The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a *Rescaled* Chi-Square

Pragya Sur* Yuxin Chen† Emmanuel J. Candès*‡

June 2017

Abstract

Logistic regression is used thousands of times a day to fit data, predict future outcomes, and assess the statistical significance of explanatory variables. When used for the purpose of statistical inference, logistic models produce p-values for the regression coefficients by using an approximation to the distribution of the likelihood-ratio test. Indeed, Wilks' theorem asserts that whenever we have a fixed number p of variables, twice the log-likelihood ratio (LLR) 2Λ is distributed as a χ^2_k variable in the limit of large sample sizes n; here, χ^2_k is a chi-square with k degrees of freedom and k the number of variables being tested. In this paper, we prove that when p is not negligible compared to n, Wilks' theorem does not hold and that the chi-square approximation is grossly incorrect; in fact, this approximation produces p-values that are far too small (under the null hypothesis).

Assume that n and p grow large in such a way that $p/n \to \kappa$ for some constant $\kappa < 1/2$. We prove that for a class of logistic models, the LLR converges to a *rescaled* chi-square, namely, $2\Lambda \stackrel{\mathrm{d}}{\to} \alpha(\kappa)\chi^2_k$, where the scaling factor $\alpha(\kappa)$ is greater than one as soon as the dimensionality ratio κ is positive. Hence, the LLR is larger than classically assumed. For instance, when $\kappa = 0.3$, $\alpha(\kappa) \approx 1.5$. In general, we show how to compute the scaling factor by solving a nonlinear system of two equations with two unknowns. Our mathematical arguments are involved and use techniques from approximate message passing theory, from non-asymptotic random matrix theory and from convex geometry. We also complement our mathematical study by showing that the new limiting distribution is accurate for finite sample sizes.

Finally, all the results from this paper extend to some other regression models such as the probit regression model.

Keywords. Logistic regression, likelihood-ratio tests, Wilks' theorem, high-dimensionality, goodness of fit, approximate message passing, concentration inequalities, convex geometry, leave-one-out analysis

1 Introduction

Logistic regression is by far the most widely used tool for relating a binary response to a family of explanatory variables. This model is used to infer the importance of variables and nearly all standard statistical softwares have inbuilt packages for obtaining p-values for assessing the significance of their coefficients. For instance, one can use the snippet of R code below to fit a logistic regression model from a vector `y` of binary responses and a matrix `X` of covariates:

```
fitted <- glm(y ~ X+0, family = 'binomial')
pvals  <- summary(fitted)$coefficients[,4]
```

*Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
†Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, U.S.A.
‡Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.

The vector `pvals` stores p-values for testing whether a variable belongs to a model or not, and it is well known that the underlying calculations used to produce these p-values can also be used to construct confidence intervals for the regression coefficients. Since logistic models are used hundreds of times every day for inference purposes, it is important to know whether these calculations—e.g. these p-values—are accurate and can be trusted.

1.1 Binary regression

Imagine we have n samples of the form $(y_i, \boldsymbol{X}_i)$, where $y_i \in \{0,1\}$ and $\boldsymbol{X}_i \in \mathbb{R}^p$. In a generalized linear model, one postulates the existence of a link function $\mu(\cdot)$ relating the conditional mean of the response variable to the linear predictor $\boldsymbol{X}_i^\top \boldsymbol{\beta}$,

$$\mathbb{E}[y_i|\boldsymbol{X}_i] = \mu(\boldsymbol{X}_i^\top \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\top \in \mathbb{R}^p$ is an unknown vector of parameters. We focus here on the two most commonly used binary regression models, namely, the logistic and the probit models for which

$$\mu(t) := \begin{cases} e^t/(1+e^t) & \text{in the logistic model,} \\ \Phi(t) & \text{in the probit model;} \end{cases} \tag{2}$$

here, Φ is the cumulative distribution function (CDF) of a standard normal random variable. In both cases, the *Symmetry Condition*

$$\mu(t) + \mu(-t) = 1 \tag{3}$$

holds, which says that the two types $y_i = 0$ and $y_i = 1$ are treated in a symmetric fashion. Assuming that the observations are independent, the negative log-likelihood function is given by [1, Section 4.1.2]

$$\ell\left(\boldsymbol{\beta}\right) := -\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log\left(1-\mu_i\right) \right\}, \qquad \mu_i := \mu(\boldsymbol{X}_i^\top \boldsymbol{\beta}).$$

Invoking the symmetry condition, a little algebra reveals an equivalent expression

$$\ell\left(\boldsymbol{\beta}\right) := \sum\nolimits_{i=1}^{n} \rho\big(-\tilde{y}_i \boldsymbol{X}_i^\top \boldsymbol{\beta}\big), \tag{4}$$

where

$$\tilde{y}_i := \begin{cases} 1 & \text{if } y_i = 1, \\ -1 & \text{if } y_i = 0, \end{cases} \qquad \text{and} \qquad \rho(t) := \begin{cases} \log\left(1+e^t\right) & \text{in the logistic case,} \\ -\log\Phi\left(-t\right) & \text{in the probit case.} \end{cases} \tag{5}$$

Throughout we refer to this function ρ as the *effective link*.

1.2 The likelihood-ratio test and Wilks' phenomenon

Researchers often wish to determine which covariates are of importance, or more precisely, to test whether the jth variable belongs to the model or not: formally, we wish to test the hypothesis

$$H_j: \quad \beta_j = 0 \quad \text{versus} \quad \beta_j \neq 0. \tag{6}$$

Arguably, one of the most commonly deployed techniques for testing H_j is the likelihood-ratio test (LRT), which is based on the log-likelihood ratio (LLR) statistic

$$\Lambda_j := \ell\big(\hat{\boldsymbol{\beta}}_{(-j)}\big) - \ell\big(\hat{\boldsymbol{\beta}}\big). \tag{7}$$

Here, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(-j)}$ denote respectively the maximum likelihood estimates (MLEs) under the full model and the reduced model on dropping the jth predictor; that is,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta}) \qquad \text{and} \qquad \hat{\boldsymbol{\beta}}_{(-j)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \beta_j=0} \ell(\boldsymbol{\beta}).$$

Inference based on such log-likelihood ratio statistics has been studied extensively in prior literature [12, 37, 52]. Arguably, one of the most celebrated results in the large-sample regime is the Wilks' theorem.

To describe the Wilk's phenomenon, imagine we have a sequence of observations $(y_i, \boldsymbol{X}_i)$ where $y_i \in \{0, 1\}$, $\boldsymbol{X}_i \in \mathbb{R}^p$ with p fixed. Since we are interested in the limit of large samples, we may want to assume that the covariates are i.i.d. drawn from some population with non-degenerate covariance matrix so that the problem is fully p-dimensional. As before, we assume a conditional logistic model for the response. In this setting, Wilks' theorem [52] calculates the asymptotic distribution of $\Lambda_j(n)$ when n grows to infinity:

> (Wilks' phenomenon) *Under suitable regularity conditions which, for instance, guarantee that the MLE exists and is unique,*[1] *the LLR statistic for testing $H_j : \beta_j = 0$ vs. $\beta_j \neq 0$ has asymptotic distribution under the null given by*
>
> $$2\Lambda_j(n) \overset{\mathrm{d}}{\to} \chi_1^2, \qquad \text{as } n \to \infty. \tag{8}$$

This fixed-p large-n asymptotic result, which is a consequence of asymptotic normality properties of the MLE [50, Theorem 5.14], applies to a much broader class of testing problems in parametric models; for instance, it applies to the probit model as well. We refer the readers to [34, Chapter 12] and [50, Chapter 16] for a thorough exposition and details on the regularity conditions under which Wilks' theorem holds. Finally, there is a well-known extension which states that if we were to drop k variables from the model, then the LLR would converge to a chi-square distribution with k degrees of freedom under the hypothesis that the reduced model is correct.

1.3 Inadequacy of Wilks' theorem in high dimensions

The chi-square approximation to the distribution of the LLR statistic is used in standard statistical softwares to provide p-values for the single or multiple coefficient likelihood ratio tests. Here, we perform a simple experiment on synthetic data to study the accuracy of the chi-square approximation when p and n are both decently large. Specifically, we set $\boldsymbol{\beta} = \boldsymbol{0}$ and test $\beta_1 = 0$ vs. $\beta_1 \neq 0$ using the LRT in a setting where $p = 1200$. In each trial, $n = 4000$ observations are produced with $y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(1/2)$, and $\boldsymbol{X} := [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n \times p}$ is obtained by generating a random matrix composed of i.i.d. $\mathcal{N}(0, 1)$ entries. We fit a logistic regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ using R, and extract the p-values for each coefficient. Figure 1 plots the pooled histogram that aggregates 4.8×10^5 p-values in total (400 trials with 1200 p-values obtained in each trial).

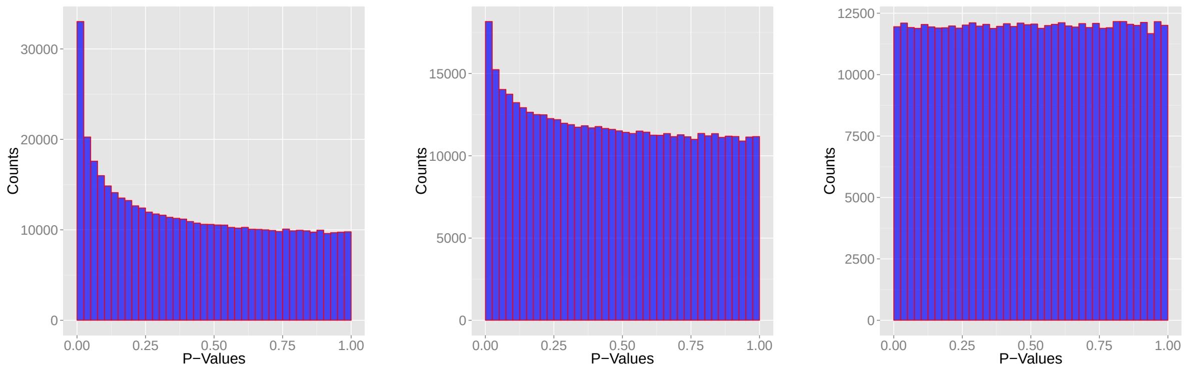
If the χ_1^2 approximation were true, then we would expect to observe uniformly distributed p-values. The histrogram from Fig. 1 is, however, far from uniform. This is an indication of the inadequacy of Wilks' theorem when p and n are both large. The same issue was also reported in [11], where the authors observed that this discrepancy is highly problematic since the distribution is skewed towards smaller values. Hence, such p-values cannot be trusted to construct level-α tests and the problem is increasingly severe when we turn attention to smaller p-values as in large-scale multiple testing applications.

1.4 The Bartlett correction?

A natural question that arises immediately is whether the observed discrepancy could be an outcome of a finite-sample effect. It has been repeatedly observed that the chi-square approximation does not yield accurate results with finite sample size. One correction to the LRT that is widely used in finite samples is the Bartlett correction, which dates back to Bartlett [5] and has been extensively studied over the past few decades (e.g. [8, 10, 13, 15, 33]). In the context of testing for a single coefficient in the logistic model, this correction can be described as follows [38]: compute the expectation of the LLR statistic up to terms of order $1/n^2$; that is, compute a parameter α such that

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

[1] Such conditions would also typically imply asymptotic normality of the MLE.



(a) (b) (c)

Figure 1: Histogram of p-values for logistic regression under i.i.d. Gaussian design, when $\boldsymbol{\beta} = \mathbf{0}$, $n = 4000$, $p = 1200$, and $\kappa = 0.3$: (a) classically computed p-values; (b) Bartlett-corrected p-values; (c) adjusted p-values.

which suggests a corrected LLR statistic

$$\frac{2\Lambda_j}{1 + \frac{\alpha_n}{n}} \tag{9}$$

with α_n being an estimator of α. With a proper choice of α_n, one can ensure

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

in the classical setting where p is fixed and n diverges. In expectation, this corrected statistic is closer to a χ^2_1 distribution than the original LLR for finite samples. Notably, the correction factor may in general be a function of the unknown $\boldsymbol{\beta}$ and, in that case, must be estimated from the null model via maximum likelihood estimation.

In the context of GLMs, Cordeiro [13] derived a general formula for the Bartlett corrected LLR statistic, see [14, 17] for a detailed survey. In the case where there is no signal ($\boldsymbol{\beta} = \mathbf{0}$), one can compute α_n for the logistic regression model following [13] and [38], which yields

$$\alpha_n = \frac{n}{2}\left[\mathrm{Tr}\left(\boldsymbol{D}_p^2\right) - \mathrm{Tr}\left(\boldsymbol{D}_{p-1}^2\right)\right]. \tag{10}$$

Here, $\boldsymbol{D}_p$ is the diagonal part of $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\boldsymbol{D}_{p-1}$ is that of $\boldsymbol{X}_{(-j)}\left(\boldsymbol{X}_{(-j)}^\top\boldsymbol{X}_{(-j)}\right)^{-1}\boldsymbol{X}_{(-j)}^\top$ in which $\boldsymbol{X}_{(-j)}$ is the design matrix $\boldsymbol{X}$ with the jth column removed. Comparing the adjusted LLRs to a χ^2_1 distribution yields adjusted p-values. In the setting of Fig. 1(a), the histogram of Bartlett corrected p-values is shown in Fig. 1(b). As we see, these p-values are still far from uniform.

If the mismatch is not due to finite sample-size effects, what is the distribution of the LLR in high dimensions? Our main contribution is to provide a very precise answer to this question; below, we derive the high-dimensional asymptotic distribution of the log-likelihood ratios, i.e. in situations where the dimension p is not necessarily negligible compared to the sample size n.

2 Main results

2.1 Modelling assumptions

In this paper, we focus on the high-dimensional regime where the sample size is not much larger than the number of parameters to be estimated—a setting which has attracted a flurry of activity in recent years. In particular, we assume that the number $p(n)$ of covariates grows proportionally with the number n of observations; that is,

$$\lim_{n\to\infty} \frac{p(n)}{n} = \kappa, \tag{11}$$

where $\kappa > 0$ is a fixed constant independent of n and $p(n)$. In fact, we shall also assume $\kappa < 1/2$ for both the logistic and the probit models, as the MLE is otherwise at ∞; see Section 2.2.

To formalize the notion of high-dimensional asymptotics when both n and $p(n)$ diverge, we consider a sequence of instances $\{\boldsymbol{X}(n), \boldsymbol{y}(n)\}_{n\geq 0}$ such that for any n,

- $\boldsymbol{X}(n) \in \mathbb{R}^{n\times p(n)}$ has i.i.d. rows $\boldsymbol{X}_i(n) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p(n)\times p(n)}$ is positive definite;
- $y_i(n) \mid \boldsymbol{X}(n) \sim y_i(n) \mid \boldsymbol{X}_i(n) \overset{\text{ind.}}{\sim} \mathsf{Bernoulli}\left(\mu(\boldsymbol{X}_i(n)^\top \boldsymbol{\beta}(n))\right)$, where μ satisfies the Symmetry Condition;
- we further assume $\boldsymbol{\beta}(n) = \boldsymbol{0}$. From the Symmetry Condition it follows that $\mu(0) = 1/2$, which directly implies that $\boldsymbol{y}(n)$ is a vector with i.i.d $\mathsf{Bernoulli}(1/2)$ entries.

The MLE is denoted by $\hat{\boldsymbol{\beta}}(n)$ and there are $p(n)$ LLR statistics $\Lambda_j(n)$ $(1 \leq j \leq p(n))$, one for each of the $p(n)$ regression coefficients. In the sequel, the dependency on n shall be suppressed whenever it is clear from the context.

