresholding function in (2.1) is almost centered at the true $\boldsymbol{\beta}$ for $t \ge 0$. So the first step (2.1) in each iteration of the generalized AMP is actually thresholding a random vector centered at the true $\boldsymbol{\beta}$. Let $f^t(\boldsymbol{z}_j) = \boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\beta}^t$, and $\boldsymbol{v}^t \in \mathbb{R}^p$ such that

$$
v_j^t \begin{cases} = \sqrt{\Sigma_{jj}^{-1}} \cdot sign(\beta_j^{t+1}) & \text{if } \beta_j^{t+1} \neq 0, \\ \le \sqrt{\Sigma_{jj}^{-1}} & \text{otherwise.} \end{cases}
$$

When $t = 0$, it is trivial that $\mathbb{E}\left[\boldsymbol{\beta}^0 + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{r}^0\right] = \boldsymbol{\beta}$.

When $t = 1$, it is to show $\mathbb{E}\left[\boldsymbol{\beta}^1 + \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{r}^1\right] \approx \boldsymbol{\beta}$. Since for each j,

$$
\begin{aligned}
\beta_j^1 + \frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j^\top\boldsymbol{r}^1 &= (1 - \frac{\Sigma_{jj}^{-1}\cdot\|\boldsymbol{z}_j\|^2}{n})\beta_j^1 + \frac{\Sigma_{jj}^{-1}\cdot\|\boldsymbol{z}_j\|^2}{n}\beta_j + \frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j\top\boldsymbol{\varepsilon} - \frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^1-\boldsymbol{\beta}) + \frac{\Sigma_{jj}^{-1}\cdot s^1}{n^2}\boldsymbol{z}_j^\top\boldsymbol{r}^0 \\
&\approx \beta_j + \frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j\top\boldsymbol{\varepsilon} - \frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^1-\boldsymbol{\beta}) + \frac{\Sigma_{jj}^{-1}\cdot s^1}{n^2}\boldsymbol{z}_j^\top\boldsymbol{r}^0,
\end{aligned}
$$

it boils down to show

$$
\mathbb{E}\left[\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^1-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] = \frac{s^1}{n}\mathbb{E}\left[\boldsymbol{z}_j^\top\boldsymbol{r}^0\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right].
$$

Based on the iteration formula, $\boldsymbol{\beta}^1 = \eta(\frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{y}; \alpha\boldsymbol{\tau}^0) = \frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \alpha\cdot\frac{\|\boldsymbol{y}\|}{n}\cdot\boldsymbol{v}^0$. As proved in the Appendix, the support of each iterate $\mathcal{S}^t$ and its size $s^t = |\mathcal{S}|$ are locally constant with respect to $\boldsymbol{z}_j$. So we are treating $\frac{\partial\boldsymbol{\beta}^t}{\partial\boldsymbol{z}_j}$ as $\frac{\partial\boldsymbol{\beta}^t_{\mathcal{S}^t}}{\partial\boldsymbol{z}_j}$ paddled with $\mathbf{0}$'s, and obtain

$$\frac{\partial\boldsymbol{\beta}^1_{\mathcal{S}^1}}{\partial\boldsymbol{z}_j^\top} = \frac{1}{n}\Sigma^{-1}_{\mathcal{S}^1,j}\boldsymbol{y}^\top + \frac{\beta_j}{n}\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top - \frac{\alpha\beta_j}{n}\boldsymbol{v}^0_{\mathcal{S}^1}\frac{\boldsymbol{y}^\top}{\|\boldsymbol{y}\|}.$$

Consequently,

$$\begin{aligned}\frac{\partial f^1(\boldsymbol{z}_j)}{\partial\boldsymbol{z}_j^\top} &= [\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\mathcal{S}^1}\frac{\partial\boldsymbol{\beta}^1_{\mathcal{S}^1}}{\partial\boldsymbol{z}_j}\\ &= \frac{1}{n}[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\mathcal{S}^1}\Sigma^{-1}_{\mathcal{S}^1,j}\boldsymbol{y}^\top + \frac{\beta_j}{n}[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\mathcal{S}^1}\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top - \frac{\alpha\beta_j}{n}[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\mathcal{S}^1}\boldsymbol{v}^0_{\mathcal{S}^1}\frac{\boldsymbol{y}^\top}{\|\boldsymbol{y}\|}\\ &= \frac{1}{n}[\boldsymbol{X} - \boldsymbol{z}_j\boldsymbol{e}_j^\top]_{\cdot,\mathcal{S}^1}\Sigma^{-1}_{\mathcal{S}^1,j}\boldsymbol{y}^\top + \frac{\beta_j}{n}[\boldsymbol{X} - \boldsymbol{z}_j\boldsymbol{e}_j^\top]_{\cdot,\mathcal{S}^1}\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top - \frac{\alpha\beta_j}{n}[\boldsymbol{X} - \boldsymbol{z}_j\boldsymbol{e}_j^\top]_{\cdot,\mathcal{S}^1}\boldsymbol{v}^0_{\mathcal{S}^1}\frac{\boldsymbol{y}^\top}{\|\boldsymbol{y}\|}.\end{aligned}$$

