Additional Supplementary Materials for "A Modern Maximum-Likelihood Theory for High-dimensional Logistic Regression"

Pragya Sur* Emmanuel J. Candès*†

June 14, 2018

Abstract

This document presents proofs of Theorems 2,3 and 4 in the paper [8]: "A modern Maximum Likelihood Theory for High-Dimensional Logistic Regression".

1 The results

We recall Theorems 2, 3 and 4 from [8] below.[1]

Theorem 1. *Assume the dimensionality and signal strength parameters κ and γ are such that $\gamma < g_{\mathrm{MLE}}(\kappa)$ (the region where the MLE exists asymptotically as characterized in [2]).[2] For any pseudo-Lipschitz function ψ of order 2, the marginal distributions of the MLE coordinates obey*

$$\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j-\alpha_\star\beta_j,\beta_j\right)\xrightarrow{\text{a.s.}}\mathbb{E}\left[\psi\left(\sigma_\star Z,\beta\right)\right],\quad Z\sim\mathcal{N}\left(0,1\right), \tag{1}$$

where $\beta\sim\Pi$, independent of Z.

Theorem 2. *Let j be any variable such that $\beta_j=0$. Then in the setting of Theorem 1, the MLE obeys*

$$\hat{\beta}_j\xrightarrow{\text{d}}\mathcal{N}\left(0,\sigma_\star^2\right). \tag{2}$$

For any finite subset of null variables $\{i_1,\ldots,i_k\}$, the components of $(\hat{\beta}_{i_1},\ldots,\hat{\beta}_{i_k})$ are asymptotically independent.

Theorem 3. *Consider the LLR $\Lambda_j=\min_{\boldsymbol{b}\,:\,b_j=0}\ell(\boldsymbol{b})-\min_{\boldsymbol{b}}\ell(\boldsymbol{b})$ for testing $\beta_j=0$, where $\ell(\boldsymbol{b})$ is the negative log-likelihood function. In the setting of Theorem 1, twice the LLR is asymptotically distributed as a multiple of a chi-square under the null,*

$$2\Lambda_j\xrightarrow{\text{d}}\frac{\kappa\,\sigma_\star^2}{\lambda_\star}\,\chi_1^2. \tag{3}$$

Also, the LLR for testing $\beta_{i_1}=\beta_{i_2}=\ldots=\beta_{i_k}=0$ for any finite k converges to the rescaled chi-square $\left(\kappa\sigma_\star^2/\lambda_\star\right)\chi_k^2$ under the null.

*Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
†Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.

[1] Notations are the same as in [8].

[2] See [2] for a definition of $g_{\mathrm{MLE}}(\gamma)$.

In the aforementioned results, $(\alpha_\star, \sigma_\star, \lambda_\star)$ is a solution to the system of equations:

$$\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'\left(Q_1\right)\left(\lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)\right)^2\right] \\ 0 = \mathbb{E}\left[\rho'(Q_1)Q_1\lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)\right] \\ 1-\kappa = \mathbb{E}\left[\dfrac{2\rho'\left(Q_1\right)}{1+\lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)}\right] \end{cases} \tag{4}$$

where (Q_1, Q_2) is a bivariate normal variable with mean $\mathbf{0}$ and covariance

$$\mathbf{\Sigma}\left(\alpha,\sigma\right) = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2+\kappa\sigma^2 \end{bmatrix}. \tag{5}$$

It can be easily checked numerically that in the regime $\gamma < g_{\mathrm{MLE}}(\kappa)$ the system (4) admits a solution. Hence, we omit proving this fact. However, we establish that in the aforementioned regime, if (4) admits a solution then the solution must be unique.[3] Thus, the parameters $(\alpha_\star, \sigma_\star, \lambda_\star)$ are well-defined in our setup.

The proximal mapping operator for any $\lambda > 0$ and convex function ρ is defined via

$$\mathsf{prox}_{\lambda\rho}\left(z\right) = \arg\min_{t\in\mathbb{R}}\left\{\lambda\rho\left(t\right) + \frac{1}{2}\left(t-z\right)^2\right\}. \tag{6}$$

In the subsequent text, it will be useful to note that the proximal mapping operator satisfies the relation:

$$\lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(z\right)\right) + \mathsf{prox}_{\lambda\rho}\left(z\right) = z. \tag{7}$$

2 Road map to the proofs

This section presents the key steps in the proofs of each theorem. Detailed proofs are provided in Sections 4–6. At a high level, the proof of Theorem 1 has the following ingredients:

1. Introduce an iterative algorithm that has iterates $\{\hat{\boldsymbol{\beta}}^t\}_{t\geq 0}$, with the aim of tracking the large sample behavior of the MLE. This was already done in [8, Section 4.1].

2. Characterize the asymptotic distribution of $\{\hat{\boldsymbol{\beta}}^t\}_{t\geq 0}$ for each fixed t, in the large sample limit. (See Theorem 6). Here, we resort to existing results in the generalized approximate message passing (G-AMP) literature [7]. However, to apply these results, one needs to establish that the algorithm introduced in the first step can be cast in the framework of a G-AMP algorithm. This is a highly non-trivial step and forms the core of the proof of Theorem 6.

3. Establish that in the large sample and large iteration limit, $\hat{\boldsymbol{\beta}}^t$ converges to the MLE $\hat{\boldsymbol{\beta}}$ in an appropriate sense (see Theorem 7). In conjunction with the previous step, this provides the desired result.

In the logistic model, the MLE is far from exhibiting any closed form expression. In fact, all information about it is contained in the optimality condition $\nabla\ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Thus, the analysis of a single null coordinate is hard. To circumvent this difficulty, we resort to the following two stage-approach:

1. Replace the MLE by a surrogate which is amenable to explicit mathematical analysis (Theorem 8). In turn, this approximation yields a convenient representation of a null coordinate.

2. Characterize the asymptotic distribution of the aforementioned representation. This is the content of the rest of the arguments in Section 5.

[3] See Remark 1 for a detailed explanation of this fact.

Finally, we arrive at Theorem 3, the proof of which can be summarized in the following two steps:

1. In Theorem 9, we establish that if $\beta_j = 0$, the quantity of interest $2\Lambda_j$ can be approximated as follows:

$$2\Lambda_j = \frac{\kappa \hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1),$$

where $\hat{\beta}_j$ denotes the j-th coordinate of the MLE, and $\lambda_{[-j]}$ defined later in (84) is a function of the Hessian of the negative log-likelihood.

2. Theorem 2 already established that $\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_\star^2)$. Thus, it suffices to show that $\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star$. This is achieved in Theorem 10, deploying techniques similar to that in [9, Appendix I].

3 Crucial building blocks

This section gathers a few important results that will be useful throughout the manuscript. Let $C_0, C_1, \ldots$, $c_0, c_1, \ldots$ denote positive universal constants, independent of n and p, whose value can change from line to line. We start by recalling a recursion from [8], and expressing it in an equivalent form.

3.1 A Useful Recursion

In [8], the authors introduced a sequence of scalar parameters: $\{\alpha_t, \sigma_t, \lambda_t\}_{t \geq 0}$, defined recursively as follows. Let (Q_1^t, Q_2^t) be a bivariate normal variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(\alpha_t, \sigma_t)$ specified by (5). Starting from an initial pair α_0, σ_0, for $t = 0, 1, \ldots$, inductively define λ_t as the solution to

$$\mathbb{E}\left[\frac{2\rho'\left(Q_1^t\right)}{1 + \lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(Q_2^t\right)\right)}\right] = 1 - \kappa, \tag{8}$$

and define $\alpha_{t+1}, \sigma_{t+1}$ as

$$\begin{aligned} \alpha_{t+1} &= \alpha_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[2\rho'\left(Q_1^t\right) Q_1^t \lambda_t \rho'\left(\mathsf{prox}_{\lambda_t\rho}\left(Q_2^t\right)\right)\right], \\ \sigma_{t+1}^2 &= \frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'\left(Q_1^t\right)\left(\lambda_t\rho'\left(\mathsf{prox}_{\lambda_t\rho}\left(Q_2^t\right)\right)\right)^2\right]. \end{aligned} \tag{9}$$

Our goal is to express the aforementioned recursive system in an equivalent form. To this end, we introduce a new sequence of scalar parameters $\{\tilde{\alpha}_t, \tilde{\sigma}_t, \tilde{\lambda}_t\}_{t \geq 0}$ defined as follows. Let $(\tilde{Q}_1^t, \tilde{Q}_2^t)$ be a bivariate normal variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(-\tilde{\alpha}_t, \tilde{\sigma}_t)$. Further, let $W \sim \mathsf{Unif}(0,1)$, independent of $(\tilde{Q}^t, \tilde{Q}_2^t)$ for all $t \geq 0$. Define the function

$$h(x, y) = \mathbf{1}_{y \leq \rho'(x)}, \text{ where } \rho'(x) = \frac{e^x}{1 + e^x}. \tag{10}$$

Starting with initial conditions $\tilde{\alpha}_0, \tilde{\sigma}_0$, for each $t \geq 0$, obtain $\tilde{\lambda}_t$ by solving

$$\mathbb{E}_{W, \tilde{Q}_1^t, \tilde{Q}_2^t}\left[\frac{1}{1 + \lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(\lambda h\left(\tilde{Q}_1^t, W\right) + \tilde{Q}_2^t\right)\right)}\right] = 1 - \kappa. \tag{11}$$

Subsequently, $\tilde{\alpha}_{t+1}, \tilde{\sigma}_{t+1}$ are updated via

$$\begin{aligned} \tilde{\alpha}_{t+1} &= \tilde{\alpha}_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[\tilde{Q}_1^t \tilde{\Psi}_t\left(\tilde{Q}_1^t, W, \tilde{Q}_2^t\right)\right], \\ \tilde{\sigma}_{t+1}^2 &= \frac{1}{\kappa^2}\,\mathbb{E}\left[\tilde{\Psi}_t^2\left(\tilde{Q}_1, W, \tilde{Q}_2^t\right)\right], \end{aligned} \tag{12}$$

where

$$\tilde{\Psi}_t\left(q_1, w, q_2\right) = \tilde{\lambda}_t\left[h\left(q_1, w\right) - \rho'\left(\mathsf{prox}_{\tilde{\lambda}_t\rho}\left(\tilde{\lambda}_t h\left(q_1, w\right) + q_2\right)\right)\right]. \tag{13}$$

We propose simplifying the right-hand side (RHS) of the first equation in (12) by first conditioning on $(\tilde{Q}_1^t, \tilde{Q}_2^t)$. This gives

$$\begin{aligned}
&\mathbb{E}_{W,\tilde{Q}_1^t,\tilde{Q}_2^t}\left[\tilde{Q}_1^t\tilde{\Psi}_t(\tilde{Q}_1^t, W, \tilde{Q}_2^t)\right] \\
&\qquad = \mathbb{E}\left[\rho'(\tilde{Q}_1^t)\tilde{Q}_1^t(\tilde{\lambda}_t - \tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{\lambda}_t\tilde{Q}_2)))\right] - \mathbb{E}\left[(1-\rho'(\tilde{Q}_1))\tilde{Q}_1\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2))\right].
\end{aligned}$$

One can easily verify the following identity

$$\mathsf{prox}_{\lambda\rho}\left(\lambda + u\right) = -\mathsf{prox}_{\lambda\rho}\left(-u\right).$$

This yields

$$\tilde{\lambda}_t - \tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{\lambda}_t + \tilde{Q}_2^t)) = \tilde{\lambda}_t - \tilde{\lambda}_t\rho'(-\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t)) = \tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t)).$$

Combining the above relations, we have

$$\begin{aligned}
&\mathbb{E}_{W,\tilde{Q}_1^t,\tilde{Q}_2^t}\left[\tilde{Q}_1^t\tilde{\Psi}_t(\tilde{Q}_1^t, W, \tilde{Q}_2^t)\right] \\
&= \mathbb{E}\left[\rho'(\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))\right] - \mathbb{E}\left[(1-\rho'(\tilde{Q}_1^t))\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] \\
&= -\mathbb{E}\left[\rho'(-\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] - \mathbb{E}\left[(1-\rho'(\tilde{Q}_1^t))\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] \\
&= -2\,\mathbb{E}\left[\rho'(-\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right].
\end{aligned} \tag{14}$$

Performing similar calculations it can be shown that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\Psi}_{\tilde{\lambda}_t}^2(\tilde{Q}_1^t, W, \tilde{Q}_2^t)\right] &= \mathbb{E}\left[\rho'(\tilde{Q}_1^t)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))\right\}^2\right] + \mathbb{E}\left[(1-\rho'\left(\tilde{Q}_1^t\right)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right\}^2\right] \\
&= \mathbb{E}\left[2\rho'(-\tilde{Q}_1^t)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right\}^2\right].
\end{aligned} \tag{15}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{\lambda}_t h(\tilde{Q}_1^t, W) + \tilde{Q}_2^t))}\right] \qquad (16) \\
&= \mathbb{E}\left[\frac{\rho'(\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(-\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))}\right] + \mathbb{E}\left[\frac{1-\rho'(\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))}\right] \\
&= \mathbb{E}\left[\frac{2\rho'(-\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))}\right]. \qquad (17)
\end{aligned}$$

Placing together (14), (15) and (16), we have effectively established that, if $\alpha_0 = \tilde{\alpha}_0, \sigma_0 = \tilde{\sigma}_0$, then for all $t \geq 0$,

$$\alpha_t \equiv \tilde{\alpha}_t, \quad \sigma_t \equiv \tilde{\sigma}_t, \quad \lambda_t \equiv \tilde{\lambda}_t. \tag{18}$$

3.2 When is the MLE bounded?

It was established in [2] that if $\gamma < g_{\text{MLE}}(\kappa)$ (resp. $\gamma > g_{\text{MLE}}(\kappa)$), the MLE exists asymptotically with probability 1 (resp. 0). [2] further characterized the width of the window in which the phase transition occurs, in terms of the sample size. However, for establishing our main results Theorems 1–3, a stronger version of the phase transition phenomenon is necessary. We require that with exponentially high probability,

$$\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} = O(1)$$

in the regime $\gamma < g_{\text{MLE}}(\kappa)$. This is the content of the theorem below.

Theorem 4. *If $\gamma < g_{\text{MLE}}(\kappa)$, there exists $N_0 \equiv N_0(\gamma, \kappa)$ such that, for all $n \geq N_0$, the norm of the MLE $\hat{\boldsymbol{\beta}}$ obeys*

$$\mathbb{P}\left(\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} \leq C_1\right) \geq 1 - C_2 n^{-\delta}, \tag{19}$$

where $\delta > 1$.

Proof: By arguments similar to that in Section 5.2.2 from [8], it can be deduced that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} \leq \frac{4\log 2}{\varepsilon^2}\right) \geq \mathbb{P}\left(\{\boldsymbol{y} \circ (\boldsymbol{X}\boldsymbol{b}) \,|\boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} = \{\boldsymbol{0}\}\right), \tag{20}$$

where $\circ$ denotes the usual Hadamard product and $\mathcal{A}$ is a cone specified by

$$\mathcal{A} := \left\{\boldsymbol{u} \in \mathbb{R}^n | \sum_{j=1}^{n} \max\{-u_j, 0\} \leq \varepsilon^2\sqrt{n}\|\boldsymbol{u}\|\right\}. \tag{21}$$

Thus, it suffices to establish that the complement of the RHS of (20) has exponentially decaying probability. This is established in the remaining proof.

By rotational invariance, we can assume that all the signal lies in the first coordinate, that is, $\boldsymbol{\beta} = \sqrt{n}(\gamma_n, 0, 0, \ldots, 0)$, where $\gamma_n = \|\boldsymbol{\beta}\|^2/n$. Letting $\boldsymbol{X}_{i\bullet}$ denote the i-th row of $\boldsymbol{X}$, we have,

$$y_i \boldsymbol{X}_{i\bullet} \stackrel{\text{d}}{=} (V, X_2, \ldots X_p),$$

where $V \stackrel{\text{d}}{=} y_i X_{i1}$, with density given by $2\rho'(\gamma_n t)\phi(t)$ ($\phi(\cdot)$ denotes the standard normal density), and $V \perp\!\!\!\perp (X_2, \ldots, X_p)$. Denote $\boldsymbol{T} = [\boldsymbol{V}, \boldsymbol{X}_{\bullet 2}, \ldots, \boldsymbol{X}_{\bullet p}]$, that is, it is the matrix with the 2 through p-th columns same as that in $\boldsymbol{X}$, and the first column given by $(V_1, \ldots, V_n)$ where V_i's are i.i.d. copies of V. Then,

$$\mathbb{P}\left(\{\boldsymbol{y} \circ (\boldsymbol{X}\boldsymbol{b}) \,|\boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right) = \mathbb{P}\left(\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right). \tag{22}$$

With $\mathcal{G}$ defined to be the event

$$\mathcal{G} := \left[\text{span}(\boldsymbol{V}) \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right],$$

we can decompose the required probability as

$$\mathbb{P}\left(\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right) = \mathbb{P}(\mathcal{G}) + \mathbb{P}\left(\mathcal{G}^c \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right).$$

The following lemma ensures that $\mathbb{P}(\mathcal{G})$ decays to zero exponentially fast in n.

Lemma 1. *Let V be a continuous random variable with density $2\rho'(\gamma_n t)\phi(t)$, where $\gamma_n = \|\boldsymbol{\beta}\|/\sqrt{n}$. Suppose $V_1, \ldots, V_n$ are i.i.d. copies of V and $\boldsymbol{V} = (V_1, \ldots, V_n)$. There exists a fixed positive constant ε_1 such that,[4] for all $\varepsilon \leq \varepsilon_1$,*

$$\mathbb{P}\left(\text{span}(\boldsymbol{V}) \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right) \leq C_0 \exp(-c_0 n).$$

[4] Recall that the definition of $\mathcal{A}$ in (21) involved a choice of ε.