2.2 When does the MLE exist?

Even though we are operating in the regime where $n > p$, the existence of the MLE cannot be guaranteed for all p and n. Interestingly, the norm of the MLE undergoes a sharp phase transition in the sense that

$$\|\hat{\boldsymbol{\beta}}\| = \infty \quad \text{if } \kappa > 1/2 \qquad \text{and} \qquad \|\hat{\boldsymbol{\beta}}\| < \infty \quad \text{if } \kappa < 1/2.$$

Here, we develop some understanding about this phenomenon. Given that $\rho(t) \geq \rho(-\infty) = 0$ for both the logistic and probit models, each summand in (4) is minimized if $\tilde{y}_i \boldsymbol{X}_i^\top \boldsymbol{\beta} = \infty$, which occurs when $\text{sign}(\boldsymbol{X}_i^\top \boldsymbol{\beta}) = \text{sign}(\tilde{y}_i)$ and $\|\boldsymbol{\beta}\| = \infty$. As a result, if there exists a nontrivial ray $\boldsymbol{\beta}$ such that

$$\boldsymbol{X}_i^\top \boldsymbol{\beta} > 0 \quad \text{if } \ \tilde{y}_i = 1 \qquad \text{and} \qquad \boldsymbol{X}_i^\top \boldsymbol{\beta} < 0 \quad \text{if } \ \tilde{y}_i = -1 \tag{12}$$

for any $1 \leq i \leq n$, then pushing $\|\boldsymbol{\beta}\|$ to infinity leads to an optimizer of (4). In other words, the solution to (4) becomes unbounded (the MLE is at ∞) whenever there is a hyperplane perfectly separating the two sets of samples $\{i \mid \tilde{y}_i = 1\}$ and $\{i \mid \tilde{y}_i = -1\}$.

Under the assumptions from Section 2.1, $\tilde{y}_i$ is independent of $\boldsymbol{X}$ and the distribution of $\boldsymbol{X}$ is symmetric. Hence, to calculate the chance that there exists a separating hyperplane, we can assume $\tilde{y}_i = 1$ $(1 \leq i \leq n)$ without loss of generality. In this case, the event (12) becomes

$$\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathbb{R}^n_{++} \neq \emptyset, \tag{13}$$

where $\mathbb{R}^n_{++}$ is the positive orthant. Write $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ so that $\boldsymbol{Z}$ is an $n \times p$ matrix with i.i.d. standard Gaussian entries, and $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$. Then the event (13) is equivalent to

$$\{\boldsymbol{Z}\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^p\} \cap \mathbb{R}^n_{++} \neq \emptyset. \tag{14}$$

Now the probability that (14) occurs is the same as that

$$\{\boldsymbol{Z}\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^p\} \cap \mathbb{R}^n_+ \neq \{\mathbf{0}\} \tag{15}$$

occurs, where $\mathbb{R}^n_+$ denotes the non-negative orthant. From the approximate kinematic formula [3, Theorem I] in the literature on convex geometry, the event (15) happens with high probability if and only if the total statistical dimension of the two closed convex cones exceeds the ambient dimension, i.e.

$$\delta\left(\{\boldsymbol{Z}\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}\right) + \delta\left(\mathbb{R}^n_+\right) > n + o(n). \tag{16}$$

Here, the statistical dimension of a closed convex cone $\mathcal{K}$ is defined as

$$\delta(\mathcal{K}) := \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\|\Pi_{\mathcal{K}}\left(\boldsymbol{g}\right)\|^2\right] \tag{17}$$

with $\Pi_{\mathcal{K}}\left(\boldsymbol{g}\right) := \arg\min_{\boldsymbol{z}\in\mathcal{K}} \|\boldsymbol{g}-\boldsymbol{z}\|$ the Euclidean projection. Recognizing that [3, Proposition 2.4]

$$\delta\left(\{\boldsymbol{Z}\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}\right) = p \quad \text{and} \quad \delta(\mathbb{R}^n_+) = n/2,$$

we reduce the condition (16) to

$$p + n/2 > n + o(n) \qquad \text{or} \qquad p/n > 1/2 + o(1),$$

thus indicating that $\|\hat{\boldsymbol{\beta}}\| = \infty$ when $\kappa = \lim p/n > 1/2$.

This argument only reveals that $\|\hat{\boldsymbol{\beta}}\| = \infty$ in the regime where $\kappa > 1/2$. If $\kappa = p/n < 1/2$, then $\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}\| = O(1)$ with high probability, a fact we shall prove in Section 5. In light of these observations we work with the additional condition

$$\kappa < 1/2. \tag{18}$$

2.3 The high-dimensional limiting distribution of the LLR

In contrast to the classical Wilks' result, our findings reveal that the LLR statistic follows a *rescaled* chi-square distribution with a rescaling factor that can be explicitly pinned down through the solution to a system of equations.

2.3.1 A system of equations

We start by setting up the crucial system of equations. Before proceeding, we first recall the proximal operator

$$\mathsf{prox}_{b\rho}(z) := \arg\min_{x\in\mathbb{R}} \left\{ b\rho(x) + \frac{1}{2}(x-z)^2 \right\} \tag{19}$$

defined for any $b > 0$ and convex function $\rho(\cdot)$. As in [18], we introduce the operator

$$\Psi(z; b) := b\rho'(\mathsf{prox}_{b\rho}(z)), \tag{20}$$

which is simply the proximal operator of the conjugate $(b\rho)^*$ of $b\rho$.[2] To see this, we note that Ψ satisfies the relation [18, Proposition 6.4]

$$\Psi(z; b) + \mathsf{prox}_{b\rho}(z) = z. \tag{21}$$

The claim that $\Psi(\cdot; b) = \mathsf{prox}_{(b\rho)^*}(\cdot)$ then follows from the Moreau decomposition

$$\mathsf{prox}_f(z) + \mathsf{prox}_{f^*}(z) = z, \qquad \forall z, \tag{22}$$

[2]The conjugate f^* of a function f is defined as $f^*(x) = \sup_{u\in\mathrm{dom}(f)}\{\langle u, x\rangle - f(u)\}$.

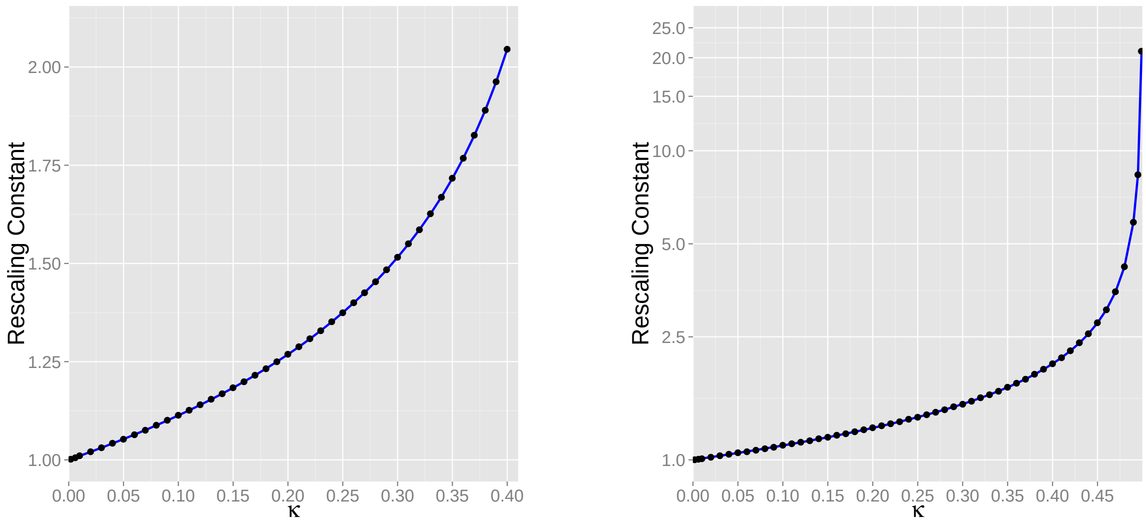


Figure 2: Rescaling constant τ_*^2/b_* as a function of κ for the logistic model. Note the logarithmic scale in the right panel. The curves for the probit model are nearly identical.

which holds for a closed convex function f [39, Section 2.5]. Interested readers are referred to [18, Appendix 1] for more properties of $\mathsf{prox}_{b\rho}$ and Ψ.

We are now in position to present the system of equations that plays a crucial role in determining the distribution of the LLR statistic in high dimensions:

$$\tau^2 = \frac{1}{\kappa}\mathbb{E}\left[(\Psi(\tau Z;\, b))^2\right], \tag{23}$$

$$\kappa = \mathbb{E}\big[\Psi'(\tau Z;\, b)\big], \tag{24}$$

where $Z \sim \mathcal{N}(0,1)$, and $\Psi'(\cdot,\cdot)$ denotes differentiation with respect to the first variable. The fact that this system of equations would admit a unique solution in $\mathbb{R}_+^2$ is not obvious *a priori*. We shall establish this for the logistic and the probit models later in Section 6.

2.3.2 Main result

Theorem 1. *Consider a logistic or probit regression model under the assumptions from Section 2.1. If $\kappa \in (0, 1/2)$, then for any $1 \le j \le p$, the log-likelihood ratio statistic Λ_j as defined in (7) obeys*

$$2\Lambda_j \ \overset{\mathrm{d}}{\to}\ \frac{\tau_*^2}{b_*}\chi_1^2, \qquad \text{as } n \to \infty, \tag{25}$$

where $(\tau_*, b_*) \in \mathbb{R}_+^2$ is the unique solution to the system of equations (23) and (24). Furthermore, the LLR statistic obtained by dropping k variables for any fixed k converges to $(\tau_*^2/b_*)\chi_k^2$. Finally, these results extend to all binary regression models with links obeying the assumptions listed in Section 2.3.3.

Hence, the limiting distribution is a rescaled chi-square with a rescaling factor τ_*^2/b_* that only depends on the aspect ratio κ. Fig. 2 illustrates the dependence of the rescaling factor on the limiting aspect ratio κ for logistic regression. The figures for the probit model are similar as the rescaling constants actually differ by very small values.

To study the quality of approximation for finite samples, we repeat the same numerical experiments as before but now obtain the p-values by comparing the LLR statistic with the rescaled chi-square suggested by

Theorem 1. For a particular run of the experiment ($n = 4000, p = 1200, \kappa = 0.3$), we compute the adjusted LLR statistic $(2b_*/\tau_*^2)\Lambda_j$ for each coefficient and obtain the p-values based on the χ_1^2 distribution. The pooled histogram that aggregates 4.8×10^5 p-values in total is shown in Fig. 1(c).

As we clearly see, the p-values are much closer to a uniform distribution now. One can compute the chi-square goodness of fit statistic to test the closeness of the above distribution to uniformity. To this end, we divide the interval $[0, 1]$ into 20 equally spaced bins of width 0.05 each. For each bin we compute the observed number of times a p-value falls in the bin out of the 4.8×10^5 values. Then a chi-square goodness of fit statistic is computed, noting that the expected frequency is 24000 for each bin. The chi-square statistic in this case is 16.049, which gives a p-value of 0.654 in comparison with a χ_{19}^2 variable. The same test when performed with the Bartlett corrected p-values (Fig. 1(b)) yields a chi-square statistic 5599 with a p-value of 0. [3] Thus, our correction gives the desired uniformity in the p-values when the true signal $\boldsymbol{\beta} = \mathbf{0}$.

Practitioners would be concerned about the validity of p-values when they are small—again, think about multiple testing applications. In order to study whether our correction yields valid results for small p-values, we compute the proportion of times the p-values (in all the three cases) lie below $5\%, 1\%, 0.5\%, 0.1\%$ out of the 4.8×10^5 times. The results are summarized in Table 1. This further illustrates the deviation from uniformity for the classical and Bartlett corrected p-values, whereas the "adjusted" p-values obtained invoking Theorem 1 are still valid.

	Classical	Bartlett-corrected	Adjusted
$\mathbb{P}\{\text{p-value} \leq 5\%\}$	11.1044%(0.0668%)	6.9592%(0.0534%)	5.0110%(0.0453%)
$\mathbb{P}\{\text{p-value} \leq 1\%\}$	3.6383%(0.038%)	1.6975%(0.0261%)	0.9944%(0.0186%)
$\mathbb{P}\{\text{p-value} \leq 0.5\%\}$	2.2477%(0.0292%)	0.9242%(0.0178%)	0.4952%(0.0116%)
$\mathbb{P}\{\text{p-value} \leq 0.1\%\}$	0.7519%(0.0155%)	0.2306%(0.0078%)	0.1008%(0.0051%)
$\mathbb{P}\{\text{p-value} \leq 0.05\%\}$	0.4669%(0.0112%)	0.124%(0.0056%)	0.0542%(0.0036%)
$\mathbb{P}\{\text{p-value} \leq 0.01\%\}$	0.1575%(0.0064%)	0.0342%(0.0027%)	0.0104%(0.0014%)

Table 1: Estimates of p-value probabilities with estimated Monte Carlo standard errors in parentheses under i.i.d. Gaussian design.

2.3.3 Extensions

As noted in Section 1.1, the Symmetry Condition (3) allows to express the negative log-likelihood in the form (4), which makes use of the effective link $\rho(\cdot)$. Theorem 1 applies to any $\rho(\cdot)$ obeying the following properties:

1. ρ is non-negative, has up to three derivatives, and obeys $\rho(t) \geq t$.
2. ρ' may be unbounded but it should grow sufficiently slowly, in particular, we assume $|\rho'(t)| = O(|t|)$ and $\rho'(\mathsf{prox}_{c\rho}(Z))$ is a sub-Gaussian random variable for any constant $c > 0$ and any $Z \sim \mathcal{N}(0, \sigma^2)$ for some finite $\sigma > 0$.
3. $\rho''(t) > 0$ for any t which implies that ρ is convex, and $\sup_t \rho''(t) < \infty$.
4. $\sup_t |\rho'''(t)| < \infty$.
5. Given any $\tau > 0$,the equation (24) has a unique solution in b.
6. The map $\mathcal{V}(\tau^2)$ as defined in (57) has a fixed point.

[3] Note that the p-values obtained at each trial are not exactly independent. However, they are exchangeable, and weakly dependent (see the proof of Corollary 1 for a formal justification of this fact). Therefore, we expect the goodness of fit test to be an approximately valid procedure in this setting.

It can be checked that the effective links for both the logistic and the probit models (5) obey *all* of the above. The last two conditions are assumed to ensure existence of a unique solution to the system of equations (23) and (24) as will be seen in Section 6; we shall justify these two conditions for the logistic and the probit models in Section 6.1.

2.4 Reduction to independent covariates

In order to derive the asymptotic distribution of the LLR statistics, it in fact suffices to consider the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

Lemma 1. *Let $\Lambda_j(\boldsymbol{X})$ be the LLR statistic based on the design matrix $\boldsymbol{X}$, where the rows of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\Lambda_j(\boldsymbol{Z})$ that where the rows are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Then*

$$\Lambda_j(\boldsymbol{X}) \stackrel{\mathrm{d}}{=} \Lambda_j(\boldsymbol{Z}).$$

Proof: Recall from (4) that the LLR statistic for testing the jth coefficient can be expressed as

$$\Lambda_j(\boldsymbol{X}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{\beta}) - \min_{\boldsymbol{\beta}:\beta_j=0} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{\beta}).$$

Write $\boldsymbol{Z}' = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ so that the rows of $\boldsymbol{Z}'$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and set $\boldsymbol{\theta}' = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$. With this reparameterization, we observe that the constraint $\beta_j = 0$ is equivalent to $\boldsymbol{a}_j^\top \boldsymbol{\theta}' = 0$ for some non-zero vector $\boldsymbol{a}_j \in \mathbb{R}^p$. This gives

$$\Lambda_j(\boldsymbol{X}) = \min_{\boldsymbol{\theta}'} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{Z}'\boldsymbol{\theta}') - \min_{\boldsymbol{\theta}':\boldsymbol{a}_j^\top\boldsymbol{\theta}'=0} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{Z}'\boldsymbol{\theta}').$$

Now let $\boldsymbol{Q}$ be an orthogonal matrix mapping $\boldsymbol{a}_j \in \mathbb{R}^p$ into the vector $\|\boldsymbol{a}_j\|\boldsymbol{e}_j \in \mathbb{R}^p$, i.e. $\boldsymbol{Q}\boldsymbol{a}_j = \|\boldsymbol{a}_j\|\boldsymbol{e}_j$. Additionally, set $\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Q}$ (the rows of $\boldsymbol{Z}$ are still i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$) and $\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\theta}'$. Since $\boldsymbol{a}_j^\top \boldsymbol{\theta}' = 0$ occurs if and only if $\theta_j = 0$, we obtain

$$\Lambda_j(\boldsymbol{X}) = \min_{\boldsymbol{\theta}} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{Z}\boldsymbol{\theta}) - \min_{\boldsymbol{\theta}:\theta_j=0} \sum_{i=1}^n \rho(-\tilde{y}_i \boldsymbol{e}_i^\top \boldsymbol{Z}\boldsymbol{\theta}) = \Lambda_j(\boldsymbol{Z}),$$

which proves the lemma. ■

In the remainder of the paper we, therefore, assume $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

2.5 Proof architecture

This section presents the main steps for proving Theorem 1. We will only prove the theorem for $\{\Lambda_j\}$, the LLR statistic obtained by dropping a single variable. The analysis for the LLR statistic obtained on dropping k variables (for some fixed k) follows very similar steps and is hence omitted for the sake of conciceness. As discussed before, we are free to work with any configuration of the y_i's. For the two steps below, we will adopt two different configurations for convenience of presentation.

