Supporting Information to: A Modern Maximum-Likelihood Theory for High-dimensional Logistic Regression

Pragya Sur* Emmanuel J. Candès*†

June, 2018

1 Main Mathematical Ideas

In the main text of this supplementary information, we describe some of the key mathematical ideas in the arguments, relying as much as possible on published results from [33]. Our detailed proofs are relatively long and we defer them to Appendix H.

1.1 The bulk distribution of the MLE

To analyze the MLE, we introduce an approximate message passing (AMP) algorithm that tracks the MLE in the limit of large n and p. Our purpose is a little different from the work in [28] which, in the context of generalized linear models, proposed AMP algorithms for Bayesian posterior inference, and whose properties have later been studied in [25] and [3]. To the best of our knowledge, an AMP algorithm for tracking the MLE from a logistic model has not yet been proposed in the literature. Our starting point is to write down a sequence of AMP iterates $\{\boldsymbol{S}^t, \hat{\boldsymbol{\beta}}^t\}_{t\geq 0}$, with $\boldsymbol{S}^t \in \mathbb{R}^n, \hat{\boldsymbol{\beta}}^t \in \mathbb{R}^p$, using the following scheme: starting with an initial guess $\boldsymbol{\beta}^0$, set $\boldsymbol{S}^0 = \boldsymbol{X}\boldsymbol{\beta}^0$ and for $t = 1, 2, \ldots$, inductively define

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

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

$$\Psi_t(y, s) = \lambda_t r_t, \qquad r_t = y - \rho'(\mathsf{prox}_{\lambda_t\rho}(\lambda_t y + s)). \tag{2}$$

Observe that the evolution (49) depends on a sequence of parameters $\{\lambda_t\}$ whose dynamics we describe next.

This description requires introducing an augmented sequence $\{\alpha_t, \sigma_t, \lambda_t\}_{t\geq 0}$. With these two extra parameters (α_t, σ_t), we let (Q_1^t, Q_2^t) be a bivariate normal variable with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(\alpha_t, \sigma_t)$ defined exactly as in Equation M-6.[1] Then starting from an initial pair α_0, σ_0, for $t = 0, 1, \ldots$, we inductively define λ_t as the solution to

$$\mathbb{E}\left[\frac{2\rho'(Q_1^t)}{1 + \lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2^t))}\right] = 1 - \kappa \tag{3}$$

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

[1] Equations, Theorems, Figures, etc. from the main text are referred to using the format 'M-' followed by the corresponding number.

www.pnas.org/cgi/doi/10.1073/pnas.1810420116

and the extra parameters $\alpha_{t+1}, \sigma_{t+1}$ as

$$\begin{aligned}\alpha_{t+1} &= \alpha_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[2\rho'(Q_1^t)Q_1^t\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right]\\ \sigma_{t+1}^2 &= \frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'(Q_1^t)\left(\lambda_t\rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right)^2\right].\end{aligned} \tag{4}$$

To repeat, we run the AMP iterations (49) using the scalar variables $\{\lambda_t\}$ calculated via the *variance map* updates (19)–(20).

In the regime where the MLE exists (see Figure M-6), the variance map updates (19)–(20) converge (as $t \to \infty$) to a unique fixed point $(\alpha_\star, \sigma_\star, \lambda_\star)$. Note that by definition, $(\alpha_\star, \sigma_\star, \lambda_\star)$ is the solution to our system Equation M-5 in three unknowns. From now on, we use $\alpha_0 = \alpha_\star$, $\sigma_0 = \sigma_\star$ so that the sequence $\{\alpha_t, \sigma_t, \lambda_t\}$ is stationary; i. e. for all $t \geq 0$,

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

With this stationary sequence of parameters, imagine now initializing the AMP iterations with a vector $\hat{\boldsymbol{\beta}}^0$ obeying

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

It is not hard to see that if the proposed AMP algorithm converges to a fixed point $\{\boldsymbol{S}_\star, \hat{\boldsymbol{\beta}}_\star\}$, then it is such that $\nabla\ell(\hat{\boldsymbol{\beta}}_\star) = 0$ (see Appendix B); that is, $\hat{\boldsymbol{\beta}}_\star$ obeys the MLE optimality conditions. This provides some intuition as to why the above algorithm would turn out to be useful in this context.

The crucial point is that we can study the properties of the MLE by studying the properties of the AMP iterates with the proviso that they converge. It turns out that the study of the sequence $\{\boldsymbol{S}^t, \hat{\boldsymbol{\beta}}^t\}$ is amenable to a rigorous analysis because several transformations reduce the above algorithm to a generalized AMP algorithm [25], which in turn yields a characterization of the limiting variance of the AMP iterates: for any function ψ as in Theorem M-2, we have as $n \to \infty$,

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j^t - \alpha_\star\beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi(\sigma_\star Z, \beta)\right], \tag{5}$$

where β is drawn from the distribution Π (see Theorem M-2) independently of $Z \sim \mathcal{N}(0,1)$, and $\sigma_\star$ is as above. To summarize, the asymptotic behavior of the AMP iterates $\hat{\boldsymbol{\beta}}^t$ can be characterized through a standard Gaussian variable, the distribution Π and the scalar quantity $\sigma_\star$ determined by the iteration (19)–(20). The description of our AMP algorithm and large sample properties of the iterates are understood only when we understand the behavior of the scalar sequences $\{\alpha_t, \sigma_t, \lambda_t\}_{t\geq 0}$, which are known as the state evolution sequence in the literature; this formalism was introduced in [5, 16–18]. From here on, an analysis similar to that in [33] establishes that in the limit of large iteration counts, the AMP iterates converge to the MLE, that is,

$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j^t - \alpha_\star\beta_j, \beta_j) = \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j - \alpha_\star\beta_j, \beta_j),$$

which is the content of Theorem M-2.

1.2 The distribution of a null coordinate

We sketch the proof of Theorem M-3 in the case where the empirical limiting distribution Π has a point mass at zero. The analysis in the general case, where the number of vanishing coefficients is arbitrary, and in particular, $o(n)$, is very different and may be found in Appendix C.

Now consider Theorem M-2 with $\psi(t,u) = t^2 1(u = 0)$. Strictly speaking, this is a discontinuous function which is not pseudo-Lipschitz. However, we can work with a smooth approximation ψ_a, instead, obtained

using standard techniques for smoothing an indicator function, such that the error $\|\psi - \psi_a\|_2$ is arbitrarily small. For simplicity, we skip the technical details underlying this approximation, and motivate the subsequent arguments using ψ directly. Theorem M-2 then yields

$$\frac{1}{p}\sum_{j\in[p]:\beta_j=0}\hat{\beta}_j^2 \xrightarrow{\text{a.s.}} \sigma_\star^2\mathbb{P}_\Pi\left[\beta=0\right] \quad\Longrightarrow\quad \frac{1}{|j\in[p]:\beta_j=0|}\sum_{j\in[p]:\beta_j=0}\hat{\beta}_j^2 \xrightarrow{\text{a.s.}} \sigma_\star^2. \tag{6}$$

Without loss of generality, assume that the first k coordinates of $\boldsymbol{\beta}$ vanish, and that $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = \left(\mathbf{0}_{[k]}, \boldsymbol{\beta}_{-[k]}\right)$ and similarly for $\hat{\boldsymbol{\beta}}$. From the rotational distributional invariance of the $\boldsymbol{X}_i$'s, it can be shown that for any fixed orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{k\times k}$, $\hat{\boldsymbol{\beta}} \overset{\text{d}}{=} \left(\boldsymbol{U}\hat{\boldsymbol{\beta}}_{[k]}, \hat{\boldsymbol{\beta}}_{-[k]}\right)$. Consequently, $\hat{\boldsymbol{\beta}}_{[k]}/\|\hat{\boldsymbol{\beta}}_{[k]}\|$ is uniformly distributed on the unit sphere $\mathbb{S}^{k-1}$ and is independent of $\|\hat{\boldsymbol{\beta}}_{[k]}\|$. Thus, any null coordinate $\hat{\beta}_j$ has the same distribution as $\|\hat{\boldsymbol{\beta}}_{[k]}\|\, Z_j/\|\boldsymbol{Z}\|$, where $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_k)$, independent of $\hat{\boldsymbol{\beta}}_{[k]}$. From (6) and the weak law of large numbers, we have $\|\hat{\boldsymbol{\beta}}_{[k]}\|/\|\boldsymbol{Z}\| \xrightarrow{\mathbb{P}} \sigma_\star$, leading to $\hat{\beta}_j \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_\star^2)$.

1.3 The distribution of the LRT

Once the distribution of $\hat{\beta}_j$ for a null j is known, the distribution of the LRT is a stone throw away, at least conceptually; that is to say, if we are willing to ignore some technical difficulties and leverage existing work. Indeed, following a reduction similar to that in [33], one can establish that

$$2\Lambda_j = \frac{\kappa}{\lambda_{[-j]}}\hat{\beta}_j^2 + o_P(1), \tag{7}$$

where $\lambda_{[-j]} := \text{Tr}\left[\nabla^2(\ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]}))^{-1}\right]/n$ in which $\ell_{[-j]}$ is the negative log-likelihood with the j-th variable removed and $\hat{\boldsymbol{\beta}}_{[-j]}$ the corresponding MLE. Put $\lambda = \text{Tr}[\nabla^2(\ell(\hat{\boldsymbol{\beta}}))^{-1}]/n$. Then following an approach similar to that in [33, Appendix I], it can be established that $\lambda_{[-j]} = \lambda + o_P(1) \xrightarrow{\mathbb{P}} \lambda_\star$. This gives that $2\Lambda_j$ is a multiple of a χ_1^2 variable with multiplicative factor given by $\kappa\sigma_\star^2/\lambda_\star$.

This rough analysis shows that the distribution of the LLR in high dimensions deviates from a χ_1^2 due to the coupled effects of two high-dimensional phenomena. The first is the inflated variance of the MLE, which is larger than classically predicted. The second comes from the term $\lambda_\star$, which is approximately equal to $\text{Tr}\left(\boldsymbol{H}^{-1}(\hat{\boldsymbol{\beta}})\right)/n$, where $\boldsymbol{H}(\hat{\boldsymbol{\beta}}) = \nabla^2\ell(\hat{\boldsymbol{\beta}})$ is the Hessian of the negative log-likelihood function. In the classical setting, this Hessian converges to a population limit. This is not the case in higher dimensions and the greater spread in the eigenvalues also contributes to the magnitude of the LRT.

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] Ahmad O Aseeri, Yu Zhuang, and Mohammed Saeed Alkatheiri. A machine learning-based security vulnerability study on xor pufs for resource-constraint internet of things. In *2018 IEEE International Congress on Internet of Things (ICIOT)*, pages 49–56. IEEE, 2018.

[3] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Phase transitions, optimal errors and optimality of message-passing in generalized linear models. *arXiv preprint arXiv:1708.03395*, 2017.

[4] Maurice S Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 268–282, 1937.

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

[6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[7] Peter J Bickel and JK Ghosh. A decomposition for the likelihood ratio statistic and the bartlett correction–a bayesian argument. *The Annals of Statistics*, pages 1070–1090, 1990.

[8] George Box. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346, 1949.

[9] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[10] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *https://arxiv.org/pdf/1804.09753.pdf*, 2018.

[11] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[12] Gauss M Cordeiro. Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 404–413, 1983.

[13] Gauss M Cordeiro, Franciso Cribari-Neto, Elisete CQ Aubin, and Silvia LP Ferrari. Bartlett corrections for one-parameter exponential family models. *Journal of Statistical Computation and Simulation*, 53(3-4):211–231, 1995.

[14] 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.

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

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

[17] David L Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing: I. motivation and construction. In *Information Theory (ITW 2010, Cairo), 2010 IEEE Information Theory Workshop on*, pages 1–5. IEEE, 2010.

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

[19] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[20] 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.

[21] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[22] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[23] David W Hosmer and Stanley Lemeshow. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[24] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

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

[26] DN Lawley. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43(3/4):295–303, 1956.

[27] Lawrence H Moulton, Lisa A Weissfeld, and Roy T St Laurent. Bartlett correction factors in logistic regression models. *Computational statistics & data analysis*, 15(1):1–11, 1993.

[28] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory*, pages 2168–2172. IEEE, 2011.

[29] Mark Rudelson, Roman Vershynin, et al. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

[30] C Sabatti, E J Candès, and M Sesia. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 08 2018.

[31] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

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

[33] 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.

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

A Fisher information

We work with the model from Section M-4 and introduce the Fisher information matrix defined as

$$I(\boldsymbol{\beta}) = \mathbb{E}\left[\sum_i \rho''(\boldsymbol{X}_i'\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{X}_i'\right] = n\,\mathbb{E}\left[\rho''(\boldsymbol{X}_i'\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{X}_i'\right].$$

With $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I})$, it is not hard to see that the (k,j)th entry of the matrix $n\rho''(\boldsymbol{X}_i'\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{X}_i'$ is distributed as

$$\rho''(\gamma X_1)X_kX_j, \quad X_1,\ldots,X_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

From here on, a reasonably straightforward calculation gives

$$I(\boldsymbol{\beta}) = \nu(\boldsymbol{I} + \delta\boldsymbol{u}\boldsymbol{u}'), \quad \boldsymbol{u} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|,$$

where

$$\nu = \mathbb{E}[\rho''(\gamma X_1)], \qquad \delta = \frac{\mathbb{E}[\rho''(\gamma X_1)X_1^2] - \mathbb{E}[\rho''(\gamma X_1)]}{\mathbb{E}[\rho''(\gamma X_1)]}.$$

This implies that

$$I^{-1}(\boldsymbol{\beta}) = \nu^{-1}\left(\boldsymbol{I} - \frac{\delta}{1+\delta}\boldsymbol{u}\boldsymbol{u}'\right),$$

which means that the classically predicted variance of $\hat{\beta}_j$ is equal to

$$\nu^{-1}\left(1 - \frac{\delta}{1+\delta}\frac{\beta_j^2}{\|\boldsymbol{\beta}\|^2}\right).$$

When $\beta_j = 0$, the predicted standard deviation is $\nu^{-1/2} = 2.66$ for $\gamma^2 = 5$.

Statistical software packages base their inferences on the approximate Fisher information defined as $\sum_i \rho''(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})\boldsymbol{X}_i\boldsymbol{X}_i'$ (or small corrections thereof). This treats the covariates as fixed and substitutes the value of the unknown regression coefficient sequence $\boldsymbol{\beta}$ with that of the MLE $\hat{\boldsymbol{\beta}}$ (plugin estimate).