Henceforth, let $\varepsilon < \varepsilon_1$. Thus,

$$\mathbb{P}\left(\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}\right)\leq\mathbb{P}\left(\mathcal{G}^c\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}]\right)+C_0\exp(-c_0 n). \tag{23}$$

Further, we restrict ourselves to a high probability event on which there is entry-wise control over the random vector $\boldsymbol{V}$ in a sense specified below. The reasons for this restriction would become evident in later parts of the analysis. To this end, note that since $\boldsymbol{V}$ has sub-Gaussian tails, for any $\zeta > 0$,

$$\begin{aligned}\mathbb{P}\left[\max_i V_i^2\geq\zeta\log n\right]&\leq n\mathbb{P}\left[|V_1|\geq\sqrt{\zeta\log n}\right]\\&\leq C_1\exp\left(\log n-c_1\frac{\zeta\log n}{K^2}\right),\end{aligned}$$

where K is the sub-Gaussian norm of the random variable V and $c > 0$ is a universal constant. We choose $\zeta > 2K^2/c$ and define the event

$$\mathcal{F}_{\mathcal{V}}:=\left\{\max_i V_i^2\leq\zeta\log n\right\}, \tag{24}$$

that satisfies

$$\mathbb{P}\left[\mathcal{F}_{\boldsymbol{V}}\right]\geq 1-C_1 n^{-\delta}, \tag{25}$$

where $\delta > 1$. Thus,

$$\mathbb{P}\left(\mathcal{G}^c\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}]\right)\leq\mathbb{P}\left(\mathcal{G}^c\cap\mathcal{F}_{\boldsymbol{V}}\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}]\right)+C_1 n^{-\delta}. \tag{26}$$

Regarding the cone $\mathcal{A}$, [9] established that, there exists a collection of $N = \exp(2\varepsilon^2 p)$ closed convex cones $\{\mathcal{B}_i|1\leq i\leq n\}$ that form a cover of $\mathcal{A}$ with probability exceeding $1-\exp(-C_1\varepsilon^2 p)$, for some universal positive constant C. Thus, by the union bound,

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c\cap\mathcal{F}_{\boldsymbol{V}}\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}]\right)&\leq\mathbb{P}\left(\mathcal{G}^c\cap\mathcal{F}_{\boldsymbol{V}}\cap\{\mathcal{B}_i|1\leq i\leq N\}\text{ does not form a cover of }\mathcal{A}\right)\\&\quad+\sum_{i=1}^{N}\mathbb{P}\left(\mathcal{G}^c\cap\mathcal{F}_{\boldsymbol{V}}\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{B}_i\neq\{\boldsymbol{0}\}]\right).\end{aligned} \tag{27}$$

For any fixed subspace $\mathcal{W}\in\mathbb{R}^n$, introduce the convex cones

$$\mathcal{C}_i\left(\mathcal{W}\right):=\left\{\boldsymbol{w}+\boldsymbol{d}|\boldsymbol{w}\in\mathcal{W},\boldsymbol{d}\in\mathcal{B}_i\right\}.$$

Denoting $\mathcal{L}=\text{span}(\boldsymbol{X}_{\bullet 2},\ldots,\boldsymbol{X}_{\bullet p})$, observe that the following events are equivalent,

$$\left[\mathcal{G}^c\cap\{\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{B}_i\neq\{\boldsymbol{0}\}\}\right]\iff\left[\mathcal{G}^c\cap\{\mathcal{L}\cap\mathcal{C}_i(\text{span}(\boldsymbol{V}))\neq\{\boldsymbol{0}\}\}\right].$$

Hence, (27) reduces to

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c\cap\mathcal{F}_{\boldsymbol{V}}\cap[\{\boldsymbol{T}\boldsymbol{b}|\boldsymbol{b}\in\mathbb{R}^p\}\cap\mathcal{A}\neq\{\boldsymbol{0}\}]\right)&\leq\mathbb{P}\left(\{\mathcal{B}_i|1\leq i\leq N\}\text{ does not form a cover of }\mathcal{A}\right)\\&\quad+\sum_{i=1}^{N}\mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}}\cap[\mathcal{L}\cap\mathcal{C}_i(\text{span}(\boldsymbol{V}))\neq\{\boldsymbol{0}\}]\right)\\&\leq\sum_{i=1}^{N}\mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}}\cap[\mathcal{L}\cap\mathcal{C}_i(\text{span}(\boldsymbol{V}))\neq\{\boldsymbol{0}\}]\right)+\exp\left(-C_1\epsilon^2 p\right).\end{aligned} \tag{28}$$

To analyze the above, we will resort to ingredients from the literature on convex geometry. Using the approximate kinematic formula [1, Theorem I], [9] argued that, for any closed convex cone $\mathcal{C}$ for which the statistical dimension[5] obeys $\delta(\mathcal{C}) < n - \delta(\mathcal{L}) = n - p + 1$,

$$\mathbb{P}\left(\mathcal{L} \cap \mathcal{C} \neq \{\mathbf{0}\}\right) \leq 4 \exp\left\{-\frac{\left(n - p - \delta(\mathcal{C})\right)^2}{8n}\right\}. \tag{29}$$

For any event $\mathcal{G}_{\boldsymbol{V}}$ measurable with respect to the sigma-algebra generated by $\boldsymbol{V}$,

$$\mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}} \cap \mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\}\right) \leq \mathbb{E}_{\boldsymbol{V}}\left[\mathbf{1}_{\mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}} \mathbb{P}\left(\mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\} | \boldsymbol{V}\right)\right] + \mathbb{P}\left(\mathcal{G}_{\boldsymbol{V}}^c\right). \tag{30}$$

Here, the following lemma will be crucial.

Lemma 2. *There exists an event $\mathcal{G}_{\boldsymbol{V}}$ in the σ-algebra generated by $\boldsymbol{V}$ and there exists a fixed constant $\nu_0 > 0$ such that for all $0 < \nu < \nu_0$, the following two properties hold:*

1. *$\mathcal{G}_{\boldsymbol{V}}$ has exponentially high probability, that is,*

$$\mathbb{P}\left(\mathcal{G}_{\boldsymbol{V}}\right) \geq 1 - C_1 \exp\left(-c_1 n\right), \tag{31}$$

   for positive universal constants C_1, c_1.

2. *For all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$,*

$$\delta\left(\mathcal{C}_i\left(\mathrm{span}(\boldsymbol{v})\right)\right) \leq n\left(1 - g_{\mathrm{MLE}}^{-1}(\gamma) + \nu + o(1)\right). \tag{32}$$

Choose $\nu < \min\{\nu_0, g_{\mathrm{MLE}}^{-1}(\gamma) - \kappa\}$ in Lemma 2. Since, we are in the regime $\gamma < g_{\mathrm{MLE}}(\kappa)$, for $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$, we then have

$$\delta\left(\mathcal{C}_i\left(\mathrm{span}(\boldsymbol{v})\right)\right) < n - p + 1.$$

Applying (29) and Lemma 2 leads to

$$\begin{aligned}
\mathbf{1}_{\mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}} \mathbb{P}\left(\mathcal{L} \cap \mathcal{C}_i\left(\mathrm{span}(\boldsymbol{V})\right) \neq \{\mathbf{0}\} | \boldsymbol{V}\right) &\leq 4 \mathbf{1}_{\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}} \exp\left[-\frac{\left\{n - p - \delta(\mathcal{C}_i(\mathrm{span}(\boldsymbol{v})))\right\}^2}{8n}\right] \\
&\leq 4 \exp\left[-\frac{n}{8}\left(g_{\mathrm{MLE}}^{-1}(\gamma) - \kappa - \nu + o(1)\right)^2\right].
\end{aligned}$$

Thus, from (30), we have

$$\mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}} \cap \left[\mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\}\right]\right) \leq 4 \exp\left[-\frac{n}{8}\left(g_{\mathrm{MLE}}^{-1}(\gamma) - \kappa - \nu + o(1)\right)^2\right] + C_1 \exp\left(-c_1 n\right).$$

Consider $n > 8 \log 4 / \left(g_{\mathrm{MLE}}^{-1}(\gamma) - \kappa - \nu + o(1)\right)^2$ and choose ε such that,

$$2\varepsilon^2 \kappa < \min\left\{c, \frac{1}{8}\left(g_{\mathrm{MLE}}^{-1}(\gamma) - \kappa - \nu + o(1)\right)^2 - \frac{\log 4}{n}\right\}.$$

Then $\sum_{i=1}^{N} \mathbb{P}\left(\mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\}\right)$ decays exponentially fast in n. Thereby, recalling (22), (23),(26) and (28) completes the proof. ■

We defer the proofs of Lemmas 1–2 until Section 7.

[5]The statistical dimension of a convex cone is defined to be $\delta(\mathcal{C}) = \mathbb{E}\|\Pi_{\mathcal{C}}(\boldsymbol{Z})\|^2$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, and $\Pi_{\mathcal{C}}$ is the projection onto $\mathcal{C}$.

3.3 Ingredients from G-AMP

As discussed in Section 2, the proof of Theorem 1 will require elements from the G-AMP literature. In this Section, we provide a brief exposition of a key result established in [7] that will be central to our analysis in Section 4. For convenience, we adhere to the same notations as in [7].

A G-AMP algorithm comprises iterates $\{\boldsymbol{x}^t\}_{t\geq 0}$, where $\boldsymbol{x}^t \in \mathcal{V}_{q\times N} \equiv (\mathbb{R}^q)^N$, for some fixed $q \in \mathbb{N}$, and N is a function of the sample size n.[6] Define $\boldsymbol{A} = \boldsymbol{G} + \boldsymbol{G}'$, where $\boldsymbol{G} \in \mathbb{R}^{N\times N}$ has i.i.d. entries from $\mathcal{N}(0, 1/2N)$. Consider a collection of mappings $\mathcal{F} = \{f^k : k \in [N]\}$, such that $f^k : \mathbb{R}^q \times N \to \mathbb{R}^q$, is locally Lipschitz in the first argument for all $k \in [N]$. Then, starting from some initial condition $\boldsymbol{x}^0 \in \mathcal{V}_{q,N}$, a G-AMP algorithm updates each element of $\boldsymbol{x}^t$ as follows:

$$\boldsymbol{x}^{t+1}_{\bullet i} = \sum_{j=1}^{N} A_{ij} f^j\left(\boldsymbol{x}^t_{\bullet j}; t\right) - \frac{1}{N}\left(\sum_{j=1}^{N} \frac{\partial f^j}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^t_{\bullet j}; t\right)\right) f^i\left(\boldsymbol{x}^{t-1}_{\bullet i}; t-1\right), \tag{33}$$

where any term with negative t-index is considered 0. Here, $\frac{\partial f^j}{\partial \boldsymbol{x}}$ denotes the Jacobian of $f^j(\cdot; t) : \mathbb{R}^q \to \mathbb{R}^q$.

The authors in [7] characterize the asymptotic variance of the iterates $\boldsymbol{x}^t$, for each t, as $n \to \infty$. To describe the characterization, we require a few additional notations which we introduce next:

1. Consider an integer q' such that for each N, a finite partition $C_1^N \cup \ldots \cup C_{q'}^N = [N]$ exists and for each $a \in [q']$,
$$\lim_{N\to\infty} \frac{C_a^N}{N} = c_a \in (0, 1).$$

2. There exists $\boldsymbol{Y} := (\boldsymbol{y}_{\bullet 1}, \ldots, \boldsymbol{y}_{\bullet N}) \in \mathcal{V}_{q,N}$ such that for each $a \in [q']$, the empirical distribution of $\{\boldsymbol{y}_{\bullet i}\}_{i\in C_a^N}$, denoted by $\widehat{P}_a$ converges weakly to P_a; that is,
$$\frac{1}{|C_a^N|} \sum_{i\in C_a^N} \delta_{\boldsymbol{y}_{\bullet i}} \xrightarrow{\mathrm{d}} P_a.$$
Further, suppose $\mathbb{E}_{P_a} \|\boldsymbol{Y}_a\|^{2k-2}$ is bounded for some $k \geq 2$, and
$$\mathbb{E}_{\widehat{P}_a}\left(\|\boldsymbol{Y}_a\|^{2k-2}\right) \to \mathbb{E}_{P_a}\left(\|\boldsymbol{Y}_a\|^{2k-2}\right).$$

3. There exists a function $g : \mathbb{R}^{q'} \times \mathbb{R}^{q'} \times [q'] \times \mathbb{N} \cup \{0\}$, such that, for each $r \in [q'], a \in [q'], t \in \mathbb{N} \cup \{0\}, g_r(..., a, t)$ is Lipschitz continuous. Further, for each $N \geq 0$, each $a \in [q']$ and each $i \in C_a^N$, $\boldsymbol{x} \in \mathbb{R}^q$,
$$f^i(\boldsymbol{x}; t) = g\left(\boldsymbol{x}, \boldsymbol{y}_{\bullet i}, a, t\right). \tag{34}$$
This requirement basically states that the functions $f^j(\cdot; t)$ in (33) can only be of the aforementioned form.

4. For each $a \in [q']$, define $\widehat{\boldsymbol{\Sigma}}$ to be the limit (in probability),
$$\lim_{N\to\infty} \frac{1}{|C_a^N|} \sum_{i\in C_a^N} g\left(\boldsymbol{x}^0_{\bullet i}, \boldsymbol{y}_{\bullet i}, a, 0\right) g\left(\boldsymbol{x}^0_{\bullet i}, \boldsymbol{y}_{\bullet i}, a, 0\right)^\top =: \widehat{\boldsymbol{\Sigma}}^{(0)}_a. \tag{35}$$
For each $t \geq 1$, define a positive semi-definite matrix $\boldsymbol{\Sigma}^{(t)} \in \mathbb{R}^{q\times q}$, obtained, by letting,
$$\boldsymbol{\Sigma}^{(t)} = \sum_{b=1}^{q'} c_b \widehat{\boldsymbol{\Sigma}}^{(t-1)}_b, \qquad \widehat{\boldsymbol{\Sigma}}^{(t)}_a = \mathbb{E}\left[g(\boldsymbol{Z}^t_a, \boldsymbol{Y}_a, a, t) g(\boldsymbol{Z}^t_a, \boldsymbol{Y}_a, a, t)^\top\right], \tag{36}$$
where $\boldsymbol{Y}_a \sim P_a, \boldsymbol{Z}^t_a \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)}\right)$ and $\boldsymbol{Y}_a \perp\!\!\!\perp \boldsymbol{Z}^t_a$.

[6] One can think of an element $\boldsymbol{x} \in \mathcal{V}_{q,N}$ as an $N-$vector $(\boldsymbol{x}_{\bullet 1}, \ldots \boldsymbol{x}_{\bullet N})$ with entries in $\mathbb{R}^q$.

Under the above assumptions, asymptotic distribution of marginals of $\boldsymbol{x}^t$ can be characterized as follows:

Theorem 5 ([7] Theorem 1)**.** *For all $t \geq 1$, each $a \in [q']$, and any pseudo-Lipschitz function $\psi : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ of order k, almost surely,*

$$\lim_{N\to\infty} \frac{1}{|C_a^N|} \sum_{j \in C_a^N} \psi\left(\boldsymbol{x}_{\bullet j}^t, \boldsymbol{y}_{\bullet j}\right) = \mathbb{E}\left\{\psi\left(\boldsymbol{Z}_a^t, \boldsymbol{Y}_a\right)\right\}, \tag{37}$$

where $\boldsymbol{Z}_a^t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)}\right)$ is independent of $\boldsymbol{Y}_a \sim P_a$.

4 Asymptotic average behavior of MLE

To begin with, we recall the iterative algorithm that [8] introduced for tracking the MLE. Starting with an initial guess $\hat{\boldsymbol{\beta}}^0$, set $\boldsymbol{S}^0 = \boldsymbol{X}\hat{\boldsymbol{\beta}}^0$ and for $t = 1, 2, \ldots$, update $\{\boldsymbol{S}^t, \hat{\boldsymbol{\beta}}^t\}_{t\geq 1}$, with $\boldsymbol{S}^t \in \mathbb{R}^n, \hat{\boldsymbol{\beta}}^t \in \mathbb{R}^p$, using the following scheme:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^t &= \hat{\boldsymbol{\beta}}^{t-1} + \kappa^{-1}\boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right) \\ \boldsymbol{S}^t &= \boldsymbol{X}\hat{\boldsymbol{\beta}}^t - \Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right) \end{aligned} \tag{38}$$

where the function Ψ_t is applied element-wise and is equal to

$$\Psi_t(y, s) = \lambda_t r_t, \qquad r_t = y - \rho'\left(\mathsf{prox}_{\lambda_t \rho}\left(\lambda_t y + s\right)\right), \tag{39}$$

and λ_t is described via the recursions (8)–(9). However, from (18), we know $\lambda_t \equiv \tilde{\lambda}_t$, where $\tilde{\lambda}_t$ is described via the update equations (11)–(12), when

$$\alpha_0 = \tilde{\alpha}_0, \quad \sigma_0 = \tilde{\sigma}_0. \tag{40}$$

Suppose we initialize the scalar sequence $(\tilde{\lambda}_0, \tilde{\sigma}_0)$ in the aforementioned way. This leads to an alternate characterization of the function Ψ_t, which will be useful in Subsection 4.1. Note that the response variables can be expressed as

$$y_i = h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right), \tag{41}$$

where $h(x, y)$ is specified via (10) and $w_1, \ldots, w_n \overset{\text{i.i.d.}}{\sim} U(0, 1)$, independent of all other random variables. Rewriting Ψ_t in terms of these quantities and recalling definition (13), we observe that

$$\Psi_t\left(y_i, S_i^t\right) \equiv \tilde{\Psi}_t\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i, S_i^t\right). \tag{42}$$

4.1 State Evolution Analysis

In this section, we characterize the asymptotic average behavior of the AMP iterates $(\hat{\boldsymbol{\beta}}^t, \boldsymbol{S}^t)$, for each fixed t, in the large sample limit. In this regard, the scalar sequence $(\alpha_t, \sigma_t, \lambda_t)$ introduced in (8)–(9) proves to be useful, as is formalized in the theorem below.