2.5.1 Step 1: characterizing the asymptotic distributions of $\hat{\beta}_j$

Without loss of generality, we assume here that $y_i = 1$ (and hence $\tilde{y}_i = 1$) for all $1 \le i \le n$ and, therefore, the MLE problem reduces to

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \sum\nolimits_{i=1}^n \rho(-\boldsymbol{X}_i^\top \boldsymbol{\beta}).$$

We would first like to characterize the marginal distribution of $\hat{\boldsymbol{\beta}}$, which is crucial in understanding the LLR statistic. To this end, our analysis follows by a reduction to the setup of [18–21], with certain modifications that are called for due to the specific choices of $\rho(\cdot)$ we deal with here. Specifically, consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tag{26}$$

and prior work [18–21] investigating the associated M-estimator

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \sum\nolimits_{i=1}^{n} \rho(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}). \tag{27}$$

Our problem reduces to (27) on setting $\boldsymbol{y} = \boldsymbol{w} = \boldsymbol{0}$ in (27). When $\rho(\cdot)$ satisfies certain assumptions (e.g. strong convexity), the asymptotic distribution of $\|\hat{\boldsymbol{\beta}}\|$ has been studied in a series of works [19–21] using a leave-one-out analysis and independently in [18] using approximate message passing (AMP) machinery. An outline of their main results is described in Section 2.7. However, the function $\rho(\cdot)$ in our cases has vanishing curvature and, therefore, lacks the essential strong convexity assumption that was utilized in both the aforementioned lines of work. To circumvent this issue, we propose to invoke the AMP machinery as in [18], in conjunction with the following critical additional ingredients:

- (*Norm Bound Condition*) We utilize results from the conic geometry literature (e.g. [3]) to establish that
$$\|\hat{\boldsymbol{\beta}}\| = O(1)$$
with high probability as long as $\kappa < 1/2$. This will be elaborated in Theorem 4.
- (*Likelihood Curvature Condition*) We establish some regularity conditions on the Hessian of the log-likelihood function, generalizing the strong convexity condition, which will be detailed in Lemma 4.
- (*Uniqueness of the Solution to* (23) *and* (24)) We establish that for both the logistic and the probit case, the system of equations (23) and (24) admits a unique solution.

We emphasize that these elements are not straightforward, require significant effort and a number of novel ideas, which form our primary technical contributions for this step.

These ingredients enable the use of the AMP machinery even in the absence of strong convexity on $\rho(\cdot)$, finally leading to the following theorem:

Theorem 2. *Under the conditions of Theorem 1,*

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}\|^2 =_{\text{a.s.}} \tau_*^2. \tag{28}$$

This theorem immediately implies that the marginal distribution of $\hat{\beta}_j$ is normal.

Corollary 1. *Under the conditions of Theorem 1, for every* $1 \le j \le p$*, it holds that*

$$\sqrt{p}\hat{\beta}_j \ \stackrel{\text{d}}{\to} \ \mathcal{N}(0, \tau_*^2), \qquad \textit{as } n \to \infty. \tag{29}$$

Proof: From the rotational invariance of our i.i.d. Gaussian design, it can be easily verified that $\hat{\boldsymbol{\beta}}/\|\hat{\boldsymbol{\beta}}\|$ is uniformly distributed on the unit sphere $\mathbb{S}^{p-1}$ and is independent of $\|\hat{\boldsymbol{\beta}}\|$. Therefore, $\hat{\beta}_j$ has the same distribution as $\|\hat{\boldsymbol{\beta}}\| Z_j/\|\boldsymbol{Z}\|$, where $\boldsymbol{Z} = (Z_1, \ldots, Z_p) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ independent of $\|\hat{\boldsymbol{\beta}}\|$. Since $\sqrt{p}\|\hat{\boldsymbol{\beta}}\|/\|\boldsymbol{Z}\|$ converges in probability to τ_*, we have, by Slutsky's theorem, that $\sqrt{p}\hat{\beta}_j$ converges to $\mathcal{N}(0, \tau_*^2)$ in distribution. ∎

2.5.2 Step 2: connecting Λ_j with $\hat{\beta}_j$

Now that we have derived the asymptotic distribution of $\hat{\beta}_j$, the next step involves a reduction of the LLR statistic to a function of the relevant coordinate of the MLE. Before continuing, we note that the distribution of Λ_j is the same for all $1 \le j \le p$ due to exchangeability. As a result, going forward we will only analyze Λ_1 without loss of generality. In addition, we introduce the following convenient notations and assumptions:

- the design matrix on dropping the first column is written as $\tilde{\boldsymbol{X}}$ and the MLE in the corresponding reduced model as $\tilde{\boldsymbol{\beta}}$;
- write $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n\times p}$ and $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1, \cdots, \tilde{\boldsymbol{X}}_n]^\top \in \mathbb{R}^{n\times(p-1)}$;
- without loss of generality, assume that $\tilde{y}_i = -1$ for all i in this subsection, and hence the MLEs under the full and the reduced models reduce to

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \ell(\boldsymbol{\beta}) := \sum\nolimits_{i=1}^n \rho(\boldsymbol{X}_i^\top \boldsymbol{\beta}), \tag{30}$$

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \quad \tilde{\ell}(\boldsymbol{\beta}) := \sum\nolimits_{i=1}^n \rho(\tilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta}). \tag{31}$$

With the above notations in place, the LLR statistic for testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ can be expressed as

$$\Lambda_1 \;:=\; \tilde{\ell}(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}) \;=\; \sum_{i=1}^n \left\{\rho(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) - \rho(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}})\right\}. \tag{32}$$

To analyze Λ_1, we invoke Taylor expansion to reach

$$\begin{aligned}\Lambda_1 \;=\; &\underbrace{\sum_{i=1}^n \rho'\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)}_{:=Q_{\mathrm{lin}}} + \frac{1}{2}\sum_{i=1}^n \rho''\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^2 \\ &+ \frac{1}{6}\sum_{i=1}^n \rho'''(\gamma_i)\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^3,\end{aligned} \tag{33}$$

where γ_i lies between $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}$. A key observation is that the linear term Q_{lin} in the above equation vanishes. To see this, note that the first-order optimality conditions for the MLE $\hat{\boldsymbol{\beta}}$ is given by

$$\sum\nolimits_{i=1}^n \rho'(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}})\boldsymbol{X}_i = \boldsymbol{0}. \tag{34}$$

Replacing $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$ with $\boldsymbol{X}_i^\top \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}$ in Q_{lin} and using the optimality condition, we obtain

$$Q_{\mathrm{lin}} \;=\; \left(\sum\nolimits_{i=1}^n \rho'\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)\boldsymbol{X}_i\right)^\top \left(\begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} - \hat{\boldsymbol{\beta}}\right) \;=\; 0.$$

Consequently, Λ_1 simplifies to the following form

$$\Lambda_1 = \frac{1}{2}\sum_{i=1}^n \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}})\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^2 + \frac{1}{6}\sum_{i=1}^n \rho'''(\gamma_i)\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right)^3. \tag{35}$$

Thus, computing the asymptotic distribution of Λ_1 boils down to analyzing $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$. Our argument is inspired by the leave-one-predictor-out approach developed in [19, 20].

We re-emphasize that our setting is not covered by that of [19, 20], due to the violation of strong convexity and some other technical assumptions. We sidestep this issue by utilizing the *Norm Bound Condition* and the *Likelihood Curvature Condition*. In the end, our analysis establishes the equivalence of Λ_1 and $\hat{\beta}_1$ up to some explicit multiplicative factors modulo negligible error terms. This is summarized as follows.

Theorem 3. *Under the assumptions of Theorem 1,*

$$2\Lambda_1 - \frac{p}{b_*}\hat{\beta}_1^2 \;\overset{\mathbb{P}}{\to}\; 0, \qquad \text{as } n \to \infty. \tag{36}$$

Theorem 3 reveals a simple yet surprising connection between the LLR statistic Λ_1 and the MLE $\hat{\boldsymbol{\beta}}$. As we shall see in the proof of the theorem, the quadratic term in (35) is $\frac{1}{2}\frac{p}{b_*}\hat{\beta}_1^2 + o(1)$, while the remaining third-order term of (35) is vanishingly small. Finally, putting Corollary 1 and Theorem 3 together directly establishes Theorem 1.

2.6 Comparisons with the classical regime

We pause to shed some light on the interpretation of the correction factor τ_*^2/b_* in Theorem 1 and understand the differences from classical results. Classical theory (e.g. [30,31]) asserts that when p is fixed and n diverges, the MLE for a fixed design $\boldsymbol{X}$ is asymptotically normal, namely,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{\mathrm{d}}{\to} \mathcal{N}(0, \mathcal{I}_{\boldsymbol{\beta}}^{-1}), \tag{37}$$

where

$$\mathcal{I}_{\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{X} \qquad \text{with} \quad \boldsymbol{D}_{\boldsymbol{\beta}} := \begin{bmatrix} \rho''\left(\boldsymbol{X}_1^\top \boldsymbol{\beta}\right) & & \\ & \ddots & \\ & & \rho''\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}\right) \end{bmatrix} \tag{38}$$

is the normalized Fisher information at the true value $\boldsymbol{\beta}$. In particular, under the global null and i.i.d. Gaussian design, this converges to

$$\mathbb{E}_{\boldsymbol{X}}[\mathcal{I}_{\boldsymbol{\beta}}] = \begin{cases} \frac{1}{4}\boldsymbol{I}, & \text{for the logistic model} \\ \frac{2}{\pi}\boldsymbol{I}, & \text{for the probit model} \end{cases}$$

as n tends to infinity [50, Example 5.40].

The behavior in high dimensions is different. In particular, Corollary 1 states that under the global null, we have

$$\sqrt{p}(\hat{\beta}_j - \beta_j) \stackrel{\mathrm{d}}{\to} \mathcal{N}(0, \tau_*^2). \tag{39}$$

Comparing the variances in the logistic model, we have that

$$\lim_{n\to\infty} \mathsf{Var}\left(\sqrt{p}\hat{\beta}_j\right) = \begin{cases} 4\kappa, & \text{in classical large-sample theory;} \\ \tau_*^2, & \text{in high dimensions.} \end{cases}$$

Fig. 3 illustrates the behavior of the ratio τ_*^2/κ as a function of κ. Two observations are immediate:

- First, in Fig. 3(a) we have $\tau_*^2 \geq 4\kappa$ for all $\kappa \geq 0$. This indicates an inflation in variance or an "extra Gaussian noise" component that appears in high dimensions, as discussed in [18]. The variance of the "extra Gaussian noise" component increases as κ grows.
- Second, as $\kappa \to 0$, we have $\tau_*^2/4\kappa \to 1$ in the logistic model, which indicates that classical theory becomes accurate in this case. In other words, our theory recovers the classical prediction in the regime where $p = o(n)$.

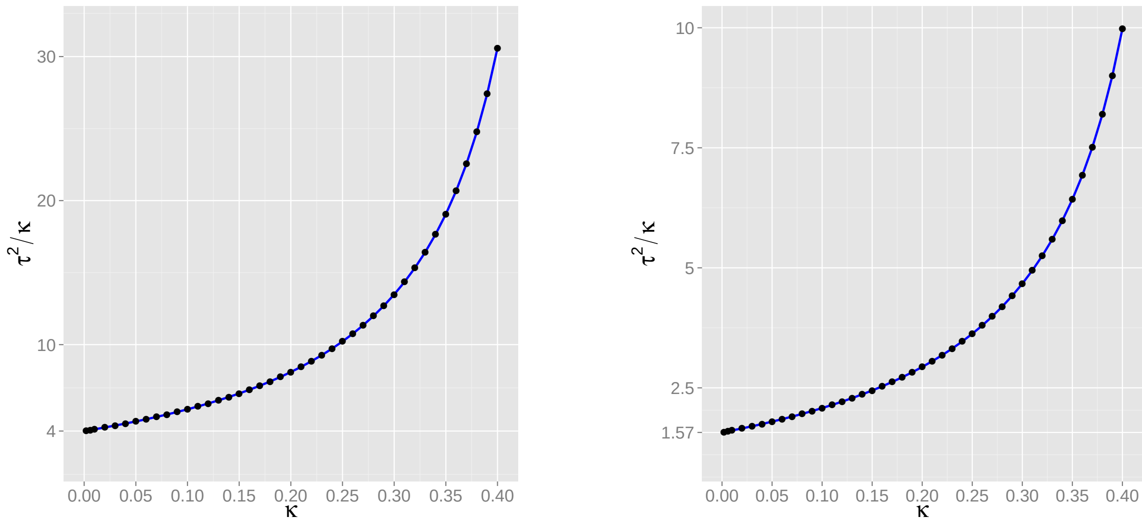
Further, for the testing problem considered here, the LLR statistic in the classical setup can be expressed, through Taylor expansion, as

$$2\Lambda_1 = n(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \mathcal{I}_{\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1), \tag{40}$$

where $\tilde{\boldsymbol{\beta}}$ is defined in (31). In the high-dimensional setting, we will also establish a quadratic approximation of the form

$$2\Lambda_1 = n(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \boldsymbol{G} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + o_P(1), \qquad \boldsymbol{G} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}_{\hat{\boldsymbol{\beta}}} \boldsymbol{X}.$$

In Theorem 7, we shall see that b_* is the limit of $\frac{1}{n}\mathrm{Tr}(\boldsymbol{G}^{-1})$, the Stieltjes transform of the empirical spectral distribution of $\boldsymbol{G}$ evaluated at 0. Thus, this quantity in some sense captures the spread in the eigenvalues of $\boldsymbol{G}$ one would expect to happen in high dimensions.



(a) logistic regression

(b) probit regression

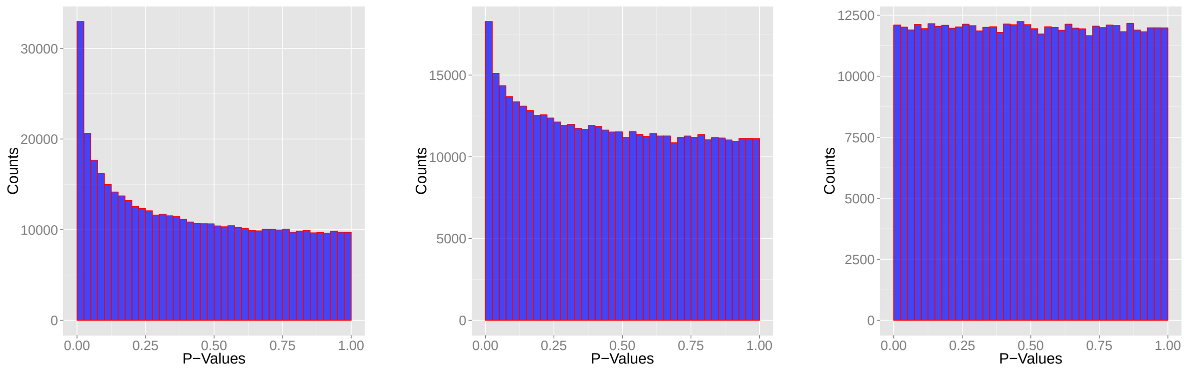
Figure 3: Ratio of asymptotic variance and dimensionality factor κ as a function of κ.