B Properties of fixed points of the AMP algorithm

In this section, we elaborate on the connection between the fixed points of (49) and the MLE $\hat{\boldsymbol{\beta}}$. From (49), we immediately see that if $(\hat{\boldsymbol{\beta}}_\star, \boldsymbol{S}_\star)$ is a fixed point, the pair satisfies

$$\begin{aligned}\boldsymbol{X}'\{\boldsymbol{y} - \rho'(\mathsf{prox}_{\lambda_\star\rho}(\lambda_\star\boldsymbol{y} + \boldsymbol{S}_\star))\} &= \boldsymbol{0}\\ (\lambda_\star\boldsymbol{y} + \boldsymbol{S}_\star) - \lambda_\star\rho'(\mathsf{prox}_{\lambda_\star\rho}(\lambda_\star\boldsymbol{y} + \boldsymbol{S}_\star)) &= \boldsymbol{X}\hat{\boldsymbol{\beta}}_\star.\end{aligned}$$

Since by definition of the proximal mapping operator, $z - \lambda\rho'(\mathsf{prox}_{\lambda\rho}(z)) = \mathsf{prox}_{\lambda\rho}(z)$, we have that $\boldsymbol{X}\hat{\boldsymbol{\beta}}_\star = \mathsf{prox}_{\lambda_\star\rho}(\lambda_\star\boldsymbol{y} + \boldsymbol{S}_\star)$ which implies

$$\boldsymbol{X}'\{\boldsymbol{y} - \rho'(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\star)\} = \boldsymbol{0}.$$

Hence, the fixed point $\hat{\boldsymbol{\beta}}_\star$ obeys $\nabla\ell(\hat{\boldsymbol{\beta}}_\star) = \boldsymbol{0}$, the optimality condition for the MLE.

C Refined analysis of the distribution of a null coordinate

The AMP analysis is useful to analyze the bulk behavior of the MLE; i.e. the expected behavior when averaging over all coordinates. It also helps in characterizing the distribution of a null coordinate when the limiting empirical cdf does not have a point mass at zero, as we have seen in Section 1.2. However, the study of the behavior of a single coordinate when there is an arbitrary number of nulls requires a more refined analysis. To this end, the proof uses a leave-one-out approach, as in [20, 21, 33]. The complete rigorous technical details are very involved and this is a reason why we only present approximate or non-rigorous heuristic calculations.

Fix j such that $\beta_j = 0$. Since the corresponding predictor plays no role in the distribution of the response, we expect that including this predictor or not in the regression will not make much difference in the fitted values, that is,

$$\boldsymbol{X}_i'\hat{\boldsymbol{\beta}} \approx \boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}; \tag{8}$$

here, $\boldsymbol{X}_{i,-j}$ is i-th row of the reduced matrix of predictors with the j-th column removed and $\hat{\boldsymbol{\beta}}_{[-j]}$ is the MLE for the reduced model. On the one hand, the approximation (8) suggests Taylor expanding $\rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})$ around the point $\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}$:

$$\rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) \approx \rho'(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}) + \rho''(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}) \left[X_{ij}\hat{\beta}_j + \boldsymbol{X}_{i,-j}'\left(\hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}_{[-j]}\right)\right],$$

where $\hat{\boldsymbol{\beta}}_{-j}$ is the full-model MLE vector, however, without the j-th coordinate. On the other hand, we can subtract the two score equations $\nabla\ell(\hat{\boldsymbol{\beta}}) = 0$ and $\nabla\ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]}) = 0$ ($\ell_{[-j]}$ is the negative log-likelihood for the reduced model), which upon separating the components corresponding to the j-th coordinate from the others, yields

$$\sum_{i=1}^n X_{ij}\left(y_i - \rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})\right) = 0$$

$$\sum_{i=1}^n \boldsymbol{X}_{i,-j}\{\rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) - \rho'(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})\} = \boldsymbol{0}.$$

Plugging in the approximation for $\rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})$ yields two equations in the two unknowns $\hat{\beta}_j$ and $(\hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}_{[-j]})$. After some algebra, solving for $\hat{\beta}_j$ yields

$$\hat{\beta}_j = \frac{\sum_{i=1}^n X_{ij}\left(y_i - \rho'(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})\right)}{\boldsymbol{X}_{\bullet j}'\boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{H}\boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{X}_{\bullet j}} + o_P(1),$$

where $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{X}_{\bullet -j}(\nabla^2\ell_{-j}(\hat{\boldsymbol{\beta}}_{[-j]}))^{-1}\boldsymbol{X}_{\bullet -j}'\boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})$ and $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})$ is an $n\times n$ diagonal matrix with $i-$th entry given by $\rho''(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})$. Above $\boldsymbol{X}_{\bullet j}$ is the j-th column of $\boldsymbol{X}$ and $\boldsymbol{X}_{\bullet -j}$ all the others. It was established in [33] that the denominator above is equal to $\kappa/\lambda_{[-j]} + o_P(1)$, where, we have see in Section 1.3 that

$$\lambda_{[-j]} := \frac{1}{n}\mathrm{Tr}[\nabla^2(\ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]}))^{-1}].$$

Note that since $\beta_j = 0$, $\boldsymbol{y}$ and $\boldsymbol{X}_{\bullet -j}, \hat{\boldsymbol{\beta}}_{[-j]}$ are independent of $\boldsymbol{X}_{\bullet j}$. This gives the approximation

$$\hat{\beta}_j = \frac{\lambda_{[-j]}s_j}{\kappa}Z + o_P(1), \quad s_j^2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \rho'(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})\right)^2, \tag{9}$$

where Z is an independent standard normal. In Section 1.3, we saw that $\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star$. It remains to understand the behavior of s_j. Looking at s_j, the complicated dependence structure between $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{y}, \boldsymbol{X})$ makes this a

potentially hard task. This is why we shall use a leave-one-out argument and seek to express the fitted values $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}$ in terms of $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}$, where $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ is the MLE when both the j-th predictor and the i-th observation are dropped. The independence between $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ will simplify matters. Denote by $\nabla\ell_{[-i],[-j]}(\tilde{\boldsymbol{\beta}}_{[-i],[-j]}) = 0$ the reduced score equation and subtract it from the score equation for $\hat{\boldsymbol{\beta}}$ to obtain

$$\boldsymbol{X}_{i,-j}\left(y_i - \rho'(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]})\right) + \sum_{k\neq i}\boldsymbol{X}_{k,-j}\left(\rho'(\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}) - \rho'(\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-j]})\right) = \boldsymbol{0}.$$

We argue that since the number of observations is large and the observations are i.i.d., dropping one observation is not expected to create much of a difference in the fitted values, hence $\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} \approx \boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-j]}$. A Taylor expansion of $\rho'(\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-j]})$ around the point $\boldsymbol{X}'_{k,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ yields

$$\boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-j]} - \hat{\boldsymbol{\beta}}_{[-i],[-j]}\right) \approx \boldsymbol{X}'_{i,-j}\left[\nabla^2\ell_{[-i],[-j]}(\hat{\boldsymbol{\beta}}_{[-i],[-j]})\right]^{-1}\boldsymbol{X}_{i,-j}\left(y_i - \rho'(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]})\right).$$

Since $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ are independent, by Hanson-Wright inequality [29, Theorem 1.1], the quadratic form above is approximately equal to $\mathrm{Tr}\left[\nabla^2\ell_{[-i],[-j]}(\hat{\boldsymbol{\beta}}_{[-i],[-j]})^{-1}\right]$. Recall that $\lambda_{[-j]} = \mathrm{Tr}[\nabla^2\ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]})^{-1}]$ and again, for a large number of i.i.d. observations, we expect these two quantities to be close. Hence, the fitted values can be approximated as

$$\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} \approx \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + \lambda_{[-j]}\left(y_i - \rho'(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]})\right).$$

Recalling the definition of the proximal mapping operator, $\mathsf{prox}_{\lambda\rho}(z) + \lambda\rho'(\mathsf{prox}_{\lambda\rho}(z)) = z$, note that the above relation gives a useful approximation for the fitted values,

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

Further, by the triangle inequality we can show that

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

It can be shown that the residuals $\{y_i - \rho'(\mathsf{prox}_{\lambda_\star\rho}(\lambda_\star y_i + \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}))\}_{i=1,\dots,n}$ are asymptotically independent among themselves, which implies that averaging over the squared residuals as in (93) should converge in probability to the corresponding expectation, leading to

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma^2), \quad \sigma^2 := \frac{\lambda_\star^2}{\kappa^2}\lim_{n\to\infty}\mathbb{E}\left[y_i - \rho'\left(\mathsf{prox}_{\lambda_\star\rho}\left(\lambda_\star y_i + \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}\right)\right)\right]^2.$$

To complete the analysis, it remains to characterize the asymptotic joint distribution of $\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ and $\boldsymbol{X}'_i\boldsymbol{\beta}$ or, equivalently, $\boldsymbol{X}'_{i,-j}\boldsymbol{\beta}_{-j}$ ($\boldsymbol{\beta}_{-j}$ is the true signal with the j-th coordinate removed) since $\beta_j = 0$. Instead, we find the joint distribution of $(\boldsymbol{X}'_{i,-j}\boldsymbol{\beta}_{-j}, \boldsymbol{X}'_{i,-j}(\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \alpha_\star\boldsymbol{\beta}_{-j}))$ and denote this pair as $(Q_1^\star, Q_2^\star)$. The asymptotic variance of $Q_1^\star$ is given by γ^2, that of $Q_2^\star$ by $\kappa\sigma_\star^2$, while the asymptotic covariance is equal to

$$\lim_{n\to\infty}\frac{\langle\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \alpha_\star\boldsymbol{\beta}_{-j}, \boldsymbol{\beta}_{-j}\rangle}{n} = \kappa\lim_{t\to\infty}\lim_{n\to\infty}\frac{\langle\hat{\boldsymbol{\beta}}^t - \alpha_\star\boldsymbol{\beta}, \boldsymbol{\beta}\rangle}{p} = 0, \tag{10}$$

by an application of (5). Hence, writing $y_i = 1(U_i \le \rho'(\boldsymbol{X}'_{i,-j}\boldsymbol{\beta}_{-j}))$, where the U_i's are i.i.d. $U(0,1)$ independent from anything else, we have

$$\lim_{n\to\infty}\mathrm{Var}(\hat{\beta}_j) = \frac{1}{\kappa^2}\lambda_\star^2\,\mathbb{E}\left[1(U_i \le \rho'(Q_1^\star)) - \rho'(\mathsf{prox}_{\lambda_\star\rho}(\alpha_\star Q_1^\star + Q_2^\star + \lambda_\star 1(U_i \le \rho'(Q_1^\star))))\right]^2.$$

Using this later fact, the above expression can be simplified to

$$\frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'(-Q_1^\star)\left(\lambda_\star\rho'(\mathsf{prox}_{\lambda_\star\rho}(\alpha_\star Q_1^\star + Q_2^\star)\right)^2\right].$$

Note that the joint distribution of $(-Q_1^\star, \alpha_\star Q_1^\star + Q_2^\star)$ is precisely the same as $\mathbf{\Sigma}(\alpha_\star, \sigma_\star)$ as specified by Equation M-6. Hence, recalling Equation M-5, we obtain the asymptotic variance of $\hat{\beta}_j$ to be $\sigma_\star^2$.

D Comparison with existing finite sample approaches

As mentioned in Section M-3, an extensive body of work has been developed to improve the accuracy of maximum-likelihood theory in finite samples. In this section, we will compare the performance of our inference procedure with two of the popular finite sample methods—the Bartlett correction method [4] for the LRT, and Firth's bias reduction method [22] for the MLE.

It has been observed in the literature that the chi-square approximation to the LRT does not yield accurate results in a finite sample setting. One correction to the LRT that is frequently used in finite samples is the Bartlett correction, which dates back to Bartlett [4] and has been extensively studied in several subsequent works (e.g. [7, 8, 12, 13, 26]). In the classical regime where p is fixed and n diverges, this correction can be described as follows [27]: compute the expectation of the LRT up to terms of order $1/n$; that is, obtain a parameter α such that

$$\mathbb{E}[2\Lambda_j] = 1 + \frac{\alpha}{n} + O\left(\frac{1}{n^2}\right).$$

If α_n is an accurate estimator of α, the aforementioned approximation suggests that the corrected LRT statistic

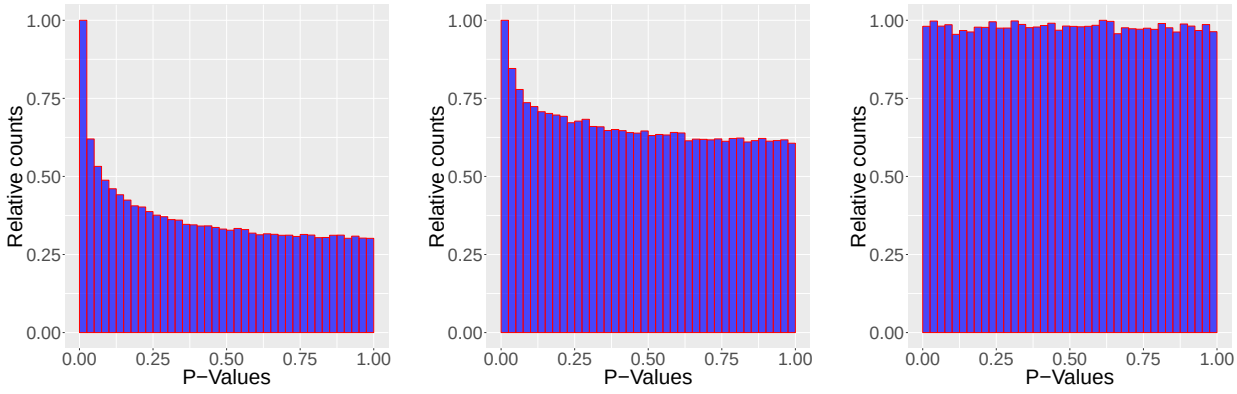
$$\frac{2\Lambda_j}{1 + \frac{\alpha_n}{n}} \tag{11}$$

is closer in expectation to a χ_1^2 distribution than the original LRT statistic for finite samples. For GLMs, Cordeiro [12] derived an explicit formula for the Bartlett correction factor α. Using this formula, we obtained p-values based on the χ_1^2 approximation to the Bartlett corrected LRT statistic in a setting where $n = 2000$ and $p = 400$. Here, half of the regression coefficients vanish while the remaining half have constant magnitude chosen such that $\gamma^2 = \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = 5$. The distribution of the covariates remains the same as in Section M-4. Figure 1(b) shows the Bartlett corrected p-values for a null variable whereas Figures 1(a) and (c) show the corresponding classical p-values and the p-values obtained on using our rescaled χ_1^2 approximation. Clearly, there is less of a spike near zero and, therefore, the Bartlett correction is indeed effective in reducing the non-uniformity. However, the Bartlett corrected p-values are still far from uniform whereas the p-values based on our theory are in perfect agreement with a uniform distribution. Based on these findings, we conclude that such first-order finite sample methods for the LRT (developed under the classical assumption that p is small and n is large) have some use. However, the corrections provided by these methods are not entirely adequate in a high-dimensional setting.