Theorem 6. *Suppose the initial conditions for the AMP iterative scheme (38), and the variance map updates (8)–(9) satisfy*

$$\alpha_0 = \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \boldsymbol{\beta}\rangle}{n}, \qquad \sigma_0^2 = \lim_{n,p\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_0\boldsymbol{\beta}\|^2. \tag{43}$$

For any pseudo-Lipshcitz function ψ of order 2,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \psi\left(\hat{\beta}_j^t - \alpha_t\beta_j, \beta_j\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\sigma_t Z, \beta\right)\right] \\ \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(\begin{bmatrix}\boldsymbol{X}_i'\boldsymbol{\beta} \\ S_i^t\end{bmatrix}, \begin{bmatrix} w_i \\ 0\end{bmatrix}\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\begin{bmatrix} Q_1^t \\ Q_2^t\end{bmatrix}, \begin{bmatrix} W \\ 0\end{bmatrix}\right)\right], \end{aligned} \tag{44}$$

where $\beta \sim \Pi, W \sim U(0,1)$ *independent of each other*[7] *and independent of*

$$\left(Q_1^t, Q_2^t\right) \sim \mathcal{N}\left(0, \begin{bmatrix} \gamma^2 & \alpha_t\gamma^2 \\ \alpha_t\gamma^2 & \kappa\sigma_t^2 + \alpha_t^2\gamma^2 \end{bmatrix}\right). \tag{45}$$

Proof: Introduce a new sequence of iterates $\{\boldsymbol{\nu}^t, \boldsymbol{R}^t\}$ defined as follows: starting with initial conditions $\boldsymbol{\nu}^0 = \hat{\boldsymbol{\beta}}^0 - \alpha_0\boldsymbol{\beta}, \boldsymbol{R}^0 = \boldsymbol{S}^0$, set:

$$\begin{aligned} \boldsymbol{\nu}^t &= q_{t-1}\left(\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right) - a_t\boldsymbol{\beta} + \kappa^{-1}\boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right) \\ \boldsymbol{R}^t &= \boldsymbol{X}\left(\boldsymbol{\nu}^t + \alpha_t\boldsymbol{\beta}\right) - \Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right), \end{aligned} \tag{46}$$

where

$$\begin{aligned} q_t &= -\frac{1}{\kappa n}\sum_{i=1}^{n}\Psi_t'\left(y_i, R_i^t\right) \\ a_0 = \alpha_0, \qquad a_t &= \frac{1}{\kappa n}\sum_{i=1}^{n}\frac{\partial}{\partial a}\Psi_{t-1}\left(h(a, W_i), R_i^{t-1}\right)\Big|_{a=\boldsymbol{X}_i'\boldsymbol{\beta}} \quad \text{for} \quad t \geq 1; \end{aligned} \tag{47}$$

Ψ_t' is the derivative w.r.t the second coordinate of Ψ_t. The difference between this recursion and that in (38) is the introduction of the new variables $\{q_t, a_t\}$, and the regression coefficients $\boldsymbol{\beta}$. It turns out that the recursive equations for $\{\boldsymbol{\nu}^t, \boldsymbol{R}^t\}$, introduced in (46), fall under the class of G-AMP algorithms. Hence, asymptotic average behavior of $\{\boldsymbol{\nu}^t, \boldsymbol{R}^t\}$ can be established by appropriately using Theorem 5. This leads to the following lemma.

Lemma 3. *For any* $t \geq 1$*, under the assumptions of Theorem 6, the recursions* $\{\boldsymbol{\nu}^t, \boldsymbol{R}^t\}$ *introduced in* (46) *satisfy*

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\nu_j^t, \beta_j\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\sigma_t Z, \beta\right)\right] \\ \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\begin{bmatrix}\boldsymbol{X}_i'\boldsymbol{\beta} \\ R_i^t\end{bmatrix}, \begin{bmatrix} w_i \\ 0\end{bmatrix}\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\begin{bmatrix} Q_1^t \\ Q_2^t\end{bmatrix}, \begin{bmatrix} W \\ 0\end{bmatrix}\right)\right]. \end{aligned}$$

Finally, Theorem 6 is established by noting the equivalence of the recursions $\{\boldsymbol{\nu}^t, \boldsymbol{R}^t\}$, and the appropriately centered versions of the original recursions, that is, $\{\hat{\boldsymbol{\beta}}^t - \alpha_t\boldsymbol{\beta}, \boldsymbol{S}^t\}$, which is formalized next.

Lemma 4. *Under the assumptions of Theorem 6, and the assumptions on the initial conditions* $\boldsymbol{\nu}^0 = \hat{\boldsymbol{\beta}}^0 - \alpha_0\boldsymbol{\beta}, \boldsymbol{R}^0 = \boldsymbol{S}^0$*, for any fixed* $t \geq 1$*,*

$$\lim_{n\to\infty}\frac{1}{p}\|\hat{\boldsymbol{\beta}}^t - \alpha_t\boldsymbol{\beta} - \boldsymbol{\nu}^t\|^2 =_{a.s.} 0, \qquad \lim_{n\to\infty}\frac{1}{n}\|\boldsymbol{S}^t - \boldsymbol{R}^t\|^2 =_{a.s.} 0.$$

[7] Recall Π is the weak limit of the empirical distribution of $\{\beta_i\}_{1\leq i\leq p}$.

Since ψ is a pseudo-Lipschitz function of order 2, we have

$$\begin{aligned}\left|\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j\right)-\frac{1}{p}\sum_{j=1}^{p}\psi\left(\nu_j^t,\beta_j\right)\right| &\le \frac{1}{p}\sum_{j=1}^{p}\left|\psi\left(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j\right)-\psi\left(\nu_j^t,\beta_j\right)\right| \\ &\le C\frac{1}{p}\sum_{j=1}^{p}\left(1+\|(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j)\|+\|(\nu_j^t,\beta_j)\|\right)\left|\hat{\beta}_j^t-\alpha_t\beta_j-\nu_j^t\right| \\ &\le C\frac{1}{p}\sqrt{\sum_{j=1}^{p}\left(1+\|(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j)\|+\|(\nu_j^t,\beta_j)\|\right)^2}\|\hat{\boldsymbol{\beta}}^t-\alpha_t\boldsymbol{\beta}-\boldsymbol{\nu}^t\|.\end{aligned}$$

By definition, $\|\boldsymbol{\beta}\|/\sqrt{p}$ is bounded. Putting together Lemma 3 and 4, we obtain $\|\hat{\boldsymbol{\beta}}^t\|/\sqrt{p}$ is bounded for all t. Hence, from Lemma 4 and the above inequality, we have

$$\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j\right)=\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\nu_j^t,\beta_j\right).$$

This establishes the first relation in (44). A similar argument holds for the other relation. ■

It remains to prove Lemmas 3–4, which we focus on next.

4.1.1 Proof of Lemma 3

Our first goal is to reduce the recursion (46) to the G-AMP form (33). Thereafter, computing the covariances $\boldsymbol{\Sigma}^t$ from (36) and an application of Theorem 5 will complete the proof.

To this end, fix $q=2k^0+1$ for some large arbitrary integer k^0, and let $N=n+p$. In the subsequent analysis, restrict $t\in\{0,\dots,q\}$. Define $\boldsymbol{x}^t\in\mathcal{V}_{q,N}$ such that $\boldsymbol{x}^0=\mathbf{0}$ and the values for other choices of t are defined as follows: for the odd iterates $t=2k+1$ $(k\ge 0)$, for each $i=1,\dots n$, define

$$\boldsymbol{x}_{\bullet i}^t:=\left[Z_i,0,R_i^0,0,R_i^1,\dots,R_i^{\frac{t-1}{2}},0,0,\dots\right]'. \tag{48}$$

For even iterates $t=2k$ $(k\ge 1)$, for each $i=n+1,\dots,n+p$, define

$$\boldsymbol{x}_{\bullet i}^t=\left[0,\nu_{i-n}^1,0,\nu_{i-n}^2,0,\nu_i^2,\dots,\nu_{i-n}^{\frac{t}{2}},0,0,\dots\right]'. \tag{49}$$

Let all other entries of $\boldsymbol{x}^t$ be 0. Let $\boldsymbol{Y}\in\boldsymbol{\mathcal{V}}_{q,N}$ have the first two rows defined via

$$\begin{bmatrix}Y_{1\bullet}\\ Y_{2\bullet}\end{bmatrix}=\begin{bmatrix}W_1, & W_2, & \dots, & W_n, & \beta_1, & \beta_2, & \dots, & \beta_p\\ 0 & & \dots & 0 & \nu_1^0, & \nu_2^0, & \dots, & \nu_p^0\end{bmatrix} \tag{50}$$

and the rest of the entries are all 0. Note that, the functions f in (34) are allowed to be functions of the elements of $\boldsymbol{Y}$. For the odd iterates $t=2k+1$ $(k\ge 0)$, let $f^i(\boldsymbol{x};2k+1)=\mathbf{0}$ for $i=n+1,\dots n+p$. Let $h=\sqrt{N/n}$. For $i=1,\dots,n$, define

$$f^i(\boldsymbol{x};2k+1)=\left[0,\frac{h}{\kappa}\Psi_0(h(x_1,Y_{1i}),x_3),0,\frac{h}{\kappa}\Psi_1(h(x_1,Y_{1i}),x_5),\dots,\frac{h}{\kappa}\Psi_{\frac{t-1}{2}}(h(x_1,Y_{1i}),x_{t+2}),0,0,\dots\right]'. \tag{51}$$

For the even iterates $t=2k$ $(k\ge 0)$, let $f^i(\boldsymbol{x};2k)=0$ for $i=1,\dots,n$ and for $i=n+1,\dots,n+p$, define

$$f^i(\boldsymbol{x};2k)=\left[hY_{1i},0,h(Y_{2i}+\alpha_0Y_{1i}),0,h(x_2+\alpha_1Y_{1i}),0,h(x_4+\alpha_2Y_{1i}),0,\dots,h(x_t+\alpha_{t/2}Y_{1i}),0,0,\dots\right]'. \tag{52}$$

Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be a symmetric matrix with $A_{ii} = 0, A_{ij} = \frac{1}{h}X_{i,j-n}$ for $1 \le i \le n$ and $n+1 \le j \le n+p$, and all other entries A_{ij} with $i < j$ are i.i.d. $\mathcal{N}(0, 1/N)$. With these definitions in place, the following result can be established.

Lemma 5. *For even iterates with column indices $i = n+1, \dots, n+p$, and for odd iterates with column indices $i = 1, \dots n$, $\boldsymbol{x}^t_{\bullet i}$ defined via (48)–(49) satisfies the recursion (33), with the collection of functions $f^i(,;t)$ given by (51)–(52) and $\boldsymbol{A}$ as described above.*

Proof: The proof follows directly from matrix multiplications and is, therefore, omitted. ■

Let $\tilde{\boldsymbol{x}}^t$ be a new sequence of iterates in $\mathcal{V}_{q,N}$ such that $\tilde{\boldsymbol{x}}^0 = \boldsymbol{0}$. For all $1 \le t \le q$, if a column i of $\boldsymbol{x}^t$ is non-zero, set the corresponding column of $\tilde{\boldsymbol{x}}^t$ as $\tilde{\boldsymbol{x}}^t_{\bullet i} = \boldsymbol{x}^t_{\bullet i}$. If a column of $\boldsymbol{x}^t$ is zero, set the corresponding column of $\tilde{\boldsymbol{x}}^t$ as follows: $\tilde{\boldsymbol{x}}^1_{\bullet i} = \sum_{j=1}^N A_{ij} f^j(\tilde{\boldsymbol{x}}^0_{\bullet j}; t)$ and for $t \ge 1$,

$$\tilde{\boldsymbol{x}}^{t+1}_{\bullet i} := \sum_{j=1}^N A_{ij} f^j\left(\tilde{\boldsymbol{x}}^t_{\bullet j}; t\right) - \frac{1}{N}\left(\sum_{j=1}^N \frac{\partial f^j}{\partial \boldsymbol{x}}\left(\tilde{\boldsymbol{x}}^t_{\bullet j}; t\right)\right) f^i\left(\tilde{\boldsymbol{x}}^{t-1}_{\bullet i}; t-1\right),$$

where any term with negative t-index is zero. Then, from Lemma 5 we trivially arrive at the following conclusion.

Lemma 6. *The sequence of iterates $\{\tilde{\boldsymbol{x}}^t\}_{1\le t\le q}$ satisfies the recursion (33) with the choice of functions f^i specified in (51) and (52).*

Thus, we have reduced the recursion in (46) to the G-AMP form (33). Theorem 5 then tells us that the asymptotic covariance structure of $\tilde{\boldsymbol{x}}^t$ can be obtained by carrying out the iterative scheme in (36), with g defined via (51) and (52). We systematically list properties of $\boldsymbol{\Sigma}^{(t)}$ that will be crucial for establishing the proof. For $t = 1$, $i = 1, \dots n$, $\tilde{\boldsymbol{x}}^1_{\bullet i}$ has first and third entries Z_i, R^0_i, with all other entries 0. From the definitions (35) and (36), it is easy to check that

$$\begin{bmatrix} \Sigma^{(1)}_{(1,1)} & \Sigma^{(1)}_{(1,3)} \\ \Sigma^{(1)}_{(3,1)} & \Sigma^{(1)}_{(3,3)} \end{bmatrix} = \begin{bmatrix} \lim_{n\to\infty} \frac{\|\boldsymbol{\beta}\|^2}{n} & \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0\rangle}{n} \\ \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0\rangle}{n} & \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0\|^2}{n} \end{bmatrix}, \tag{53}$$

which is consistent with the asymptotic covariance structure we expect to see in this case, since $Z_i = \boldsymbol{X}_i'\boldsymbol{\beta}, R^0_i = S^0_i = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}^0$. Computing $\boldsymbol{\Sigma}^{(2)}$, using the formula (36) and applying Theorem 5 yields,

$$\frac{1}{p}\sum_{j=1}^p \psi\left(\nu^1_j, \beta_j\right) \to \mathbb{E}\left[\psi\left(\tau_1 Z, \beta\right)\right], \quad \text{where} \quad \tau_1^2 = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_0^2\left(h\left(Q^0_1, U\right), Q^0_2\right)\right], \tag{54}$$

and (Q^0_1, Q^0_2) is multivariate normal with mean $\boldsymbol{0}$ and covariance matrix specified in (53).

Note that, for $\boldsymbol{\Sigma}^{(3)}$, the first 3×3 sub-block would be the same as in (53). Among the rest, the distinct non-trivial entries are $\Sigma^{(3)}_{(1,5)}, \Sigma^{(3)}_{(3,5)}, \Sigma^{(3)}_{(5,5)}$, given by

$$\Sigma^{(3)}_{(1,5)} = \alpha_1\gamma^2, \qquad \Sigma^{(3)}_{(3,5)} = \alpha_1 \lim_{n\to\infty}\frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0\rangle}{n}, \qquad \Sigma^{(3)}_{(5,5)} = \kappa\tau_1^2 + \alpha_1^2\gamma^2.$$

From Theorem 5, this immediately yields,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \psi\left(\begin{bmatrix}\boldsymbol{X}_i'\boldsymbol{\beta} \\ R^0_i \\ R^1_i\end{bmatrix}, \begin{bmatrix} w_i \\ 0 \\ 0\end{bmatrix}\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(3)}, \begin{bmatrix} W \\ 0 \\ 0\end{bmatrix}\right)\right], \tag{55}$$

where

$$\boldsymbol{Z}^{(3)} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \gamma^2 & \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \alpha_1\gamma^2 \\ \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0\|^2}{n} & \alpha_1 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} \\ \alpha_1\gamma^2 & \alpha_1 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \kappa\tau_1^2 + \alpha_1^2\gamma^2 \end{bmatrix}\right),$$

$W \sim U(0,1) \perp\!\!\!\perp \boldsymbol{Z}^{(3)}$.