2.7 Prior art

Wilks' type of phenomenon in the presence of a diverging dimension p has received much attention in the past. For instance, Portnoy [43] investigated simple hypotheses in regular exponential families, and established the asymptotic chi-square approximation for the LLR test statistic as long as $p^{3/2}/n \to 0$. This phenomenon was later extended in [46] to accommodate the MLE with a quadratic penalization, and in [53] to account for parametric models underlying several random graph models. Going beyond parametric inference, Fan et al. [22, 24] explored extensions to infinite-dimensional non-parametric inference problems, for which the MLE might not even exist or might be difficult to derive. While the classical Wilks' phenomenon fails to hold in such settings, Fan et al. [22, 24] proposed a generalization of the likelihood ratio statistics based on suitable non-parametric estimators and characterized the asymptotic distributions. Such results have further motivated Boucheron and Massart [9] to investigate the non-asymptotic Wilks' phenomenon or, more precisely, the concentration behavior of the difference between the excess empirical risk and the true risk, from a statistical learning theory perspective. The Wilks' phenomenon for penalized empirical likelihood has also been established [48]. However, the precise asymptotic behavior of the LLR statistic in the regime that permits p to grow proportional to n is still beyond reach.

On the other hand, as demonstrated in Section 2.5.1, the MLE here under the global null can be viewed as an M-estimator for a linear regression problem. Questions regarding the behavior of robust linear regression estimators in high dimensions—where p is allowed to grow with n—were raised in Huber [30], and have been extensively studied in subsequent works, e.g. [36, 40–42]. When it comes to logistic regression, the behavior of the MLE was studied for a diverging number of parameters by [28], which characterized the squared estimation error of the MLE if $(p \log p)/n \to 0$. In addition, the asymptotic normality properties of the MLE and the penalized MLE for logistic regression have been established by [35] and [23], respectively. A very recent paper by Fan et al. [25] studied the logistic model under the global null $\boldsymbol{\beta} = \mathbf{0}$, and investigated the classical asymptotic normality as given in (37). It was discovered in [25] that the convergence property (37) breaks down even in terms of the marginal distribution, namely,

$$\frac{\sqrt{n}\hat{\beta}_i}{\left(\mathcal{I}_{\boldsymbol{\beta}}\right)_{i,i}^{-1/2}} \overset{\mathrm{d}}{\nrightarrow} \mathcal{N}(0,1), \qquad \mathcal{I}_{\boldsymbol{\beta}} = \frac{1}{4n}\boldsymbol{X}^\top \boldsymbol{X},$$



(a) (b) (c)

Figure 4: Histogram of p-values for logistic regression under i.i.d. Bernoulli design, when $\boldsymbol{\beta} = \mathbf{0}$, $n = 4000$, $p = 1200$, and $\kappa = 0.3$: (a) classically computed p-values; (b) Bartlett corrected p-values; (c) adjusted p-values.

as soon as p grows at a rate exceeding $n^{2/3}$. This result, however, does not imply the asymptotic distribution of the likelihood-ratio statistic in this regime. In fact, our theorem implies that LLR statistic $2\Lambda_j$ goes to χ^2_1 (and hence Wilks phenomenon remains valid) when $\kappa = p/n \to 0$.

The line of work that is most relevant to the present paper was initially started by El Karoui et al. [21]. Focusing on the regime where p is comparable to n, the authors uncovered, via a non-rigorous argument, that the asymptotic ℓ_2 error of the MLE could be characterized by a system of nonlinear equations. This seminal result was later made rigorous independently by Donoho et al. [18] under i.i.d. Gaussian design and by El Karoui [19, 20] under more general i.i.d. random design as well as certain assumptions on the error distribution. Both approaches rely on strong convexity on the function $\rho(\cdot)$ that defines the M-estimator, which does not hold in the models considered herein.

2.8 Notations

We adopt the standard notation $f(n) = O\left(g(n)\right)$ or $f(n) \lesssim g(n)$ which means that there exists a constant $c > 0$ such that $|f(n)| \leq c|g(n)|$. Likewise, $f(n) = \Omega\left(g(n)\right)$ or $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \geq c\,|g(n)|$, $f(n) \asymp g(n)$ means that there exist constants $c_1, c_2 > 0$ such that $c_1|g(n)| \leq |f(n)| \leq c_2|g(n)|$, and $f(n) = o(g(n))$ means that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. Any mention of C, C_i, c, c_i for $i \in \mathbb{N}$ refers to some positive universal constants whose value may change from line to line. For a square symmetric matrix $\boldsymbol{M}$, the minimum eigenvalue is denoted by $\lambda_{\min}(\boldsymbol{M})$. Logarithms are base e.

3 Numerics

3.1 Non-Gaussian covariates

In this section we first study the sensitivity of our result to the Gaussianity assumption on the design matrix. To this end, we consider a high dimensional binary regression set up with a Bernoulli design matrix. We simulate $n = 4000$ i.i.d. observations $(y_i, \boldsymbol{X}_i)$ with $y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(1/2)$, and $\boldsymbol{X}_i$ generated independent of y_i, such that each entry takes on values in $\{1, -1\}$ w.p. $1/2$. At each trial, we fit a logistic regression model to the data and obtain the classical, Bartlett corrected and adjusted p-values (using the rescaling factor τ_*^2/b_*). Figure 4 plots the histograms for the pooled p-values, obtained across 400 trials.

It is instructive to compare the histograms to that obtained in the Gaussian case (Figure 1). The classical and Bartlett corrected p-values exhibit similar deviations from uniformity as in the Gaussian design case, whereas our adjusted p-values continue to have an approximate uniform distribution. We test for deviations from uniformity using a formal chi-squared goodness of fit test as in Section 2.3.2. For the Bartlett corrected p-values, the chi-squared statistic turns out to be 5885, with a p-value 0. For the adjusted p-values,the chi-squared statistic is 24.1024, with a p-value 0.1922.[4]

Once again, the Bartlett correction fails to provide valid p-values whereas the adjusted p-values are consistent with a uniform distribution. These findings indicate that the distribution of the LLR statistic under the i.i.d. Bernoulli design is in agreement to the rescaled χ_1^2 derived under the Gaussian design in Theorem 1, suggesting that the distribution is not too sensitive to the Gaussianity assumption. Estimates of p-value probabilities for our method are provided in Table 2.

	Adjusted
$\mathbb{P}\{\text{p-value} \leq 5\%\}$	5.0222%(0.0412%)
$\mathbb{P}\{\text{p-value} \leq 1\%\}$	1.0048%(0.0174%)
$\mathbb{P}\{\text{p-value} \leq 0.5\%\}$	0.5123%(0.0119%)
$\mathbb{P}\{\text{p-value} \leq 0.1\%\}$	0.1108%(0.005%)
$\mathbb{P}\{\text{p-value} \leq 0.05\%\}$	0.0521%(0.0033%)
$\mathbb{P}\{\text{p-value} \leq 0.01\%\}$	0.0102%(0.0015%)

Table 2: Estimates of p-value probabilities with estimated Monte Carlo standard errors in parentheses under i.i.d. Bernoulli design.

3.2 Quality of approximations for finite sample sizes

In the rest of this section, we report some numerical experiments which study the applicability of our theory in finite sample setups.

Validity of tail approximation The first experiment explores the efficacy of our correction for extremely small p-values. This is particularly important in the context of multiple comparisons, where practitioners care about the validity of exceedingly small p-values. To this end, the empirical cumulative distribution of the adjusted p-values is estimated under a standard Gaussian design with $n = 4000$, $p = 1200$ and 4.8×10^5 p-values. The range $[0.1/p, 12/p]$ is divided into points which are equi-spaced with a distance of $1/p$ between any two consecutive points. The estimated empirical CDF at each of these points is represented in blue in Figure 5. The estimated CDF is in near-perfect agreement with the diagonal, suggesting that the adjusted p-values computed using the rescaled chi-square distribution are remarkably close to a uniform, even when we zoom in at very small resolutions as would be the case when applying Bonferroni-style corrections.

Moderate sample sizes The final experiment studies the accuracy of our asymptotic result for moderately large samples. This is especially relevant for applications where the sample sizes are not too large. We repeat our numerical experiments with $n = 200$, $p = 60$ for i.i.d. Gaussian design, and 4.8×10^5 p-values. The empirical CDF for these p-values are estimated and Figure 6 shows that the adjusted p-values are nearly uniformly distributed even for moderate sample sizes such as $n = 200$.

4 Preliminaries

This section gathers a few preliminary results that will be useful throughout the paper. We start by collecting some facts regarding i.i.d. Gaussian random matrices.

[4]Recall our earlier footnote about the use of a χ^2 test.

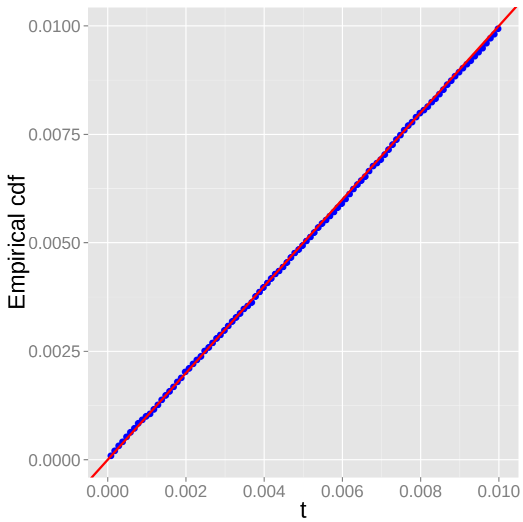


Figure 5: Empirical CDF of adjusted pvalues for logistic regression when $\boldsymbol{\beta} = \mathbf{0}$, $n = 4000$, $p = 1200$. Here, the blue points represent the empirical CDF (t vs. the fraction of p-values below t), and the red line is the diagonal.

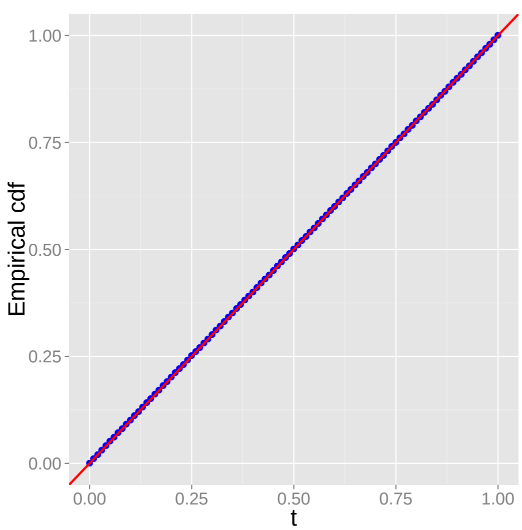


Figure 6: Empirical CDF of adjusted pvalues for logistic regression when $\boldsymbol{\beta} = \mathbf{0}$, $n = 200$, $p = 60$. Here, the blue points represent the empirical CDF (t vs. the fraction of p-values below t), and the red line is the diagonal.

Lemma 2. *Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n]^\top$ be an $n \times p$ matrix with i.i.d. standard Gaussian entries. Then*

$$\mathbb{P}\left(\|\boldsymbol{X}^\top \boldsymbol{X}\| \leq 9n\right) \geq 1 - 2\exp(-n/2); \tag{41}$$

$$\mathbb{P}\left(\sup_{1\leq i\leq n} \|\boldsymbol{X}_i\| \leq 2\sqrt{p}\right) \geq 1 - 2n\exp(-(\sqrt{p}-1)^2/2). \tag{42}$$

Proof: This is a straighforward application of [51, Corollary 5.35] and the union bound. ■

Lemma 3. *Suppose $\boldsymbol{X}$ is an $n \times p$ matrix with entries i.i.d $\mathcal{N}(0,1)$, then there exists a constant ϵ_0 such that whenever $0 \leq \epsilon \leq \epsilon_0$ and $0 \leq t \leq \sqrt{1-\epsilon} - \sqrt{p/n}$,*

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i\in S} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \geq \left(\sqrt{1-\epsilon} - \sqrt{\frac{p}{n}} - t\right)^2, \quad \forall S \subseteq [n] \text{ with } |S| = (1-\epsilon)n \tag{43}$$

with probability exceeding $1 - 2\exp\left(-\left(\frac{(1-\epsilon)t^2}{2} - H(\epsilon)\right)n\right)$. Here, $H(\epsilon) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$.

Proof: See Appendix A.1. ■

The above facts are useful in establishing an eigenvalue lower bound on the Hessian of the log-likelihood function. Specifically, recall that

$$\nabla^2 \ell(\boldsymbol{\beta}) = \sum_{i=1}^n \rho''\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right) \boldsymbol{X}_i \boldsymbol{X}_i^\top, \tag{44}$$

and the result is this:

Lemma 4 (Likelihood Curvature Condition)**.** *Suppose that $p/n < 1$ and that $\rho''(\cdot) \geq 0$. Then there exists a constant ϵ_0 such that whenever $0 \leq \epsilon \leq \epsilon_0$, with probability at least $1 - 2\exp(-nH(\epsilon)) - 2\exp(-n/2)$, the matrix inequality*

$$\frac{1}{n}\nabla^2 \ell(\boldsymbol{\beta}) \succeq \left(\inf_{z:|z|\leq \frac{3\|\boldsymbol{\beta}\|}{\sqrt{\epsilon}}} \rho''(z)\right)\left(\sqrt{1-\epsilon} - \sqrt{\frac{p}{n}} - 2\sqrt{\frac{H(\epsilon)}{1-\epsilon}}\right)^2 \boldsymbol{I} \tag{45}$$

holds simultaneously for all $\boldsymbol{\beta} \in \mathbb{R}^p$.

Proof: See Appendix A.2. ■

The message of Lemma 4 is this: take $\epsilon > 0$ to be a sufficiently small constant. Then

$$\frac{1}{n}\nabla^2 \ell(\boldsymbol{\beta}) \succeq \omega(\|\boldsymbol{\beta}\|)\, \boldsymbol{I}$$

for some non-increasing and positive function $\omega(\cdot)$ independent of n. This is a generalization of the strong convexity condition.

5 When is the MLE bounded?

5.1 Phase transition

In Section 2.2, we argued that the MLE is at infinity if we have less than two observations per dimension or $\kappa > 1/2$. In fact, a stronger version of the phase transition phenemonon occurs in the sense that

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

as soon as $\kappa < 1/2$. This is formalized in the following theorem.

Theorem 4 (Norm Bound Condition)**.** *Fix any small constant $\epsilon > 0$, and let $\hat{\boldsymbol{\beta}}$ be the MLE for a model with effective link satisfying the conditions from Section 2.3.3.*

(i) If $p/n \geq 1/2 + \epsilon$, then

$$\|\hat{\boldsymbol{\beta}}\| = \infty$$

with probability exceeding $1 - 4\exp\left(-\epsilon^2 n/8\right)$.

(ii) There exist universal constants $c_1, c_2, C_2 > 0$ such that if $p/n < 1/2 - c_1\epsilon^{3/4}$, then[5]

$$\|\hat{\boldsymbol{\beta}}\| < \frac{4\log 2}{\epsilon^2}$$

with probability at least $1 - C_2\exp(-c_2\epsilon^2 n)$.

These conclusions clearly continue to hold if $\hat{\boldsymbol{\beta}}$ is replaced by $\tilde{\boldsymbol{\beta}}$ (the MLE under the restricted model obtained on dropping the first predictor).

The rest of this section is devoted to proving this theorem. As we will see later, the fact that $\|\hat{\boldsymbol{\beta}}\| = O(1)$ is crucial for utilizing the AMP machinery in the absence of strong convexity.

5.2 Proof of Theorem 4

As in Section 2.5.1, we assume $\tilde{y}_i \equiv 1$ throughout this section, and hence the MLE reduces to

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \ell_0\left(\boldsymbol{\beta}\right) := \sum\nolimits_{i=1}^{n} \rho(-\boldsymbol{X}_i^\top \boldsymbol{\beta}). \tag{46}$$

5.2.1 Proof of Part (i)

Invoking [3, Theorem I] yields that if

$$\delta\left(\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}\right) + \delta\left(\mathbb{R}_+^n\right) \geq \left(1+\epsilon\right)n,$$

or equivalently, if $p/n \geq 1/2 + \epsilon$, then

$$\mathbb{P}\left\{\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathbb{R}_+^n \neq \{\mathbf{0}\}\right\} \geq 1 - 4\exp\left(-\epsilon^2 n/8\right).$$

As is seen in Section 2.2, $\|\hat{\boldsymbol{\beta}}\| = \infty$ when $\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathbb{R}_+^n \neq \{\mathbf{0}\}$, establishing Part (i) of Theorem 4.