In several practical applications, the MLE has been observed to be biased in finite samples and, historically, this bias has been attributed to a small sample effect. To address this issue, Firth [22] proposed a general approach for reducing the finite sample bias of the MLE. This approach is also a first-order method in the following sense: in the classical setting where p is fixed and n diverges, the asymptotic bias of the MLE may be expressed as

$$\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] =: b(\boldsymbol{\beta}) = \frac{b_1(\boldsymbol{\beta})}{n} + \frac{b_2(\boldsymbol{\beta})}{n^2} + \dots.$$

The work [22] proposed a general approach for bias reduction with the aim of removing the $O(1/n)$ term above. Figure 2 compares Firth's bias reduction procedure with the bias corrected MLE proposed in [32]—that is, $\hat{\boldsymbol{\beta}}/\alpha_\star$—in the setting of Figure M-2 for two different dimensions. Observe that for many of the points, the blue circles (bias corrected MLE) are almost masked by the red circles (Firth coefficient estimates).

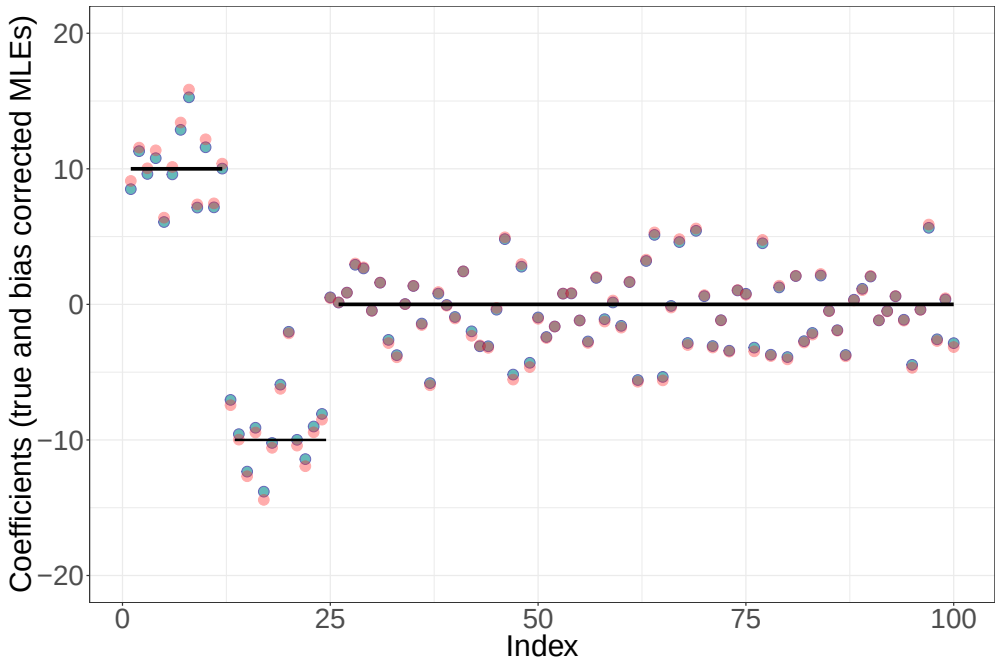


(a) (b) (c)

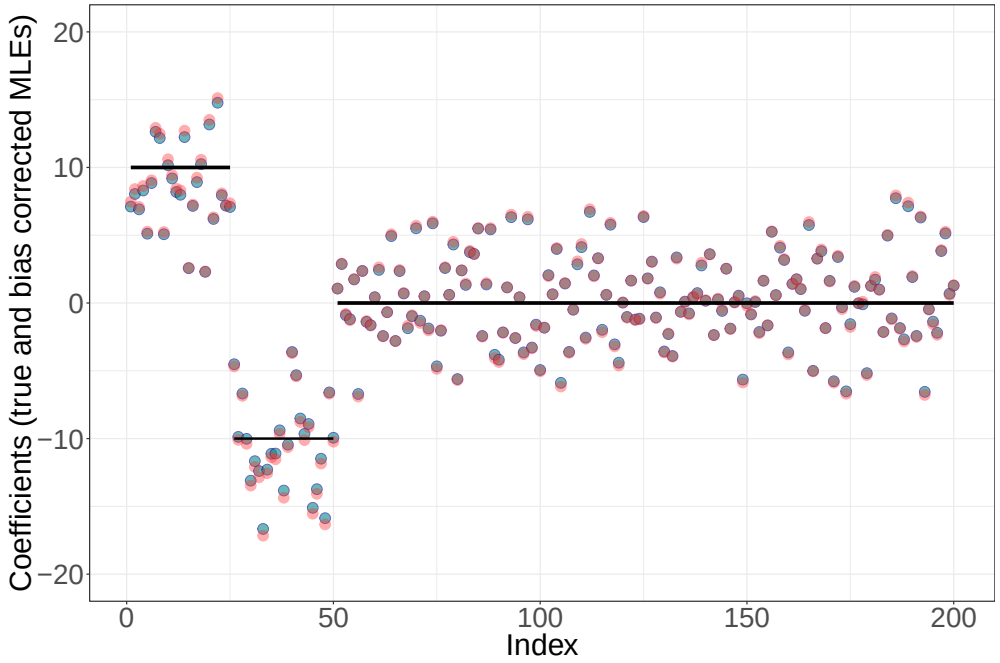
Figure 1: Histograms of p-values for logistic regression under i.i.d. Gaussian design, when $n = 2000$, $p = 400$ and $\kappa = 0.2$. (a) classically computed p-values; (b) Bartlett-corrected p-values; (c) adjusted p-values by comparing the LRT statistic to the rescaled chi square distribution from Theorem M-4.

Thus, Firth's correction works well in this setup, with only some of the blue circles being closer to the true regression coefficients than the corresponding Firth corrected estimates. However, Firth's approach is computationally infeasible for higher dimensions. For instance, the runtime for Firth's approach for Figure 2(b) was approximately 10 minutes and a similar implementation of the method for $n = 2000$ and $p = 400$ required over 2.5 hours. Thus, although Firth's approach has performance comparable to our proposed bias correction method, it appears not scalable to high-dimensional datasets.

In conclusion, the finite sample correction methods are certainly useful for bias correction and improve the validity of p-values. However, either these methods are not scalable to higher dimensions or the improvement they provide over classical theory is not sufficiently adequate in a high-dimensional setting.



(a)



(b)

Figure 2: True signal values β_j (black lines), scaled ML estimates $\hat{\beta}_j/\alpha_\star$ (blue circles) and Firth corrected ML coordinates (red circles). (a) Dimensions: $n = 500, p = 100, \kappa = 0.2$. (b) Dimensions: $n = 1000, p = 200, \kappa = 0.2$.

E Misspecified models

The theory ensuring the validity of our proposed p-values relies on a logistic model for the response given the covariates, which may fail to hold in practice. Hence, it is important to study the robustness of our procedure to model misspecification. We first consider a setting where the response is drawn from a logistic model, but we do not observe all of the relevant predictors. To this end, consider $n = 2000$ i.i.d. draws $\{(y_i, \boldsymbol{X}_i)\}_{1 \le i \le n}$ such that $y_i \sim \mathsf{Bernoulli}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$, where $\sigma(\cdot)$ is the usual sigmoid function, $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p\times p}/n)$ and $p = 300$; the first 100 entries of $\boldsymbol{\beta}$ are i.i.d. draws from $\mathcal{N}(6, 8)$ while the remaining are zero, and hence the signal strength $\gamma^2 := \mathrm{Var}(\boldsymbol{X}_i^\top \boldsymbol{\beta}) = 5$. Suppose we fail to observe one-fourth of the relevant predictors, so that the sample we obtain is of the form $(y_i, \tilde{\boldsymbol{X}}_i)_{1 \le i \le n}$, where $\tilde{\boldsymbol{X}}_i \in \mathbb{R}^{p-20}$ and contains all but the first 20 coordinates of $\boldsymbol{X}_i$. We then fit a logistic regression to this reduced sample and compute the log-likelihood ratio (LLR) statistic corresponding to a specified null coordinate. Using the observed LLR, we calculate p-values based on both the classical Wilks' theorem and the rescaled χ^2 distribution proposed in Theorem M-4, which we refer to as adjusted p-values. This experiment is repeated 5×10^5 times. Note that calculating the rescaling constant requires solving the system of equations in Eq. M-5, which takes as input two parameters—the aspect ratio κ and the signal strength γ. To determine the latter, we apply the ProbeFrontier (Section M-5) method on two independent samples generated from the aforementioned model, and obtain an estimate of the signal strength in each case. We solve the system of equations where γ is set to be the average of these two estimates and $\kappa = (p - 20)/n$. The resulting rescaling constant is then used to correct the p-values in all the 5×10^5 replicates. The results are displayed in Figure 3(a). Notice that the classical p-values visibly deviate from uniformity, whereas the adjusted p-values remain approximately valid.

Next, we investigate the empirical performance of the adjusted p-values in a setting where the link function is misspecified. Consider the scenario $y_i \sim \mathsf{Bernoulli}(\rho(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$, where $\rho(x) = 1 - \exp(-\exp(x))$, the complementary log-log link. We draw $n = 2000$ i.i.d. observations $\{(y_i, \boldsymbol{X}_i)\}_{1 \le i \le n}$ from this model, with $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p\times p}/n)$ and $p = 300$, using the same regression coefficients as above. As before, we then fit a logistic model to the observed sample and calculate the classical and adjusted p-values for a given null coefficient, based on the LLR. This experiment is repeated 5×10^5 times. To calculate the rescaling constant required for the adjusted p-values, we solve the system of equations Eq. M-5 with $\kappa = 0.15$ and γ estimated using the ProbeFrontier approach, as described in the preceding experiment. Figure 3(b) depicts the empirical CDFs of the p-values. The classical p-values deviate from a uniform distribution whereas the adjusted p-values remain in close agreement, despite the fact that the truth differs from a logistic model. Here, our theory clearly provides a better approximation to the true distribution of null p-values than the standard approximation.

F Real data inspired designs

Our theoretical results assume i.i.d. Gaussian entries for the covariate matrix. However, these may apply to a broader class of covariate distributions, as demonstrated in Section M-4.g. In this section, we investigate the performance of our method for non-Gaussian designs generated from real data. Our experiments are based on three datasets:

- A genome-wide association study (GWAS) on Crohn's disease (Source: Wellcome Trust Case Control Consortium, WTCCC, [11])
- A subset of the million song dataset [6], called the YearPredictionMSD data (Source: UCI machine learning repository [19])
- The Physical Unclonable Functions (PUFs) dataset (Source: UCI machine learning repository [19])

The first dataset comprises genetic information on cases and controls for Crohn's disease (CD), with $n = 4913$, $p = 377,749$, and has been used previously to benchmark variable selection methods (see for

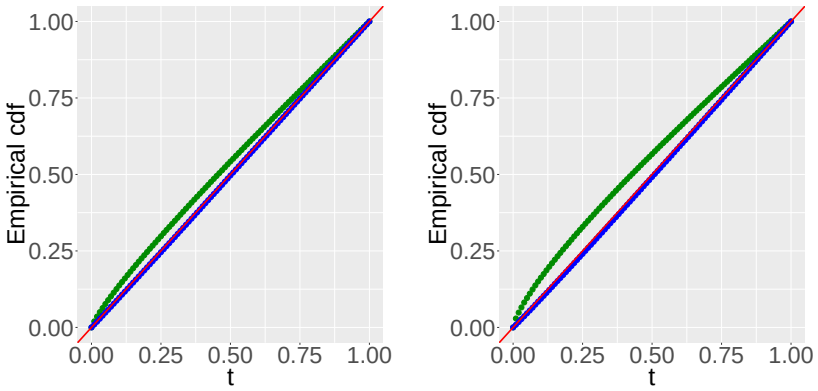


Figure 3: Empirical CDFs of classical p-values (green) and adjusted p-values (blue), based on the LLR statistic from a logistic regression fitted to the sampled data in two settings. (a) The true model is a logistic, but we fail to observe one-fourth of the non-null variables. (b) The true model has a complementary log-log link. The red line represents the diagonal.

instance, [9, 30]). In this case, the design matrix contains information on single nucleotide polymorphisms (SNPs) and all entries lie in {0, 1, 2}. Further, neighboring sites are highly correlated. We will now generate data sets that mimic this CD data in order to study the performance of our proposed p-values. To this end, we first obtain a subsample from the CD data with dimensions $n = 3930, p = 514$, adopting the same processing step as in [30, Section 7.1]. The processing step allows us to pre-specify a threshold c such that the correlations among the features in the created subsample will lie below c; we chose $c = 0.1$. The resulting subsample contains two parts: a 3930×1 response vector that we will call $\tilde{\boldsymbol{y}}$ and a 3930×514 design matrix that we will call $\tilde{\boldsymbol{X}}$. Next, we approximate the distribution of $\tilde{\boldsymbol{X}}$ using a Hidden Markov Model (HMM), as has been widely done in GWAS (see [30] and the references cited therein). In particular, we use the HMM implemented in the fastPHASE [31] fitting algorithm, which is parametrized by three vectors $(\boldsymbol{r}, \boldsymbol{\alpha}, \boldsymbol{\theta})$, and we denote the corresponding distribution by $\mathrm{HMM}(\boldsymbol{r}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. Applying fastPHASE to $\tilde{\boldsymbol{X}}$ provides estimates $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ of the parameters, so that $\mathrm{HMM}(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ approximates the distribution of $\tilde{\boldsymbol{X}}$. Turning to the response, it is harder to mimic the generation of $\tilde{\boldsymbol{y}}$, since the underlying relation between $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}$ is unknown. For the purpose of this simulation, we fit a logistic regression of $\tilde{\boldsymbol{y}}$ onto $\tilde{\boldsymbol{X}}$ to obtain the MLE $\hat{\boldsymbol{\beta}}$. We then set one-fourth of the coordinates of $\hat{\boldsymbol{\beta}}$ to be zero and keep the remaining coefficients intact; denote the resulting vector by $\boldsymbol{\beta}_0$. Putting these two parts together, we generate $n = 2500$ independent observations $(y_i, \boldsymbol{X}_i)_{1 \le i \le n}$ by first taking $\boldsymbol{X}_i \sim \mathrm{HMM}(\hat{\boldsymbol{r}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$, with $\boldsymbol{X}_i \in \mathbb{R}^{514}$, and then sampling $y_i \sim \mathsf{Bernoulli}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0))$. Finally, we center and scale the design matrix so that each column has zero mean and unit norm. With this synthetic sample, we fit a logistic regression model to obtain classical and adjusted p-values based on the LLR for a given null variable, and then we repeat this procedure for a total of 1000 replicates. Note that calculating the adjusted p-values requires an estimate of the signal strength $\gamma^2 = \mathrm{Var}(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0)$, which we obtain via the ProbeFrontier method, applied in the same manner as in Section E. The empirical CDFs of the p-values are shown in Figure 4(a). Observe that the classical p-values are far from uniformly distributed, but the adjusted p-values remain reasonably close to uniform.