Computing $\boldsymbol{\Sigma}^{(4)}$, we obtain

$$\frac{1}{p}\sum_{j=1}^{p} \psi\left(\begin{bmatrix} \nu_j^1 \\ \nu_j^2 \end{bmatrix}, \begin{bmatrix} \beta_j \\ 0 \end{bmatrix}\right) \to \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(4)}, \begin{bmatrix} \beta \\ 0 \end{bmatrix}\right)\right], \tag{56}$$

where $\boldsymbol{Z}^{(4)} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tau_1^2 & \rho_{12} \\ \rho_{12} & \tau_2^2 \end{bmatrix}\right)$, with

$$\tau_2^2 = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_1^2\left(h\left(Z_1^{(3)}, U\right), Z_3^{(3)}\right)\right], \qquad \rho_{12} = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_0\left(h\left(Z_1^{(3)}, U\right), Z_2^{(3)}\right)\Psi_1\left(h\left(Z_1^{(3)}, U\right), Z_3^{(3)}\right)\right]. \tag{57}$$

We continue similar calculations to obtain $\boldsymbol{\Sigma}^{(5)}$ and $\boldsymbol{\Sigma}^{(6)}$. The 5×5 principal sub-matrix of $\boldsymbol{\Sigma}^{(5)}$, is identical to $\boldsymbol{\Sigma}^{(3)}$. Other distinct non-zero entries are listed below:

$$\Sigma^{(5)}_{(1,7)} = \alpha_2\gamma^2, \qquad \Sigma^{(5)}_{(5,7)} = \kappa\rho_{12} + \alpha_1\alpha_2\gamma^2,$$
$$\Sigma^{(5)}_{(3,7)} = \alpha_2 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n}, \qquad \Sigma^{(5)}_{(7,7)} = \kappa\tau_2^2 + \alpha_2^2\gamma^2.$$

Hence, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(\begin{bmatrix} \boldsymbol{X}_i'\boldsymbol{\beta} \\ R_i^0 \\ R_i^1 \\ R_i^2 \end{bmatrix}, \begin{bmatrix} w_i \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(5)}, \begin{bmatrix} W \\ 0 \\ 0 \\ 0 \end{bmatrix}\right)\right], \tag{58}$$

where $W \sim U(0,1) \perp\!\!\!\perp \boldsymbol{Z}^{(5)}$ and

$$\boldsymbol{Z}^{(5)} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \gamma^2 & \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \alpha_1\gamma^2 & \alpha_2\gamma^2 \\ \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0\|^2}{n} & \alpha_1 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \alpha_2 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} \\ \alpha_1\gamma^2 & \alpha_1 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \kappa\tau_1^2 + \alpha_1^2\gamma^2 & \kappa\rho_{12} + \alpha_1\alpha_2\gamma^2 \\ \alpha_2\gamma^2 & \alpha_2 \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0 \rangle}{n} & \kappa\rho_{12} + \alpha_1\alpha_2\gamma^2 & \kappa\tau_2^2 + \alpha_2^2\gamma^2 \end{bmatrix}\right).$$

Computing $\boldsymbol{\Sigma}^{(6)}$, we obtain

$$\frac{1}{p}\sum_{j=1}^{p} \psi\left(\begin{bmatrix} \nu_j^1 \\ \nu_j^2 \\ \nu_j^3 \end{bmatrix}, \begin{bmatrix} \beta_j \\ 0 \\ 0 \end{bmatrix}\right) \to \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(6)}, \begin{bmatrix} \beta \\ 0 \end{bmatrix}\right)\right], \tag{59}$$

where $\boldsymbol{Z}^{(6)} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tau_1^2 & \rho_{12} & \rho_{13} \\ \rho_{12} & \tau_2^2 & \rho_{23} \\ \rho_{13} & \rho_{23} & \tau_3^2 \end{bmatrix}\right)$, with

$$\tau_3^2 = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_2^2\left(h\left(Z_1^{(5)}, U\right), Z_4^{(5)}\right)\right], \qquad \rho_{lm} = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_{l-1}\left(h\left(Z_1^{(5)}, U\right), Z_{l+1}^{(5)}\right)\Psi_{m-1}\left(h\left(Z_1^{(5)}, U\right), Z_{m+1}^{(5)}\right)\right]. \tag{60}$$

Repeating the above procedure and reading off the relevant entries in the covariance matrices, we arrive at the following results: for all $t \leq q$,

$$\lim_{n\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi\left(\nu_j^t, \beta_j\right) \overset{\text{a.s.}}{=} \mathbb{E}\left[\psi\left(\tau_t Z, \beta\right)\right]$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(\begin{bmatrix} \boldsymbol{X}_i'\boldsymbol{\beta} \\ R_i^t \end{bmatrix}, \begin{bmatrix} w_i \\ 0 \end{bmatrix}\right) \overset{\text{a.s.}}{=} \mathbb{E}\left[\psi\left(\begin{bmatrix} Z_1^{(2t+1)} \\ Z_{t+2}^{(2t+1)} \end{bmatrix}, \begin{bmatrix} W \\ 0 \end{bmatrix}\right)\right],$$

where $\left(Z_1^{(2t+1)}, Z_{t+2}^{(2t+1)}\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\left(-\alpha_t, \tau_t\right)\right)$ and τ_t^2 is defined by the relation

$$\tau_t^2 = \frac{1}{\kappa^2}\, \mathbb{E}\left[\Psi_{t-1}^2\left(h\left(Z_1^{(2t-1)}, U\right), Z_{t+1}^{(2t-1)}\right)\right],$$

with $\left(Z_1^{(2t-1)}, Z_{t+1}^{(2t-1)}\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\left(-\alpha_{t-1}, \tau_{t-1}\right)\right)$ and $\boldsymbol{\Sigma}(\alpha, \sigma)$ as in (5). The final step is to relate the scalar sequence $\{\tau_t\}$, first to the sequence $\{\tilde{\sigma}_t\}$ defined in (12), and thereafter to the sequence $\{\sigma_t\}$ in the statement of Theorem 6. To this end, recall the initial conditions on $\{\alpha_t, \sigma_t\}$ imposed via the relations

$$\alpha_0 = \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \boldsymbol{\beta}\rangle}{n}, \qquad \sigma_0^2 = \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0 - \alpha_0\boldsymbol{\beta}\|^2}{p}. \tag{61}$$

It is easy to check that, with this choice, the covariance in (53) is precisely $\boldsymbol{\Sigma}(-\alpha_0, \sigma_0) = \boldsymbol{\Sigma}\left(-\tilde{\alpha}_0, \tilde{\sigma}_0\right)$, since $(\tilde{\alpha}_0, \tilde{\sigma}_0)$ was initialized to (α_0, σ_0) (recall (40)).

The equivalence between the functions Ψ_t and $\tilde{\Psi}_t$ from (42), and the definition of $\tilde{\sigma}_t$ from (12) then leads to $\tau_1^2 = \tilde{\sigma}_1^2$, which subsequently yields $\tau_t^2 \equiv {\tilde{\sigma}_t}^2$. The equivalence between $\{\tilde{\sigma}_t\}$ and $\{\sigma_t\}$ established in (18), then completes the proof.

4.1.2 Proof of Lemma 4

The proof partly follows along lines similar to [3, Lemma 6.7], but has some additional ingredients which we detail here. Denote $\boldsymbol{\theta}^t = \hat{\boldsymbol{\beta}}^t - \alpha_t\boldsymbol{\beta}$. Comparing the recursive equations in (46) and (38), and using the triangle inequality we obtain,

$$\left\|\boldsymbol{R}^t - \boldsymbol{S}^t\right\| \leq \|\boldsymbol{X}\| \left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| + \left\|\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right) - \Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right)\right\|.$$

Applying [3, Proposition 6.3], we obtain

$$\frac{\partial \Psi_t(y,s)}{\partial s} = \frac{-\lambda_t\, \rho''(x)|_{x=\mathsf{prox}(\lambda_t y+s)}}{1+\lambda_t\, \rho''(x)|_{\mathsf{prox}(\lambda_t y+s)}}. \tag{62}$$

Hence, $\Psi(\boldsymbol{y}, \cdot)$ is Lipschitz continuous with Lipschitz constant at most 1, which yields

$$\left\|\boldsymbol{R}^t - \boldsymbol{S}^t\right\| \leq \|\boldsymbol{X}\| \left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| + \left\|\boldsymbol{R}^{t-1} - \boldsymbol{S}^{t-1}\right\|. \tag{63}$$

Similarly, comparing (46) and (38) again, we obtain

$$\begin{aligned}
&\boldsymbol{\nu}^t - \boldsymbol{\theta}^t = q_{t-1}\left(\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right) - a_t\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{t-1} + \alpha_t\boldsymbol{\beta} + \kappa^{-1}\left(\boldsymbol{X}'\Psi_{t-1}(\boldsymbol{y}, \boldsymbol{R}^{t-1}) - \boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right)\right) \\
&= \left(\boldsymbol{\nu}^{t-1} - \boldsymbol{\theta}^{t-1}\right) + \left(q_{t-1} - 1\right)\left(\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right) + \left(\alpha_t - a_t\right)\boldsymbol{\beta} + \kappa^{-1}\left(\boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right) - \boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right)\right),
\end{aligned}$$

where the second equality is obtained after appropriate rearranging. Using the triangle inequality,

$$\left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| \leq \left\|\boldsymbol{\nu}^{t-1} - \boldsymbol{\theta}^{t-1}\right\| + |q_{t-1} - 1| \left\|\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right\| + |\alpha_t - a_t| \|\boldsymbol{\beta}\| + \frac{1}{\kappa} \|\boldsymbol{X}\| \left\|\boldsymbol{R}^{t-1} - \boldsymbol{S}^{t-1}\right\|. \tag{64}$$

Since $\boldsymbol{\nu}^0 = \boldsymbol{\theta}^0$, iterating (63) and (64), it can be established that there exists a constant C, depending on κ, such that

$$\left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| \leq (C\,\|\boldsymbol{X}\|)^{2t} \left(\sum_{l=0}^{t-1} |q_l - 1| \left\|\boldsymbol{\nu}^l + \alpha_l \boldsymbol{\beta}\right\| + \sum_{l=0}^{t-1} |\alpha_l - a_l|\,\|\boldsymbol{\beta}\| \right). \tag{65}$$

Using Lemma 3, the definition of q_t and (62), we have,

$$\begin{aligned} \lim_{n\to\infty} q_t &= \lim_{n\to\infty} -\frac{1}{\kappa n} \sum_{i=1}^{n} \left\{ \frac{-\lambda_t \rho'' \left(\mathsf{prox}\left(\lambda_t h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right) + R_i^t\right)\right)}{1 + \lambda_t \rho'' \left(\mathsf{prox}\left(\lambda_t h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right) + R_i^t\right)\right)} \right\} \\ &= \mathbb{E}\left[\frac{1}{\kappa} \left\{ 1 - \frac{1}{1 + \lambda_t \rho'' \left(\mathsf{prox}\left(\lambda_t h\left(Q_1^t, U\right) + Q_2^t\right)\right)} \right\} \right], \end{aligned} \tag{66}$$

where $(Q_1^t, Q_2^t) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(-\alpha_t, \sigma_t)\right)$. The equivalence (18) yields

$$\lim_{n\to\infty} q_t = \mathbb{E}\left[\frac{1}{\kappa} \left\{ 1 - \frac{1}{1 + \tilde{\lambda}_t \rho'' \left(\mathsf{prox}\left(\tilde{\lambda}_t h\left(\tilde{Q}_1^t, U\right) + \tilde{Q}_2^t\right)\right)} \right\} \right] = 1,$$

where $\left(\tilde{Q}_1^t, \tilde{Q}_2^t\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\left(-\tilde{\alpha}_t, \tilde{\sigma}_t\right)\right)$, and the last equality follows from the definition of $\tilde{\lambda}_t$ in (11). Note that to obtain (66), we applied Lemma 3 to the function $\partial\Psi(y,s)/\partial s$ which is not necessarily continuous, but a smoothing argument similar to that in the proof of [3, Lemma 6.7] helps circumvent this technicality. Now, recall that for each n, we have a matrix of covariates $\boldsymbol{X} \equiv \boldsymbol{X}(n)$ that has dimension $n \times p$ and i.i.d. $\mathcal{N}(0, 1/n)$ entries. Since, $\lim_{n\to\infty} \|\boldsymbol{X}\| < \infty$ and $\|\boldsymbol{\nu}^t\|/\sqrt{p}$ is bounded for all t, we arrive at

$$\lim_{n\to\infty} \frac{1}{\sqrt{p}} (C\,\|\boldsymbol{X}\|)^{2t} \sum_{l=0}^{t-1} |q_l - 1| \left\|\boldsymbol{\nu}^l + \alpha_l \boldsymbol{\beta}\right\| = 0. \tag{67}$$

It remains to analyze the second term in the RHS of (65). To analyze the large sample limit of a_t defined in (47), we invoke Lemma 3 once again, in conjunction with the smoothing techniques from [3, Lemma 6.7], which yields

$$\lim_{n\to\infty} a_t = \frac{1}{\kappa}\,\mathbb{E}\left[\frac{\partial}{\partial a} \Psi_{t-1}\left(h(a, W), s\right)\Big|_{a=Q_1^{t-1}, s=Q_2^{t-1}} \right]. \tag{68}$$

In order to analyze (68), we will invoke Stein's lemma, which states that if $X \sim \mathcal{N}(\mu, \sigma^2)$ and h is a function for which $\mathbb{E}\,h(X)(X-\mu)$ and $\sigma^2\,\mathbb{E}\,h'(X)$ both exist,

$$\mathbb{E}\,h(X)(X-\mu) = \sigma^2\,\mathbb{E}\,h'(X). \tag{69}$$

To this end, it will be useful to express Q_2^{t-1} in terms of Q_1^{t-1} and an independent standard Gaussian Z, as shown below

$$Q_2^{t-1} = \alpha_{t-1} Q_1^{t-1} + \sqrt{\kappa \sigma_{t-1}^2}\, Z =: f\left(Q_1^{t-1}, Z\right),$$

since $\left(Q_1^{t-1}, Q_2^{t-1}\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\left(-\alpha_{t-1}, \sigma_{t-1}\right)\right)$. Thus, one can represent Ψ_{t-1} as

$$\Psi_{t-1}\left(h\left(Q_1^{t-1}, W\right), Q_2^{t-1}\right) = \Psi_{t-1}\left(h\left(Q_1^{t-1}, W\right), f\left(Q_1^{t-1}, Z\right)\right).$$

Obviously,

$$\mathbb{E}\left[\frac{\partial}{\partial a} \Psi_{t-1}\left(h(a, W), s\right)\Big|_{a=Q_1^{t-1}, s=Q_2^{t-1}} \right] = \mathbb{E}_{W,Z}\left[\mathbb{E}_{Q_1^{t-1}}\left[\frac{\partial}{\partial a} \Psi_{t-1}\left(h(a, W), s\right)\Big|_{a=Q_1^{t-1}, s=f\left(Q_1^{t-1}, Z\right)} \Big| W, Z \right] \right].$$

Since Q_1^{t-1} is independent of (W, Z), (69) immediately gives

$$\gamma^2 \, \mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial Q_1^{t-1}} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), f \left(Q_1^{t-1}, Z \right) \right) \middle| W, Z \right] = \mathbb{E} \left[Q_1^{t-1} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), f \left(Q_1^{t-1}, Z \right) \right) \middle| W, Z \right].$$

The LHS can be decomposed using the chain rule as follows

$$\begin{aligned}
&\mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial Q_1^{t-1}} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), f \left(Q_1^{t-1}, Z \right) \right) \middle| W, Z \right] \\
&\qquad = \mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial a} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = f\left(Q_1^{t-1}, Z\right)} \middle| W, Z \right] \\
&\qquad\qquad + \alpha_{t-1} \, \mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial s} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = f\left(Q_1^{t-1}, Z\right)} \middle| W, Z \right].
\end{aligned}$$

Putting these together,

$$\begin{aligned}
&\mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial a} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = f\left(Q_1^{t-1}, Z\right)} \middle| W, Z \right] \\
&\qquad = \frac{1}{\gamma^2} \, \mathbb{E}_{Q_1^{t-1}} \left[Q_1^{t-1} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), f \left(Q_1^{t-1}, Z \right) \right) \middle| W, Z \right] \\
&\qquad\qquad - \alpha_{t-1} \, \mathbb{E}_{Q_1^{t-1}} \left[\frac{\partial}{\partial s} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = f\left(Q_1^{t-1}, Z\right)} \middle| W, Z \right].
\end{aligned}$$

Marginalizing over W, Z and recalling (68), we have

$$\lim_{n \to \infty} a_t = \frac{1}{\kappa \gamma^2} \, \mathbb{E} \left[Q_1^{t-1} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), Q_2^{t-1} \right) \right] - \frac{\alpha_{t-1}}{\kappa} \, \mathbb{E} \left[\frac{\partial}{\partial s} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = Q_2^{t-1}} \right].$$

Combining (16) and (62), we obtain

$$\frac{1}{\kappa} \left[1 - \mathbb{E} \left[\frac{2\rho' \left(-Q_1^{t-1} \right)}{1 + \lambda \rho'' \left(\mathsf{prox}_{\lambda \rho} \left(Q_2^{t-1} \right) \right)} \right] \right] = -\frac{1}{\kappa} \, \mathbb{E} \left[\frac{\partial}{\partial a} \Psi_{t-1} \left(h(a, W), s \right) \Big|_{a = Q_1^{t-1}, s = Q_2^{t-1}} \right].$$

Since $\left(-Q_1^{t-1}, Q_2^{t-1} \right) \sim \mathcal{N} \left(\mathbf{0}, \mathbf{\Sigma} \left(\alpha_{t-1}, \sigma_{t-1} \right) \right)$, comparing with (8), we obtain that the LHS equals 1. Further, from (14), we have

$$\mathbb{E} \left[2\rho' \left(-Q_1^{t-1} \right) \left(-Q_1^{t-1} \right) \lambda_t \rho' \left(\mathsf{prox}_{\lambda_t \rho} \left(Q_2^{t-1} \right) \right) \right] = \mathbb{E} \left[Q_1^{t-1} \Psi_{t-1} \left(h \left(Q_1^{t-1}, W \right), Q_2^{t-1} \right) \right].$$

Thus,

$$\lim_{n \to \infty} a_t = \alpha_{t-1} + \frac{1}{\kappa \gamma^2} \, \mathbb{E} \left[2\rho' \left(-Q_1^{t-1} \right) \left(-Q_1^{t-1} \right) \lambda_t \rho' \left(\mathsf{prox}_{\lambda_t \rho} \left(Q_2^{t-1} \right) \right) \right] = \alpha_t,$$

where the last equality follows directly from the definition of α_t in (9). Hence, for any finite t,

$$\lim_{n \to \infty} |\alpha_t - a_t| = 0,$$

which leads to

$$\lim_{n \to \infty} \frac{1}{\sqrt{p}} (C \, \|\boldsymbol{X}\|)^{2t} \sum_{l=0}^{t-1} |\alpha_l - a_l| \, \|\boldsymbol{\beta}\| = 0.$$

Combining this with (65) and (67), we obtain

$$\lim_{n\to\infty} \frac{1}{\sqrt{p}} \left\| \boldsymbol{\nu}^t - \boldsymbol{\theta}^t \right\| = 0.$$

The scaled norm of $\boldsymbol{R}^t - \boldsymbol{S}^t$ is then controlled using (63) and the fact that $\lim_{n\to\infty} \|\boldsymbol{X}\|$ is finite almost surely. This completes the proof.