5.2.2 Proof of Part (ii)

We now turn to the regime in which $p/n \leq 1/2 - O(\epsilon^{3/4})$, where $0 < \epsilon < 1$ is any fixed constant. Begin by observing that the least singular value of $\boldsymbol{X}$ obeys

$$\sigma_{\min}\left(\boldsymbol{X}\right) \geq \sqrt{n}/4 \tag{47}$$

with probability at least $1 - 2\exp\left(-\frac{1}{2}\left(\frac{3}{4} - \frac{1}{\sqrt{2}}\right)^2 n\right)$ (this follows from Lemma 3 using $\epsilon = 0$). Then for any $\boldsymbol{\beta} \in \mathbb{R}^p$ obeying

$$\ell_0(\boldsymbol{\beta}) = \sum\nolimits_{j=1}^{n} \rho\left(-\boldsymbol{X}_j^\top \boldsymbol{\beta}\right) \leq n\log 2 = \ell_0(\mathbf{0}) \tag{48}$$

$$\text{and} \qquad \|\boldsymbol{\beta}\| \geq \frac{4\log 2}{\epsilon^2}, \tag{49}$$

[5]When $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for a general $\boldsymbol{\Sigma} \succ \mathbf{0}$, one has $\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}\| \lesssim 1/\epsilon^2$ with high probability.

we must have

$$\sum_{j=1}^{n} \max\left\{-\boldsymbol{X}_j^\top \boldsymbol{\beta},\ 0\right\} = \sum_{j:\ \boldsymbol{X}_j^\top \boldsymbol{\beta}<0} \left(-\boldsymbol{X}_j^\top \boldsymbol{\beta}\right) \overset{\text{(a)}}{\leq} \sum_{j:\ \boldsymbol{X}_j^\top \boldsymbol{\beta}<0} \rho\left(-\boldsymbol{X}_j^\top \boldsymbol{\beta}\right) \overset{\text{(b)}}{\leq} n\log 2;$$

(a) follows since $t \leq \rho(t)$ and (b) is a consequence of (48). Continuing, (47) and (49) give

$$n \log 2 \leq 4\sqrt{n}\frac{\|\boldsymbol{X}\boldsymbol{\beta}\|}{\|\boldsymbol{\beta}\|}\log 2 \leq \epsilon^2 \sqrt{n}\|\boldsymbol{X}\boldsymbol{\beta}\|.$$

This implies the following proposition: if the solution $\hat{\boldsymbol{\beta}}$—which necessarily satisfies $\ell_0(\hat{\boldsymbol{\beta}}) \leq \ell_0(\mathbf{0})$—has norm exceeding $\|\hat{\boldsymbol{\beta}}\| \geq \frac{4\log 2}{\epsilon^2}$, then $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ must fall within the cone

$$\mathcal{A} := \left\{\boldsymbol{u} \in \mathbb{R}^n \;\middle|\; \sum\nolimits_{j=1}^{n} \max\left\{-u_j, 0\right\} \leq \epsilon^2\sqrt{n}\|\boldsymbol{u}\| \right\}. \tag{50}$$

Therefore, if one wishes to rule out the possibility of having $\|\hat{\boldsymbol{\beta}}\| \geq \frac{4\log 2}{\epsilon^2}$, it suffices to show that with high probability,

$$\left\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\right\} \cap \mathcal{A} = \{\mathbf{0}\}. \tag{51}$$

This is the content of the remaining proof.

We would like to utilize tools from conic geometry [3] to analyze the probability of the event (51). Note, however, that $\mathcal{A}$ is not convex, while the theory developed in [3] applies only to convex cones. To bypass the non-convexity issue, we proceed in the following three steps:

1. Generate a set of $N = \exp\left(2\epsilon^2 p\right)$ closed *convex* cones $\{\mathcal{B}_i \mid 1 \leq i \leq N\}$ such that it forms a cover of $\mathcal{A}$ with probability exceeding $1 - \exp\left(-\Omega(\epsilon^2 p)\right)$.

2. Show that if $p < \left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 2H(2\sqrt{\epsilon})\right)n$ and if n is sufficiently large, then

   $$\mathbb{P}\left\{\left\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\right\} \cap \mathcal{B}_i \neq \{\mathbf{0}\}\right\} \leq 4\exp\left\{-\frac{1}{8}\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \frac{p}{n}\right)^2 n\right\}$$

   for each $1 \leq i \leq N$.

3. Invoke the union bound to reach

   $$\begin{aligned} \mathbb{P}\left\{\left\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\right\} \cap \mathcal{A} \neq \{\mathbf{0}\}\right\} &\leq \mathbb{P}\left\{\{\mathcal{B}_i \mid 1 \leq i \leq N\} \text{ does not form a cover of } \mathcal{A}\right\} \\ &\quad + \sum_{i=1}^{N} \mathbb{P}\left\{\left\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\right\} \cap \mathcal{B}_i \neq \{\mathbf{0}\}\right\} \\ &\leq \exp\left(-\Omega(\epsilon^2 p)\right), \end{aligned}$$

   where we have used the fact that

   $$\begin{aligned} \sum_{i=1}^{N} \mathbb{P}\left\{\left\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\right\} \cap \mathcal{B}_i \neq \{\mathbf{0}\}\right\} &\leq 4N\exp\left\{-\frac{1}{8}\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \frac{p}{n}\right)^2 n\right\} \\ &< 4\exp\left\{-\left(\frac{1}{8}\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \frac{p}{n}\right)^2 - 2\epsilon^2\right)n\right\} \\ &< 4\exp\left\{-\epsilon^2 n\right\}. \end{aligned}$$

   Here, the last inequality holds if $\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \frac{p}{n}\right)^2 > 24\epsilon^2$, or equivalently, $\frac{p}{n} < \frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \sqrt{24}\epsilon$.

Taken collectively, these steps establish the following claim: if $\frac{p}{n} < \frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \sqrt{24}\epsilon$, then

$$\mathbb{P}\left\{\|\hat{\boldsymbol{\beta}}\| > \frac{4\log 2}{\epsilon^2}\right\} \quad < \quad \exp\left\{-\Omega(\epsilon^2 n)\right\},$$

thus establishing Part (ii) of Theorem 4. We defer the complete details of the preceding steps to Appendix D.

6 Asymptotic ℓ_2 error of the MLE

This section aims to establish Theorem 2, which characterizes precisely the asymptotic squared error of the MLE $\hat{\boldsymbol{\beta}}$ under the global null $\boldsymbol{\beta} = \mathbf{0}$. As described in Section 2.5.1, it suffices to assume that $\hat{\boldsymbol{\beta}}$ is the solution to the following problem

$$\text{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \sum\nolimits_{i=1}^{n} \rho(-\boldsymbol{X}_i^\top \boldsymbol{\beta}). \tag{52}$$

In what follows, we derive the asymptotic convergence of $\|\hat{\boldsymbol{\beta}}\|$ under the assumptions from our main theorem.

Theorem 5. *Under the assumptions of Theorem 1, the solution $\hat{\boldsymbol{\beta}}$ to (52) obeys*

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}\|^2 =_{\text{a.s.}} \tau_*^2. \tag{53}$$

Theorem 5 is derived by invoking the AMP machinery [6, 7, 32]. The high-level idea is the following: in order to study $\hat{\boldsymbol{\beta}}$, one introduces an iterative algorithm (called AMP) where a sequence of iterates $\hat{\boldsymbol{\beta}}^t$ is formed at each time t. The algorithm is constructed so that the iterates asymptotically converge to the MLE in the sense that

$$\lim_{t\to\infty}\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|^2 =_{\text{a.s.}} 0. \tag{54}$$

On the other hand, the asymptotic behavior (asymptotic in n) of $\hat{\boldsymbol{\beta}}^t$ for each t can be described accurately by a scalar sequence $\{\tau_t\}$—called *state evolution* (SE)—following certain update equations [6]. This, in turn, provides a characterization of the ℓ_2 loss of $\hat{\boldsymbol{\beta}}$.

Further, in order to prove Theorem 2, one still needs to justify

(a) the existence of a solution to the system of equations (23) and (24),

(b) and the existence of a fixed point for the iterative map governing the SE sequence updates.

We will elaborate on these steps in the rest of this section.

6.1 State evolution

We begin with the SE sequence $\{\tau_t\}$ introduced in [18]. Starting from some initial point τ_0, we produce two sequences $\{b_t\}$ and $\{\tau_t\}$ following a two-step procedure.

- For $t = 0, 1, \ldots$:
  - Set b_t to be the solution in b to

$$\kappa = \mathbb{E}\left[\Psi'(\tau_t Z; b)\right]; \tag{55}$$

  - Set τ_{t+1} to be

$$\tau_{t+1}^2 = \frac{1}{\kappa}\,\mathbb{E}\left[(\Psi^2(\tau_t Z; b_t))\right]. \tag{56}$$

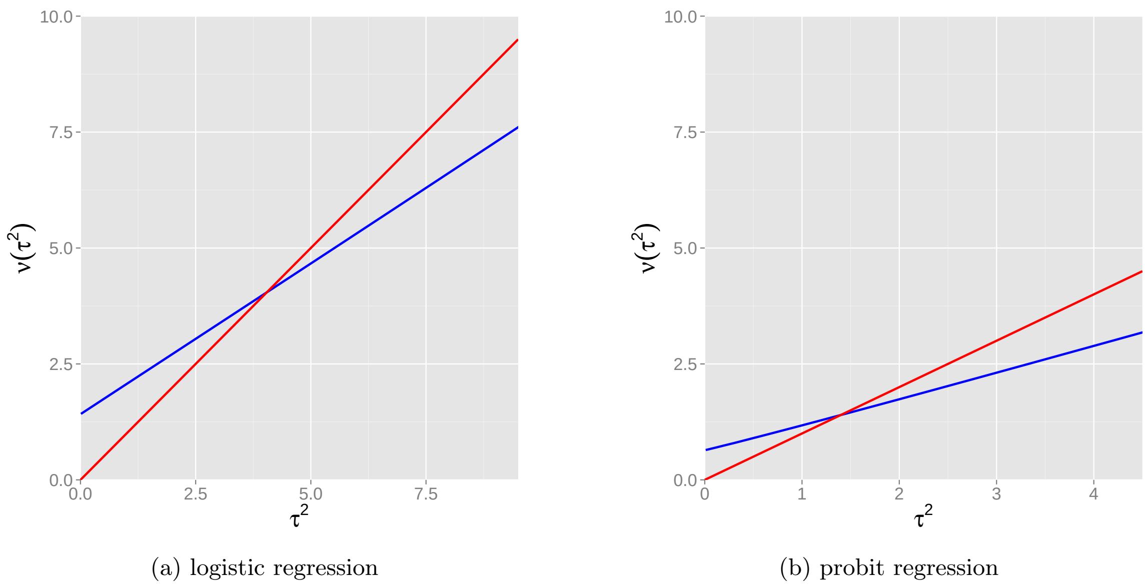


Figure 7: The variance map for both the logistic and the probit models when $\kappa = 0.3$: (blue line) variance map $\mathcal{V}(\tau^2)$ as a function of τ^2; (red line) diagonal.

Suppose that for given any $\tau > 0$, the solution in b to (55) with $\tau_t = \tau$ exists and is unique, then one can denote the solution as $b(\tau)$, which in turn allows one to write the sequence $\{\tau_t\}$ as

$$\tau_{t+1}^2 = \mathcal{V}(\tau_t^2)$$

with the variance map

$$\mathcal{V}(\tau^2) = \frac{1}{\kappa}\,\mathbb{E}\left[\Psi^2(\tau Z; b(\tau))\right]. \tag{57}$$

As a result, if there exists a fixed point τ_* obeying $\mathcal{V}(\tau_*^2) = \tau_*^2$ and if we start with $\tau_0 = \tau_*$, then by induction,

$$\tau_t \equiv \tau_* \quad \text{and} \quad b_t \equiv b_* := b(\tau_*), \qquad t = 0, 1, \ldots$$

Notably, (τ_*, b_*) solves the system of equations (23) and (24). We shall work with this choice of initial condition throughout our proof.

The preceding arguments hold under two conditions: (i) the solution to (24) exists and is unique for any $\tau_t > 0$; (ii) the variance map (57) admits a fixed point. To verify these two conditions, we make two observations.

- Condition (i) holds if one can show that the function

$$G(b) := \mathbb{E}\left[\Psi'(\tau Z; b)\right], \qquad b > 0 \tag{58}$$

  is strictly monotone for any given $\tau > 0$, and that $\lim_{b\to 0} G(b) < \kappa < \lim_{b\to\infty} G(b)$.

- Since $\mathcal{V}(\cdot)$ is a continuous function, Condition (ii) becomes self-evident once we show that $\mathcal{V}(0) > 0$ and that there exists $\tau > 0$ obeying $\mathcal{V}(\tau^2) < \tau^2$. The behavior of the variance map is illustrated in Figure 7 for the logistic and probit regression when $\kappa = 0.3$. One can in fact observe that the fixed point is unique. For other values of κ, the variance map shows the same behavior.

In fact, the aforementioned properties can be proved for a certain class of effective links, as summarized in the following lemmas. In particular, they can be shown for the logistic and the probit models.

Lemma 5. *Suppose the effective link ρ satisfies the following two properties:*

(a) ρ' *is log-concave.*

(b) For any fixed $\tau > 0$ *and any fixed* z, $b\rho''(\mathsf{prox}_{b\rho}(\tau z)) \to \infty$ *when* $b \to \infty$.

Then for any $\tau > 0$, *the function* $G(b)$ *defined in (58) is an increasing function in* b ($b > 0$), *and the equation*

$$G(b) = \kappa$$

has a unique positive solution.

Proof: See Appendix B. ■

Lemma 6. *Suppose that* $0 < \kappa < 1/2$ *and that* $\rho = \log(1 + e^t)$ *or* $\rho = -\log \Phi(-t)$. *Then*

(i) $\mathcal{V}(0) > 0$;

(ii) $\mathcal{V}(\tau^2) < \tau^2$ *for some sufficiently large* τ^2.

Proof: See Appendix C and the supplemental material [47]. ■

Remark 1. A byproduct of the proof is that the following relations hold for any constant $0 < \kappa < 1/2$:

- In the logistic case,

$$\begin{cases} \lim_{\tau\to\infty} \frac{\mathcal{V}(\tau^2)}{\tau^2} & = \left.\frac{x^2\mathbb{P}\{Z>x\}+\mathbb{E}\left[Z^2\mathbf{1}_{\{0<Z<x\}}\right]}{\mathbb{P}\{0<Z<x\}}\right|_{x=\Phi^{-1}(\kappa+0.5)}; \\ \lim_{\tau\to\infty} \frac{b(\tau)}{\tau} & = \Phi^{-1}(\kappa + 0.5). \end{cases}$$

- In the probit case,

$$\lim_{\tau\to\infty} b(\tau) = \frac{2\kappa}{1-2\kappa} \qquad \text{and} \qquad \lim_{\tau\to\infty} \frac{\mathcal{V}(\tau^2)}{\tau^2} = 2\kappa. \tag{59}$$

Remark 2. Lemma 6 is proved for the two special effective link functions, the logistic and the probit cases. However, the proof sheds light on general conditions on the effective link that suffice for the lemma to hold. Such general sufficient conditions are also discussed in the supplemental material [47].

6.2 AMP recursion

In this section, we construct the AMP trajectory tracked by two sequences $\{\hat{\boldsymbol{\beta}}^t(n) \in \mathbb{R}^p\}$ and $\{\boldsymbol{\eta}^t(n) \in \mathbb{R}^n\}$ for $t \geq 0$. Going forward we suppress the dependence on n to simplify presentation. Picking $\hat{\boldsymbol{\beta}}^0$ such that

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}^0\|^2 = \tau_0^2 = \tau_*^2$$

and taking $\boldsymbol{\eta}^{-1} = \mathbf{0}$ and $b_{-1} = 0$, the AMP path is obtained via Algorithm 1, which is adapted from the algorithm in [18, Section 2.2].

Algorithm 1 Approximate message passing.

For $t = 0, 1, \cdots$**:**

1. Set
$$\boldsymbol{\eta}^t = \boldsymbol{X}\hat{\boldsymbol{\beta}}^t + \Psi\left(\boldsymbol{\eta}^{t-1}; b_{t-1}\right); \tag{60}$$

2. Let b_t be the solution to
$$\kappa = \mathbb{E}\left[\Psi'(\tau_t Z; b)\right], \tag{61}$$
where τ_t is the SE sequence value at that time.