Therefore,

$$\begin{aligned}trace\left\{\frac{\partial f^1(\boldsymbol{z}_j)}{\partial\boldsymbol{z}_j^\top}\right\} &\approx trace\left\{\frac{\beta_j}{n}\boldsymbol{X}_{\cdot,\mathcal{S}^1}\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top\right\}\\ &= trace\left\{\frac{\beta_j}{n}\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top\boldsymbol{X}_{\cdot,\mathcal{S}^1}\right\}\\ &\approx \beta_j\cdot s^1\end{aligned}$$

Here, the first approximation is due to the ignorance of rank-one matrices as their traces are of smaller order than the full-rank matrices. The last approximate is due to the fact $\Sigma^{-1}_{\mathcal{S}^1,\cdot}\boldsymbol{X}^\top\boldsymbol{X}_{\cdot,\mathcal{S}^1} \approx n\cdot\boldsymbol{I}_{s^1}$. On another hand,

$$\begin{aligned}\frac{\partial g^0(\boldsymbol{z}_j)}{\partial\boldsymbol{z}_j^\top} &= \frac{\partial\boldsymbol{y}}{\partial\boldsymbol{z}_j\top}\\ &= \beta_j\cdot\boldsymbol{I}_n\\ trace\left\{\frac{\partial g^0(\boldsymbol{z}_j)}{\partial\boldsymbol{z}_j^\top}\right\} &= \beta_j\cdot n.\end{aligned}$$

It follows

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^1 - \boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right] &= \beta_j\cdot s^1 = \frac{s^1}{n}\mathbb{E}\left[\boldsymbol{z}_j^\top\boldsymbol{r}^0\middle|\boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon}\right],\\ \mathbb{E}\left[\beta_j^1 + \frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top\boldsymbol{r}^1\right] &\approx \beta_j.\end{aligned}$$

When $t \geq 2$, the term inside the thresholding function is

$$\begin{aligned}&\beta_j^t + \frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t + \frac{s^t}{n}\boldsymbol{r}^{t-1})\\ =&\beta_j + \frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top\boldsymbol{\varepsilon} - \frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t - \boldsymbol{\beta}) + \frac{s^t\Sigma^{-1}_{jj}}{n^2}\boldsymbol{z}_j^\top\boldsymbol{r}^{t-1}\end{aligned}$$

And we need to show

$$\mathbb{E}\left[\boldsymbol{z}_j^\top \boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \approx \frac{s^t}{n}\mathbb{E}\left[\boldsymbol{z}_j^\top \boldsymbol{r}^{t-1}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right].$$

Again, the coordinates outside $\mathcal{S}^t$ are locally constant with respect to $\boldsymbol{z}_j$, and hence

$$\begin{aligned}\frac{\partial \boldsymbol{\beta}^t}{\partial \boldsymbol{z}_j^\top} &= \frac{\partial \boldsymbol{\beta}^t_{\mathcal{S}^t}}{\partial \boldsymbol{z}_j^\top} \\ &= \left(\boldsymbol{I}-\Sigma^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)_{\mathcal{S}^t,\cdot}\frac{\partial \boldsymbol{\beta}^{t-1}}{\partial \boldsymbol{z}_j^\top}+\frac{\beta_j-\beta_j^{t-1}}{n}\Sigma^{-1}_{\mathcal{S}^t,\cdot}\boldsymbol{X}^\top+\frac{s^{t-1}}{n^2}\Sigma^{-1}_{\mathcal{S}^t,\cdot}\boldsymbol{X}^\top\frac{\partial \boldsymbol{r}^{t-2}}{\partial \boldsymbol{z}_j^\top}.\end{aligned}$$

On another hand, $trace\left\{[\boldsymbol{X}\boldsymbol{Q}_j]_{\cdot,\mathcal{S}^t}\frac{\partial \boldsymbol{\beta}^t_{\mathcal{S}^t}}{\partial \boldsymbol{z}_j^\top}\right\} \approx trace\left\{\boldsymbol{X}_{\cdot,\mathcal{S}^t}\frac{\partial \boldsymbol{\beta}^t_{\mathcal{S}^t}}{\partial \boldsymbol{z}_j^\top}\right\}$. Then we have