4.2 Convergence to the MLE

In this subsection, we establish that the AMP iterates $\{\hat{\boldsymbol{\beta}}^t\}$ converge to the MLE $\hat{\boldsymbol{\beta}}$, in the large n and t limit. As mentioned earlier, it can be checked numerically that the system of equations (4) admits a solution in the regime $\gamma < g_{\mathrm{MLE}}(\kappa)$. In addition, we can establish the following result.

Lemma 7. *Given a pair (α, σ), the equation*

$$1 - \kappa = \mathbb{E}\left[\frac{2\rho'\left(Q_1\right)}{1 + \lambda \rho''\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)} \right]$$

has a unique solution in λ, where $(Q_1, Q_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\alpha, \sigma))$, with the covariance function specified in (5).

We defer the proof of Lemma 7 to Section 7, and proceed with the rest of the proof here. The aforementioned results together establish that if the variance map updates (8)–(9) are initialized using $\alpha_0 = \alpha_\star, \sigma_0 = \sigma_\star$, the iterates $(\alpha_t, \sigma_t, \lambda_t)$ remain stationary, that is, for all t,

$$\alpha_t = \alpha_\star, \quad \sigma_t = \sigma_\star, \quad \lambda_t = \lambda_\star,$$

where, recall from Section 1 that $(\alpha_\star, \sigma_\star, \lambda_\star)$ refers to a solution of (4). In the subsequent theorem, we adhere to this particular initialization.

Theorem 7. *Suppose $\gamma < g_{\mathrm{MLE}}(\kappa)$ and assume that the AMP iterates are initialized using*

$$\alpha_0 = \frac{1}{\gamma^2}, \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \boldsymbol{\beta} \rangle}{n}, \qquad \lim_{n,p\to\infty} \frac{1}{p} \|\hat{\boldsymbol{\beta}}^0 - \alpha_\star \boldsymbol{\beta}\|^2 = \sigma_\star^2,$$

where $(\alpha_\star, \sigma_\star, \lambda_\star)$ is a solution to (4). Then the AMP trajectory and the MLE can be related as

$$\lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{\sqrt{p}} \|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\| =_{a.s.} 0. \tag{70}$$

Proof: The proof can be established using techniques similar to that in [9, Theorem 6]. The details are therefore omitted. The crucial point is that, invoking these techniques requires that the following three properties are satisfied:

- Almost surely, the MLE obeys

$$\lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} < \infty. \tag{71}$$

  This follows from Theorem 4, and an application of Borel-Cantelli.

- There exists some non-increasing continuous function $0 < \omega(.) < 1$ independent of n such that

$$\mathbb{P}\left[\nabla^2 \ell(\boldsymbol{\beta}) \succeq \omega\left(\frac{\|\boldsymbol{\beta}\|}{\sqrt{n}} \right) \cdot \boldsymbol{I} \ \text{ for all } \ \boldsymbol{\beta} \right] = 1 - c_1 e^{-c_2 n},$$

  where c_1, c_2 are positive universal constants. This was established in [9, Lemma 4].

- The AMP iterates satisfy a form of Cauchy property:

$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{1}{\sqrt{p}}\left\|\hat{\boldsymbol{\beta}}^{t+1}-\hat{\boldsymbol{\beta}}^{t}\right\| =_{\text{a.s.}} 0,$$
$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{1}{\sqrt{p}}\left\|\boldsymbol{S}^{t+1}-\boldsymbol{S}^{t}\right\| =_{\text{a.s.}} 0.$$

  This can be established by straightforward modifications of [3, Lemma 6.8, Lemma 6.9], using the covariances for $\boldsymbol{Z}^t$ derived in the proof of Lemma 3.

■

Finally, we are in a position to complete the proof of Theorem 1.

Proof of Theorem 1: Start the variance map updates at $\alpha_0 = \alpha_\star, \sigma_0 = \sigma_\star$, so that $\alpha_t \equiv \alpha_0, \sigma_t \equiv \sigma_0$. Choosing $\psi(x,y) = x^2$ in Theorem 6, it directly follows that for every $t \geq 0$,

$$\lim_{n\to\infty}\frac{\|\hat{\boldsymbol{\beta}}^t\|}{\sqrt{p}} \leq \lim_{n\to\infty}\frac{\|\hat{\boldsymbol{\beta}}^t-\alpha_\star\boldsymbol{\beta}\|}{\sqrt{p}} + \lim_{n\to\infty}\frac{\|\alpha_\star\boldsymbol{\beta}\|}{\sqrt{p}} = \sigma_\star + \frac{\gamma\alpha_\star}{\sqrt{\kappa}}$$

$$\implies \lim_{t\to\infty}\lim_{n\to\infty}\frac{\|\hat{\boldsymbol{\beta}}^t\|}{\sqrt{p}} < \infty. \tag{72}$$

Since ψ is a pseudo-Lipschitz function of order 2, by the triangle inequality and Cauchy-Schwartz,

$$\left|\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j-\alpha_t\beta_j,\beta_j\right)-\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j\right)\right|$$
$$\leq C\frac{1}{p}\sqrt{\sum_{j=1}^{p}\left(1+\left\|(\hat{\beta}_j-\alpha_t\beta_j,\beta_j)\right\|+\left\|(\hat{\beta}_j^t-\alpha_t\beta_j,\beta_j)\right\|\right)^2}\left\|\hat{\boldsymbol{\beta}}^t-\hat{\boldsymbol{\beta}}\right\|. \tag{73}$$

Using (72), (71) and invoking Theorem 7, we arrive at

$$\lim_{n\to\infty}\frac{1}{p}\sum_{i=1}^{p}\psi\left(\hat{\beta}_j-\alpha_\star\beta_j,\beta_j\right) = \lim_{t\to\infty}\lim_{n\to\infty}\frac{1}{p}\sum_{i=1}^{p}\psi\left(\hat{\beta}_j^t-\alpha_\star\beta_j,\beta_j\right) = \mathbb{E}\left[\psi\left(\sigma_\star Z,\beta\right)\right].$$

This completes the proof. ■

Remark 1. Theorem 1 in conjunction with Lemma 7 leads to the following crucial result: in the regime $\gamma < g_{\text{MLE}}(\kappa)$, the system of equations (4) admits a *unique* solution. To see this, note that Theorem 1 tells us that for any solution $(\alpha_\star, \sigma_\star, \lambda_\star)$,

$$\alpha_\star = \frac{\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\hat{\beta}_i}{\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\beta_i}.$$

Since for each n, p the MLE $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is unique, the RHS above must be unique. Hence, $\alpha_\star$ has to be unique. Similarly, since

$$\sigma_\star^2 = \lim_{n\to\infty}\frac{1}{p}\left\|\hat{\boldsymbol{\beta}}-\alpha_\star\boldsymbol{\beta}\right\|^2,$$

and the RHS above must be unique, we obtain that $\sigma_\star$ is unique. Then, Lemma 7 establishes that $\lambda_\star$ must also be unique.

5 Asymptotic behavior of the null MLE coordinates

This section presents the proof of Theorem 2. To begin with, we introduce a few notations that will be useful throughout.The reduced MLE, obtained on dropping the j-th predictor is denoted by $\hat{\boldsymbol{\beta}}_{[-j]}$. Define $\boldsymbol{X}_{\bullet j}, \boldsymbol{X}_{\bullet -j}$ to be the j-th column and all the other columns of $\boldsymbol{X}$ respectively. Set $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})$, $\boldsymbol{D}(\boldsymbol{\beta})$ to be the $n \times n$ diagonal matrices with the i-th entry given by $\rho''(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]})$ and $\rho''(\boldsymbol{X}'_i\boldsymbol{\beta})$ respectively, where $\boldsymbol{X}_{i,-j}, \boldsymbol{X}_i$ denote the i-th row of $\boldsymbol{X}_{\bullet -j}$ and $\boldsymbol{X}$ respectively. Suppose the negative log-likelihood obtained on removing the j-th predictor be represented by ℓ_{-j}. Introduce the Gram matrices

$$\boldsymbol{G} = \nabla^2 \ell(\hat{\boldsymbol{\beta}}), \qquad \boldsymbol{G}_{[-j]} = \nabla^2 \ell_{-j}\left(\hat{\boldsymbol{\beta}}_{[-j]}\right). \tag{74}$$

Further, let $\hat{\boldsymbol{\beta}}_{[-i][-j]}$ be the MLE obtained on dropping the i-th observation and the j-th predictor and $\ell_{-i,-j}$ denote the corresponding negative log-likelihood function. Analogously, denote $\hat{\boldsymbol{\beta}}_{[-ik][-j]}$ to be the MLE obtained when both the i-th and the k-th ($i \neq k$) observations are dropped, and in addition, the j-th predictor is removed. Suppose $\ell_{-ik,-j}$ is the corresponding negative log-likelihood function. Define the respective versions of the Gram matrices

$$\boldsymbol{G}_{[-i],[-j]} = \nabla^2 \ell_{[-i],[-j]}\left(\hat{\boldsymbol{\beta}}_{[-i][-j]}\right), \qquad \boldsymbol{G}_{[-ik],[-j]} = \nabla^2 \ell_{[-ik],[-j]}\left(\hat{\boldsymbol{\beta}}_{[-ik][-j]}\right). \tag{75}$$

Before proceeding, it is useful to record a few observations regarding the differences and similarities between our setup here and that in [9]. Analogues of Theorems 2 and 3 were proved in [9] under the global null, that is, $\boldsymbol{\beta} = \boldsymbol{0}$ and under the assumption that the matrix of covariates $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Along the way, [9] established some important generic properties of the logistic link function $\rho(x)$ and the Hessian of the negative log-likelihood function. The logistic link is naturally the same in both the cases, while the Hessian of the negative log-likelihood here has the same distribution as the scaled Hessian $\nabla^2\ell(\boldsymbol{\beta})/n$ from [9].[8] Thus, the properties of these objects established there will be extremely useful in the subsequent discussion. Moreover, as we go along the proofs here, we will see that sometimes it is necessary to generalize certain results in [9] to the $\boldsymbol{\beta} \neq \boldsymbol{0}$ setup. In such scenarios, often the proof techniques from [9] will go through verbatim when particular terms defined therein are replaced by more complicated terms that we will define here. In these cases, we explain the appropriate mapping between the quantities in [9] and those defined here. We leave it to the meticulous reader to check that after such a mapping, the proofs of the corresponding results here indeed go through similarly.

In addition, note that [8, Appendix C] described the skeleton of the proofs for Theorems 2 and 3. In the aforementioned outline, the authors provide a brief sketch of some of the intermediate steps and prove some others rigorously. In Sections 5–7 of this manuscript, we will only provide rigorous proofs of the steps for which the details were left out from [8, Appendix C]. Thus, it may be convenient for the reader to proceed with the rest of this manuscript with [8] and [9] by her side.

The mathematical analyses in this and the subsequent section crucially hinge on the following fact: the minimum eigenvalues of these different versions of $\boldsymbol{G}$ are bounded away from 0 with very high probability. This is established in the following lemma.

Lemma 8. *There exist positive universal constants λ_{lb}, C such that*

$$\mathbb{P}\left[\lambda_{\min}\left(\boldsymbol{G}\right) \geq \lambda_{\mathrm{lb}}\right] \geq 1 - Cn^{-\delta},$$

where $\delta > 1$. The same result holds for $\boldsymbol{G}_{[-j]}, \boldsymbol{G}_{[-i],[-j]}, \boldsymbol{G}_{[-ik],[-j]}$ for any $j \in [p]$ and for all $i, k \in [n], i \neq k$.

Proof: In [9, Lemma 4], it was established that with exponentially high probability, for all sufficiently small $\varepsilon > 0$, the Hessian of the negative log-likelihood satisfies

$$\lambda_{\min}\left(\nabla^2 \ell\left(\boldsymbol{\beta}\right)\right) \geq \left(\inf_{z:|z| \leq \frac{3\|\boldsymbol{\beta}\|}{\sqrt{n\varepsilon}}} \rho''(z)\right) C(\varepsilon),$$

[8] This is simply due to the difference in the variance of the entries of $\boldsymbol{X}$ in the two setups.

where $C(\varepsilon)$ is a positive constant depending on ε and independent of n. This, in conjunction with Theorem 4 completes the proof. ■

Through the rest of this manuscript, for any given n, we restrict ourselves to the event:

$$\mathcal{D}_n := \left\{\lambda_{\min}(\boldsymbol{G}) > \lambda_{\text{lb}}\right\} \cap \left\{\lambda_{\min}(\boldsymbol{G}_{[-j]}) > \lambda_{\text{lb}}\right\} \cap \left\{\cap_{i=1}^{n} \lambda_{\min}(\boldsymbol{G}_{[-i],[-j]}) > \lambda_{\text{lb}}\right\} \cap \left\{\lambda_{\min}(\boldsymbol{G}_{[-12],[-j]}) > \lambda_{\text{lb}}\right\}. \tag{76}$$

By Lemma 8, $\mathcal{D}_n$ occurs with high probability; to be precise,

$$\mathbb{P}\left[\mathcal{D}_n\right] \geq 1 - Cn^{-(\delta-1)}.$$

Later, in Lemma 13 we will use the fact that for any given pair $k, l \in [n]$ with $k \neq l$, $\mathbb{P}\left[\lambda_{\min}(\boldsymbol{G}_{[-kl],[-j]}) > \lambda_{\text{lb}}\right] \geq 1 - Cn^{-\delta}$. In this context, without loss of generality, one can choose $k = 1, l = 2$ and this explains the choice of the last event in (76).

We are now in a position to begin the proof of Theorem 2. To this end, note that the MLE has an implicit description via the KKT conditions and is, therefore, potentially intractable mathematically. To circumvent this barrier, we introduce a surrogate $\boldsymbol{b}_{[-j]}$ for $\hat{\boldsymbol{\beta}}$ that would be more amenable to mathematical analysis. Define

$$\boldsymbol{b}_{[-j]} = \begin{bmatrix} 0 \\ \hat{\boldsymbol{\beta}}_{[-j]} \end{bmatrix} + b_{[-j],1} \begin{bmatrix} 1 \\ -\boldsymbol{G}_{[-j]}^{-1}\boldsymbol{w} \end{bmatrix}, \tag{77}$$

where

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{X}'_{\bullet -j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{X}_{\bullet j}, \\ b_{[-j],1} &= \frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}}, \end{aligned} \tag{78}$$

with the convention that ρ' is applied element-wise and $\boldsymbol{H} := \boldsymbol{I} - \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet -j}\boldsymbol{G}_{[-j]}^{-1}\boldsymbol{X}'_{\bullet -j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}$. Inspired by [5], an analogous surrogate was introduced in [9] for studying the MLE when $\boldsymbol{\beta} = \boldsymbol{0}$, and the choice was motivated in detail. Although the surrogate has a different definition here, the same insight is applicable. Thus, we refer the readers to [9] for the reasoning behind this particular choice. As mentioned earlier, the surrogate is constructed with the hope that $\hat{\boldsymbol{\beta}} \approx \boldsymbol{b}_{[-j]}$. This is formalized in the subsequent theorem.

Theorem 8. *The MLE $\hat{\boldsymbol{\beta}}$ and the surrogate $\boldsymbol{b}_{[-j]}$ defined in (77) satisfy*

$$\begin{aligned} \mathbb{P}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{b}_{[-j]}\| \lesssim n^{-1/2+o(1)}\right] &= 1 - o(1), \\ \mathbb{P}\left[\sup_{1\leq i\leq n}\left|\boldsymbol{X}_i'\boldsymbol{b}_{[-j]} - \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right| \lesssim n^{-1/2+o(1)}\right] &= 1 - o(1). \end{aligned} \tag{79}$$

The fitted values satisfy

$$\mathbb{P}\left[\sup_{1\leq i\leq n}|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \tag{80}$$

Further, we have

$$\mathbb{P}\left[|\hat{\beta}_j - b_{[-j],1}| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \tag{81}$$

Proof: The proof of (79) follows upon tracing out the steps in [9, Theorem 8] verbatim using $\boldsymbol{b}_{[-j]}, b_{[-j],1}$ and $\hat{\boldsymbol{\beta}}_{[-j]}$ defined in (77) instead of $\tilde{\boldsymbol{b}}, \tilde{b}_1$ and $\tilde{\boldsymbol{\beta}}$ respectively. In [9], $\tilde{\boldsymbol{b}}$ is the surrogate for the MLE and $\tilde{b}_1$ is the first coordinate of the surrogate, whereas $\tilde{\boldsymbol{\beta}}$ refers to the MLE obtained on dropping the first predictor. Now, note that the terms $\boldsymbol{G}_{[-j]}, \boldsymbol{w}$ and $b_{[-j],1}$ involve $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})$ and $\rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]})$. They differ from their corresponding counterparts since $\hat{\boldsymbol{\beta}}_{[-j]}$ and $\tilde{\boldsymbol{\beta}}$ have different distributions. However, the only properties pertaining to these objects that are used in the proof of [9, Theorem 8] are the following:

1. $\rho'(x), \rho''(x)$ are bounded, a property we have by virtue of the logistic link,
2. the minimum eigenvalue of $\boldsymbol{G}_{[-j]}$ is strictly positive with very high probability, a fact we have established in our setup in Lemma 8.