The YearPredictionMSD data contains measurements on the timbre of $n = 515,345$ audio files with $p = 90$ features. The features are continuous, and the empirical average absolute correlation between the features is 0.1093. To study the performance of our procedure, we split the rows of $\boldsymbol{X}$ and obtain 858 subsamples $\{\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_{858}\}$, each containing $n = 600$ rows. We center and scale each subsample so that the columns have zero mean and unit norm. Next, we generate regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^{90}$ such that one-third of the coordinates are i.i.d. draws from $\mathcal{N}(6, 8)$ and the remaining are zero. For each subsample $\tilde{\boldsymbol{X}}_j$,

we then generate a synthetic response vector $\tilde{\boldsymbol{y}}_j$ from a logistic model with design matrix $\tilde{\boldsymbol{X}}_j$ and regression vector $\boldsymbol{\beta}$. Lastly, we fit a logistic regression to each $(\tilde{\boldsymbol{y}}_j, \tilde{\boldsymbol{X}}_j)$, and compute the classical and adjusted LLR-based p-values for a specified null coordinate. As with the previous data set, we use the ProbeFrontier method to estimate the signal strength needed to compute the adjusted p-values. The results are displayed in Figure 4(b); again, the classical p-values deviate from uniformity whereas the adjusted p-values show closer agreement.

Lastly, we consider the PUFs data, which is generated from 'XOR Arbiter PUFs' simulations (see [2] for details). The design matrix is tall and skinny with $n \approx 2.4$ million, $p = 64$, and entries taking values ± 1. The average absolute correlation between the features in this case is 0.0005. We split the rows of the design matrix to obtain 1000 subsamples, each containing $n = 427$ rows. For each subsample we conduct experiments in the same fashion as in the preceding example, and the results are presented in Figure 4(c). The classical p-values again deviate from the uniform distribution. For the adjusted p-values, the empirical CDF fluctuates mildly around the diagonal, but remains reasonably close throughout.

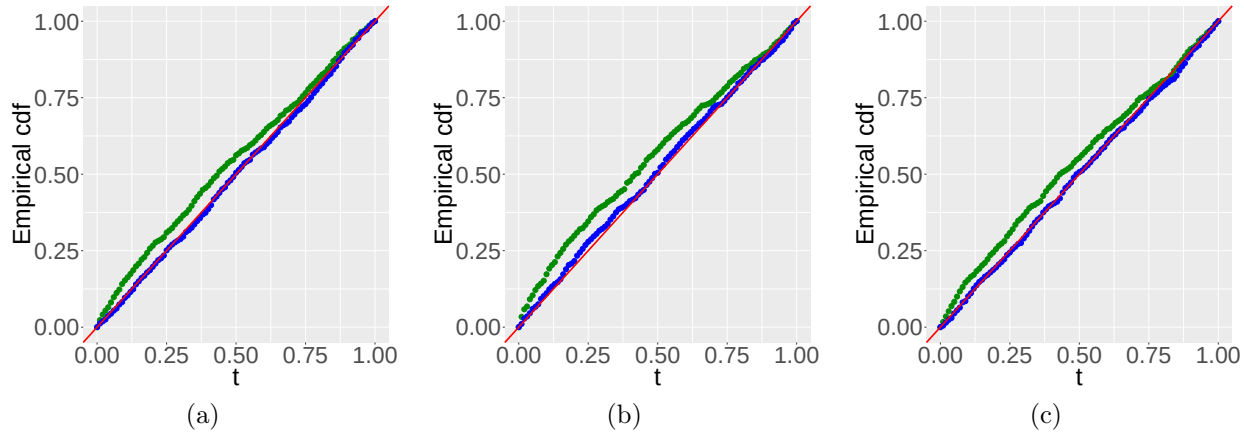


Figure 4: Empirical CDFs of classical p-values (green) and adjusted p-values (blue) for the (a) CD-data-based experiment (b) YearPredictionMSD-data-based experiment (c)PUFs-data-based experiment. The red line is the diagonal.

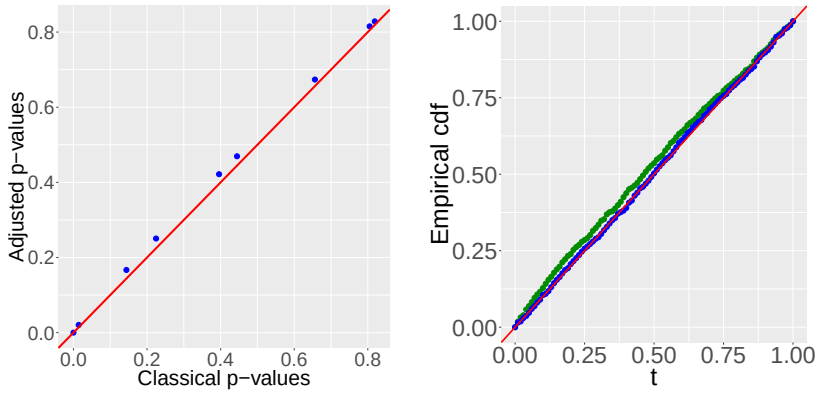
To conclude, we emphasize that the covariates considered in the aforementioned numerical experiments arise from widely different applications; in particular, they do not fall within the framework covered by our theory. Nonetheless, the p-values based on the calculations we presented exhibit a close agreement with the uniform distribution, and better approximate the true distribution of null p-values than their classical counterparts, across all three settings.

G Small aspect ratio and moderate sample size

We have observed that for large n and p, classical null p-values depart from uniformity in a predictable way. However, Figure M-7 suggests that, when the number of features is very small compared to the sample size, the classical and adjusted p-values should be approximately equal. Here, we demonstrate this with a traditional data set. We consider the low birthweight data from [23] with $n = 189, p = 9$, and we center and scale each feature to have zero mean and unit norm. We then calcuate the classical and adjusted p-values for each variable; the adjusted p-values are calculated using the ProbeFrontier method. The results are displayed in Figure 5(a), and as suggested in Figure M-7, the adjusted p-values are very similar to the classical p-values in this setting.

Next, we compare the classical and adjusted p-values in a simulated setting with moderate n and p.

We choose $n = 200$ and $p = 20$ and then generate samples from a logistic model with the same covariate distribution as in Figure 3: half of the regression coefficients are i.i.d. draws from $\mathcal{N}(6, 8)$ and the remaining half are equal to zero, so that the signal strength is $\gamma^2 = 5$. Figure 5(b) displays the empirical CDFs of the classical and adjusted p-values for a null variable across 1000 replicates. Unlike the settings with larger n and p (Section M-1, Appendix E), here the classical p-values are close to a uniform, albeit with minor deviations. The adjusted p-values are in close agreement with a uniform distribution. Thus, although our theory provides asymptotic guarantees only, the corresponding p-values may serve as a rather good approximation to null p-values under moderate sample size and feature dimension.



(a) (b)

Figure 5: (a) Scatterplot of adjusted p-values versus classical p-values, computed based on the LLR statistics for the nine variables in the low birthweight data [23]. (b) Empirical CDFs of classical p-values (green) and adjusted p-values (blue) in a setting with $n = 200, p = 20$. The red line is the diagonal.

H Detailed Proofs

For the convenience of the reader, the subsequent sections are intended to be as self-contained as possible. We first recall our main results, that is, Theorems M-2, M-3, M-4, and subsequently provide detailed proofs. Along the way, we delineate all the mathematical ingredients we build our results upon.

H.1 The results

We recall Theorems M-2, M-3 and M-4 below.[2]

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 [10]).*[3] *For any pseudo-Lipschitz function ψ of order 2, the marginal distributions of the MLE coordinates obey*

$$\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j - \alpha_\star\beta_j, \beta_j\right) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi\left(\sigma_\star Z, \beta\right)\right], \quad Z \sim \mathcal{N}(0,1), \tag{12}$$

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

[2] Notations are the same as in [32].
[3] See [10] for a definition of $g_{\mathrm{MLE}}(\gamma)$.

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

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma_\star^2\right). \tag{13}$$

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,*

$$2\Lambda_j \xrightarrow{\mathrm{d}} \frac{\kappa\, \sigma_\star^2}{\lambda_\star}\, \chi_1^2. \tag{14}$$

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.

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

$$\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\, \mathbb{E}\left[2\rho'\left(Q_1\right)\left(\lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)\right)^2\right] \\ 0 = \mathbb{E}\left[\rho'(Q_1) Q_1 \lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)\right] \\ 1 - \kappa = \mathbb{E}\left[\dfrac{2\rho'\left(Q_1\right)}{1 + \lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(Q_2\right)\right)}\right] \end{cases} \tag{15}$$

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

$$\boldsymbol{\Sigma}\left(\alpha, \sigma\right) = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{bmatrix}. \tag{16}$$

It can be easily checked numerically that in the regime $\gamma < g_{\mathrm{MLE}}(\kappa)$ the system (15) admits a solution. Hence, we omit proving this fact. However, we establish that in the aforementioned regime, if (15) admits a solution then the solution must be unique.[4] 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

$$\mathsf{prox}_{\lambda\rho}\left(z\right) = \arg\min_{t \in \mathbb{R}} \left\{\lambda\rho\left(t\right) + \frac{1}{2}\left(t - z\right)^2\right\}. \tag{17}$$

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

$$\lambda\rho'\left(\mathsf{prox}_{\lambda\rho}\left(z\right)\right) + \mathsf{prox}_{\lambda\rho}\left(z\right) = z. \tag{18}$$

H.2 Road map to the proofs

This section presents the key steps in the proofs of each theorem. Detailed proofs are provided in Appendices H.4–H.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 Section M-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 [25]. 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.

[4] See Remark 2 for a detailed explanation of this fact.

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. This step is based on the leave-one-out techniques introduced in [20, 21] for studying such high-dimensional estimators.

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

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 (95) 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 [33, Appendix I] and [20].

H.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 [32], and expressing it in an equivalent form.

H.3.1 A Useful Recursion

In [32], 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 (16). Starting from an initial pair α_0, σ_0, for $t = 0, 1, \ldots$, inductively define λ_t as the solution to

$$\mathbb{E}\left[\frac{2\rho'\left(Q_1^t\right)}{1 + \lambda \rho''\left(\mathsf{prox}_{\lambda\rho}\left(Q_2^t\right)\right)}\right] = 1 - \kappa, \tag{19}$$

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

$$\begin{aligned} \alpha_{t+1} &= \alpha_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[2\rho'\left(Q_1^t\right) Q_1^t \lambda_t \rho'\left(\mathsf{prox}_{\lambda_t \rho}\left(Q_2^t\right)\right)\right], \\ \sigma_{t+1}^2 &= \frac{1}{\kappa^2}\,\mathbb{E}\left[2\rho'\left(Q_1^t\right)\left(\lambda_t \rho'\left(\mathsf{prox}_{\lambda_t \rho}\left(Q_2^t\right)\right)\right)^2\right]. \end{aligned} \tag{20}$$

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

$$h(x, y) = \mathbf{1}_{y \leq \rho'(x)}, \text{ where } \rho'(x) = \frac{e^x}{1 + e^x}. \tag{21}$$

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

$$\mathbb{E}_{W,\tilde{Q}_1^t,\tilde{Q}_2^t}\left[\frac{1}{1+\lambda\rho''\left(\mathsf{prox}_{\lambda\rho}\left(\lambda h\left(\tilde{Q}_1^t, W\right)+\tilde{Q}_2^t\right)\right)}\right] = 1-\kappa. \tag{22}$$

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

$$\begin{aligned}\tilde{\alpha}_{t+1} &= \tilde{\alpha}_t + \frac{1}{\kappa\gamma^2}\,\mathbb{E}\left[\tilde{Q}_1^t \tilde{\Psi}_t\left(\tilde{Q}_1^t, W, \tilde{Q}_2^t\right)\right], \\ \tilde{\sigma}_{t+1}^2 &= \frac{1}{\kappa^2}\,\mathbb{E}\left[\tilde{\Psi}_t^2\left(\tilde{Q}_1, W, \tilde{Q}_2^t\right)\right],\end{aligned} \tag{23}$$

where

$$\tilde{\Psi}_t\left(q_1, w, q_2\right) = \tilde{\lambda}_t\left[h\left(q_1, w\right) - \rho'\left(\mathsf{prox}_{\tilde{\lambda}_t\rho}\left(\tilde{\lambda}_t h\left(q_1, w\right)+q_2\right)\right)\right]. \tag{24}$$