$$\begin{aligned}trace\left\{\boldsymbol{X}_{\cdot,\mathcal{S}^t}\frac{\partial \boldsymbol{\beta}^t_{\mathcal{S}^t}}{\partial \boldsymbol{z}_j^\top}\right\} &\approx trace\left\{\boldsymbol{X}_{\cdot,\mathcal{S}^t}\left(\boldsymbol{I}-\Sigma^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)_{\mathcal{S}^t,\cdot}\frac{\partial \boldsymbol{\beta}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\}+\frac{\beta_j-\beta_j^{t-1}}{n}trace\left\{\Sigma^{-1}_{\mathcal{S}^t,\cdot}\boldsymbol{X}^\top\boldsymbol{X}_{\cdot,\mathcal{S}^t}\right\} \\ &\quad+\frac{s^{t-1}}{n^2}trace\left\{\boldsymbol{X}_{\cdot,\mathcal{S}^t}\Sigma^{-1}_{\mathcal{S}^t,\cdot}\boldsymbol{X}^\top\frac{\partial \boldsymbol{r}^{t-2}}{\partial \boldsymbol{z}_j^\top}\right\} \qquad (2.17)\\ &\approx(\beta_j-\beta_j^{t-1})s^t+\frac{s^{t-1}s^t}{n^2}trace\left\{\frac{\partial \boldsymbol{r}^{t-2}}{\partial \boldsymbol{z}_j^\top}\right\} \qquad (2.18)\end{aligned}$$

Using similar techniques, we get

$$trace\left\{\frac{\partial \boldsymbol{r}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\} \approx -trace\left\{\boldsymbol{X}\frac{\partial \boldsymbol{\beta}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\}+\frac{s^{t-1}}{n}trace\left\{\frac{\partial \boldsymbol{r}^{t-2}}{\partial \boldsymbol{z}_j^\top}\right\}+(\beta_j-\beta_j^{t-1})n \qquad (2.19)$$

It follows that

$$trace\left\{\boldsymbol{X}\boldsymbol{Q}_j\frac{\partial \boldsymbol{\beta}^t}{\partial \boldsymbol{z}_j^\top}\right\}-\frac{s^t}{n}trace\left\{\frac{\partial \boldsymbol{r}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\} \approx \frac{s^t}{n}trace\left\{\boldsymbol{X}\frac{\partial \boldsymbol{\beta}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\}$$

$$\frac{\Sigma^{-1}_{jj}}{n}\mathbb{E}\left[\boldsymbol{z}_j^\top \boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t-\boldsymbol{\beta})\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right]-\frac{s^t\Sigma^{-1}_{jj}}{n^2}\mathbb{E}\left[\boldsymbol{z}_j^\top \boldsymbol{r}^{t-1}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right] \approx \frac{s^t\Sigma^{-1}_{jj}}{n^2}trace\left\{\boldsymbol{X}\frac{\partial \boldsymbol{\beta}^{t-1}}{\partial \boldsymbol{z}_j^\top}\right\}=o(1)$$

2.5 Approximate thresholding value for each iteration

A natural choice for the thresholding value $(\tau_j^t)^2$ is the variance of $\frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top\boldsymbol{\varepsilon}-\frac{\Sigma^{-1}_{jj}}{n}\boldsymbol{z}_j^\top\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t-\boldsymbol{\beta})+\frac{s^t\Sigma^{-1}_{jj}}{n^2}\boldsymbol{z}_j^\top\boldsymbol{r}^{t-1}$. Let

$$g^t(\boldsymbol{z}_j)=\boldsymbol{\varepsilon}-\boldsymbol{X}\boldsymbol{Q}_j(\boldsymbol{\beta}^t-\boldsymbol{\beta})+\frac{s^t}{n}\boldsymbol{r}^{t-1}.$$

By Theorem 1.1 in [3], we derive the variance of $\boldsymbol{z}_j^\top g^t(\boldsymbol{z}_j)$ as

$$
\begin{aligned}
&Var\left[\frac{\Sigma_{jj}^{-1}}{n}\boldsymbol{z}_j^\top g^t(\boldsymbol{z}_j)\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right]\\
=&\frac{\Sigma_{jj}^{-1}}{n^2}\mathbb{E}\left[\|g^t(\boldsymbol{z}_j)\|^2+\frac{1}{\Sigma_{jj}^{-1}}trace\left\{[\frac{\partial g^t}{\partial \boldsymbol{z}_j^\top}]^2\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right]\\
=&\frac{\Sigma_{jj}^{-1}}{n^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^t+\frac{s^t}{n}\boldsymbol{r}^{t-1}+(\beta_j^t-\beta_j)\boldsymbol{z}_j\|^2+\frac{1}{n^2}\mathbb{E}\left[trace\left\{[\frac{\partial g^t}{\partial \boldsymbol{z}_j^\top}]^2\right\}\middle|\boldsymbol{X}\boldsymbol{Q}_j,\boldsymbol{\varepsilon}\right]\\
\approx&\frac{\Sigma_{jj}^{-1}}{n^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^t+\frac{s^t}{n}\boldsymbol{r}^{t-1}+(\beta_j^t-\beta_j)\boldsymbol{z}_j\|^2+O\left(\frac{(s^{t-1})^2(s^t)^2}{n^4}\right)\\
\approx&\frac{\Sigma_{jj}^{-1}}{n^2}\|\boldsymbol{r}^t\|^2
\end{aligned}
$$