3. Set
$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t - \frac{1}{p}\boldsymbol{X}^\top \Psi\left(\boldsymbol{\eta}^t; b_t\right). \tag{62}$$

Here, $\Psi(\cdot)$ is applied in an entrywise manner, and $\Psi'(.,.)$ denotes derivative w.r.t the first variable.

As asserted by [18], the SE sequence $\{\tau_t\}$ introduced in Section 6.1 proves useful as it offers a formal procedure for predicting operating characteristics of the AMP iterates at any fixed iteration. In particular it assigns predictions to two types of observables: observables which are functions of the $\hat{\boldsymbol{\beta}}^t$ sequence and those which are functions of $\boldsymbol{\eta}^t$. Repeating identical argument as in [18, Theorem 3.4], we obtain

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}^t\|^2 =_{\text{a.s.}} \tau_t^2 \equiv \tau_*^2, \qquad t = 0, 1, \ldots. \tag{63}$$

6.3 AMP converges to the MLE

We are now in position to show that the AMP iterates $\{\hat{\boldsymbol{\beta}}^t\}$ converge to the MLE in the large n and t limit. Before continuing, we state below two properties that are satisfied under our assumptions.

- The MLE $\hat{\boldsymbol{\beta}}$ obeys
$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}\| < \infty \tag{64}$$
almost surely.

- And there exists some non-increasing continuous function $0 < \omega(\cdot) < 1$ independent of n such that
$$\mathbb{P}\left\{\frac{1}{n}\nabla^2 \ell(\boldsymbol{\beta}) \succeq \omega(\|\boldsymbol{\beta}\|)\cdot \boldsymbol{I},\ \forall \boldsymbol{\beta}\right\} \geq 1 - c_1 e^{-c_2 n}. \tag{65}$$

In fact, the norm bound (64) follows from Theorem 4 together with Borel-Cantelli, while the likelihood curvature condition (65) is an immediate consequence of Lemma 4. With this in place, we have:

Theorem 6. *Suppose (64) and (65) hold. Let (τ_*, b_*) be a solution to the system (23) and (24), and assume that $\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}^0\|^2 = \tau_*^2$. Then the AMP trajectory as defined in Algorithm 1 obeys*

$$\lim_{t\to\infty}\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\| =_{\text{a.s.}} 0.$$

Taken collectively, Theorem 6 and Eqn. (63) imply that

$$\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}\| =_{\text{a.s.}} \lim_{t\to\infty}\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}^t\| =_{\text{a.s.}} \tau_*, \tag{66}$$

thus establishing Theorem 5. In addition, an upshot of these theorems is a uniqueness result:

Corollary 2. *The solution to the system of equations (23) and (24) is unique.*

Proof: When the AMP trajectory $\hat{\boldsymbol{\beta}}^t$ is started with the initial condition from Theorem 6, $\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}\|^2 =_{\text{a.s.}} \tau_*^2$. This holds for any τ_* such that (τ_*, b_*) is a solution to (23) and (24). However, since the MLE problem is strongly convex and hence admits a unique solution $\hat{\boldsymbol{\beta}}$, this implies that τ_* must be unique, which together with the monotonicity of $G(\cdot)$ (cf. (58)) implies that b_* is unique as well. ■

Proof of Theorem 6: To begin with, repeating the arguments in [18, Lemma 6.9] we reach

$$\lim_{t\to\infty}\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}^{t+1}-\hat{\boldsymbol{\beta}}^t\|^2 =_{\text{a.s.}} 0; \tag{67}$$

$$\lim_{t\to\infty}\lim_{n\to\infty}\frac{1}{n}\|\boldsymbol{\eta}^{t+1}-\boldsymbol{\eta}^t\|^2 =_{\text{a.s.}} 0. \tag{68}$$

To show that the AMP iterates converge to the MLE, we shall analyze the log-likelihood function. Recall from Taylor's theorem that

$$\ell(\hat{\boldsymbol{\beta}}) = \ell(\hat{\boldsymbol{\beta}}^t) + \left\langle \nabla\ell(\hat{\boldsymbol{\beta}}^t), \hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\right\rangle + \frac{1}{2}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\right)^\top \nabla^2\ell\left(\hat{\boldsymbol{\beta}}^t+\lambda(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t)\right)\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\right)$$

holds for some $0<\lambda<1$. To deal with the quadratic term, we would like to control the Hessian of the likelihood at a point between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^t$. Invoking the likelihood curvature condition (65), one has

$$\ell(\hat{\boldsymbol{\beta}}^t) \geq \ell(\hat{\boldsymbol{\beta}}) \geq \ell(\hat{\boldsymbol{\beta}}^t) + \left\langle \nabla\ell(\hat{\boldsymbol{\beta}}^t), \hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\right\rangle + \frac{1}{2}n\omega\Big(\max\left\{\|\hat{\boldsymbol{\beta}}\|, \|\hat{\boldsymbol{\beta}}^t\|\right\}\Big)\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\|^2 \tag{69}$$

with high probability. Apply Cauchy-Schwarz to yield that with exponentially high probability,

$$\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\| \leq \frac{2}{\omega\Big(\max\left\{\|\hat{\boldsymbol{\beta}}\|,\|\hat{\boldsymbol{\beta}}^t\|\right\}\Big)}\Big\|\frac{1}{n}\nabla\ell(\hat{\boldsymbol{\beta}}^t)\Big\| \leq \frac{2}{\omega\big(\|\hat{\boldsymbol{\beta}}\|\big)\omega\big(\|\hat{\boldsymbol{\beta}}^t\|\big)}\Big\|\frac{1}{n}\nabla\ell(\hat{\boldsymbol{\beta}}^t)\Big\|,$$

where the last inequality follows since $0<\omega(\cdot)<1$ and $\omega(\cdot)$ is non-decreasing.

It remains to control $\|\nabla\ell(\hat{\boldsymbol{\beta}}^t)\|$. The identity $\Psi(z;b_*) = z - \mathsf{prox}_{b_*\rho}(z)$ and (60) give

$$\mathsf{prox}_{b_*\rho}\left(\boldsymbol{\eta}^{t-1}\right) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^t + \boldsymbol{\eta}^{t-1} - \boldsymbol{\eta}^t. \tag{70}$$

In addition, substituting $\Psi(z;b) = b\rho'(\mathsf{prox}_{\rho b}(z))$ into (62) yields

$$\frac{p}{b_*}(\hat{\boldsymbol{\beta}}^t-\hat{\boldsymbol{\beta}}^{t-1}) = -\boldsymbol{X}^\top\rho'\left(\mathsf{prox}_{b_*\rho}(\boldsymbol{\eta}^{t-1})\right) = -\boldsymbol{X}^\top\rho'\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t+\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\right).$$

We are now ready to bound $\|\nabla\ell(\hat{\boldsymbol{\beta}}^t)\|$. Recalling that

$$\nabla\ell(\hat{\boldsymbol{\beta}}^t) = \boldsymbol{X}^\top\rho'(\boldsymbol{X}^\top\hat{\boldsymbol{\beta}}^t) = \boldsymbol{X}^\top\rho'\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t+\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\right) + \boldsymbol{X}^\top\left(\rho'(\boldsymbol{X}^\top\hat{\boldsymbol{\beta}}^t) - \rho'\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t+\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\right)\right)$$

and that $\sup_z \rho''(z)<\infty$, we have

$$\begin{aligned}\left\|\nabla\ell(\hat{\boldsymbol{\beta}}^t)\right\| &\leq \left\|-\boldsymbol{X}^\top\rho'\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t+\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\right)\right\| + \|\boldsymbol{X}\|\left|\rho'\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t+\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\right)-\rho'(\boldsymbol{X}\hat{\boldsymbol{\beta}}^t)\right| \\ &\leq \frac{p}{b_*}\|\hat{\boldsymbol{\beta}}^t-\hat{\boldsymbol{\beta}}^{t-1}\| + \|\boldsymbol{X}\|\left(\sup_z\rho''(z)\right)\|\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\|.\end{aligned}$$

This establishes that with probability at least $1-c_1e^{-c_2n}$,

$$\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\| \leq \frac{2}{\omega\left(\|\hat{\boldsymbol{\beta}}\|\right)\omega\left(\|\hat{\boldsymbol{\beta}}^t\|\right)}\left\{\frac{p}{b_*n}\|\hat{\boldsymbol{\beta}}^t-\hat{\boldsymbol{\beta}}^{t-1}\| + \frac{1}{n}\left(\sup_z\rho''(z)\right)\|\boldsymbol{X}\|\|\boldsymbol{\eta}^{t-1}-\boldsymbol{\eta}^t\|\right\}. \tag{71}$$

Using (42) together with Borel-Cantelli yields $\lim_{n\to\infty}\|\boldsymbol{X}\|/\sqrt{n}<\infty$ almost surely. Further, it follows from (63) that $\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}^t\|$ is finite almost surely as $\tau_*<\infty$. These taken together with (64), (67) and (68) yield

$$\lim_{t\to\infty}\lim_{n\to\infty}\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^t\| =_{\text{a.s.}} 0 \tag{72}$$

as claimed. ■

7 Likelihood ratio analysis

This section presents the analytical details for Section 2.5.2, which relates the log-likelihood ratio statistic Λ_i with $\hat{\beta}_i$. Recall from (35) that the LLR statistic for testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ is given by

$$\Lambda_1 = \frac{1}{2}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \boldsymbol{D}_{\hat{\boldsymbol{\beta}}}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \frac{1}{6}\sum_{i=1}^{n}\rho'''(\gamma_i)\left(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}\right)^3, \tag{73}$$

where

$$\boldsymbol{D}_{\hat{\boldsymbol{\beta}}} := \begin{bmatrix} \rho''\left(\boldsymbol{X}_1^\top\hat{\boldsymbol{\beta}}\right) & & \\ & \ddots & \\ & & \rho''\left(\boldsymbol{X}_n^\top\hat{\boldsymbol{\beta}}\right) \end{bmatrix} \tag{74}$$

and γ_i lies between $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}$. The asymptotic distribution of Λ_1 claimed in Theorem 3 immediately follows from the result below, whose proof is the subject of the rest of this section.

Theorem 7. *Let (τ_*, b_*) be the unique solution to the system of equations (23) and (24), and define*

$$\tilde{\boldsymbol{G}} = \frac{1}{n}\tilde{\boldsymbol{X}}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{X}} \qquad \text{and} \qquad \tilde{\alpha} = \frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1}). \tag{75}$$

Suppose $p/n \to \kappa \in (0, 1/2)$. Then

(a) the log-likelihood ratio statistic obeys

$$2\Lambda_1 - p\hat{\beta}_1^2/\tilde{\alpha} \overset{\mathbb{P}}{\to} 0; \tag{76}$$

(b) and the scalar $\tilde{\alpha}$ converges,

$$\tilde{\alpha} \overset{\mathbb{P}}{\to} b_*. \tag{77}$$

7.1 More notations and preliminaries

Before proceeding, we introduce some notations that will be used throughout. For any matrix $\boldsymbol{X}$, denote by X_{ij} and $\boldsymbol{X}_{\cdot j}$ its (i, j)-th entry and jth column, respectively. We denote an analogue $\boldsymbol{r} = \{r_i\}_{1\le i\le n}$ (resp. $\tilde{\boldsymbol{r}} = \{\tilde{r}_i\}_{1\le i\le n}$) of residuals in the full (resp. reduced) model by

$$r_i := -\rho'\big(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}\big) \qquad \text{and} \qquad \tilde{r}_i := -\rho'\big(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}\big). \tag{78}$$

As in (74), set

$$\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} := \begin{bmatrix} \rho''\big(\tilde{\boldsymbol{X}}_1^\top\tilde{\boldsymbol{\beta}}\big) & & \\ & \ddots & \\ & & \rho''\big(\tilde{\boldsymbol{X}}_n^\top\tilde{\boldsymbol{\beta}}\big) \end{bmatrix} \quad \text{and} \quad \boldsymbol{D}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}} := \begin{bmatrix} \rho''\left(\gamma_1^*\right) & & \\ & \ddots & \\ & & \rho''(\gamma_n^*), \end{bmatrix}, \tag{79}$$

where γ_i^* is between $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}$, and $\tilde{\boldsymbol{b}}$ is to be defined later in Section 7.2. Further, as in (75), introduce the Gram matrices

$$\boldsymbol{G} := \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{D}_{\hat{\boldsymbol{\beta}}}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{G}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{D}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}}\boldsymbol{X}. \tag{80}$$

Let $\tilde{\boldsymbol{G}}_{(i)}$ denote the version of $\tilde{\boldsymbol{G}}$ without the term corresponding to the ith observation, that is,

$$\tilde{\boldsymbol{G}}_{(i)} = \frac{1}{n}\sum\nolimits_{j:j\neq i}\rho''(\tilde{\boldsymbol{X}}_j^\top\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}_j\tilde{\boldsymbol{X}}_j^\top. \tag{81}$$

Additionally, let $\hat{\boldsymbol{\beta}}_{[-i]}$ be the MLE when the ith observation is dropped and let $\boldsymbol{G}_{[-i]}$ be the corresponding Gram matrix,

$$\boldsymbol{G}_{[-i]} = \frac{1}{n} \sum\nolimits_{j:j\neq i} \rho''(\boldsymbol{X}_j^\top \hat{\boldsymbol{\beta}}_{[-i]}) \boldsymbol{X}_j \boldsymbol{X}_j^\top. \tag{82}$$

Further, let $\tilde{\boldsymbol{\beta}}_{[-i]}$ be the MLE when the first predictor and ith observation are removed, i.e.

$$\tilde{\boldsymbol{\beta}}_{[-i]} := \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \sum\nolimits_{j:j\neq i} \rho(\tilde{\boldsymbol{X}}_j^\top \boldsymbol{\beta}).$$

Below $\tilde{\boldsymbol{G}}_{[-i]}$ is the corresponding version of $\tilde{\boldsymbol{G}}$,

$$\tilde{\boldsymbol{G}}_{[-i]} = \frac{1}{n} \sum\nolimits_{j:j\neq i} \rho''(\tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{\beta}}_{[-i]}) \tilde{\boldsymbol{X}}_j \tilde{\boldsymbol{X}}_j^\top. \tag{83}$$

For these different versions of $\boldsymbol{G}$, their least eigenvalues are all bounded away from 0, as asserted by the following lemma.

Lemma 7. *There exist some absolute constants $\lambda_{\mathrm{lb}}, C, c > 0$ such that*

$$\mathbb{P}(\lambda_{\min}(\boldsymbol{G}) > \lambda_{\mathrm{lb}}) \geq 1 - Ce^{-cn}.$$

Moreover, the same result holds for $\tilde{\boldsymbol{G}}$, $\boldsymbol{G}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}}$, $\tilde{\boldsymbol{G}}_{(i)}$, $\boldsymbol{G}_{[-i]}$ and $\tilde{\boldsymbol{G}}_{[-i]}$ for all $i \in [n]$.

Proof: This result follows directly from Lemma 2, Lemma 4, and Theorem 4. ■

Throughout the rest of this section, we restrict ourselves (for any given n) to the following event:

$$\begin{aligned}\mathcal{A}_n := &\{\lambda_{\min}(\tilde{\boldsymbol{G}}) > \lambda_{\mathrm{lb}}\} \cap \{\lambda_{\min}(\boldsymbol{G}) > \lambda_{\mathrm{lb}}\} \cap \{\lambda_{\min}(\boldsymbol{G}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}}) > \lambda_{\mathrm{lb}}\} \\ &\cap \{\cap_{i=1}^n \lambda_{\min}(\tilde{\boldsymbol{G}}_{(i)}) > \lambda_{\mathrm{lb}}\} \cap \{\cap_{i=1}^n \lambda_{\min}(\tilde{\boldsymbol{G}}_{[-i]}) > \lambda_{\mathrm{lb}}\} \cap \{\cap_{i=1}^n \lambda_{\min}(\boldsymbol{G}_{[-i]}) > \lambda_{\mathrm{lb}}\}.\end{aligned} \tag{84}$$

By Lemma 7, $\mathcal{A}_n$ arises with exponentially high probability, i.e.

$$\mathbb{P}(\mathcal{A}_n) \geq 1 - \exp(-\Omega(n)). \tag{85}$$

7.2 A surrogate for the MLE

In view of (73), the main step in controlling Λ_1 consists of characterizing the differences $\boldsymbol{X}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}} - \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}$. Since the definition of $\hat{\boldsymbol{\beta}}$ is implicit and not amenable to direct analysis, we approximate $\hat{\boldsymbol{\beta}}$ by a more amenable surrogate $\tilde{\boldsymbol{b}}$, an idea introduced in [19–21]. We collect some properties of the surrogate which will prove valuable in the subsequent analysis.