Thus, the techniques from [9, Theorem 8] are applicable here for establishing (79). Next, note that by the triangle inequality,

$$\sup_{1\le i\le n} |\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}'_i\hat{\boldsymbol{\beta}}| \le \sup_{1\le i\le n} \left|\boldsymbol{X}'_i\boldsymbol{b}_{[-j]} - \boldsymbol{X}'_i\hat{\boldsymbol{\beta}}\right| + \sup_{1\le i\le n} \left|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}'_i\boldsymbol{b}_{[-j]}\right|.$$

Combining this inequality with (79) and the fact that $\sup_i \|\boldsymbol{X}_i\| = O(1)$ with high probability, we have the required result (80). Finally, (81) follows trivially from (79). ■

Henceforth, whenever necessary, we describe suitable correspondences between the terms here and their appropriate analogues in [9], for the convenience of the reader. To keep the subsequent discussion concise, we will no longer recall the definitions of the relevant terms in the context of [9].

Applying Theorem 8 we have the approximation

$$\hat{\beta}_j = \frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}} + o_P(1). \tag{82}$$

At this point, recall from [8, Appendix C, Equation 20] that the above expression can be simplified to the following form:

$$\frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}} = \frac{\lambda_{[-j]}s_j}{\kappa}Z + o_P(1), \tag{83}$$

where

$$s_j^2 = \frac{\|\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]})\|^2}{n} \qquad \text{and} \qquad \lambda_{[-j]} = \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{G}_{[-j]}^{-1}\right). \tag{84}$$

Later, in Theorem 10, we will establish that $\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star$, where $\lambda_\star$ is part of the solution to the system of equations (4). Hence, it remains to analyze the terms s_j. For convenience of notation, denote the residuals as

$$r_i := y_i - \rho'(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}), \tag{85}$$

which implies

$$s_j^2 = \frac{1}{n}\sum_{i=1}^n r_i^2. \tag{86}$$

Since $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-j]}$ are dependent, the analysis of s_j hard. To circumvent this issue, we express the fitted values $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ as a function of $y_i, \boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$, where recall that $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ is the MLE obtained on removing the i-th observation and the j-th predictor. Such a representation of the fitted values makes things more tractable since $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ are independent. This reduction relies heavily on a leave-one-observation out approach [5,9], in which one constructs a surrogate for $\hat{\boldsymbol{\beta}}_{[-j]}$, starting from $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$, as is done below.

Lemma 9. *Suppose $\hat{\boldsymbol{\beta}}_{[-j]}$ is the MLE obtained on dropping the j-th predictor, and $\hat{\boldsymbol{\beta}}_{[-i],[-j]]}$ is the MLE obtained on further removing the i-th observation. Define $q_i, \hat{\boldsymbol{b}}_{[-j]}$ as follows:*

$$\begin{aligned} q_i &:= \boldsymbol{X}'_{i,-j}\boldsymbol{G}^{-1}_{[-i],[-j]}\boldsymbol{X}_{i,-j}, \\ \hat{\boldsymbol{b}}_{[-j]} &:= \hat{\boldsymbol{\beta}}_{[-i],[-j]} + \boldsymbol{G}^{-1}_{[-i],[-j]}\boldsymbol{X}_{i,-j}\left(y_i - \rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i)\right)\right), \end{aligned} \tag{87}$$

where $\boldsymbol{G}_{[-i],[-j]}$ *is specified by* (75). *Then* $\hat{\boldsymbol{\beta}}_{[-j]}, \hat{\boldsymbol{b}}_{[-j]}$ *satisfy*

$$\mathbb{P}\left[\|\hat{\boldsymbol{\beta}}_{[-j]} - \hat{\boldsymbol{b}}_{[-j]}\| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1).$$

Proof: The proof follows using techniques from [9, Lemma 18], with the choice of q_i specified in (87) and $\hat{\boldsymbol{b}}_{[-j]}$ in place of $\hat{\boldsymbol{b}}$ in [9]. Note that, $\boldsymbol{G}_{[-i],[-j]}$ and $\rho'\left(\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right)\right)$ differ in distribution from the corresponding quantities there, but once again, the properties required in the proof are simply boundedness of ρ' and the eigenvalue bound for $\boldsymbol{G}_{[-i],[-j]}$ established in Lemma 8. ■

We are now in a position to express the fitted values $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ in a more convenient form.

Lemma 10. *The fitted values* $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ *are uniformly close to a function of* $\{y_i, \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]}\}_{i=1,\ldots n}$, *in the following sense:*

$$\sup_{i=1,\ldots,n}\left|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{88}$$

Further, the residuals can be simultaneously approximated using

$$\sup_{i=1,\ldots,n}\left|r_i - \left\{y_i - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right)\right\}\right| \xrightarrow{\mathbb{P}} 0. \tag{89}$$

Proof: Since ρ'' is bounded, (89) follows from (88) trivially. Thus, it suffices to show (88). From the definition of $\hat{\boldsymbol{b}}_{[-j]}$ in (87), it directly follows that

$$\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{b}}_{[-j]} = \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i - q_i\rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i)\right).$$

Comparing the above with relation (7) that involves the proximal mapping operator, we obtain

$$\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{b}}_{[-j]} = \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right).$$

Applying Lemma 9, since $\sup_i \|\boldsymbol{X}_{i,-j}\| = O(1)$ with high probability (see [9, Lemma 2] for a formal statement), we have

$$\sup_i\left|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right)\right| \lesssim n^{-1/2+o(1)}, \tag{90}$$

with high probability. For (88), it then suffices to establish that

$$\sup_i\left|\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right) - \mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{91}$$

To this end, we first examine the differences $|q_i - \lambda_\star|$. By the triangle inequality,

$$\sup_i |q_i - \lambda_\star| \le \sup_i |q_i - \lambda_{[-j]}| + |\lambda_{[-j]} - \lambda_\star|. \tag{92}$$

Using $q_i, \lambda_{[-j]}$ instead of $\tilde{q}_i, \tilde{\alpha}$ in [9, Lemma 19] and following the proof line by line in conjunction with Lemma 8, we have

$$\sup_i |q_i - \lambda_{[-j]}| \lesssim n^{-1/2+o(1)} \tag{93}$$

with high probability. Further, it can be shown that $\lambda_{[-j]} = \lambda_\star + o_P(1)$. This is established later in Theorem 10. For now, we assume this result and proceed with the rest of the arguments. Thus, we have

$$\sup_i |q_i - \lambda_\star| \xrightarrow{\mathbb{P}} 0.$$

The partial derivatives of the proximal mapping operator are given by [4, Proposition 6.3]

$$\frac{\partial}{\partial z}\mathsf{prox}_{b\rho}(z) = \left.\frac{1}{1+b\rho''(x)}\right|_{x=\mathsf{prox}_{b\rho}(z)}, \quad \frac{\partial}{\partial b}\mathsf{prox}_{b\rho}(z) = -\left.\frac{\rho'(x)}{1+b\rho''(x)}\right|_{x=\mathsf{prox}_{b\rho}(z)}, \tag{94}$$

for $b > 0$. By repeated application of the triangle inequality,

$$\begin{aligned}
&\sup_i \left|\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right) - \mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&\leq \sup_i \left|\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right) - \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&+ \sup_i \left|\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + \lambda_\star y_i\right) - \mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&\leq \sup_i |q_i - \lambda_\star| \left\{ \left| \left.\frac{\partial}{\partial z}\mathsf{prox}_{q_i\rho}(z)\right|_{z=\tilde{q}_i y_i} \right| + \left| \left.\frac{\partial}{\partial b}\mathsf{prox}_{b\rho}(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i)\right|_{b=\tilde{\lambda}_i} \right| \right\},
\end{aligned} \tag{95}$$

where $\tilde{q}_i$ lies between $q_i y_i$,$\lambda_\star y_i$ and $\tilde{\lambda}_i$ lies between q_i and $\lambda_\star$. From (94), note that the partial derivatives are both bounded by 1 since $q_i, \tilde{\lambda}_i > 0$. This establishes (91). Combining with (90), we have the required result (88).

■

Recall from (83) and (86) that in order to analyze $\hat{\beta}_j$, we require to study the average of the squared residuals, that is, $\sum_{i=1}^n r_i^2/n$. Note that the residuals are identically distributed. Hence, we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n r_i^2\right) &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(r_i^2) + \frac{1}{n^2}\sum_{i\neq j}\mathrm{Cov}\ (r_i^2, r_j^2) \\
&= \frac{1}{n}\mathrm{Var}(r_1^2) + \frac{(n-1)}{n}\mathrm{Cov}\ (r_1^2, r_2^2).
\end{aligned}$$

The first term above is $o(1)$. From (85), observe that each residual r_i implicitly depends on n. We argue in the subsequent text that

$$\lim_{n\to\infty} \mathrm{Cov}\ (r_1^2, r_2^2) = 0. \tag{96}$$

From (89), we know that r_1, r_2 are close in probability to functions of $\{y_1, \boldsymbol{X}'_{1,-j}\hat{\boldsymbol{\beta}}_{[-1],[-j]}\}$ and $\{y_2, \boldsymbol{X}'_{2,-j}\hat{\boldsymbol{\beta}}_{[-2],[-j]}\}$ respectively. Thus, the entire dependence between r_1 and r_2 seeps in through the dependence between $\hat{\boldsymbol{\beta}}_{[-1],[-j]}$ and $\hat{\boldsymbol{\beta}}_{[-2],[-j]}$. To tackle this dependence structure, we will use a leave-two-observation out approach, that is inspired by [5, 9]. To this end, we establish a crucial result below.

Lemma 11. *For any pair $(i,k) \in [n]$, let $\hat{\boldsymbol{\beta}}_{[-i],[-j]}, \hat{\boldsymbol{\beta}}_{[-k],[-j]}$ denote the MLEs obtained on dropping the i-th and k-th observations respectively, and, in addition, removing the j-th predictor. Further, denote $\hat{\boldsymbol{\beta}}_{[-ik],[-j]}$ to be the MLE obtained on dropping both the i-th, k-th observations and the j-th predictor. Then the following relation holds*

$$\mathbb{P}\left[\max\left\{\left|\boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{\beta}}_{[-ik],[-j]}\right)\right|, \left|\boldsymbol{X}'_{k,-j}\left(\hat{\boldsymbol{\beta}}_{[-k],[-j]} - \hat{\boldsymbol{\beta}}_{[-ik],[-j]}\right)\right|\right\} \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \tag{97}$$

Proof: We focus on one of the indices, say i. To this end, we will rely heavily on Lemma 9. Define $\hat{\boldsymbol{b}}_{[-ik],[-j]}$ analogously to (87) as follows:

$$\hat{\boldsymbol{b}}_{[-i],[-j]} := \hat{\boldsymbol{\beta}}_{[-ik],[-j]} + \boldsymbol{G}^{-1}_{[-ik],[-j]}\boldsymbol{X}_{k,-j}\left(y_k - \rho'\left(\mathsf{prox}_{\tilde{q}_k\rho}(\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-ik],[-j]} + \tilde{q}_k y_k)\right)\right),$$

where $\tilde{q}_k = \boldsymbol{X}'_{k,-j}\boldsymbol{G}^{-1}_{[-ik],[-j]}\boldsymbol{X}_{k,-j}$. An application of Lemma 9 establishes that with high probability

$$\|\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{b}}_{[-i],[-j]}\| \lesssim n^{-1/2+o(1)}. \tag{98}$$

Hence,

$$\left|\boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{\beta}}_{[-ik],[-j]}\right)\right| \leq \|\boldsymbol{X}_{i,-j}\|\|\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{b}}_{[-i],[-j]}\| + \left|\boldsymbol{X}'_{i,-j}\boldsymbol{G}^{-1}_{[-ik],[-j]}\boldsymbol{X}_{k,-j}\right|.$$

The first term is controlled using (98) and the fact that $\|\boldsymbol{X}_{i,-j}\| = O(1)$ with high probability. For the second term note that, conditional on $\boldsymbol{X}_{i,-j}, \boldsymbol{G}^{-1}_{[-ik],[-j]}$, it is a Gaussian random variable with mean zero and variance $\boldsymbol{X}'_{i,-j}\boldsymbol{G}^{-2}_{[-ik],[-j]}\boldsymbol{X}_{i,-j}/n \lesssim 1/n$. Hence, the second term is $O(n^{-1/2+o(1)})$ with high probability. This completes the proof for index i. A similar argument works for index k, hence the result. ■

We are now in a position to establish (96). From (89), we have that for each $\vartheta, \delta > 0$, there exists N such that for all $n \geq N$,

$$\mathbb{P}\left[\sup_{i=1,\ldots,n}\left|r_i - \left\{y_i - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right)\right\}\right| \leq \vartheta\right] \geq 1 - \delta. \tag{99}$$

Let $\mathcal{E}_1, \mathcal{E}_2$ denote the high probability event in (97) and the event in (99) respectively. Denote $\mathcal{H} = \mathcal{D}_n \cap \mathcal{E}_1 \cap \mathcal{E}_2$, where $\mathcal{D}_n$ is defined via (76). Then,

$$|\mathrm{Cov}(r_1^2, r_2^2)| \leq \left|\mathbb{E}\left(r_1^2 - \mathbb{E}\, r_1^2\right)\left(r_2^2 - \mathbb{E}\, r_2^2\right)\mathbf{1}_{\mathcal{H}}\right| + \mathbb{P}(\mathcal{H}^c),$$

since $|r_i^2 - \mathbb{E}\, r_i^2|$ is at most 1. Define for $l = 1, 2$,

$$f(M_l, y_l) := \left(y_l - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(M_l + \lambda_\star y_l\right)\right)\right)^2 - \mathbb{E}\left(y_l - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(M_l + \lambda_\star y_l\right)\right)\right)^2,$$

where $M_l := \boldsymbol{X}'_{l,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}$. Combining (97) and (99) we obtain that for any $\vartheta, \delta > 0$, for every $n \geq N$,

$$|\mathrm{Cov}(r_1^2, r_2^2)| \leq \mathbb{E}\, f(M_1, y_1) f(M_2, y_2) + C\vartheta^2 + \delta, \tag{100}$$

where $C > 0$ is an absolute constant. By arguments similar to that in [5, Lemma 3.23], one can show that

$$\mathbb{E}\, e^{i\boldsymbol{t}'(M_1,y_1) + i\boldsymbol{w}'(M_2,y_2)} - \mathbb{E}\, e^{i\boldsymbol{t}'(M_1,y_1)}\, \mathbb{E}\, e^{i\boldsymbol{w}'(M_2,y_2)} \to 0.$$

Thereafter, repeated applications of the multivariate inversion theorem to obtain densities from characteristic functions yields

$$\mathbb{E}\, f(M_1, y_1) f(M_2, y_2) - \mathbb{E}\, f(M_1, y_1)\, \mathbb{E}\, f(M_2, y_2) \to 0.$$

From (100), we have $\mathbb{E}\, f(M_l, y_l) = 0$, by definition. Then (100) leads to the required result (96). By Chebyshev's inequality, we have effectively established that

$$\frac{1}{n}\sum_{i=1}^n r_i^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, r_i^2 \xrightarrow{\mathbb{P}} 0. \tag{101}$$

Since, the residuals are identically distributed, the approximation to $\hat{\beta}_j$ derived in (82) and (83) yields that for a null j and any $m \in [n]$,

$$\hat{\beta}_j = \frac{\lambda_\star\sqrt{\mathbb{E}\, r_m^2}\, Z}{\kappa} + o_P(1).$$

Appealing to (99) and using arguments similar to that for establishing (96), we have

$$\lim_{n\to\infty} \mathbb{E}\, r_m^2 = \lim_{n\to\infty} \mathbb{E}\left\{y_m - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(\boldsymbol{X}'_{m,-j}\hat{\boldsymbol{\beta}}_{[-m][-j]} + \lambda_\star y_m\right)\right)\right\}^2.$$

Now, the discussion at the end of [8, Appendix C] rigorously established that

$$\frac{\lambda_\star^2 \lim_{n\to\infty} \mathbb{E}\left\{y_m - \rho'\left(\mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{m,-j}\hat{\boldsymbol{\beta}}_{[-m][-j]} + \lambda_\star y_m\right)\right)\right\}^2}{\kappa^2} = \sigma_\star^2,$$

which leads to $\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_\star^2)$ by Slutsky's theorem. This completes the first part of the proof of Theorem 2.