2.6 The convergence of the generalized AMP iteration

In this part, we show that the generalized AMP iteration actually converges, and meanwhile obtain its convergence rate. Let $\boldsymbol{h}^t \coloneqq \boldsymbol{\beta}^t-\boldsymbol{\beta}$, and $\mathcal{S}\coloneqq supp(\boldsymbol{\beta})$. The iterative relationship between $\boldsymbol{h}^{t+1}$ and $\boldsymbol{h}^t$ is

$$
\boldsymbol{h}^{t+1}=\big(\boldsymbol{I}-\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{X}/n\big)\boldsymbol{h}^t+\frac{s^t}{n^2}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{r}^{t-1}+\frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}-\alpha\cdot\frac{\|\boldsymbol{r}^t\|}{n}\boldsymbol{v}^t, \tag{2.20}
$$

where

$$
v_j^t\begin{cases}
=-\sqrt{\Sigma_{jj}^{-1}} & \text{if } \beta_j^{t+1}>0\\
\in[-\sqrt{\Sigma_{jj}^{-1}},\sqrt{\Sigma_{jj}^{-1}}] & \text{if } \beta_j^{t+1}=0\\
=\sqrt{\Sigma_{jj}^{-1}} & \text{if } \beta_j^{t+1}<0.
\end{cases}
$$

Hereinafter, we abbreviate $\frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{r}^{t-1}$ to $\boldsymbol{a}^{t-1}$, and $\frac{1}{n}\Sigma^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$ to $\boldsymbol{b}$. Suppose we can prove that the support of each $\boldsymbol{h}^t$ is inside a certain set $\mathcal{S}^*$ such that $|\mathcal{S}^*|=O(|\mathcal{S}|)\ll n$. Then (2.20) reduces to

$$
\boldsymbol{h}_{\mathcal{S}^*}^{t+1}=\big(\boldsymbol{I}_{\mathcal{S}^*,\mathcal{S}^*}-\boldsymbol{Z}_{\mathcal{S}^*,\cdot}^\top\boldsymbol{X}_{\cdot,\mathcal{S}^*}/n\big)\boldsymbol{h}_{\mathcal{S}^*}^t+\frac{s^t}{n}\boldsymbol{a}_{\mathcal{S}^*}^{t-1}+\boldsymbol{b}_{\mathcal{S}^*}-\alpha\cdot\frac{\|\boldsymbol{r}^t\|}{n}\boldsymbol{v}_{\mathcal{S}^*}^t \tag{2.21}
$$

And it is easy to show that $\left\|\boldsymbol{I}_{\mathcal{S}^*,\mathcal{S}^*}-\boldsymbol{Z}_{\mathcal{S}^*,\cdot}^\top\boldsymbol{X}_{\cdot,\mathcal{S}^*}/n\right\|_{op}=O(\sqrt{|\mathcal{S}^*|/n})\ll 1$. So $\boldsymbol{h}^t$ is contracting.

3 Numerical simulation

We conduct the simulation experiments under the setting: p=1200, n=800, s=50, $\Sigma_{ij}=\rho^{|i-j|}$, $\boldsymbol{x}_i\sim N(\boldsymbol{0},\Sigma)$, noise $\sim N(0,1)$. The following simulation results will show that the iterative algorithm (2.1)-(2.3) converges fast (within about 7 iterations) and the stationary solution enjoys reasonable error rates, sparsity and false negative rate, where "false negative" means the proportion of the true support that is not on the support of the stationary solution. Besides, we explore the varying properties of the stationary solution under different combinations of ρ and α.

3.1 Convergence and basic properties

Our experiment consists the following procedure:
1. generating a true $\boldsymbol{\beta}$ by sampling its support from N(0,1); computing the covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.2$.
2. repeating for 500 times: drawing $\boldsymbol{x}_i \sim N(0, \Sigma)$ and generating $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, 2, \cdots, n$; run the generalized AMP iteration with $(\boldsymbol{X}, \boldsymbol{y})$ for T=20 steps; record the relative $l2$ error $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_2^2 / \|\boldsymbol{\beta}\|^2$ and the relative $l1$ error $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_1 / \|\boldsymbol{\beta}\|_1$ for each $t = 1, 2, \cdots, T$; also record the sparsity and the "false negative number" of the stationary solution.
3. Each dot in Fig 1. (A) represents the average of 500 relative $l2$ errors at tth iteration; Each dot in Fig 1. (B) represents the average of 500 relative $l1$ errors at tth iteration; Fig 1. (C) depicts the dispersion of the sparsity of 500 repeats; Fig 1. (D) shows the "false negative numbers" of 500 repeats.

The "false negative" coordinates which are missed by the stationary solution are small coordinates, with the most of them $\beta_j < 0.1$ and the rest lying within $(0.1, 0.25)$.

3.2 Normality of the term inside the soft-thresholding function

We conjecture that the term $\hat{\beta}_j + \frac{\Sigma_{j\cdot}^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - \hat{s}}$ inside the soft-thresholding function after reaching stationary status will distribute approximately Gaussian centered around the true β_j. Without rigorous proof, the simulation results are supporting this conjecture (Fig 2).