To begin with, our surrogate is

$$\tilde{\boldsymbol{b}} = \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} + \tilde{b}_1 \begin{bmatrix} 1 \\ -\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w} \end{bmatrix}, \tag{86}$$

where $\tilde{\boldsymbol{G}}$ is defined in (80),

$$\boldsymbol{w} := \frac{1}{n} \sum\nolimits_{i=1}^n \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) X_{i1} \tilde{\boldsymbol{X}}_i = \frac{1}{n} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1}, \tag{87}$$

and $\tilde{b}_1$ is a scalar to be specified later. This vector is constructed in the hope that

$$\hat{\boldsymbol{\beta}} \approx \tilde{\boldsymbol{b}}, \qquad \text{or equivalently,} \qquad \begin{cases} \hat{\beta}_1 \approx \tilde{b}_1, \\ \hat{\boldsymbol{\beta}}_{2:p} - \tilde{\boldsymbol{\beta}} \approx -\tilde{b}_1 \tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}, \end{cases} \tag{88}$$

where $\hat{\boldsymbol{\beta}}_{2:p}$ contains the 2nd through pth components of $\hat{\boldsymbol{\beta}}$.

Before specifying $\tilde{b}_1$, we shall first shed some insights into the remaining terms in $\tilde{\boldsymbol{b}}$. By definition,

$$\nabla^2 \ell\left(\begin{bmatrix}0\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right) = \boldsymbol{X}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1} & \boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1} & \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1} & n\boldsymbol{w}^\top \\ n\boldsymbol{w} & n\tilde{\boldsymbol{G}}\end{bmatrix}.$$

Employing the first-order approximation of $\nabla \ell(\cdot)$ gives

$$\nabla^2 \ell\left(\begin{bmatrix}0\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right)\left(\hat{\boldsymbol{\beta}} - \begin{bmatrix}0\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right) \approx \nabla \ell(\hat{\boldsymbol{\beta}}) - \nabla \ell\left(\begin{bmatrix}0\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right). \tag{89}$$

Suppose $\hat{\boldsymbol{\beta}}_{2:p}$ is well approximated by $\tilde{\boldsymbol{\beta}}$. Then all but the first coordinates of $\nabla \ell(\tilde{\boldsymbol{\beta}})$ and $\nabla \ell\left(\begin{bmatrix}0\\ \hat{\boldsymbol{\beta}}\end{bmatrix}\right)$ are also very close to each other. Therefore, taking the 2nd through pth components of (89) and approximating them by zero give

$$\left[\boldsymbol{w}, \tilde{\boldsymbol{G}}\right]\left(\hat{\boldsymbol{\beta}} - \begin{bmatrix}0\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}\right) \approx \boldsymbol{0}.$$

This together with a little algebra yields

$$\hat{\boldsymbol{\beta}}_{2:p} - \tilde{\boldsymbol{\beta}} \;\approx\; -\hat{\beta}_1 \tilde{\boldsymbol{G}}^{-1}\boldsymbol{w} \;\approx\; -\tilde{b}_1 \tilde{\boldsymbol{G}}^{-1}\boldsymbol{w},$$

which coincides with (88). In fact, for all but the 1st entries, $\tilde{\boldsymbol{b}}$ is constructed by moving $\tilde{\boldsymbol{\beta}}$ one-step in the direction which takes it closest to $\hat{\boldsymbol{\beta}}$.

Next, we come to discussing the scalar $\tilde{b}_1$. Introduce the projection matrix

$$\boldsymbol{H} := \boldsymbol{I} - \frac{1}{n}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}, \tag{90}$$

and define $\tilde{b}_1$ as

$$\tilde{b}_1 := \frac{\boldsymbol{X}_{\cdot 1}^\top \tilde{\boldsymbol{r}}}{\boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{X}_{\cdot 1}}, \tag{91}$$

where $\tilde{\boldsymbol{r}}$ comes from (78). In fact, the expression $\tilde{b}_1$ is obtained through similar (but slightly more complicated) first-order approximation as for $\tilde{\boldsymbol{b}}_{2:p}$, in order to make sure that $b_1 \approx \hat{\beta}_1$; see [21, Pages 14560-14561] for a detailed description.

We now formally justify that the surrogate $\tilde{\boldsymbol{b}}$ and the MLE $\hat{\boldsymbol{\beta}}$ are close to each other.

Theorem 8. *The MLE $\hat{\boldsymbol{\beta}}$ and the surrote $\tilde{\boldsymbol{b}}$ (86) obey*

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \;\lesssim\; n^{-1+o(1)}, \tag{92}$$

$$|\tilde{b}_1| \;\lesssim\; n^{-1/2+o(1)}, \tag{93}$$

and

$$\sup_{1\leq i\leq n} |\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}| \;\lesssim\; n^{-1/2+o(1)} \tag{94}$$

with probability tending to one as $n \to \infty$.

Proof: See Section 7.4. ∎

The global accuracy (92) immediately leads to a coordinate-wise approximation result between $\hat{\beta}_1$ and $\tilde{b}_1$.

Corollary 3. *With probability tending to one as $n \to \infty$,*

$$\sqrt{n}|\tilde{b}_1 - \hat{\beta}_1| \;\lesssim\; n^{-1/2+o(1)}. \tag{95}$$

Another consequence from Theorem 8 is that the value $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}$ in the full model and its counterpart $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$ in the reduced model are uniformly close.

Corollary 4. *The values $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$ are uniformly close in the sense that*

$$\sup_{1\le i\le n} \left|\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right| \;\lesssim\; n^{-1/2+o(1)} \tag{96}$$

holds with probability approaching one as $n \to \infty$.

Proof: Note that

$$\sup_{1\le i\le n} \left|\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right| \;\le\; \sup_{1\le i\le n} \left|\boldsymbol{X}_i^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}})\right| + \sup_{1\le i\le n} \left|\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right|.$$

The second term in the right-hans side is upper bounded by $n^{-1/2+o(1)}$ with probability $1-o(1)$ according to Theorem 8. Invoking Lemma 2 and Theorem 8 and applying Cauchy-Schwarz inequality yield that the first term is $O(n^{-1/2+o(1)})$ with probability $1-o(1)$. This establishes the claim. ■

7.3 Analysis of the likelihood-ratio statistic

We are now positioned to use our surrogate $\tilde{\boldsymbol{b}}$ to analyze the likelihood-ratio statistic. In this subsection we establish Theorem 7(a). The proof for Theorem 7(b) is deferred to Appendix I.

Recall from (35) that

$$2\Lambda_1 = (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{D}_{\hat{\boldsymbol{\beta}}} (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \underbrace{\frac{1}{3}\sum_{i=1}^n \rho'''(\gamma_i)(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}})^3}_{:=I_3}.$$

To begin with, Corollary 4 together with the assumption $\sup_z \rho'''(z) < \infty$ implies that

$$I_3 \;\lesssim\; n^{-1/2+o(1)}$$

with probability $1-o(1)$. Hence, I_3 converges to zero in probability.

Reorganize the quadratic term as follows:

$$\begin{aligned}
(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{D}_{\hat{\boldsymbol{\beta}}} (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \sum_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right)^2 \\
&= \sum_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \left[\boldsymbol{X}_i^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}) + (\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\right]^2 \\
&= \sum_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) (\boldsymbol{X}_i^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}))^2 + 2\sum_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \boldsymbol{X}_i^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}})(\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) \\
&\quad + \sum_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) \left(\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right)^2.
\end{aligned} \tag{97}$$

We control each of the three terms in the right-hand side of (97).

- Since $\sup_z \rho''(z) < \infty$, the first term in (97) is bounded by

  $$\sum\nolimits_i \rho''(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}) (\boldsymbol{X}_i^\top (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}))^2 \;\lesssim\; ||\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}||^2 \left\|\sum\nolimits_i \boldsymbol{X}_i \boldsymbol{X}_i^\top\right\| \;\lesssim\; n^{-1+o(1)}$$

  with probability $1-o(1)$, by an application of Theorem 8 and Lemma 2.

- From the definition of $\tilde{\boldsymbol{b}}$, the second term can be upper bounded by

$$2\sum_i \rho''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{b}})^\top \boldsymbol{X}_i\boldsymbol{X}_i^\top \tilde{b}_1 \begin{bmatrix} 1 \\ -\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w} \end{bmatrix} \le |\tilde{b}_1|\cdot\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{b}}\|\cdot\left\|\sum\nolimits_i \boldsymbol{X}_i\boldsymbol{X}_i^\top\right\|\cdot\sqrt{1+\boldsymbol{w}^\top\tilde{\boldsymbol{G}}^{-2}\boldsymbol{w}}$$
$$\lesssim n^{-\frac{1}{2}+o(1)}$$

with probability $1-o(1)$, where the last line follows from a combination of Theorem 8, Lemma 2 and the following lemma.

Lemma 8. *Let $\tilde{\boldsymbol{G}}$ and $\boldsymbol{w}$ be as defined in (80) and (87), respectively. Then*

$$\mathbb{P}\left(\boldsymbol{w}^\top\tilde{\boldsymbol{G}}^{-2}\boldsymbol{w}\lesssim 1\right) \;\ge\; 1-\exp(-\Omega(n)). \tag{98}$$

Proof: See Appendix E. ■

- The third term in (97) can be decomposed as

$$\begin{aligned}
&\sum_i \rho''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}})(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}))^2\\
&\quad=\sum_i\left(\rho''(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}})-\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\right)(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}))^2+\sum_i\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})^2\\
&\quad=\sum_i\rho'''(\tilde{\gamma}_i)(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\left(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}\right)^2+\sum_i\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\left(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}\right)^2
\end{aligned}\tag{99}$$

for some $\tilde{\gamma}_i$ between $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}$. From Theorem 8 and Corollary 4, the first term in (99) is $O(n^{-1/2+o(1)})$ with probability $1-o(1)$. Hence, the only remaining term is the second.

In summary, we have

$$2\Lambda_1-\underbrace{\sum_i\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\left(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}\right)^2}_{=\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}\boldsymbol{v}}\;\overset{\mathbb{P}}{\to}\;0, \tag{100}$$

where $\boldsymbol{v}:=\tilde{b}_1\begin{bmatrix}1\\-\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}\end{bmatrix}$ according to (86). On simplification, the quadratic form reduces to

$$\begin{aligned}
\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}\boldsymbol{v}&=\tilde{b}_1^2\left(\boldsymbol{X}_{\cdot 1}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}\right)^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\left(\boldsymbol{X}_{\cdot 1}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}\right)\\
&=\tilde{b}_1^2\left(\boldsymbol{X}_{\cdot 1}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}_{\cdot 1}-2\boldsymbol{X}_{\cdot 1}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}+\boldsymbol{w}^\top\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}\right)\\
&=\tilde{b}_1^2\left(\boldsymbol{X}_{\cdot 1}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}_{\cdot 1}-n\boldsymbol{w}^\top\tilde{\boldsymbol{G}}^{-1}\boldsymbol{w}\right)\\
&=n\tilde{b}_1^2\underbrace{\left(\frac{1}{n}\boldsymbol{X}_{\cdot 1}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{X}_{\cdot 1}\right)}_{:=\xi},
\end{aligned}$$

recalling the definitions (80), (87), and (90). Hence, the log-likelihood ratio $2\Lambda_1$ simplifies to $n\tilde{b}_1^2\xi+o_P(1)$ on $\mathcal{A}_n$.

Finally, rewrite $\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}\boldsymbol{v}$ as $n(\tilde{b}_1^2-\hat{\beta}_1^2)\xi+n\hat{\beta}_1^2\xi$. To analyze the first term, note that

$$n|\tilde{b}_1^2-\hat{\beta}_1^2|=n|\tilde{b}_1-\hat{\beta}_1|\cdot|\tilde{b}_1+\hat{\beta}_1|\le n|\tilde{b}_1-\hat{\beta}_1|^2+2n|\tilde{b}_1|\cdot|\tilde{b}_1-\hat{\beta}_1|\lesssim n^{-\frac{1}{2}+o(1)} \tag{101}$$

with probability $1-o(1)$ in view of Theorem 8 and Corollary 3. It remains to analyze ξ. Recognize that $\boldsymbol{X}_{\cdot 1}$ is independent of $\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}$. Applying the Hanson-Wright inequality [27, 44] and the Sherman-Morrison-Woodbury formula (e.g. [26]) leads to the following lemma:

Lemma 9. *Let $\tilde{\alpha} = \frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1})$, where $\tilde{\boldsymbol{G}} = \frac{1}{n}\tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{X}}$. Then one has*

$$\left|\frac{p-1}{n} - \tilde{\alpha}\frac{1}{n}\boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{X}_{\cdot 1}\right| \lesssim n^{-1/2+o(1)} \tag{102}$$

with probability approaching one as $n \to \infty$.

Proof: See Appendix F. ■

In addition, if one can show that $\tilde{\alpha}$ is bounded away from zero with probability $1-o(1)$, then it is seen from Lemma 9 that

$$\xi - \frac{p}{n\tilde{\alpha}} \overset{\mathbb{P}}{\to} 0. \tag{103}$$

To justify the above claim, we observe that since ρ'' is bounded, $\lambda_{\max}(\tilde{\boldsymbol{G}}) \lesssim \lambda_{\max}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})/n \lesssim 1$ with exponentially high probability (Lemma 2). This yields

$$\tilde{\alpha} \;=\; \mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1})/n \;\gtrsim\; p/n$$

with probability $1-o(1)$. On the other hand, on $\mathcal{A}_n$ one has

$$\tilde{\alpha} \;\le\; p/(n\lambda_{\min}(\tilde{\boldsymbol{G}})) \;\lesssim\; p/n.$$

Hence, it follows that $\xi = \Omega(1)$ with probability $1-o(1)$. Putting this together with (101) gives the approximation

$$\boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\boldsymbol{X}\boldsymbol{v} = n\hat{\beta}_1^2\xi + o(1). \tag{104}$$

Taken collectively (100), (103) and (104) yields the desired result

$$2\Lambda_1 - p\hat{\beta}_1^2/\tilde{\alpha} \;\overset{\mathbb{P}}{\to}\; 0.$$

7.4 Proof of Theorem 8

This subsection outlines the main steps for the proof of Theorem 8. To begin with, we shall express the difference $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}$ in terms of the gradient of the negative log-likelihood function. Note that $\nabla\ell(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, and hence

$$\begin{aligned}\nabla\ell(\tilde{\boldsymbol{b}}) = \nabla\ell(\tilde{\boldsymbol{b}}) - \nabla\ell(\hat{\boldsymbol{\beta}}) &= \sum\nolimits_{i=1}^n \boldsymbol{X}_i[\rho'(\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}) - \rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})] \\ &= \sum\nolimits_{i=1}^n \rho''(\gamma_i^*)\boldsymbol{X}_i\boldsymbol{X}_i^\top(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{\beta}}),\end{aligned}$$

where γ_i^* is between $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^\top\tilde{\boldsymbol{b}}$. Recalling the notation introduced in (80), this can be rearranged as

$$\tilde{\boldsymbol{b}} - \hat{\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{G}_{\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{b}}}^{-1}\nabla\ell(\tilde{\boldsymbol{b}}).$$

Hence, on $\mathcal{A}_n$, this yields

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \le \frac{\|\nabla\ell(\tilde{\boldsymbol{b}})\|}{\lambda_{\mathrm{lb}} n}. \tag{105}$$

The next step involves expressing $\nabla\ell(\tilde{\boldsymbol{b}})$ in terms of the difference $\tilde{\boldsymbol{b}} - \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}$.

Lemma 10. *On the event $\mathcal{A}_n$ (84), the negative log-likelihood evaluated at the surrogate $\tilde{\boldsymbol{b}}$ obeys*

$$\nabla \ell(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n} \left[\rho''(\gamma_i^*) - \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\right] \boldsymbol{X}_i \boldsymbol{X}_i^\top \left(\tilde{\boldsymbol{b}} - \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}\right),$$

where γ_i^* is some quantity between $\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}$.