We propose simplifying the right-hand side (RHS) of the first equation in (23) 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

$$\begin{aligned}&\mathbb{E}_{W,\tilde{Q}_1^t,\tilde{Q}_2^t}\left[\tilde{Q}_1^t\tilde{\Psi}_t(\tilde{Q}_1^t, W, \tilde{Q}_2^t)\right] \\ &= \mathbb{E}\left[\rho'(\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))\right] - \mathbb{E}\left[(1-\rho'(\tilde{Q}_1^t))\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] \\ &= -\mathbb{E}\left[\rho'(-\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] - \mathbb{E}\left[(1-\rho'(\tilde{Q}_1^t))\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right] \\ &= -2\,\mathbb{E}\left[\rho'(-\tilde{Q}_1^t)\tilde{Q}_1^t\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right].\end{aligned} \tag{25}$$

Performing similar calculations it can be shown that

$$\begin{aligned}\mathbb{E}\left[\tilde{\Psi}_{\tilde{\lambda}_t}^2(\tilde{Q}_1^t, W, \tilde{Q}_2^t)\right] &= \mathbb{E}\left[\rho'(\tilde{Q}_1^t)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))\right\}^2\right] + \mathbb{E}\left[(1-\rho'\left(\tilde{Q}_1^t\right)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right\}^2\right] \\ &= \mathbb{E}\left[2\rho'(-\tilde{Q}_1^t)\left\{\tilde{\lambda}_t\rho'(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))\right\}^2\right].\end{aligned} \tag{26}$$

Similarly,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{\lambda}_t h(\tilde{Q}_1^t, W)+\tilde{Q}_2^t))}\right] \\ &= \mathbb{E}\left[\frac{\rho'(\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(-\mathsf{prox}_{\tilde{\lambda}_t\rho}(-\tilde{Q}_2^t))}\right] + \mathbb{E}\left[\frac{1-\rho'(\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))}\right]\end{aligned} \tag{27}$$

$$= \mathbb{E}\left[\frac{2\rho'(-\tilde{Q}_1^t)}{1+\tilde{\lambda}_t\rho''(\mathsf{prox}_{\tilde{\lambda}_t\rho}(\tilde{Q}_2^t))}\right]. \tag{28}$$

Placing together (25), (26) and (27), we have effectively established that, if $\alpha_0 = \tilde{\alpha}_0, \sigma_0 = \tilde{\sigma}_0$, then for all $t \geq 0$,

$$\alpha_t \equiv \tilde{\alpha}_t, \quad \sigma_t \equiv \tilde{\sigma}_t, \quad \lambda_t \equiv \tilde{\lambda}_t. \tag{29}$$

Remark 1. We remark that (22) and the second equation in (23) can be related to the equation system derived in [21, Equations S1,S2] in the context of M-estimation for linear models.

H.3.2 When is the MLE bounded?

It was established in [10] that if $\gamma < g_{\mathrm{MLE}}(\kappa)$ (resp. $\gamma > g_{\mathrm{MLE}}(\kappa)$), the MLE exists asymptotically with probability 1 (resp. 0). [10] 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 $\gamma < g_{\mathrm{MLE}}(\kappa)$. This is the content of the theorem below.

Theorem 4. *If $\gamma < g_{\mathrm{MLE}}(\kappa)$, 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*

$$\mathbb{P}\left(\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} \leq C_1\right) \geq 1 - C_2 n^{-\delta}, \tag{30}$$

where $\delta > 1$.

Proof: By arguments similar to that in Section 5.2.2 from [32], it can be deduced that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} \leq \frac{4\log 2}{\varepsilon^2}\right) \geq \mathbb{P}\left(\{\boldsymbol{y} \circ (\boldsymbol{X}\boldsymbol{b}) \,|\boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} = \{\boldsymbol{0}\}\right), \tag{31}$$

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

$$\mathcal{A} := \left\{\boldsymbol{u} \in \mathbb{R}^n | \sum_{j=1}^{n} \max\{-u_j, 0\} \leq \varepsilon^2 \sqrt{n}\|\boldsymbol{u}\|\right\}. \tag{32}$$

Thus, it suffices to establish that the complement of the RHS of (31) 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,

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

where $V \stackrel{\mathrm{d}}{=} y_i X_{i1}$, 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,

$$\mathbb{P}\left(\{\boldsymbol{y} \circ (\boldsymbol{X}\boldsymbol{b}) \,|\boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right) = \mathbb{P}\left(\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right). \tag{33}$$

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

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

we can decompose the required probability as

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

The following lemma ensures that $\mathbb{P}\left(\mathcal{G}\right)$ 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,*[5] *for all $\varepsilon \leq \varepsilon_1$,*

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

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

$$\mathbb{P}\left(\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}\right) \leq \mathbb{P}\left(\mathcal{G}^c \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right) + C_0 \exp(-c_0 n). \tag{34}$$

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

$$\mathcal{F}_{\mathcal{V}} := \left\{\max_i V_i^2 \leq \zeta \log n\right\}, \tag{35}$$

that satisfies

$$\mathbb{P}\left[\mathcal{F}_{\boldsymbol{V}}\right] \geq 1 - C_1 n^{-\delta}, \tag{36}$$

where $\delta > 1$. Thus,

$$\mathbb{P}\left(\mathcal{G}^c \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right) \leq \mathbb{P}\left(\mathcal{G}^c \cap \mathcal{F}_{\boldsymbol{V}} \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right) + C_1 n^{-\delta}. \tag{37}$$

Regarding the cone $\mathcal{A}$, [33] 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,

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c \cap \mathcal{F}_{\boldsymbol{V}} \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right) &\leq \mathbb{P}\left(\mathcal{G}^c \cap \mathcal{F}_{\boldsymbol{V}} \cap \{\mathcal{B}_i | 1 \leq i \leq N\} \text{ does not form a cover of } \mathcal{A}\right) \\ &\quad + \sum_{i=1}^{N} \mathbb{P}\left(\mathcal{G}^c \cap \mathcal{F}_{\boldsymbol{V}} \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{B}_i \neq \{\boldsymbol{0}\}]\right).\end{aligned} \tag{38}$$

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

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

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, (38) reduces to

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}^c \cap \mathcal{F}_{\boldsymbol{V}} \cap [\{\boldsymbol{T}\boldsymbol{b} | \boldsymbol{b} \in \mathbb{R}^p\} \cap \mathcal{A} \neq \{\boldsymbol{0}\}]\right) &\leq \mathbb{P}\left(\{\mathcal{B}_i | 1 \leq i \leq N\} \text{ does not form a cover of } \mathcal{A}\right) \\ &\quad + \sum_{i=1}^{N} \mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}} \cap [\mathcal{L} \cap \mathcal{C}_i(\text{span}(\boldsymbol{V})) \neq \{\boldsymbol{0}\}]\right) \\ &\leq \sum_{i=1}^{N} \mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}} \cap [\mathcal{L} \cap \mathcal{C}_i(\text{span}(\boldsymbol{V})) \neq \{\boldsymbol{0}\}]\right) + \exp\left(-C_1\epsilon^2 p\right).\end{aligned} \tag{39}$$

[5] Recall that the definition of $\mathcal{A}$ in (32) involved a choice of ε.

To analyze the above, we will resort to ingredients from the literature on convex geometry. Using the approximate kinematic formula [1, Theorem I], [33] argued that, for any closed convex cone $\mathcal{C}$ for which the statistical dimension[6] obeys $\delta(\mathcal{C}) < n - \delta(\mathcal{L}) = n - p + 1$,

$$\mathbb{P}\left(\mathcal{L} \cap \mathcal{C} \neq \{\mathbf{0}\}\right) \leq 4 \exp\left\{-\frac{\left(n - p - \delta(\mathcal{C})\right)^2}{8n}\right\}. \tag{40}$$

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

$$\mathbb{P}\left(\mathcal{F}_{\boldsymbol{V}} \cap \mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\}\right) \leq \mathbb{E}_{\boldsymbol{V}}\left[\mathbf{1}_{\mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}} \mathbb{P}\left(\mathcal{L} \cap \mathcal{C}_i(\mathrm{span}(\boldsymbol{V})) \neq \{\mathbf{0}\} | \boldsymbol{V}\right)\right] + \mathbb{P}\left(\mathcal{G}_{\boldsymbol{V}}^c\right). \tag{41}$$

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,*

$$\mathbb{P}\left(\mathcal{G}_{\boldsymbol{V}}\right) \geq 1 - C_1 \exp\left(-c_1 n\right), \tag{42}$$

   for positive universal constants C_1, c_1.

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

$$\delta\left(\mathcal{C}_i\left(\mathrm{span}(\boldsymbol{v})\right)\right) \leq n\left(1 - g_{\mathrm{MLE}}^{-1}(\gamma) + \nu + o(1)\right). \tag{43}$$

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

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

Applying (40) 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 (41), 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 (33), (34),(37) and (39) completes the proof. ■

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

H.3.3 Ingredients from G-AMP

As discussed in Appendix H.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 [25] that will be central to our analysis in Appendix H.4. For convenience, we adhere to the same notations as in [25].

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.[7] Define $\boldsymbol{A} = \boldsymbol{G} + \boldsymbol{G}'$, where $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ has i.i.d. entries from $\mathcal{N}(0, 1/2N)$.

[6] 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 $\mathcal{C}$.

[7] 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 $\mathbb{R}^q$.

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:

$$\boldsymbol{x}_{\bullet i}^{t+1} = \sum_{j=1}^{N} A_{ij} f^j\left(\boldsymbol{x}_{\bullet j}^t; t\right) - \frac{1}{N}\left(\sum_{j=1}^{N} \frac{\partial f^j}{\partial \boldsymbol{x}}\left(\boldsymbol{x}_{\bullet j}^t; t\right)\right) f^i\left(\boldsymbol{x}_{\bullet i}^{t-1}; t-1\right), \tag{44}$$

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 [25] 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 $i \in C_a^N$, $\boldsymbol{x} \in \mathbb{R}^q$,
$$f^i(\boldsymbol{x}; t) = g\left(\boldsymbol{x}, \boldsymbol{y}_{\bullet i}, a, t\right). \tag{45}$$
This requirement basically states that the functions $f^j(\cdot; t)$ in (44) can only be of the aforementioned form.

4. For each $a \in [q']$, define $\widehat{\boldsymbol{\Sigma}}$ to be the limit (in probability),
$$\lim_{N\to\infty} \frac{1}{|C_a^N|} \sum_{i \in C_a^N} g\left(\boldsymbol{x}_{\bullet i}^0, \boldsymbol{y}_{\bullet i}, a, 0\right) g\left(\boldsymbol{x}_{\bullet i}^0, \boldsymbol{y}_{\bullet i}, a, 0\right)^\top =: \widehat{\boldsymbol{\Sigma}}_a^{(0)}. \tag{46}$$
For each $t \geq 1$, define a positive semi-definite matrix $\boldsymbol{\Sigma}^{(t)} \in \mathbb{R}^{q \times q}$, obtained, by letting,
$$\boldsymbol{\Sigma}^{(t)} = \sum_{b=1}^{q'} c_b \widehat{\boldsymbol{\Sigma}}_b^{(t-1)}, \qquad \widehat{\boldsymbol{\Sigma}}_a^{(t)} = \mathbb{E}\left[g(\boldsymbol{Z}_a^t, \boldsymbol{Y}_a, a, t) g(\boldsymbol{Z}_a^t, \boldsymbol{Y}_a, a, t)^\top\right], \tag{47}$$
where $\boldsymbol{Y}_a \sim P_a$, $\boldsymbol{Z}_a^t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)}\right)$ and $\boldsymbol{Y}_a \perp\!\!\!\perp \boldsymbol{Z}_a^t$.

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

Theorem 5 ([25] 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,*
$$\lim_{N\to\infty} \frac{1}{|C_a^N|} \sum_{j \in C_a^N} \psi\left(\boldsymbol{x}_{\bullet j}^t, \boldsymbol{y}_{\bullet j}\right) = \mathbb{E}\left\{\psi\left(\boldsymbol{Z}_a^t, \boldsymbol{Y}_a\right)\right\}, \tag{48}$$
where $\boldsymbol{Z}_a^t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)}\right)$ is independent of $\boldsymbol{Y}_a \sim P_a$.

H.4 Asymptotic average behavior of MLE

To begin with, we recall the iterative algorithm that [32] 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:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^t &= \hat{\boldsymbol{\beta}}^{t-1} + \kappa^{-1}\boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right) \\ \boldsymbol{S}^t &= \boldsymbol{X}\hat{\boldsymbol{\beta}}^t - \Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{S}^{t-1}\right) \end{aligned} \tag{49}$$

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

$$\Psi_t\left(y, s\right) = \lambda_t r_t, \qquad r_t = y - \rho'\left(\mathsf{prox}_{\lambda_t \rho}\left(\lambda_t y + s\right)\right), \tag{50}$$

and λ_t is described via the recursions (19)–(20). However, from (29), we know $\lambda_t \equiv \tilde{\lambda}_t$, where $\tilde{\lambda}_t$ is described via the update equations (22)–(23), when

$$\alpha_0 = \tilde{\alpha}_0, \quad \sigma_0 = \tilde{\sigma}_0. \tag{51}$$

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 Appendix H.4.1. Note that the response variables can be expressed as

$$y_i = h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right), \tag{52}$$

where $h(x, y)$ is specified via (21) 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 (24), we observe that

$$\Psi_t\left(y_i, S_i^t\right) \equiv \tilde{\Psi}_t\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i, S_i^t\right). \tag{53}$$

H.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 (19)–(20) proves to be useful, as is formalized in the theorem below.