The experiment consists:
1. generating a true $\boldsymbol{\beta}$ by sampling its support from N(0,1); computing the covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.2$.
2. repeating for 500 times: drawing $\boldsymbol{x}_i \sim N(0, \Sigma)$ and generating $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, 2, \cdots, n$; run the generalized AMP iteration with $(\boldsymbol{X}, \boldsymbol{y})$ for T=20 steps; record the errors $\hat{\beta}_j + \frac{\Sigma_{j\cdot}^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - \hat{s}} - \beta_j$ for each coordinate j. 3. each plot in Fig 2. shows the distribution of 500 errors for a single coordinate. The first row shows the error distributions of four large coordinates ($\beta_j \geq 1.8$); the second row shows the error distributions of four median coordinates ($0.9 \leq \beta_j \leq 1.1$), *etc.*

3.3 Second order Stein's method accurately predicts the variance

We used second order Stein's method to derive (2.3), which is expected to be approximately the variance of the term $\hat{\beta}_j + \frac{\Sigma_{j\cdot}^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - \hat{s}}$. The following simulation results further supports our theoretical derivation. With similar experiment as in section 3.2, we compute the sample variance of $\hat{\beta}_j + \frac{\Sigma_{j\cdot}^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - \hat{s}} - \beta_j$ over 500 repeats, and then compare the sample variance with the predicted variance $(\hat{\tau}_j)^2$. As $(\hat{\tau}_j)^2$ itself is a random variable with varying $\boldsymbol{X}$'s, we use its average value over 500 repeats. In fact, $(\hat{\tau}_j)^2$ does not vary much over 500 repeats, it concentrates tightly around its mean (result not shown). For those coordinates with the same predicted variance value $(\hat{\tau}_j)^2$, we compare the sample variance and the predicted variance.

Judging from Fig 3, the sample variance matches exactly the predicted variance with a little dispersion.