Next, we investigate the joint distribution of multiple null MLE coordinates. Without loss of generality assume $\beta_j = \beta_l = 0$ for some $j, l \in [p]$. From the relations in (82) and (83) in conjunction with Theorem 10, it follows that

$$\begin{bmatrix}\hat{\beta}_j \\ \hat{\beta}_l\end{bmatrix} = \begin{bmatrix}\frac{\lambda_\star}{\kappa}\sum_i X_{ij}\left(y_i - \rho'\left(X'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right)\right) \\ \frac{\lambda_\star}{\kappa}\sum_i X_{il}\left(y_i - \rho'\left(X'_{i,-l}\hat{\boldsymbol{\beta}}_{[-l]}\right)\right)\end{bmatrix} + o_P(1). \tag{102}$$

Let $X_{i,[-jl]}$ be the i-th row of $\boldsymbol{X}$ without the j and l-th entries. Further define $\hat{\boldsymbol{\beta}}_{[-jl]}$ to be the MLE obtained on dropping the j-th and l-th predictors. In (80) we established that if any one of p predictors is dropped, the fitted values before and after are close with high probability. Applying this result to the $p-1$ predictors in $[p]\backslash\{j\}$ we obtain that on further dropping the l-th predictor, the fitted values satisfy

$$\mathbb{P}\left[\sup_i \left|X'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - X'_{i,-[jl]}\hat{\boldsymbol{\beta}}_{[-jl]}\right| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1).$$

Similarly, we have

$$\mathbb{P}\left[\sup_i \left|X'_{i,-l}\hat{\boldsymbol{\beta}}_{[-l]} - X'_{i,-[jl]}\hat{\boldsymbol{\beta}}_{[-jl]}\right| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1).$$

Combining with (102), this implies that

$$\begin{aligned}\begin{bmatrix}\hat{\beta}_j \\ \hat{\beta}_l\end{bmatrix} &= \begin{bmatrix}\frac{\lambda_\star}{\kappa}\sum_i X_{ij}\left(y_i - \rho'\left(X'_{i,-[jl]}\hat{\boldsymbol{\beta}}_{[-jl]}\right)\right) \\ \frac{\lambda_\star}{\kappa}\sum_i X_{il}\left(y_i - \rho'\left(X'_{i,-[jl]}\hat{\boldsymbol{\beta}}_{[-jl]}\right)\right)\end{bmatrix} + o_P(1) \\ &= \frac{\lambda_\star s_{[jl]}}{\kappa}\begin{bmatrix}Z_j \\ Z_l\end{bmatrix} + o_P(1),\end{aligned}$$

where Z_j, Z_l are independent standard normals and

$$s^2_{[jl]} = \frac{1}{n}\sum_{i=1}^n \left(y_i - \rho'\left(X'_{i,-[jl]}\hat{\boldsymbol{\beta}}_{[-jl]}\right)\right)^2.$$

By arguments similar to that for establishing (101), one can establish that $s^2_{[jl]} \xrightarrow{\mathbb{P}} \mathbb{E}\, s^2_{[jl]} =: s_\star$. Then we have

$$\begin{bmatrix}\hat{\beta}_j \\ \hat{\beta}_l\end{bmatrix} = \frac{\lambda_\star s_\star}{\kappa}\begin{bmatrix}Z_j \\ Z_l\end{bmatrix} + o_P(1),$$

which in turn implies that

$$\begin{bmatrix}\hat{\beta}_j \\ \hat{\beta}_l\end{bmatrix} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\boldsymbol{0}, \sigma_\star^2 \boldsymbol{I}\right).$$

For any finite subset of null coordinates, say $i_1, \ldots, i_k$, similar calculations can be carried out as above to obtain that $(\hat{\beta}_{i_1}, \ldots, \hat{\beta}_{i_k}) \xrightarrow{\mathrm{d}} \mathcal{N}(\boldsymbol{0}, \sigma_\star^2 \boldsymbol{I})$.

6 Asymptotic distribution of the LRT

Finally we turn to the proof of Theorem 3. To this end, the following approximation to the LLR is extremely useful.

Theorem 9. *Suppose j is null, that is, $\beta_j = 0$. If $\gamma < g_{\mathrm{MLE}}(\kappa)$, the log-likelihood ratio statistic $\Lambda_j = \ell(\hat{\boldsymbol{\beta}}_{[-j]}) - \ell(\hat{\boldsymbol{\beta}})$ can be approximated as follows:*

$$2\Lambda_j = \frac{\kappa \hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1), \tag{103}$$

where $\lambda_{[-j]}$ is defined in (84).

Proof: Using the KKT condition $\nabla \ell(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ and Taylor expansion, we arrive at

$$2\Lambda_j = \left(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}} \boldsymbol{D}(\hat{\boldsymbol{\beta}}) \left(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \frac{1}{3}\sum_{i=1}^{n} \rho'''(\gamma_i) \left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}'_i\hat{\boldsymbol{\beta}}\right)^3, \tag{104}$$

where γ_i lies between $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ and $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$. Invoking Theorem 8 and the fact that $|\rho'''|_\infty$ is bounded, we obtain that the cubic term in (104) is $o_P(1)$. Subsequently, it can be checked that calculations similar to those in [9, Section 7.3] go through in this setup on using Theorem 8. This completes the proof. ■

To establish Theorem 3, it remains to analyze $\lambda_{[-j]}$. To this end, the following lemma and an application of Slutsky's theorem completes the proof.

Theorem 10. *If $\gamma < g_{\mathrm{MLE}}(\kappa)$, the random variable $\lambda_{[-j]}$ defined in (84) converges in probability to a constant. In fact,*

$$\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star,$$

where $\lambda_\star$ is part of the solution to the system (4).

Proof: The proof follows by arguments similar to that in [9, Appendix I] with some modifications. First, we establish that $\lambda_{[-j]}$ is an approximate zero of a random function $\delta_n(x)$, in a sense that is formalized below.

Lemma 12. *Define $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ to be the MLE obtained when the i-th observation and the j-th predictors are removed and $\boldsymbol{X}'_{i,-j}$ to be the i-th row of the matrix $\boldsymbol{X}$, with the j-th column removed. Let $\delta_n(x)$ be the random function*

$$\delta_n(x) := \frac{p}{n} - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + xy_i\right)\right)}. \tag{105}$$

Then, $\lambda_{[-j]}$ obeys

$$\delta_n\left(\lambda_{[-j]}\right) \xrightarrow{\mathbb{P}} 0.$$

Proof of Lemma 12: Upon replacing $\tilde{\alpha}$ by $\lambda_{[-j]}$ in the proof of [9, Proposition 2], we obtain

$$\frac{p}{n} - 1 + \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{1 + \rho''\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right)\lambda_{[-j]}} \right\} \xrightarrow{\mathbb{P}} 0. \tag{106}$$

We claim that the fitted values $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ can be approximated as follows:

$$\sup_i \left| \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{\lambda_{[-j]}\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + \lambda_{[-j]}y_i\right) \right| \lesssim n^{-1/2+o(1)}, \tag{107}$$

with high probability. The claim is established by comparing (90), (93), and by arguments similar to that in (95), with $\lambda_\star$ replaced by $\lambda_{[-j]}$.

Using the fact that $\left|\frac{1}{1+x}-\frac{1}{1+y}\right| \le |x-y|$ for $x, y \ge 0$, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{1+\rho''\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right)\lambda_{[-j]}}\right\} - \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{1+\rho''\left(\mathsf{prox}_{\lambda_{[-j]}\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}+\lambda_{[-j]}y_i\right)\right)\lambda_{[-j]}}\right\}\right| \le |\lambda_{[-j]}|\,|\rho'''|_\infty \sup_i \left|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{\lambda_{[-j]}\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}+\lambda_{[-j]}y_i\right)\right|.$$

On the event $\mathcal{D}_n$ defined in (76), $|\lambda_{[-j]}| \le p/(n\lambda_{\mathrm{lb}})$. Further, ρ''' is bounded. Hence, from (107) we have the desired result.

■

The next stage is to show that the random function $\delta_n(x)$ converges in a uniform sense to a deterministic function $\Delta(x)$.

Lemma 13. *Define $\Delta(x)$ to be the deterministic function*

$$\Delta(x) = \kappa - 1 + \mathbb{E}\left[\frac{1}{1+x\rho''\left(\mathsf{prox}_{x\rho}\left(xh(\tilde{Q}_1, W)+\tilde{Q}_2\right)\right)}\right], \tag{108}$$

where $(\tilde{Q}_1, \tilde{Q}_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(-\alpha_\star, \sigma_\star))$, $W \sim U(0,1) \perp\!\!\!\perp (\tilde{Q}_1, \tilde{Q}_2)$, $\boldsymbol{\Sigma}$ is specified via (5) and $(\alpha_\star, \sigma_\star)$ form part of the solution to the system (4). Then, for any $B > 0$,

$$\sup_{x\in[0,B]} |\delta_n(x) - \Delta(x)| \xrightarrow{\mathbb{P}} 0. \tag{109}$$

Proof of Lemma 13: As a first step, using compactness of the interval $[0, B]$ and the definitions of $\delta_n(x)$ and $\Delta(x)$ in (105) and (108) respectively, it can be established that for (109), it suffices to show the following: for any given $x \in [0, B]$

$$|\delta_n(x) - G_n(x)| \xrightarrow{\mathbb{P}} 0, \tag{110}$$
$$|G_n(x) - \Delta(x)| \to 0, \tag{111}$$

where $G_n(x) = \mathbb{E}\left(\delta_n(x)\right)$. (We refer the interested reader to the proof of [9, Proposition 3] for a detailed analogous computation in the simpler setup $\boldsymbol{\beta} = \mathbf{0}$).

We first establish (111). To this end, we seek to express $G_n(x)$ in an alternative, more convenient form. Denote by $\boldsymbol{\beta}_{-j}$ the vector of regression coefficients without the j-th coordinate. Recall that the discussion at the end of [8, Appendix C] rigorously established the following fact:

$$\begin{bmatrix} Q_1^\star \\ Q_2^\star \end{bmatrix} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \gamma^2 & 0 \\ 0 & \kappa\sigma_\star^2 \end{bmatrix}\right), \quad \text{where} \quad Q_1^\star := \boldsymbol{X}'_{i,-j}\boldsymbol{\beta}_{-j}, \quad Q_2^\star = \boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-i][-j]} - \alpha_\star\boldsymbol{\beta}_{-j}\right). \tag{112}$$

As mentioned in (41), the responses can be expressed as $y_i = h(Q_1^\star, w_i)$, since we operate under the null $\beta_j = 0$, where $w_i \sim U(0,1)$ is independent of both $Q_1^\star$ and $Q_2^\star$. In terms of these random variables, $G_n(x)$ can be expressed as

$$G_n(x) = \frac{p}{n} - 1 + \mathbb{E}\left[\frac{1}{1+x\rho''\left(\mathsf{prox}_{x\rho}\left(Q_2^\star + \alpha_\star Q_1^\star + xh\left(Q_1^\star, w_i\right)\right)\right)}\right].$$

Now, the function

$$(t, l, w) \mapsto \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(l + \alpha_\star t + xh\left(t, w\right)\right)\right)}$$

is bounded with the discontinuity points having Lebesgue measure zero. Note that $(Q_1^\star, Q_2^\star, w_i)$ arise from a continuous joint distribution. Hence, from (112) we can conclude that

$$\mathbb{E}\left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(Q_2^\star + \alpha_\star Q_1^\star + xh\left(Q_1^\star, w_i\right)\right)\right)}\right] \to \mathbb{E}\left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\tilde{Q}_2 + xh\left(\tilde{Q}_1, W\right)\right)\right)}\right],$$

where $\tilde{Q}_1, \tilde{Q}_2, W$ are as in the statement of the lemma. This completes the proof of (111).

To analyze (110), note that

$$\delta_n(x) - G_n(x) = \frac{1}{n}\sum_{i=1}^{n} f(M_i, y_i), \text{ where } M_i = \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]},$$

$$f(M_i, y_i) = \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(M_i + xy_i\right)\right)} - \mathbb{E}\left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(M_i + xy_i\right)\right)}\right].$$

Since $\{f(M_i, y_i)\}_{i=1\ldots n}$ are identically distributed this immediately gives,

$$\begin{aligned}\mathrm{Var}(\delta_n(x)) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[f^2\left(M_i, y_i\right)\right] + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[f(M_i, y_i)f(M_j, y_j)\right] \\ &= \frac{\mathbb{E}\left[f^2(M_1, y_1)\right]}{n} + \frac{n(n-1)}{n^2}\mathbb{E}\left[f(M_1, y_1)f(M_2, y_2)\right].\end{aligned}$$

It suffices to establish that $\mathbb{E}\left[f(M_1, y_1)f(M_2, y_2)\right] \to 0$, since it ensures $\delta_n(x) \xrightarrow{L_2} G_n(x)$. To this end, we resort to the leave-two-observation-out approach discussed in Section 5. By routine arguments using the triangle inequality, properties of the partial derivatives of the proximal mapping operator (94), the fact that $\|f\|_\infty \leq 1$, and invoking the approximations in Lemma 11, we arrive at

$$f(M_1, y_1)f(M_2, y_2) - f(\boldsymbol{X}'_{1,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_1)f(\boldsymbol{X}'_{2,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_2) \xrightarrow{L_1} 0.$$

From arguments similar to [5, Lemma 3.23] and the multivariate inversion theorem, we obtain

$$\mathbb{E}\left[f(\boldsymbol{X}'_{1,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_1)f(\boldsymbol{X}'_{2,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_2)\right] - \mathbb{E}\left[f(\boldsymbol{X}'_{1,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_1)\right]\mathbb{E}\left[f(\boldsymbol{X}'_{2,-j}\hat{\boldsymbol{\beta}}_{[-12],[-j]}, y_2)\right] \to 0,$$

which yields the desired result, since f is centered. ■

Putting together Lemmas 12 and 13, since $\lambda_{[-j]} \leq p/n\lambda_{\mathrm{lb}}$ on the high probability event $\mathcal{D}_n$ defined in (76), we obtain that

$$\Delta(\lambda_{[-j]}) \xrightarrow{\mathbb{P}} 0.$$

To complete the proof, recall from (16) that $\Delta(x)$ can be alternatively expressed as

$$\Delta(x) = \kappa - 1 + \mathbb{E}\left[\frac{2\rho'(-\tilde{Q}_1)}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\tilde{Q}_2\right)\right)}\right].$$

From Lemma 7, we know that $\Delta(x) = 0$ has a unique solution. Comparing with the system of equations in (4) and noting that $(-\tilde{Q}_1, \tilde{Q}_2) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\alpha_\star, \sigma_\star))$, we obtain that $\lambda_\star$ is the unique solution to $\Delta(x) = 0$. Hence, $\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star$.

■

7 Proof of Supporting Lemmas

In this section, we provide proofs of Lemmas 7, 2 and 1.

7.1 Proof of Lemma 7

Let $a = -\alpha, b = \sqrt{\kappa}\sigma$ and denote the function

$$G(\lambda) = \mathbb{E}\left[\frac{2\rho'(Q_1)}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(aQ_1+bZ))}\right],$$

where $Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp Q_1$ and $\lambda > 0$. It is required to show that

$$1 - G(\lambda) = \kappa \tag{113}$$

has a unique solution. Note that $\lambda \mapsto G(\lambda)$ is continuous. To prove the lemma, it suffices to show that G is strictly increasing and that

$$\lim_{\lambda\to 0}(1-G(\lambda)) = 0 \tag{114}$$

$$\lim_{\lambda\to\infty}(1-G(\lambda)) = 1. \tag{115}$$

To this end, define the function

$$K_\lambda(p,s) := \lambda\rho'\left(\mathsf{prox}_{\lambda\rho}(p+s)\right).$$

The partial derivative of the above with respect to the second argument is given by [3, Proposition 6.4]

$$K'_\lambda(p,s) := \frac{\partial K_\lambda(p,s)}{\partial s} = \frac{\lambda\rho''(\mathsf{prox}_{\lambda\rho}(p+s))}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(p+s))}. \tag{116}$$

Hence, $G(\lambda)$ can be expressed as

$$G(\lambda) = \mathbb{E}\left[2\rho'(Q_1)\left(1-K'_\lambda(aQ_1,bZ)\right)\right]. \tag{117}$$

Applying Stein's formula (69), one can check that

$$\mathbb{E}\left[K'_\lambda(aQ_1,bZ)|Q_1\right] = -\frac{1}{b}\int_{-\infty}^{\infty} K_\lambda(aQ_1,bz)\phi'(z)dz,$$

where $\phi(\cdot)$ is the standard normal density. Plugging this back in (117) and differentiating with respect to λ, we obtain

$$G'(\lambda) = \frac{1}{b}\,\mathbb{E}_{Q_1}\left[2\rho'(Q_1)\int_{-\infty}^{\infty}\frac{\partial K_\lambda(aQ_1,bz)}{\partial\lambda}\phi'(z)dz\right].$$

Define $f(\cdot)$ to be the function

$$f(q) = \frac{1}{b}\int_{-\infty}^{\infty}\frac{\partial K_\lambda(aq,bz)}{\partial\lambda}\phi'(z)dz.$$

A result analogous to Lemma 7 was proved in [9, Lemma 5] for a different choice of the function G. In the proof, it was established that $f(0) < 0$. One can check that, in order to study $f(q)$ for any fixed $q \in \mathbb{R}$, the same arguments go through and we have $f(q) < 0$ for all $q \in \mathbb{R}$. Since $\rho'(\cdot) > 0$, this implies $G'(\lambda) < 0$. Hence, the function $1 - G(\lambda)$ is strictly increasing.

To show (114), note that for $\lambda > 0$, $x \mapsto \lambda x/(1+\lambda x)$ is strictly increasing in x. Hence, for any (q_1, z),

$$0 \le K'_\lambda(aq_1,bz) \le \frac{\lambda\|\rho''\|_\infty}{1+\lambda\|\rho''\|_\infty} \le 1.$$

This implies that for any (q_1, z), when $\lambda \to 0$, $K'_\lambda(aq_1, bz) \to 0$. Further, since ρ' is bounded, by the dominated convergence theorem, we have

$$\lim_{\lambda \to 0} (1 - G(\lambda)) = 1 - \mathbb{E}\left[2\rho'\left(Q_1\right)\right],$$

recalling the expression for $G(\lambda)$ provided in (117). Now, we know that $\rho'(x) = 1 - \rho'(-x)$ and $Q_1 \overset{\mathrm{d}}{=} -Q_1$, which yields

$$\begin{aligned} 2\,\mathbb{E}\,\rho'(Q_1) &= \mathbb{E}\,\rho'(Q_1) + 1 - \mathbb{E}\,\rho'(-Q_1) \\ &= \mathbb{E}\,\rho'(Q_1) + 1 - \mathbb{E}\,\rho'(Q_1) \\ \implies 1 - 2\,\mathbb{E}\,\rho'(Q_1) &= 0, \end{aligned}$$

thus establishing (114).