Theorem 6. *Suppose the initial conditions for the AMP iterative scheme (49), and the variance map updates (19)–(20) satisfy*

$$\alpha_0 = \frac{1}{\gamma^2}\lim_{n\to\infty}\frac{\langle\hat{\boldsymbol{\beta}}^0, \boldsymbol{\beta}\rangle}{n}, \qquad \sigma_0^2 = \lim_{n,p\to\infty}\frac{1}{p}\|\hat{\boldsymbol{\beta}}^0 - \alpha_0\boldsymbol{\beta}\|^2. \tag{54}$$

For any pseudo-Lipshcitz function ψ of order 2,

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\psi\left(\hat{\beta}_j^t - \alpha_t\beta_j, \beta_j\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\sigma_t Z, \beta\right)\right] \\ \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\begin{bmatrix}\boldsymbol{X}_i'\boldsymbol{\beta} \\ S_i^t\end{bmatrix}, \begin{bmatrix}w_i \\ 0\end{bmatrix}\right) &\overset{a.s.}{=} \mathbb{E}\left[\psi\left(\begin{bmatrix}Q_1^t \\ Q_2^t\end{bmatrix}, \begin{bmatrix}W \\ 0\end{bmatrix}\right)\right], \end{aligned} \tag{55}$$

where $\beta \sim \Pi, W \sim U(0, 1)$ independent of each other[8] *and independent of*

$$\left(Q_1^t, Q_2^t\right) \sim \mathcal{N}\left(0, \begin{bmatrix}\gamma^2 & \alpha_t\gamma^2 \\ \alpha_t\gamma^2 & \kappa\sigma_t^2 + \alpha_t^2\gamma^2\end{bmatrix}\right). \tag{56}$$

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

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:

$$\begin{aligned}\boldsymbol{\nu}^t &= q_{t-1}\left(\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right) - a_t\boldsymbol{\beta} + \kappa^{-1}\boldsymbol{X}'\Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right)\\ \boldsymbol{R}^t &= \boldsymbol{X}\left(\boldsymbol{\nu}^t + \alpha_t\boldsymbol{\beta}\right) - \Psi_{t-1}\left(\boldsymbol{y}, \boldsymbol{R}^{t-1}\right),\end{aligned} \tag{57}$$

where

$$\begin{aligned}q_t &= -\frac{1}{\kappa n}\sum_{i=1}^{n}\Psi_t'\left(y_i, R_i^t\right)\\ a_0 = \alpha_0, \qquad a_t &= \frac{1}{\kappa n}\sum_{i=1}^{n}\frac{\partial}{\partial a}\Psi_{t-1}\left(h(a, W_i), R_i^{t-1}\right)\Big|_{a=\boldsymbol{X}_i'\boldsymbol{\beta}} \quad \text{for} \quad t \geq 1;\end{aligned} \tag{58}$$

Ψ_t' is the derivative w.r.t the second coordinate of Ψ_t. The difference between this recursion and that in (49) 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 (57), 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 (57) 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 $t \geq 1$,*

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

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

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

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 (55). A similar argument holds for the other relation. ■

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

H.4.2 Proof of Lemma 3

Our first goal is to reduce the recursion (57) to the G-AMP form (44). Thereafter, computing the covariances $\boldsymbol{\Sigma}^t$ from (47) 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 $t = 2k + 1$ $(k \geq 0)$, for each $i = 1, \dots n$, define

$$\boldsymbol{x}_{\bullet i}^t := \left[Z_i, 0, R_i^0, 0, R_i^1, \dots, R_i^{\frac{t-1}{2}}, 0, 0, \dots\right]'. \tag{59}$$

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

$$\boldsymbol{x}_{\bullet i}^t = \left[0, \nu_{i-n}^1, 0, \nu_{i-n}^2, 0, \nu_i^2, \dots, \nu_{i-n}^{\frac{t}{2}}, 0, 0, \dots\right]'. \tag{60}$$

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

$$\begin{bmatrix} Y_{1\bullet} \\ Y_{2\bullet} \end{bmatrix} = \begin{bmatrix} W_1, & W_2, & \dots, & W_n, & \beta_1, & \beta_2, & \dots, & \beta_p \\ 0 & & \dots & 0 & \nu_1^0, & \nu_2^0, & \dots, & \nu_p^0 \end{bmatrix} \tag{61}$$

and the rest of the entries are all 0. Note that, the functions f in (45) are allowed to be functions of the elements of $\boldsymbol{Y}$. For the odd iterates $t = 2k + 1$ $(k \geq 0)$, 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

$$f^i(\boldsymbol{x}; 2k+1) = \left[0, \frac{h}{\kappa}\Psi_0(h(x_1, Y_{1i}), x_3), 0, \frac{h}{\kappa}\Psi_1(h(x_1, Y_{1i}), x_5), \dots, \frac{h}{\kappa}\Psi_{\frac{t-1}{2}}(h(x_1, Y_{1i}), x_{t+2}), 0, 0, \dots\right]'. \tag{62}$$

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

$$f^i(\boldsymbol{x}; 2k) = \left[hY_{1i}, 0, h(Y_{2i} + \alpha_0 Y_{1i}), 0, h(x_2 + \alpha_1 Y_{1i}), 0, h(x_4 + \alpha_2 Y_{1i}), 0, \dots, h(x_t + \alpha_{t/2} Y_{1i}), 0, 0, \dots\right]'. \tag{63}$$

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 $1 \leq i \leq n$ and $n+1 \leq j \leq n+p$, 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 $i = 1, \dots n$, $\boldsymbol{x}_{\bullet i}^t$ defined via (59)–(60) satisfies the recursion (44), with the collection of functions $f^i(,;t)$ given by (62)–(63) 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 $\tilde{\boldsymbol{x}}^0 = \mathbf{0}$. 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 (44) with the choice of functions f^i specified in (62) and (63).*

Thus, we have reduced the recursion in (57) to the G-AMP form (44). 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 (47), with g defined via (62) and (63). 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_i^0, with all other entries 0. From the definitions (46) and (47), it is easy to check that

$$\begin{bmatrix} \Sigma^{(1)}_{(1,1)} & \Sigma^{(1)}_{(1,3)} \\ \Sigma^{(1)}_{(3,1)} & \Sigma^{(1)}_{(3,3)} \end{bmatrix} = \begin{bmatrix} \lim_{n\to\infty} \frac{\|\boldsymbol{\beta}\|^2}{n} & \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0\rangle}{n} \\ \lim_{n\to\infty} \frac{\langle \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^0\rangle}{n} & \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0\|^2}{n} \end{bmatrix}, \tag{64}$$

which is consistent with the asymptotic covariance structure we expect to see in this case, since $Z_i = \boldsymbol{X}_i'\boldsymbol{\beta}, R_i^0 = S_i^0 = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}^0$. Computing $\boldsymbol{\Sigma}^{(2)}$, using the formula (47) and applying Theorem 5 yields,

$$\frac{1}{p}\sum_{j=1}^p \psi\left(\nu_j^1, \beta_j\right) \to \mathbb{E}\left[\psi\left(\tau_1 Z, \beta\right)\right], \quad \text{where} \quad \tau_1^2 = \frac{1}{\kappa^2}\mathbb{E}\left[\Psi_0^2\left(h\left(Q_1^0, U\right), Q_2^0\right)\right], \tag{65}$$

and (Q_1^0, Q_2^0) is multivariate normal with mean $\mathbf{0}$ and covariance matrix specified in (64).

Note that, for $\boldsymbol{\Sigma}^{(3)}$, the first 3×3 sub-block would be the same as in (64). 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,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \psi\left(\begin{bmatrix}\boldsymbol{X}_i'\boldsymbol{\beta}\\ R_i^0\\ R_i^1\end{bmatrix}, \begin{bmatrix} w_i\\ 0\\ 0\end{bmatrix}\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(3)}, \begin{bmatrix} W\\ 0\\ 0\end{bmatrix}\right)\right], \tag{66}$$

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

$$\frac{1}{p}\sum_{j=1}^p \psi\left(\begin{bmatrix}\nu_j^1\\ \nu_j^2\end{bmatrix}, \begin{bmatrix}\beta_j\\ 0\end{bmatrix}\right) \to \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(4)}, \begin{bmatrix}\beta\\ 0\end{bmatrix}\right)\right], \tag{67}$$

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

$$\tau_2^2 = \frac{1}{\kappa^2}\,\mathbb{E}\left[\Psi_1^2\left(h\left(Z_1^{(3)}, U\right), Z_3^{(3)}\right)\right], \qquad \rho_{12} = \frac{1}{\kappa^2}\,\mathbb{E}\left[\Psi_0\left(h\left(Z_1^{(3)}, U\right), Z_2^{(3)}\right)\Psi_1\left(h\left(Z_1^{(3)}, U\right), Z_3^{(3)}\right)\right]. \tag{68}$$

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

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(\begin{bmatrix} \boldsymbol{X}_i'\boldsymbol{\beta} \\ R_i^0 \\ R_i^1 \\ R_i^2 \end{bmatrix}, \begin{bmatrix} w_i \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) \overset{a.s.}{=} \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(5)}, \begin{bmatrix} W \\ 0 \\ 0 \\ 0 \end{bmatrix}\right)\right], \tag{69}$$

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

$$\frac{1}{p}\sum_{j=1}^{p} \psi\left(\begin{bmatrix} \nu_j^1 \\ \nu_j^2 \\ \nu_j^3 \end{bmatrix}, \begin{bmatrix} \beta_j \\ 0 \\ 0 \end{bmatrix}\right) \to \mathbb{E}\left[\psi\left(\boldsymbol{Z}^{(6)}, \begin{bmatrix} \beta \\ 0 \end{bmatrix}\right)\right], \tag{70}$$

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

$$\tau_3^2 = \frac{1}{\kappa^2}\,\mathbb{E}\left[\Psi_2^2\left(h\left(Z_1^{(5)}, U\right), Z_4^{(5)}\right)\right], \qquad \rho_{lm} = \frac{1}{\kappa^2}\,\mathbb{E}\left[\Psi_{l-1}\left(h\left(Z_1^{(5)}, U\right), Z_{l+1}^{(5)}\right)\Psi_{m-1}\left(h\left(Z_1^{(5)}, U\right), Z_{m+1}^{(5)}\right)\right]. \tag{71}$$

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

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

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 $\left(Z_1^{(2t-1)}, Z_{t+1}^{(2t-1)}\right) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\left(-\alpha_{t-1}, \tau_{t-1}\right)\right)$ and $\mathbf{\Sigma}(\alpha, \sigma)$ as in (16). The final step is to relate the scalar sequence $\{\tau_t\}$, first to the sequence $\{\tilde{\sigma}_t\}$ defined in (23), 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

$$\alpha_0 = \frac{1}{\gamma^2} \lim_{n\to\infty} \frac{\langle \hat{\boldsymbol{\beta}}^0, \boldsymbol{\beta} \rangle}{n}, \qquad \sigma_0^2 = \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^0 - \alpha_0 \boldsymbol{\beta}\|^2}{p}. \tag{72}$$

It is easy to check that, with this choice, the covariance in (64) is precisely $\mathbf{\Sigma}(-\alpha_0, \sigma_0) = \mathbf{\Sigma}\left(-\tilde{\alpha}_0, \tilde{\sigma}_0\right)$, since $(\tilde{\alpha}_0, \tilde{\sigma}_0)$ was initialized to (α_0, σ_0) (recall (51)).

The equivalence between the functions Ψ_t and $\tilde{\Psi}_t$ from (53), and the definition of $\tilde{\sigma}_t$ from (23) 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 (29), then completes the proof.

H.4.3 Proof of Lemma 4

The proof partly follows along lines similar to [14, 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 (57) and (49), 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 [14, Proposition 6.3], we obtain

$$\frac{\partial \Psi_t(y,s)}{\partial s} = \frac{-\lambda_t\, \rho''(x)|_{x=\mathsf{prox}(\lambda_t y + s)}}{1 + \lambda_t\, \rho''(x)|_{\mathsf{prox}(\lambda_t y + s)}}. \tag{73}$$

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

$$\left\|\boldsymbol{R}^t - \boldsymbol{S}^t\right\| \leq \|\boldsymbol{X}\| \left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| + \left\|\boldsymbol{R}^{t-1} - \boldsymbol{S}^{t-1}\right\|. \tag{74}$$

Similarly, comparing (57) and (49) again, we obtain

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

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

$$\left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| \leq \left\|\boldsymbol{\nu}^{t-1} - \boldsymbol{\theta}^{t-1}\right\| + |q_{t-1} - 1| \left\|\boldsymbol{\nu}^{t-1} + \alpha_{t-1}\boldsymbol{\beta}\right\| + |\alpha_t - a_t| \|\boldsymbol{\beta}\| + \frac{1}{\kappa} \|\boldsymbol{X}\| \left\|\boldsymbol{R}^{t-1} - \boldsymbol{S}^{t-1}\right\|. \tag{75}$$

Since $\boldsymbol{\nu}^0 = \boldsymbol{\theta}^0$, iterating (74) and (75), it can be established that there exists a constant C, depending on κ, such that

$$\left\|\boldsymbol{\nu}^t - \boldsymbol{\theta}^t\right\| \leq (C\,\|\boldsymbol{X}\|)^{2t} \left(\sum_{l=0}^{t-1} |q_l - 1| \left\|\boldsymbol{\nu}^l + \alpha_l\boldsymbol{\beta}\right\| + \sum_{l=0}^{t-1} |\alpha_l - a_l| \|\boldsymbol{\beta}\|\right). \tag{76}$$

Using Lemma 3, the definition of q_t and (73), we have,

$$\begin{aligned}
\lim_{n\to\infty} q_t &= \lim_{n\to\infty} -\frac{1}{\kappa n} \sum_{i=1}^{n} \left\{ \frac{-\lambda_t \rho''\left(\mathsf{prox}\left(\lambda_t h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right) + R_i^t\right)\right)}{1 + \lambda_t \rho''\left(\mathsf{prox}\left(\lambda_t h\left(\boldsymbol{X}_i'\boldsymbol{\beta}, w_i\right) + R_i^t\right)\right)} \right\} \\
&= \mathbb{E}\left[\frac{1}{\kappa}\left\{1 - \frac{1}{1 + \lambda_t \rho''\left(\mathsf{prox}\left(\lambda_t h\left(Q_1^t, U\right) + Q_2^t\right)\right)}\right\}\right],
\end{aligned} \tag{77}$$

where $(Q_1^t, Q_2^t) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}(-\alpha_t, \sigma_t)\right)$. The equivalence (29) 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 $\left(\tilde{Q}_1^t, \tilde{Q}_2^t\right) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\left(-\tilde{\alpha}_t, \tilde{\sigma}_t\right)\right)$, and the last equality follows from the definition of $\tilde{\lambda}_t$ in (22). Note that to obtain (77), 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 [14, 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

$$\lim_{n\to\infty} \frac{1}{\sqrt{p}} (C\,\|\boldsymbol{X}\|)^{2t} \sum_{l=0}^{t-1} |q_l - 1| \left\|\boldsymbol{\nu}^l + \alpha_l \boldsymbol{\beta}\right\| = 0. \tag{78}$$

It remains to analyze the second term in the RHS of (76). To analyze the large sample limit of a_t defined in (58), we invoke Lemma 3 once again, in conjunction with the smoothing techniques from [14, Lemma 6.7], which yields

$$\lim_{n\to\infty} a_t = \frac{1}{\kappa}\,\mathbb{E}\left[\frac{\partial}{\partial a}\Psi_{t-1}\left(h(a, W), s\right)\Big|_{a=Q_1^{t-1}, s=Q_2^{t-1}}\right]. \tag{79}$$

In order to analyze (79), 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,

$$\mathbb{E}\,h(X)(X-\mu) = \sigma^2\,\mathbb{E}\,h'(X). \tag{80}$$

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 $\left(Q_1^{t-1}, Q_2^{t-1}\right) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\left(-\alpha_{t-1}, \sigma_{t-1}\right)\right)$. 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), (80) immediately gives

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

The LHS can be decomposed using the chain rule as follows

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

Putting these together,

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

Marginalizing over W, Z and recalling (79), 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 (27) and (73), 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 (19), we obtain that the LHS equals 1. Further, from (25), 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 (20). Hence, for any finite t,

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

which leads to

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

Combining this with (76) and (78), 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 (74) and the fact that $\lim_{n\to\infty}\left\|\boldsymbol{X}\right\|$ is finite almost surely. This completes the proof.