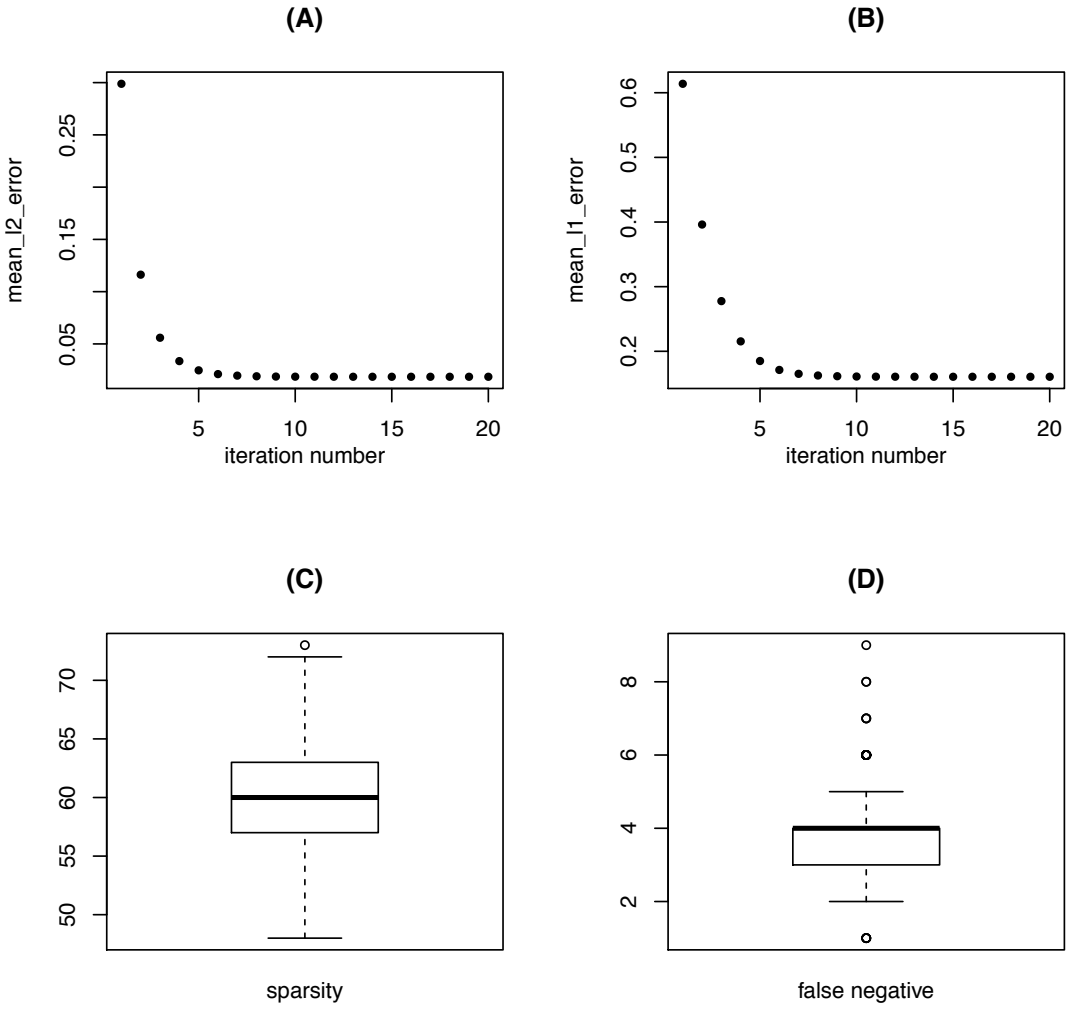

Figure 1: basic properties of the stationary solution

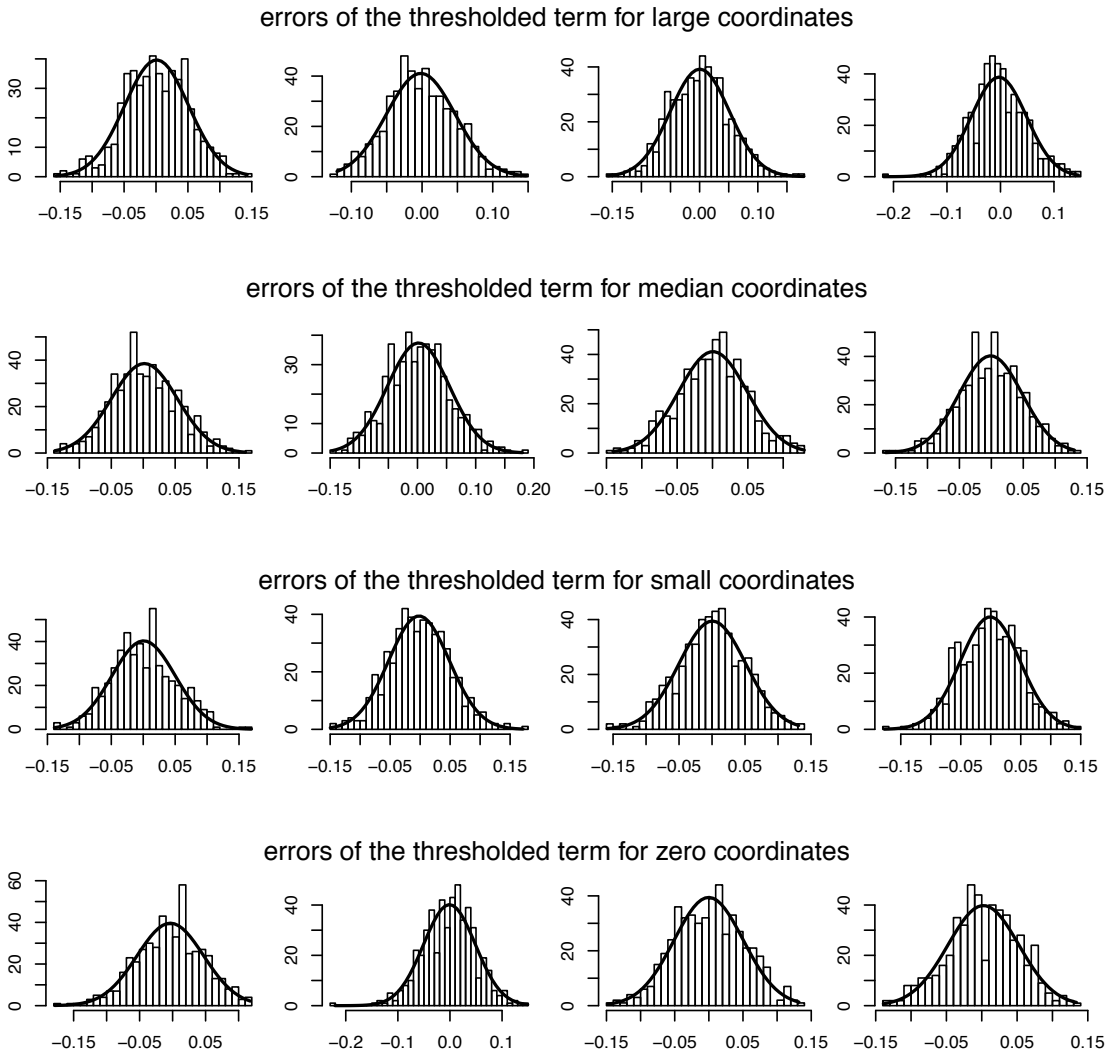

Figure 2: approximate normal distribution of the term to be thresholded on after stationary status

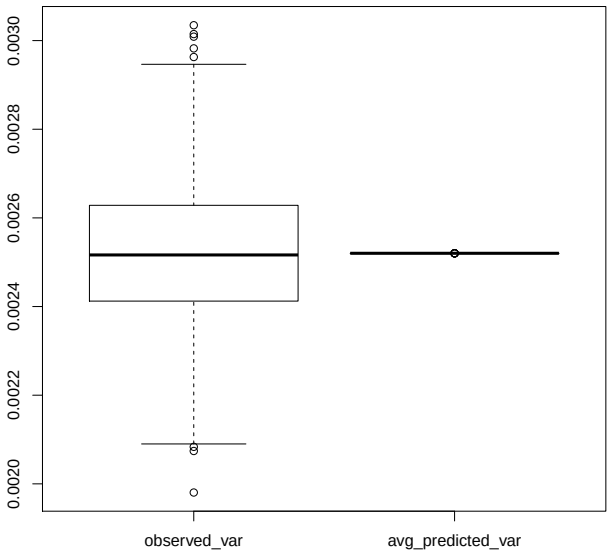

Figure 3: observed variance vs. predicted variance

References

[1] Mohsen Bayati and Andrea Montanari. "The dynamics of message passing on dense graphs, with applications to compressed sensing". In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 764–785.