Finally, we turn to the proof of (115). To this end, note that [9, Remark 3] established the following crucial property regarding the logistic link function ρ: for any $(q_1, z) \in \mathbb{R}^2$,

$$\lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(aq_1 + bz\right)\right) \to \infty \text{ when } \lambda \to \infty.$$

Hence, for any (q_1, z) recalling (116), we obtain that $K'_\lambda(aq_1, bz) \to 1$ when $\lambda \to \infty$. Again, by the dominated convergence theorem, we have

$$\mathbb{E}\left[2\rho'(Q_1)\left(1 - K'_\lambda(aQ_1, bZ)\right)\right] \to 0 \text{ when } \lambda \to \infty,$$

proving (115).

7.2 Proof of Lemma 2

Proof: For any $\boldsymbol{v} \in \mathbb{R}^n$, denote $\mathcal{C}_i(\mathrm{span}(\boldsymbol{v})) = \mathcal{C}_i^{\boldsymbol{v}}$. From the definition of the statistical dimension,

$$\delta(\mathcal{C}_i^{\boldsymbol{v}}) = \mathbb{E}\left[\|\Pi_{\mathcal{C}_i^{\boldsymbol{v}}}\|^2\right] = \mathbb{E}\left[\|\boldsymbol{g}\|^2 - \min_{t\in\mathbb{R}} \min_{\boldsymbol{u}\in\mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2\right], \tag{118}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. It can be checked that an approach similar to that in [9, Appendix D.2] leads to the lower bound

$$\begin{aligned} &\min_{t\in\mathbb{R}} \min_{\boldsymbol{u}\in\mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2 \\ &\geq \min_t \left\{ \sum_{i:(g_i - tv_i)<0} (g_i - tv_i)^2 - \max_{S\subset[n]:|S|=2\sqrt{\epsilon}n} \sum_{i\in S} (g_i - tv_i)^2 - 2\sqrt{2}\varepsilon^{3/4}\|\boldsymbol{g} - t\boldsymbol{v}\|^2 \right\}, \end{aligned} \tag{119}$$

where $\varepsilon > 0$ is a small constant. In the remaining proof, we carefully analyze the RHS of (119). To this end, define $G^{\boldsymbol{v}}(t) = F^{\boldsymbol{v}}(t) - \varepsilon^{\boldsymbol{v}}(t)$, where

$$\begin{aligned} F^{\boldsymbol{v}}(t) &= \sum_{i:(g_i - tv_i)<0} (g_i - tv_i)^2 \\ \varepsilon^{\boldsymbol{v}}(t) &= \max_{S\subset[n]:|S|=2\sqrt{\varepsilon}n} \sum_{i\in S} (g_i - tv_i)^2 + 2\sqrt{2}\varepsilon^{3/4}\|\boldsymbol{g} - t\boldsymbol{v}\|^2. \end{aligned} \tag{120}$$

Further, define $f^{\boldsymbol{v}}(t) = \mathbb{E}[F^{\boldsymbol{v}}(t)]$ and let t_0 and $t_\star$ be the minimizers of $f^{\boldsymbol{v}}(t)$ and $F^{\boldsymbol{v}}(t)$ respectively. At this point, it is useful to record a crucial observation that follows from [2, Section 3.3]:

$$\frac{1}{n} F^{\boldsymbol{v}}(t_\star) \overset{\mathbb{P}}{\to} g_{\mathrm{MLE}}^{-1}(\gamma). \tag{121}$$

We require the following lemma to complete the proof.

Lemma 14. *There exists a fixed positive constant ε_0 such that for all $\epsilon \le \epsilon_0$, there exists an event $\mathcal{G}_{\boldsymbol{V}}$ in the σ-algebra generated by $\boldsymbol{V}$ satisfying condition (31) and the following property: for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$, with high probability,*[9]

$$\sup_{t \in [t_0 - M, t_0 + M]} |\varepsilon^{\boldsymbol{v}}(t)| \le n f(\varepsilon), \tag{122}$$

$$\forall \quad t \notin [t_0 - M, t_0 + M] \qquad G^{\boldsymbol{v}}(t) > n g_{\mathrm{MLE}}^{-1}(\gamma), \tag{123}$$

where $\varepsilon^{\boldsymbol{v}}(t), G^{\boldsymbol{v}}(t)$ are defined via (120). Above, $M \equiv M(\varepsilon)$ is a positive constant independent of n, $\mathcal{F}_{\boldsymbol{V}}$ is the event defined in (24), $f(x)$ is a smooth function such that $\lim_{x \to 0} f(x) = 0$ and $f(x)$ is increasing on $[0, \varepsilon_0]$.

Let $\nu_0 = f(\varepsilon_0)$. Then for all $0 < \nu < \nu_0$, applying Lemma 14 it can be established that, with high probability for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$,

$$\begin{aligned} \min_{t \in [t_0 - M, t_0 + M]} G^{\boldsymbol{v}}(t) &\ge \min_{t \in [t_0 - M, t_0 + M]} F^{\boldsymbol{v}}(t) - \sup_{t \in [t_0 - M, t_0 + M]} \varepsilon^{\boldsymbol{v}}(t) \\ &\ge F^{\boldsymbol{v}}(t_\star) - n\nu + o_P(1) \ge n(g_{\mathrm{MLE}}^{-1}(\gamma) - \nu + o_P(1)), \end{aligned}$$

where the last inequality follows from (121). Here, $o_P(1)$ denotes a random variable that converges to zero in probability as $n \to \infty$, under the law of $\boldsymbol{g}$. Combining this with the high probability lower bound for $G^{\boldsymbol{v}}(t)$ on the complement of $[t_0 - M, t_0 + M]$ obtained from Lemma 14 yields that, for all t and for all $0 < \nu < \nu_0$,

$$G^{\boldsymbol{v}}(t) \ge n(g_{\mathrm{MLE}}^{-1}(\gamma) - \nu + o_P(1)).$$

In conjunction with (119), this yields that with high probability,

$$\min_{t \in \mathbb{R}} \min_{\boldsymbol{u} \in \mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2 \ge n(g_{\mathrm{MLE}}^{-1}(\gamma) - \nu + o(1)).$$

Denote this high probability event by $\mathcal{M}$. Since

$$\mathbb{E}\left[\min_{t \in \mathbb{R}} \min_{\boldsymbol{u} \in \mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2\right] \ge \mathbb{E}\left[\min_{t \in \mathbb{R}} \min_{\boldsymbol{u} \in \mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2 \mathbf{1}_{\mathcal{M}}\right],$$

recalling (118), we have

$$\delta(\mathcal{C}_i^{\boldsymbol{v}}) \le n - n(g_{\mathrm{MLE}}^{-1}(\gamma) - \nu + o(1)),$$

thus completing the proof. ∎

It remains to prove Lemma 14, which is the focus of the rest of this subsection.

Proof of Lemma 14: To begin with, we will specify the event $\mathcal{G}_{\boldsymbol{V}}$. Since $\boldsymbol{V}$ has sub-Gaussian tails, by an application of [6] and the union bound, for $a(\varepsilon) = 2\max\{2\sqrt{\varepsilon}H(2\sqrt{\varepsilon})\}$,

$$\mathbb{P}_{\boldsymbol{V}}\left[\max_{S : |S| = 2\sqrt{\varepsilon}n} \sum_{i \in S} V_i^2 \le C_1 n a(\epsilon)\right] \ge 1 - e^{-H(2\sqrt{\varepsilon})n}, \tag{124}$$

where $H(x) = -x\log x - (1-x)\log(1-x)$ and $\mathbb{P}_{\boldsymbol{V}}$ denotes the probability under the law of $\boldsymbol{V}$. From results on the norm of a random vector with independent sub-Gaussian entries, [11, Sec 3.1], it can be established that

$$\mathbb{P}_{\boldsymbol{V}}\left[\|\boldsymbol{V}\|_2 \le C_1\sqrt{n}\right] \ge 1 - 2\exp\left(-c_1 n\right), \tag{125}$$

[9] Here, the probability is over the law of $\boldsymbol{g}$.

where $\boldsymbol{V}$ denotes the random vector $\boldsymbol{V} = (V_1, \ldots, V_n)$. Since, $|V| - \mathbb{E}\,|V|$ is sub-Gaussian, applying the Hoeffding-type inequality [10, Proposition 5.10], we have

$$\mathbb{P}_{\boldsymbol{V}}\left[\sum_{i=1}^{n} |V_i| \leq C_1 n\right] \geq 1 - C_1 \exp\left(-c_1 n\right). \tag{126}$$

Next, note that $V^2 \mathbf{1}_{V>0} - \mathbb{E}\, V^2 \mathbf{1}_{V>0}$ is sub-exponential and from the Bernstein-type inequality [10, Proposition 5.16], it can be shown that

$$\mathbb{P}_{\boldsymbol{V}}\left[\sum_{i=1}^{n} V_i^2 \mathbf{1}_{V_i>0} \geq C_1 n\right] \geq 1 - 2\exp\left(-c_1 n\right). \tag{127}$$

Let $\mathcal{G}_{\boldsymbol{V}}$ denote the high probability event formed by the intersection of the events in (124),(125),(126) and (127). Thus, any $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$ satisfies the following properties:

$$\max_{S\subset[n]:|S|=2\sqrt{\epsilon}n} \sum_{i\in S} v_i^2 \leq C_1 n a(\varepsilon), \quad \|\boldsymbol{v}\|^2 \leq C_2 n, \quad \sum_{i=1}^{n} |v_i| \leq C_3 n, \quad \sum_{i:v_i>0} v_i^2 \geq C_4 n, \quad \max_i v_i^2 \leq \zeta \log n. \tag{128}$$

We are now in a position to establish (122) and (123). To this end, recall that,

$$\begin{aligned}
\varepsilon^{\boldsymbol{v}}(t) &= \max_{S\subset[n]:|S|=2\sqrt{\varepsilon}n} \sum_{i\in S} (g_i - t v_i)^2 + 2\sqrt{2}\varepsilon^{3/4}\|\boldsymbol{g} - t\boldsymbol{v}\|^2 \\
&\leq \max_{S\subset[n]:|S|=2\sqrt{\varepsilon}n} 2\sum_{i\in S} g_i^2 + \max_{S\subset[n]:|S|=2\sqrt{\varepsilon}n} 2t^2 \sum_{i\in S} v_i^2 + 2\sqrt{2}\varepsilon^{3/4}\{\|\boldsymbol{g}\|^2 + t^2\|\boldsymbol{v}\|^2\}.
\end{aligned}$$

To control the above, note that similar to (124) and (125), we have

$$\max_{S\subset[n]:|S|=2\sqrt{\epsilon}n} \sum_{i\in S} g_i^2 \leq C_1 n a(\varepsilon), \qquad \|\boldsymbol{g}\|^2 \leq C_2 n,$$

with high probability. Putting these together, for all t,

$$\varepsilon^{\boldsymbol{v}}(t) \leq n\left(1 + t^2\right)\left(C_1 a(\varepsilon) + C_2 \varepsilon^{3/4}\right), \tag{129}$$

with high probability. Hence, for any positive universal constant M, for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$, with high probability,

$$\sup_{t\in[t_0-M,t_0+M]} \varepsilon^{\boldsymbol{v}}(t) \leq n f(\varepsilon),$$

where $f(x)$ is specified in the statement of the lemma.

It remains to lower bound $G^{\boldsymbol{v}}(t)$ outside the finite interval $[t_0 - M, t_0 + M]$ where M is any positive constant independent of n. Consider $t > 1$. In this case, invoking (129) we have for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$,

$$G^{\boldsymbol{v}}(t) \geq \sum_{i:g_i<tv_i} (g_i - t v_i)^2 - nt^2(C_1 a(\varepsilon) + C_2\varepsilon^{3/4}) \tag{130}$$

with high probability. Observe that $\{i : v_i > 0, g_i \leq 0\} \subset \{i : g_i - t v_i < 0\}$. Thus,

$$\sum_{i:g_i<tv_i} (g_i - t v_i)^2 \geq t^2 \sum_{i:v_i>0,g_i\leq 0} v_i^2 - 2t \sum_{i:v_i>0,g_i<0} |v_i g_i| \geq t^2 \sum_{i:v_i>0,g_i\leq 0} v_i^2 - 2t\sum_{i=1}^{n} |v_i g_i|. \tag{131}$$

Since $\boldsymbol{g} \to \sum_{i=1}^{n} |g_i v_i|$ is Lipschitz with Lipschitz constant at most $\|\boldsymbol{v}\|$, by Gaussian concentration of Lipschitz functions and from the properties of $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$ described in (128), we have

$$\sum_{i=1}^{n} |g_i v_i| \leq C_2 n \tag{132}$$

with high probability.

Thus, it only remains to analyze the first term in the RHS of (131). Note that the v_i's are deterministic in this term and $\boldsymbol{g}$ is the random variable. So, $v_i^2 \mathbf{1}_{v_i>0, g_i \leq 0} - v_i^2 \mathbf{1}_{v_i>0}/2$ is a centered multiple of a Bernoulli random variable and from [10, Proposition 5.10] we have,

$$\mathbb{P}_{\boldsymbol{g}} \left[\left| \sum_{i:v_i>0} v_i^2 \mathbf{1}_{g_i \leq 0} - \frac{1}{2} \sum_{i:v_i>0} v_i^2 \right| \geq t \right] \leq C_1 \exp\left(- \frac{c_1 t^2}{n \left(\max_i v_i^2\right)^2} \right),$$

where $\mathbb{P}_{\boldsymbol{g}}$ denotes the probability under the law of $\boldsymbol{g}$. This is where the control over $\max_i v_i^2$, that is ensured by restricting $\boldsymbol{v}$ to $\mathcal{F}_{\boldsymbol{V}}$ defined via (24), is crucial. Recalling the properties of $\boldsymbol{v}$ from (128), we can choose $t = C_1 n$ such that

$$\sum_{i:v_i>0} v_i^2 \mathbf{1}_{g_i \leq 0} \geq C_2 n \tag{133}$$

with high probability. Combining (132), (133) and recalling (131), we finally arrive at

$$G^{\boldsymbol{v}}(t) \geq C_1 t^2 n - 2t C_2 n - n t^2 (C_3 a(\varepsilon) + C_4 \varepsilon^{3/4})$$

for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$ with high probability, when $t > 1$. If ε is sufficiently small, one can choose a positive constant M such that $t_0 + M > 1$ and for all $t > t_0 + M$ the RHS in the above inequality exceeds $n g_{\text{MLE}}^{-1}(\gamma)$. This establishes the desired result for all $t > t_0 + M$. The case of $t < t_0 - M$ can be analyzed similarly and is, therefore, omitted.

■

7.3 Proof. of Lemma 1

The event $\{\text{span}\,(\boldsymbol{V}) \cap \mathcal{A} \neq \{\mathbf{0}\}\}$ occurs if and only if

$$\exists\ a \neq 0 \text{ such that } a\boldsymbol{V} \in \mathcal{A}.$$

Hence,

$$\mathbb{P}\left[\text{span}\,(\boldsymbol{V}) \cap \mathcal{A} \neq \{\mathbf{0}\}\right] \leq \mathbb{P}\left[\exists\ a > 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right] + \mathbb{P}\left[\exists\ a < 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right]. \tag{134}$$

From the definition of $\mathcal{A}$ in (21), it follows that

$$\mathbb{P}\left[\exists\ a > 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right] = \mathbb{P}\left[\sum_{j=1}^{n} |V_j| \mathbf{1}_{V_j<0} \leq \varepsilon^2 \sqrt{n} \|\boldsymbol{V}\|\right] \leq \mathbb{P}\left[\sum_{j=1}^{n} |V_j| \mathbf{1}_{V_j<0} \leq \varepsilon^2 n\right] + C_1 \exp\left(-c_1 n\right), \tag{135}$$

where the last inequality follows from (125). Since $|V_i| \mathbf{1}_{V_i<0} - \mathbb{E}\,|V_i| \mathbf{1}_{V_i<0}$ is sub-gaussian, applying [10, Proposition 5.10] we obtain

$$\mathbb{P}\left[\frac{\left(\mathbb{E}\,|V_1| \mathbf{1}_{V_1<0}\right) n}{2} \leq \sum_{i=1}^{n} |V_i| \mathbf{1}_{V_i<0} \leq \frac{3\left(\mathbb{E}\,|V_1| \mathbf{1}_{V_1<0}\right) n}{2}\right] \geq 1 - C_1 \exp\left(-c_1 n\right). \tag{136}$$

Combining (135) and (136) yields that for sufficiently small ε,

$$\mathbb{P}\left[\exists\ a > 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right] \leq C_1 \exp\left(-c_1 n\right).$$

The second term in the RHS of (134) can be analyzed similarly.

References

[1] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005, 2014.

[2] E. J. Candes and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, April 2018.

[3] David Donoho and Andrea Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, pages 1–35, 2013.

[4] David Donoho and Andrea Montanari. Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. *Technical report*, 2015.

[5] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2017.

[6] Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

[7] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, page iat004, 2013.

[8] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

[9] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv preprint arXiv:1706.01191*, 2017.

[10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, pages 210 – 268, 2012.

[11] Roman Vershynin. High-dimensional probability. *An Introduction with Applications*, 2016.