H.4.4 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 (15) 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 (16).

We defer the proof of Lemma 7 to Appendix H.7, and proceed with the rest of the proof here. The aforementioned results together establish that if the variance map updates (19)–(20) 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 Appendix H.1 that $(\alpha_\star, \sigma_\star, \lambda_\star)$ refers to a solution of (15). 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 (15). Then the AMP trajectory and the MLE can be related as

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{1}{\sqrt{p}}\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\| =_{a.s.} 0. \tag{81}$$

Proof: The proof can be established using techniques similar to that in [33, 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
$$\lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{n}} < \infty. \tag{82}$$
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 [33, 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 [14, 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 $t \geq 0$,

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

$$\implies \lim_{t\to\infty} \lim_{n\to\infty} \frac{\|\hat{\boldsymbol{\beta}}^t\|}{\sqrt{p}} < \infty. \tag{83}$$

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

$$\begin{aligned}
&\left| \frac{1}{p} \sum_{j=1}^{p} \psi\left(\hat{\beta}_j - \alpha_t \beta_j, \beta_j\right) - \frac{1}{p} \sum_{j=1}^{p} \psi\left(\hat{\beta}_j^t - \alpha_t \beta_j, \beta_j\right) \right| \\
&\qquad \leq C \frac{1}{p} \sqrt{\sum_{j=1}^{p} \left(1 + \left\|(\hat{\beta}_j - \alpha_t \beta_j, \beta_j)\right\| + \left\|(\hat{\beta}_j^t - \alpha_t \beta_j, \beta_j)\right\|\right)^2} \left\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\right\|.
\end{aligned} \tag{84}$$

Using (83), (82) 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 2. 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 (15) 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.

H.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

$$\boldsymbol{G} = \nabla^2 \ell(\hat{\boldsymbol{\beta}}), \qquad \boldsymbol{G}_{[-j]} = \nabla^2 \ell_{-j}\left(\hat{\boldsymbol{\beta}}_{[-j]}\right). \tag{85}$$

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

$$\boldsymbol{G}_{[-i],[-j]} = \nabla^2 \ell_{[-i],[-j]} \left(\hat{\boldsymbol{\beta}}_{[-i][-j]} \right), \qquad \boldsymbol{G}_{[-ik],[-j]} = \nabla^2 \ell_{[-ik],[-j]} \left(\hat{\boldsymbol{\beta}}_{[-ik][-j]} \right). \tag{86}$$

Before proceeding, it is useful to record a few observations regarding the differences and similarities between our setup here and that in [33]. Analogues of Theorems 2 and 3 were proved in [33] 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, [33] 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 [33].[9] 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 [33] to the $\boldsymbol{\beta} \neq \boldsymbol{0}$ setup. In such scenarios, often the proof techniques from [33] 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 [33] 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 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 Appendices H.5–H.7 of this manuscript, we will only provide rigorous proofs of the steps for which the details were left out from Appendix C. Thus, it may be convenient for the reader to proceed with the rest of this manuscript with [32] and [33] 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 [33, Lemma 4], it was established that with exponentially high probability, for all sufficiently small $\varepsilon > 0$, the Hessian of the negative log-likelihood satisfies

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

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:

$$\mathcal{D}_n := \left\{\lambda_{\min}(\boldsymbol{G}) > \lambda_{\mathrm{lb}}\right\} \cap \left\{\lambda_{\min}(\boldsymbol{G}_{[-j]}) > \lambda_{\mathrm{lb}}\right\} \cap \left\{\cap_{i=1}^{n} \lambda_{\min}(\boldsymbol{G}_{[-i],[-j]}) > \lambda_{\mathrm{lb}}\right\} \cap \left\{\lambda_{\min}(\boldsymbol{G}_{[-12],[-j]}) > \lambda_{\mathrm{lb}}\right\}. \tag{87}$$

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

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

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

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 (87).

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

$$\boldsymbol{b}_{[-j]} = \begin{bmatrix} 0 \\ \hat{\boldsymbol{\beta}}_{[-j]} \end{bmatrix} + b_{[-j],1} \begin{bmatrix} 1 \\ -\boldsymbol{G}_{[-j]}^{-1}\boldsymbol{w} \end{bmatrix}, \tag{88}$$

where

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{X}'_{\bullet -j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{X}_{\bullet j}, \\ b_{[-j],1} &= \frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}}, \end{aligned} \tag{89}$$

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 [20], an analogous surrogate was introduced in [33] 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 [33] 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 (88) satisfy*

$$\begin{aligned} \mathbb{P}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{b}_{[-j]}\| \lesssim n^{-1/2+o(1)}\right] &= 1 - o(1), \\ \mathbb{P}\left[\sup_{1\leq i\leq n}\left|\boldsymbol{X}_i'\boldsymbol{b}_{[-j]} - \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right| \lesssim n^{-1/2+o(1)}\right] &= 1 - o(1). \end{aligned} \tag{90}$$

The fitted values satisfy

$$\mathbb{P}\left[\sup_{1\leq i\leq n}|\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \tag{91}$$

Further, we have

$$\mathbb{P}\left[|\hat{\beta}_j - b_{[-j],1}| \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \tag{92}$$

Proof: The proof of (90) follows upon tracing out the steps in [33, Theorem 8] verbatim using $\boldsymbol{b}_{[-j]}, b_{[-j],1}$ and $\hat{\boldsymbol{\beta}}_{[-j]}$ defined in (88) instead of $\tilde{\boldsymbol{b}}, \tilde{b}_1$ and $\tilde{\boldsymbol{\beta}}$ respectively. In [33], $\tilde{\boldsymbol{b}}$ 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 [33, 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 [33, Theorem 8] are applicable here for establishing (90). Next, note that by the triangle inequality,

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

Combining this inequality with (90) and the fact that $\sup_i \|\boldsymbol{X}_i\| = O(1)$ with high probability, we have the required result (91). Finally, (92) follows trivially from (90). ■

Henceforth, whenever necessary, we describe suitable correspondences between the terms here and their appropriate analogues in [33], 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 [33].

Applying Theorem 8 we have the approximation

$$\hat{\beta}_j = \frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}} + o_P(1). \tag{93}$$

At this point, recall from (93) that the above expression can be simplified to the following form:

$$\frac{\boldsymbol{X}'_{\bullet j}(\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]}))}{\boldsymbol{X}'_{\bullet j}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{H}\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})^{1/2}\boldsymbol{X}_{\bullet j}} = \frac{\lambda_{[-j]} s_j}{\kappa} Z + o_P(1), \tag{94}$$

where

$$s_j^2 = \frac{\|\boldsymbol{y} - \rho'(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]})\|^2}{n} \qquad \text{and} \qquad \lambda_{[-j]} = \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{G}_{[-j]}^{-1}\right). \tag{95}$$

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 (15). Hence, it remains to analyze the terms s_j. For convenience of notation, denote the residuals as

$$r_i := y_i - \rho'(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}), \tag{96}$$

which implies

$$s_j^2 = \frac{1}{n}\sum_{i=1}^n r_i^2. \tag{97}$$

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 [20, 33], 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:*

$$\begin{aligned} q_i &:= \boldsymbol{X}'_{i,-j}\boldsymbol{G}^{-1}_{[-i],[-j]}\boldsymbol{X}_{i,-j}, \\ \hat{\boldsymbol{b}}_{[-j]} &:= \hat{\boldsymbol{\beta}}_{[-i],[-j]} + \boldsymbol{G}^{-1}_{[-i],[-j]}\boldsymbol{X}_{i,-j}\left(y_i - \rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i)\right)\right), \end{aligned} \tag{98}$$

where $\boldsymbol{G}_{[-i],[-j]}$ is specified by (86). 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 [33, Lemma 18], with the choice of q_i specified in (98) and $\hat{\boldsymbol{b}}_{[-j]}$ in place of $\hat{\boldsymbol{b}}$ in [33]. 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 $\{y_i, \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]}\}_{i=1,\dots n}$, in the following sense:*

$$\sup_{i=1,\dots,n} \left| \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i \right) \right| \xrightarrow{\mathbb{P}} 0. \tag{99}$$

Further, the residuals can be simultaneously approximated using

$$\sup_{i=1,\dots,n} \left| r_i - \left\{ y_i - \rho'\left(\mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i \right) \right) \right\} \right| \xrightarrow{\mathbb{P}} 0. \tag{100}$$

Proof: Since ρ'' is bounded, (100) follows from (99) trivially. Thus, it suffices to show (99). From the definition of $\hat{\boldsymbol{b}}_{[-j]}$ in (98), 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 (18) 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 [33, Lemma 2] for a formal statement), we have

$$\sup_i \left| \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{q_i \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i \right) \right| \lesssim n^{-1/2+o(1)}, \tag{101}$$

with high probability. For (99), it then suffices to establish that

$$\sup_i \left| \mathsf{prox}_{q_i \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i \right) - \mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i \right) \right| \xrightarrow{\mathbb{P}} 0. \tag{102}$$

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

$$\sup_i |q_i - \lambda_\star| \le \sup_i |q_i - \lambda_{[-j]}| + |\lambda_{[-j]} - \lambda_\star|. \tag{103}$$

Using $q_i, \lambda_{[-j]}$ instead of $\tilde{q}_i, \tilde{\alpha}$ in [33, Lemma 19] and following the proof line by line in conjunction with Lemma 8, we have

$$\sup_i |q_i - \lambda_{[-j]}| \lesssim n^{-1/2+o(1)} \tag{104}$$

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 [15, Proposition 6.3]

$$\frac{\partial}{\partial z}\mathsf{prox}_{b\rho}(z) = \left. \frac{1}{1 + b\rho''(x)} \right|_{x = \mathsf{prox}_{b\rho}(z)}, \quad \frac{\partial}{\partial b}\mathsf{prox}_{b\rho}(z) = -\left. \frac{\rho'(x)}{1 + b\rho''(x)} \right|_{x = \mathsf{prox}_{b\rho}(z)}, \tag{105}$$

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

$$\begin{aligned}
&\sup_i \left| \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right) - \mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&\leq \sup_i \left| \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + q_i y_i\right) - \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&+ \sup_i \left| \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + \lambda_\star y_i\right) - \mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right| \\
&\leq \sup_i |q_i - \lambda_\star| \left\{ \left| \frac{\partial}{\partial z} \mathsf{prox}_{q_i\rho}(z)\Big|_{z=\tilde{q}_i y_i} \right| + \left| \frac{\partial}{\partial b} \mathsf{prox}_{b\rho}(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i)\Big|_{b=\tilde{\lambda}_i} \right| \right\}, \qquad (106)
\end{aligned}$$

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 (105), note that the partial derivatives are both bounded by 1 since $q_i, \tilde{\lambda}_i > 0$. This establishes (102). Combining with (101), we have the required result (99).

■

Recall from (94) and (97) 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 (96), observe that each residual r_i implicitly depends on n. We argue in the subsequent text that

$$\lim_{n\to\infty} \mathrm{Cov}\ (r_1^2, r_2^2) = 0. \qquad (107)$$

From (100), 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 [20, 33]. 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*

$$\mathbb{P}\left[\max\left\{\left|\boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{\beta}}_{[-ik],[-j]}\right)\right|, \left|\boldsymbol{X}'_{k,-j}\left(\hat{\boldsymbol{\beta}}_{[-k],[-j]} - \hat{\boldsymbol{\beta}}_{[-ik],[-j]}\right)\right|\right\} \lesssim n^{-1/2+o(1)}\right] = 1 - o(1). \qquad (108)$$

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 (98) 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

$$\|\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \hat{\boldsymbol{b}}_{[-i],[-j]}\| \lesssim n^{-1/2+o(1)}. \qquad (109)$$

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 (109) 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 (107). From (100), we have that for each $\vartheta, \delta > 0$, there exists N such that for all $n \geq N$,

$$\mathbb{P}\left[\sup_{i=1,\ldots,n}\left|r_i - \left\{y_i - \rho'\left(\mathsf{prox}_{\lambda_\star \rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + \lambda_\star y_i\right)\right)\right\}\right| \leq \vartheta\right] \geq 1-\delta. \tag{110}$$

Let $\mathcal{E}_1, \mathcal{E}_2$ denote the high probability event in (108) and the event in (110) respectively. Denote $\mathcal{H} = \mathcal{D}_n \cap \mathcal{E}_1 \cap \mathcal{E}_2$, where $\mathcal{D}_n$ is defined via (87). 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 (108) and (110) we obtain that for any $\vartheta, \delta > 0$, for every $n \geq N$,

$$|\mathrm{Cov}(r_1^2, r_2^2)| \leq \mathbb{E}\, f(M_1, y_1) f(M_2, y_2) + C\vartheta^2 + \delta, \tag{111}$$

where $C > 0$ is an absolute constant. By arguments similar to that in [20, 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 (111), we have $\mathbb{E}\, f(M_l, y_l) = 0$, by definition. Then (111) leads to the required result (107). By Chebyshev's inequality, we have effectively established that

$$\frac{1}{n}\sum_{i=1}^n r_i^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, r_i^2 \xrightarrow{\mathbb{P}} 0. \tag{112}$$

Since, the residuals are identically distributed, the approximation to $\hat{\beta}_j$ derived in (93) and (94) 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 (110) and using arguments similar to that for establishing (107), 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 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 (93) and (94) in conjunction with Theorem 10, it follows that

$$\begin{bmatrix} \hat{\beta}_j \\ \hat{\beta}_l \end{bmatrix} = \begin{bmatrix} \frac{\lambda_\star}{\kappa} \sum_i X_{ij} \left(y_i - \rho' \left(X'_{i,-j} \hat{\boldsymbol{\beta}}_{[-j]} \right) \right) \\ \frac{\lambda_\star}{\kappa} \sum_i X_{il} \left(y_i - \rho' \left(X'_{i,-l} \hat{\boldsymbol{\beta}}_{[-l]} \right) \right) \end{bmatrix} + o_P(1). \tag{113}$$

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 (91) 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 (113), 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 (112), 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

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

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

H.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:*

$$2\Lambda_j = \frac{\kappa \hat{\beta}_j^2}{\lambda_{[-j]}} + o_P(1), \tag{114}$$

where $\lambda_{[-j]}$ is defined in (95).