[2] Mohsen Bayati and Andrea Montanari. "The LASSO risk for Gaussian matrices". In: *IEEE Transactions on Information Theory* 58.4 (2012), pp. 1997–2017.

[3] Pierre C Bellec and Cun-Hui Zhang. "Second order Stein: SURE for SURE and other applications in high-dimensional inference". In: *arXiv preprint arXiv:1811.04121* (2018).

[4] Pierre C Bellec and Cun-Hui Zhang. "De-Biasing The Lasso With Degrees-of-Freedom Adjustment". In: *arXiv preprint arXiv:1902.08885* (2019).

[5] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. "Atomic decomposition by basis pursuit". In: *SIAM review* 43.1 (2001), pp. 129–159.

[6] David L Donoho, Arian Maleki, and Andrea Montanari. "Message-passing algorithms for compressed sensing". In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919.

[7] David L Donoho, Arian Maleki, and Andrea Montanari. "How to design message passing algorithms for compressed sensing". In: *preprint* (2011).

[8] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[9] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[10] Adel Javanmard and Andrea Montanari. "Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory". In: *IEEE Transactions on Information Theory* 60.10 (2014), pp. 6522–6554.

[11] Arian Maleki and David L Donoho. "Optimally tuned iterative reconstruction algorithms for compressed sensing". In: *IEEE Journal of Selected Topics in Signal Processing* 4.2 (2010), pp. 330–341.

[12] Mark Rudelson and Roman Vershynin. "On sparse reconstruction from Fourier and Gaussian measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61.8 (2008), pp. 1025–1045.

[13] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

Appendices

We use induction to show that $\mathcal{S}^t$ and s^t stay constant in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$ for all $t = 1, 2, \cdots$

InductionAssumption **:**
1. Assume $\boldsymbol{\beta}^t(\boldsymbol{z}_j)$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$ for $t = 1, 2, \cdots, t^*$
2. Assume s^t does not change in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$ for $t = 1, 2, \cdots, t^*$ i.e. $\frac{\partial s^t}{\partial \boldsymbol{z}_j} = \mathbf{0}$.

The goal is to show these two assumptions also hold for $t = t^* + 1$ provided they hold for all $1 \leq t \leq t^*$

First, we go through the first iteration (t=1) to verify that the above induction assumptions hold for t=1. With initialization $\boldsymbol{\beta}^0 = \mathbf{0}$

Let $\mathcal{S}^1$ denote the support of $\boldsymbol{\beta}^1$, conditional on $\boldsymbol{X}\boldsymbol{Q}_j,\ \boldsymbol{\varepsilon}$, for all $j \in [p]$:

$$s^1 = |\mathcal{S}^1| = \sum_{j=1}^{p} \mathbb{1}\{\left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| > \alpha\tau_j^0\}$$

The part $\boldsymbol{z}_j^T\boldsymbol{y} = \boldsymbol{z}_j^T(\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\beta} + \beta_j\boldsymbol{z}_j)$ is continuously differentiable with respect to $\boldsymbol{z}_j$. When $\boldsymbol{z}_j$ lies in a compact set of $\mathbb{R}^n$, which is true with probability one under the assumption $\boldsymbol{X} \sim N(0, \Sigma)$, $\boldsymbol{z}_j^T\boldsymbol{y}$ is Lipschitz continuous w.r.t. $\boldsymbol{z}_j$. Similarly,

$$\tau_j^0 = \frac{\sqrt{\Sigma_{jj}^{-1}}}{n}\|\boldsymbol{y}\| = \frac{\sqrt{\Sigma_{jj}^{-1}}}{n}\|\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\beta} + \beta_j\boldsymbol{z}_j\|,$$

is Lipschitz continuous with respect to $\boldsymbol{z}_j$. Hence, any j such that $\left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| > \alpha\tau_j^0$ remains in $\mathcal{S}^1$ when $\boldsymbol{z}_j$ is varying in a small enough non-trivial neighborhood. And all j's such that $\left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| < \alpha\tau_j^0$ remain in the complement of $\mathcal{S}^1$. Trouble rises in those j's with $\left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| = \alpha\tau_j^0$. We need to ensure that such event happens with probability zero. *i.e.* the following condition holds with probability one, which should be true when $(\boldsymbol{X}, \boldsymbol{y})$ admits a density with respect to Lebesgue measure.

$$j \in \mathcal{S}^1 \iff \left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| > \alpha\tau_j^0 \tag{3.1}$$

$$j \in \overline{\mathcal{S}^1} \iff \left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| < \alpha\tau_j^0 \tag{3.2}$$

where $\overline{\mathcal{S}^1}$ is the complement of $\mathcal{S}^1$. If (3.1) and (3.2) hold, $\mathcal{S}^1$ and s^1 stay constant in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$, i.e. $\frac{\partial s^1}{\partial \boldsymbol{z}_j} = \mathbf{0}$.