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

$$2\Lambda_j = \left(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}} \boldsymbol{D}(\hat{\boldsymbol{\beta}}) \left(\boldsymbol{X}_{\bullet -j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \frac{1}{3}\sum_{i=1}^{n} \rho'''(\gamma_i) \left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \boldsymbol{X}'_i\hat{\boldsymbol{\beta}}\right)^3, \tag{115}$$

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 (115) is $o_P(1)$. Subsequently, it can be checked that calculations similar to those in [33, 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 (95) 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 (15).

Proof: The proof follows by arguments similar to that in [33, 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*

$$\delta_n(x) := \frac{p}{n} - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i][-j]} + xy_i\right)\right)}. \tag{116}$$

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 [33, Proposition 2], we obtain

$$\frac{p}{n} - 1 + \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{1 + \rho''\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]}\right)\lambda_{[-j]}} \right\} \xrightarrow{\mathbb{P}} 0. \tag{117}$$

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

$$\sup_i \left| \boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-j]} - \mathsf{prox}_{\lambda_{[-j]}\rho}\left(\boldsymbol{X}'_{i,-j}\hat{\boldsymbol{\beta}}_{[-i],[-j]} + \lambda_{[-j]}y_i\right) \right| \lesssim n^{-1/2+o(1)}, \tag{118}$$

with high probability. The claim is established by comparing (101), (104), and by arguments similar to that in (106), with $\lambda_\star$ replaced by $\lambda_{[-j]}$.

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

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

On the event $\mathcal{D}_n$ defined in (87), $|\lambda_{[-j]}| \leq p/(n\lambda_{\mathrm{lb}})$. Further, ρ''' is bounded. Hence, from (118) 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*

$$\Delta(x) = \kappa - 1 + \mathbb{E}\left[\frac{1}{1+x\rho''\left(\mathsf{prox}_{x\rho}\left(xh(\tilde{Q}_1,W)+\tilde{Q}_2\right)\right)}\right], \tag{119}$$

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 (16) and $(\alpha_\star,\sigma_\star)$ form part of the solution to the system (15). Then, for any $B>0$,

$$\sup_{x\in[0,B]}|\delta_n(x)-\Delta(x)| \xrightarrow{\mathbb{P}} 0. \tag{120}$$

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 (116) and (119) respectively, it can be established that for (120), it suffices to show the following: for any given $x \in [0,B]$

$$|\delta_n(x)-G_n(x)| \xrightarrow{\mathbb{P}} 0, \tag{121}$$
$$|G_n(x)-\Delta(x)| \to 0, \tag{122}$$

where $G_n(x) = \mathbb{E}\left(\delta_n(x)\right)$. (We refer the interested reader to the proof of [33, Proposition 3] for a detailed analogous computation in the simpler setup $\boldsymbol{\beta}=\mathbf{0}$).

We first establish (122). 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 Appendix C rigorously established the following fact:

$$\begin{bmatrix}Q_1^\star\\ Q_2^\star\end{bmatrix} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0},\begin{bmatrix}\gamma^2 & 0\\ 0 & \kappa\sigma_\star^2\end{bmatrix}\right), \quad \text{where} \quad Q_1^\star := \boldsymbol{X}'_{i,-j}\boldsymbol{\beta}_{-j}, \quad Q_2^\star = \boldsymbol{X}'_{i,-j}\left(\hat{\boldsymbol{\beta}}_{[-i][-j]}-\alpha_\star\boldsymbol{\beta}_{-j}\right). \tag{123}$$

As mentioned in (52), 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 (123) 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 (122).

To analyze (121), 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 $\delta_n(x)\overset{L_2}{\to}G_n(x)$. To this end, we resort to the leave-two-observation-out approach discussed in Appendix H.5. By routine arguments using the triangle inequality, properties of the partial derivatives of the proximal mapping operator (105), the fact that $\|f\|_\infty\leq 1$, and invoking the approximations in Lemma 11, we arrive at

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

From arguments similar to [20, 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 (87), we obtain that

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

To complete the proof, recall from (27) 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 (15) 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, $\lambda_{[-j]}\overset{\mathbb{P}}{\to}\lambda_\star$.

■

H.7 Proof of Supporting Lemmas

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

H.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

$$1 - G(\lambda) = \kappa \tag{124}$$

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

$$\lim_{\lambda\to 0}(1 - G(\lambda)) = 0 \tag{125}$$

$$\lim_{\lambda\to\infty}(1 - G(\lambda)) = 1. \tag{126}$$

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 [14, Proposition 6.4]

$$K'_\lambda(p,s) := \frac{\partial K_\lambda(p,s)}{\partial s} = \frac{\lambda\rho''(\mathsf{prox}_{\lambda\rho}(p+s))}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(p+s))}. \tag{127}$$

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

$$G(\lambda) = \mathbb{E}\left[2\rho'(Q_1)\left(1 - K'_\lambda(aQ_1, bZ)\right)\right]. \tag{128}$$

Applying Stein's formula (80), 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 (128) 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 [33, 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 (125), 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 (128). 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 (125).

Finally, we turn to the proof of (126). To this end, note that [33, 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 (127), 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 (126).

H.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,

$$\delta(\mathcal{C}_i^{\boldsymbol{v}}) = \mathbb{E}\left[\|\Pi_{\mathcal{C}_i^{\boldsymbol{v}}}\|^2\right] = \mathbb{E}\left[\|\boldsymbol{g}\|^2 - \min_{t\in\mathbb{R}} \min_{\boldsymbol{u}\in\mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2\right], \tag{129}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. It can be checked that an approach similar to that in [33, Appendix D.2] leads to the lower bound

$$\begin{aligned} &\min_{t\in\mathbb{R}} \min_{\boldsymbol{u}\in\mathcal{C}_i^{\boldsymbol{v}}} \|\boldsymbol{g} - t\boldsymbol{v} - \boldsymbol{u}\|^2 \\ &\geq \min_t \left\{ \sum_{i:(g_i - tv_i)<0} (g_i - tv_i)^2 - \max_{S\subset[n]:|S|=2\sqrt{\epsilon}n} \sum_{i\in S} (g_i - tv_i)^2 - 2\sqrt{2}\varepsilon^{3/4}\|\boldsymbol{g} - t\boldsymbol{v}\|^2 \right\}, \end{aligned} \tag{130}$$

where $\varepsilon > 0$ is a small constant. In the remaining proof, we carefully analyze the RHS of (130). To this end, define $G^{\boldsymbol{v}}(t) = F^{\boldsymbol{v}}(t) - \varepsilon^{\boldsymbol{v}}(t)$, where

$$\begin{aligned} F^{\boldsymbol{v}}(t) &= \sum_{i:(g_i - tv_i)<0} (g_i - tv_i)^2 \\ \varepsilon^{\boldsymbol{v}}(t) &= \max_{S\subset[n]:|S|=2\sqrt{\varepsilon}n} \sum_{i\in S} (g_i - tv_i)^2 + 2\sqrt{2}\varepsilon^{3/4}\|\boldsymbol{g} - t\boldsymbol{v}\|^2. \end{aligned} \tag{131}$$

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 [10, Section 3.3]:

$$\frac{1}{n} F^{\boldsymbol{v}}(t_\star) \xrightarrow{\mathbb{P}} g_{\mathrm{MLE}}^{-1}(\gamma). \tag{132}$$

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 (42) and the following property: for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$, with high probability,*[10]

$$\sup_{t \in [t_0 - M, t_0 + M]} |\varepsilon^{\boldsymbol{v}}(t)| \le n f(\varepsilon), \tag{133}$$

$$\forall \quad t \notin [t_0 - M, t_0 + M] \qquad G^{\boldsymbol{v}}(t) > n g_{\mathrm{MLE}}^{-1}(\gamma), \tag{134}$$

where $\varepsilon^{\boldsymbol{v}}(t), G^{\boldsymbol{v}}(t)$ are defined via (131). Above, $M \equiv M(\varepsilon)$ is a positive constant independent of n, $\mathcal{F}_{\boldsymbol{V}}$ is the event defined in (35), $f(x)$ 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 (132). 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 (130), 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 (129), 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 [24] and the union bound, for $a(\varepsilon) = 2\max\{2\sqrt{\varepsilon}H(2\sqrt{\varepsilon})\}$,

$$\mathbb{P}_{\boldsymbol{V}}\left[\max_{S : |S| = 2\sqrt{\varepsilon}n} \sum_{i \in S} V_i^2 \le C_1 n a(\epsilon)\right] \ge 1 - e^{-H(2\sqrt{\varepsilon})n}, \tag{135}$$

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, [**?**, Sec 3.1], it can be established that

$$\mathbb{P}_{\boldsymbol{V}}\left[\|\boldsymbol{V}\|_2 \le C_1\sqrt{n}\right] \ge 1 - 2\exp\left(-c_1 n\right), \tag{136}$$

[10] 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 [34, Proposition 5.10], we have

$$\mathbb{P}_{\boldsymbol{V}}\left[\sum_{i=1}^{n} |V_i| \leq C_1 n\right] \geq 1 - C_1 \exp\left(-c_1 n\right). \tag{137}$$

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 [34, Proposition 5.16], it can be shown that

$$\mathbb{P}_{\boldsymbol{V}}\left[\sum_{i=1}^{n} V_i^2 \mathbf{1}_{V_i>0} \geq C_1 n\right] \geq 1 - 2\exp\left(-c_1 n\right). \tag{138}$$

Let $\mathcal{G}_{\boldsymbol{V}}$ denote the high probability event formed by the intersection of the events in (135),(136),(137) and (138). Thus, any $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$ satisfies the following properties:

$$\max_{S\subset[n]:|S|=2\sqrt{\epsilon}n} \sum_{i\in S} v_i^2 \leq C_1 n a(\varepsilon), \quad \|\boldsymbol{v}\|^2 \leq C_2 n, \quad \sum_{i=1}^{n} |v_i| \leq C_3 n, \quad \sum_{i:v_i>0} v_i^2 \geq C_4 n, \quad \max_i v_i^2 \leq \zeta \log n. \tag{139}$$

We are now in a position to establish (133) and (134). 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 (135) and (136), 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,

$$\varepsilon^{\boldsymbol{v}}(t) \leq n\left(1 + t^2\right)\left(C_1 a(\varepsilon) + C_2 \varepsilon^{3/4}\right), \tag{140}$$

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 (140) we have for all $\boldsymbol{v} \in \mathcal{G}_{\boldsymbol{V}} \cap \mathcal{F}_{\boldsymbol{V}}$,

$$G^{\boldsymbol{v}}(t) \geq \sum_{i:g_i<tv_i} (g_i - t v_i)^2 - nt^2(C_1 a(\varepsilon) + C_2\varepsilon^{3/4}) \tag{141}$$

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

$$\sum_{i:g_i<tv_i} (g_i - t v_i)^2 \geq t^2 \sum_{i:v_i>0,g_i\leq 0} v_i^2 - 2t \sum_{i:v_i>0,g_i<0} |v_i g_i| \geq t^2 \sum_{i:v_i>0,g_i\leq 0} v_i^2 - 2t\sum_{i=1}^{n} |v_i g_i|. \tag{142}$$

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 (139), we have

$$\sum_{i=1}^{n} |g_i v_i| \leq C_2 n \tag{143}$$

with high probability.

Thus, it only remains to analyze the first term in the RHS of (142). 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 [34, 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 (35), is crucial. Recalling the properties of $\boldsymbol{v}$ from (139), we can choose $t = C_1 n$ such that

$$\sum_{i:v_i>0} v_i^2 \mathbf{1}_{g_i \leq 0} \geq C_2 n \tag{144}$$

with high probability. Combining (143), (144) and recalling (142), we finally arrive at

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

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.

■

H.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,

$$\mathbb{P}\left[\text{span}\,(\boldsymbol{V}) \cap \mathcal{A} \neq \{\mathbf{0}\}\right] \leq \mathbb{P}\left[\exists\ a > 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right] + \mathbb{P}\left[\exists\ a < 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right]. \tag{145}$$

From the definition of $\mathcal{A}$ in (32), it follows that

$$\mathbb{P}\left[\exists\ a > 0 \text{ s.t. } a\boldsymbol{V} \in \mathcal{A}\right] = \mathbb{P}\left[\sum_{j=1}^{n} |V_j| \mathbf{1}_{V_j<0} \leq \varepsilon^2 \sqrt{n} \|\boldsymbol{V}\|\right] \leq \mathbb{P}\left[\sum_{j=1}^{n} |V_j| \mathbf{1}_{V_j<0} \leq \varepsilon^2 n\right] + C_1 \exp\left(-c_1 n\right), \tag{146}$$

where the last inequality follows from (136). Since $|V_i| \mathbf{1}_{V_i<0} - \mathbb{E}\,|V_i| \mathbf{1}_{V_i<0}$ is sub-gaussian, applying [34, Proposition 5.10] we obtain

$$\mathbb{P}\left[\frac{\left(\mathbb{E}\,|V_1| \mathbf{1}_{V_1<0}\right) n}{2} \leq \sum_{i=1}^{n} |V_i| \mathbf{1}_{V_i<0} \leq \frac{3\left(\mathbb{E}\,|V_1| \mathbf{1}_{V_1<0}\right) n}{2}\right] \geq 1 - C_1 \exp\left(-c_1 n\right). \tag{147}$$

Combining (146) and (147) 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 (145) can be analyzed similarly.