As for $\boldsymbol{\beta}^1$:

$$\begin{aligned}\beta_j^1 &= \eta(\beta_j^0 + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^0)}{n-s^0}, \alpha\tau_j^0) = \eta(\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}, \alpha\tau_j^0) \\ &= \mathbb{1}\{\left|\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}\right| > \alpha\tau_j^0\}\left(\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n} - sign(\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n})\cdot\alpha\tau_j^0\right)\end{aligned}$$

We know that $\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n}$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$, and hence, $sign(\frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{y}}{n})$ does not change in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$. Combined with the fact that τ_j^0 is also Lipschitz continuous with respect to $\boldsymbol{z}_j$, we obtain that β_j^1 is Lipschitz continuous with respect to $\boldsymbol{z}_j$ and $\frac{\partial\beta_i^1}{\partial\boldsymbol{z}_j^T}$ exists almost everywhere for $j = 1, \cdots, p$. For $i \neq j$,

$$\frac{\partial\beta_i^1}{\partial\boldsymbol{z}_j^T} = \frac{\partial\beta_i^1}{\partial\boldsymbol{z}_i^T}\cdot\frac{\partial\boldsymbol{z}_i}{\partial\boldsymbol{z}_j^T} = \frac{\Sigma_{ji}^{-1}}{\Sigma_{ii}^{-1}}\cdot\frac{\partial\beta_i^1}{\partial\boldsymbol{z}_i^T}$$

Therefore, β_i^1 is Lipschitz continuous with respect to $\boldsymbol{z}_j$ for all $i = 1, \cdots, p$, which further implies $\boldsymbol{\beta}^1$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$.

Next, we will show the two induction assumptions also hold for t^*+1 given that they hold for $t = 1, 2, \cdots, t^*$.

$$s^{t^*+1} = |\mathcal{S}^{t^*+1}| = \sum_{j=1}^{p}\mathbb{1}\{\left|\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}\right| > \alpha\tau_j^{t^*}\}$$

We know that $\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$ because $\beta_j^{t^*}$ is Lipschitz continuous (by the induction assumption) and $\boldsymbol{r}^{t^*}$ is Lipschitz continuous (by the iteration formula 2.2). We also know that $\tau_j^{t^*}$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$ by the iteration formula (2.3). Agian if the following condition holds strictly (without equality), *i.e.*

$$j \in \mathcal{S}^{t^*+1} \iff \left|\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}\right| > \alpha\tau_j^{t^*} \tag{3.3}$$

$$j \in \overline{\mathcal{S}^{t^*+1}} \iff \left|\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}\right| < \alpha\tau_j^{t^*} \tag{3.4}$$

which should be true when $(\boldsymbol{X}, \boldsymbol{y})$ admits a density with respect to Lebesgue measure. Then $\mathcal{S}^{t^*+1}$ and s^{t^*+1} remain constant in a small enough non-trivial neighborhood of $\boldsymbol{z}_j$.

$$\begin{aligned}\beta_j^{t^*+1} &= \eta\left(\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}; \alpha\tau_j^{t^*}\right) \\ &= \mathbb{1}\{\left|\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n}\right| > \alpha\tau_j^{t^*}\}\left(\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n} - sign(\beta_j^{t^*} + \frac{\Sigma_{jj}^{-1}\boldsymbol{z}_j^T\boldsymbol{r}^{t^*}}{n})\cdot\alpha\tau_j^{t^*}\right)\end{aligned}$$

Using the similar argument as for β_j^1, we obtain that $\beta_j^{t^*+1}$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$. And for $i \neq j$,

$$\frac{\partial \beta_i^{t^*+1}}{\partial \boldsymbol{z}_j^T} = \frac{\partial \beta_i^{t^*+1}}{\partial \boldsymbol{z}_i^T} \cdot \frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{z}_j^T} = \frac{\Sigma_{ji}^{-1}}{\Sigma_{ii}^{-1}} \cdot \frac{\partial \beta_i^{t^*+1}}{\partial \boldsymbol{z}_i^T}$$

which means $\beta_i^{t^*+1}$ is Lipschitz with respect to $\boldsymbol{z}_j$ for all $i = 1, 2, \cdots, p$. Therefore, $\boldsymbol{\beta}^{t^*+1}$ is Lipschitz continuous with respect to $\boldsymbol{z}_j$.

So far, we have shown that the induction assumptions hold for all $t = 1, 2, \cdots$.