Proof: The proof follows exactly the same argument as in the proof of [20, Proposition 3.11], and is thus omitted. ■

The point of expressing $\nabla \ell(\tilde{\boldsymbol{b}})$ in this way is that the difference $\tilde{\boldsymbol{b}} - \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}$ is known explicitly from the definition of $\tilde{\boldsymbol{b}}$. Invoking Lemma 10 and the definition (86) allows one to further upper bound (105) as

$$\begin{aligned} \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \;&\lesssim\; \frac{1}{n}\left\|\nabla \ell(\tilde{\boldsymbol{b}})\right\| \;\lesssim\; \sup_i \left|\rho''(\gamma_i^*) - \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\right| \left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right\| \left\|\tilde{\boldsymbol{b}} - \begin{bmatrix} 0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix}\right\| \\ &\lesssim\; \sup_i \left|\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right| |\rho'''|_\infty \left\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right\| \cdot |\tilde{b}_1| \sqrt{1 + \boldsymbol{w}^\top \tilde{\boldsymbol{G}}^{-2} \boldsymbol{w}} \\ &\lesssim\; |\tilde{b}_1| \sup_i \left|\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right| \end{aligned} \tag{106}$$

with probability at least $1 - \exp(-\Omega(n))$. The last inequality here comes from our assumption that $\sup_z |\rho'''(z)| < \infty$ together with Lemmas 2 and 8.

In order to bound (106), we first make use of the definition of $\tilde{\boldsymbol{b}}$ to reach

$$\sup_i \left|\boldsymbol{X}_i^\top \tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}\right| = |\tilde{b}_1| \sup_i |X_{i1} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}|. \tag{107}$$

The following lemma provides an upper bound on $\sup_i |X_{i1} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}|$.

Lemma 11. *With $\tilde{\boldsymbol{G}}$ and $\boldsymbol{w}$ as defined in (80) and (87),*

$$\mathbb{P}\left(\sup_{1 \le i \le n} \left|X_{i1} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}\right| \le n^{o(1)}\right) \ge 1 - o(1). \tag{108}$$

Proof: See Appendix G. ■

In view of Lemma 11, the second term in the right-hand side of (107) is bounded above by $n^{o(1)}$ with high probability. Thus, in both the bounds (106) and (107), it only remains to analyze the term $\tilde{b}_1$. To this end, we control the numerator and the denominator of $\tilde{b}_1$ separately.

- Recall from the definition (91) that the numerator of $\tilde{b}_1$ is given by $\boldsymbol{X}_{\cdot 1}^\top \tilde{\boldsymbol{r}}$ and that $\tilde{\boldsymbol{r}}$ is independent of $\boldsymbol{X}_{\cdot 1}$. Thus, conditional on $\tilde{\boldsymbol{X}}$, the quantity $\boldsymbol{X}_{\cdot 1}^\top \tilde{\boldsymbol{r}}$ is distributed as a Gaussian with mean zero and variance
$$\sigma^2 = \sum\nolimits_{i=1}^{n} \left(\rho'(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\right)^2.$$
Since $|\rho'(x)| = O(|x|)$, the variance is bounded by
$$\sigma^2 \;\lesssim \tilde{\boldsymbol{\beta}}^\top \left(\sum\nolimits_{i=1}^{n} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i^\top\right) \tilde{\boldsymbol{\beta}} \;\lesssim\; n\|\tilde{\boldsymbol{\beta}}\|^2 \lesssim n \tag{109}$$
with probability at least $1 - \exp(-\Omega(n)))$, a consequence from Theorem 4 and Lemma 2. Therefore, with probability $1 - o(1)$, we have
$$\frac{1}{\sqrt{n}} \boldsymbol{X}_{\cdot 1}^\top \tilde{\boldsymbol{r}} \;\lesssim\; n^{o(1)}. \tag{110}$$

- We now move on to the denominator of $\tilde{b}_1$ in (91). In the discussion following Lemma 9 we showed $\frac{1}{n}\boldsymbol{X}_{\cdot 1}^{\top}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{X}_{\cdot 1} = \Omega(1)$ with probability $1 - o(1)$.

Putting the above bounds together, we conclude

$$\mathbb{P}\left(|\tilde{b}_1| \lesssim n^{-\frac{1}{2}+o(1)}\right) = 1 - o(1). \tag{111}$$

Substitution into (106) and (107) yields

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\| \;\lesssim\; n^{-1+o(1)} \qquad \text{and} \qquad \sup_i \left|\boldsymbol{X}_i^{\top}\tilde{\boldsymbol{b}} - \tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}}\right| \;\lesssim\; n^{-1/2+o(1)}$$

with probability $1 - o(1)$ as claimed.

8 Discussion

In this paper, we derived the high-dimensional asymptotic distribution of the LLR under our modelling assumptions. In particular, we showed that the LLR is inflated vis a vis the classical Wilks' approximation and that this inflation grows as the dimensionality κ increases. This inflation is typical of high-dimensional problems, and one immediate practical consequence is that it explains why classically computed p-values are completely off since they tend to be far too small under the null hypothesis. In contrast, we have shown in our simulations that our new limiting distribution yields reasonably accurate p-values in finite samples. Having said this, our work raises a few important questions that we have not answered and we conclude this paper with a couple of them.

- We expect that our results continue to hold when the covariates are not normally distributed, see Section 3 for some numerical evidence in this direction. To be more precise, we expect the same limiting distribution to hold when the variables are simply sub-Gaussian. If this were true, then this would imply that our rescaled chi-square has a form of universal validity.

- The major limitation of our work is arguably the fact that our limiting distribution holds under the global null; that is, under the assumption that all the regression coefficients vanish. It is unclear to us how the distribution would change in the case where the coefficients are not all zero. In particular, would the limiting distribution depend upon the unknown values of these coefficients? Are there assumptions under which it would not? Suppose for instance that we model the regression coefficients as i.i.d. samples from the mixture model

  $$(1-\epsilon)\delta_0 + \epsilon\Pi^{\star},$$

  where $0 < \epsilon < 1$ is a mixture parameter, δ_0 is a point mass at zero and $\Pi^{\star}$ is a distribution with vanishing mass at zero. Then what would we need to know about ϵ and $\Pi^{\star}$ to compute the asymptotic distribution of the LLR under the null?

Acknowledgements

E. C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, and by the Math + X Award from the Simons Foundation. Y. C. and P. S. are grateful to Andrea Montanari for his help in understanding AMP and [18]. Y. C. thanks Kaizheng Wang and Cong Ma for helpful discussion about [20], and P. S. thanks Subhabrata Sen for several helpful discussions regarding this project. E. C. would like to thank Iain Johnstone for a helpful discussion as well.

A Proofs for Eigenvalue Bounds

A.1 Proof of Lemma 3

Fix $\epsilon \geq 0$ sufficiently small. For any given $S \subseteq [n]$ obeying $|S| = (1-\epsilon)n$ and $0 \leq t \leq \sqrt{1-\epsilon} - \sqrt{p/n}$ it follows from [51, Corollary 5.35] that

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i\in S}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right) < \frac{1}{n}\left(\sqrt{|S|}-\sqrt{p}-t\sqrt{n}\right)^2 = \left(\sqrt{1-\epsilon}-\sqrt{\frac{p}{n}}-t\right)^2$$

holds with probability at most $2\exp\left(-\frac{t^2|S|}{2}\right) = 2\exp\left(-\frac{(1-\epsilon)t^2 n}{2}\right)$. Taking the union bound over all possible subsets S of size $(1-\epsilon)n$ gives

$$\begin{aligned}
&\mathbb{P}\left\{\exists S \subseteq [n] \text{ with } |S| = (1-\epsilon)n \quad \text{s.t.} \quad \frac{1}{n}\lambda_{\min}\left(\sum_{i\in S}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right) < \left(\sqrt{1-\epsilon}-\sqrt{\frac{p}{n}}-t\right)^2\right\} \\
&\leq \binom{n}{(1-\epsilon)n} 2\exp\left(-\frac{(1-\epsilon)\,t^2 n}{2}\right) \\
&\leq 2\exp\left(nH(\epsilon) - \frac{(1-\epsilon)\,t^2}{2}n\right),
\end{aligned}$$

where the last line is a consequence of the inequality $\binom{n}{(1-\epsilon)n} \leq e^{nH(\epsilon)}$ [16, Example 11.1.3].

A.2 Proof of Lemma 4

Define

$$S_B(\boldsymbol{\beta}) := \left\{i : \ |\boldsymbol{X}_i^\top\boldsymbol{\beta}| \leq B\|\boldsymbol{\beta}\|\right\}$$

for any $B > 0$ and any $\boldsymbol{\beta}$. Then

$$\sum_{i=1}^n \rho''\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\boldsymbol{X}_i\boldsymbol{X}_i^\top \ \succeq \ \sum_{i\in S_B(\boldsymbol{\beta})}\rho''\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\boldsymbol{X}_i\boldsymbol{X}_i^\top \succeq \inf_{z:|z|\leq B\|\boldsymbol{\beta}\|}\rho''(z)\sum_{i\in S_B(\boldsymbol{\beta})}\boldsymbol{X}_i\boldsymbol{X}_i^\top.$$

If one also has $|S_B(\boldsymbol{\beta})| \geq (1-\epsilon)n$ (for $\epsilon \geq 0$ sufficiently small), then this together with Lemma 3 implies that

$$\frac{1}{n}\sum_{i=1}^n \rho''\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\boldsymbol{X}_i\boldsymbol{X}_i^\top \succeq \inf_{z:|z|\leq B\|\boldsymbol{\beta}\|}\rho''(z)\left(\sqrt{1-\epsilon}-\sqrt{\frac{p}{n}}-t\right)^2\boldsymbol{I}$$

with probability at least $1-2\exp\left(-\left(\frac{(1-\epsilon)t^2}{2}-H(\epsilon)\right)n\right)$.

Thus if we can ensure that with high probability, $|S_B(\boldsymbol{\beta})| \geq (1-\epsilon)n$ holds simultaneously for all $\boldsymbol{\beta}$, then we are done. From Lemma 2 we see that $\frac{1}{n}\left\|\boldsymbol{X}^\top\boldsymbol{X}\right\| \leq 9$ with probability exceeding $1-2\exp(-n/2)$. On this event,

$$\|\boldsymbol{X}\boldsymbol{\beta}\|^2 \leq 9n\|\boldsymbol{\beta}\|^2, \qquad \forall\boldsymbol{\beta}. \tag{112}$$

On the other hand, the definition of $S_B(\boldsymbol{\beta})$ gives

$$\|\boldsymbol{X}\boldsymbol{\beta}\|^2 \geq \sum_{i\notin S_B(\boldsymbol{\beta})}\left|\boldsymbol{X}_i^\top\boldsymbol{\beta}\right|^2 \geq \left(n-|S_B(\boldsymbol{\beta})|\right)(B\|\boldsymbol{\beta}\|)^2 = n\left(1-\frac{|S_B(\boldsymbol{\beta})|}{n}\right)B^2\|\boldsymbol{\beta}\|^2. \tag{113}$$

Taken together, (112) and (113) yield

$$|S_B(\boldsymbol{\beta})| \geq \left(1-\frac{9}{B^2}\right)n, \qquad \forall\boldsymbol{\beta}$$

with probability at least $1-2\exp(-n/2)$. Therefore, with probability $1-2\exp(-n/2)$, $\left|S_{3/\sqrt{\epsilon}}(\boldsymbol{\beta})\right| \geq (1-\epsilon)\,n$ holds simultaneously for all $\boldsymbol{\beta}$. Putting the above results together and setting $t = 2\sqrt{\frac{H(\epsilon)}{1-\epsilon}}$ give

$$\sum_{i=1}^{n} \rho''\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right) \boldsymbol{X}_i \boldsymbol{X}_i^\top \succeq \inf_{z:|z|\leq \frac{3\|\boldsymbol{\beta}\|}{\sqrt{\epsilon}}} \rho''(z) \left(\sqrt{1-\epsilon} - \sqrt{\frac{p}{n}} - 2\sqrt{\frac{H(\epsilon)}{1-\epsilon}}\right)^2 \boldsymbol{I}$$

simultaneously for all $\boldsymbol{\beta}$ with probability at least $1-2\exp\left(-nH\left(\epsilon\right)\right)-2\exp\left(-n/2\right)$.

B Proof of Lemma 5

Applying an integration by parts leads to

$$\begin{aligned}
\mathbb{E}\left[\Psi'(\tau Z; b)\right] &= \int_{-\infty}^{\infty} \Psi'(\tau z; b)\phi(z)\mathrm{d}z = \frac{1}{\tau}\Psi(\tau z; b)\phi(z)\Big|_{-\infty}^{\infty} - \frac{1}{\tau}\int_{-\infty}^{\infty} \Psi(\tau z; b)\phi'(z)\mathrm{d}z \\
&= -\frac{1}{\tau}\int_{-\infty}^{\infty} \Psi(\tau z; b)\phi'(z)\mathrm{d}z
\end{aligned}$$

with $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$. This reveals that

$$\begin{aligned}
G'(b) &= -\frac{1}{\tau}\int_{-\infty}^{\infty} \frac{\partial \Psi(\tau z; b)}{\partial b}\phi'(z)\mathrm{d}z = -\frac{1}{\tau}\int_{-\infty}^{\infty} \frac{\rho'\left(\mathsf{prox}_{b\rho}(\tau z)\right)}{1 + b\rho''\left(\mathsf{prox}_{b\rho}(\tau z)\right)}\phi'(z)\mathrm{d}z \\
&= \frac{1}{\tau}\int_{0}^{\infty} \left(\frac{\rho'\left(\mathrm{prox}_{b\rho}(-\tau z)\right)}{1 + x\rho''\left(\mathrm{prox}_{b\rho}(-\tau z)\right)} - \frac{\rho'\left(\mathrm{prox}_{b\rho}(\tau z)\right)}{1 + x\rho''\left(\mathrm{prox}_{b\rho}(\tau z)\right)}\right)\phi'(z)\mathrm{d}z, \qquad (114)
\end{aligned}$$

where the second identity comes from [18, Proposition 6.4], and the last identity holds since $\phi'(z) = -\phi'(-z)$.

Next, we claim that

(a) The function $h\left(z\right) := \frac{\rho'(z)}{1+b\rho''(z)}$ is increasing in z;

(b) $\mathrm{prox}_{b\rho}(z)$ is increasing in z.

These two claims imply that

$$\frac{\rho'\left(\mathrm{prox}_{b\rho}(-\tau z)\right)}{1 + b\rho''\left(\mathrm{prox}_{b\rho}(-\tau z)\right)} - \frac{\rho'\left(\mathrm{prox}_{b\rho}(\tau z)\right)}{1 + b\rho''\left(\mathrm{prox}_{b\rho}(\tau z)\right)} < 0, \quad \forall z > 0,$$

which combined with the fact $\phi'(z) < 0$ for $z > 0$ reveals

$$\mathrm{sign}\left(\left(\frac{\rho'\left(\mathrm{prox}_{b\rho}(-\tau z)\right)}{1 + b\rho''\left(\mathrm{prox}_{b\rho}(-\tau z)\right)} - \frac{\rho'\left(\mathrm{prox}_{b\rho}(\tau z)\right)}{1 + b\rho''\left(\mathrm{prox}_{b\rho}(\tau z)\right)}\right)\phi'(z)\right) = 1, \quad \forall z > 0.$$

In other words, the integrand in (114) is positive, which allows one to conclude that $G'(b) > 0$.

We then move on to justify (a) and (b). For the first, the derivative of h is given by

$$h'(z) = \frac{\rho''(z) + b(\rho''(z))^2 - b\rho'(z)\rho'''(z)}{\left(1 + b\rho''(z)\right)^2}.$$

Since ρ' is log concave, this directly yields $(\rho'')^2 - \rho'\rho''' > 0$. As $\rho'' > 0$ and $b \geq 0$, the above implies $h'(z) > 0$ for all z. The second claim follows from $\frac{\partial \mathrm{prox}_{b\rho}(z)}{\partial z} \geq \frac{1}{1+b\|\rho''\|_\infty} > 0$ (cf. [18, Equation (56)]).

It remains to analyze the behavior of G in the limits when $b \to 0$ and $b \to \infty$. From [18, Proposition 6.4], $G(b)$ can also be expressed as

$$G(b) = 1 - \mathbb{E}\left[\frac{1}{1 + b\rho''(\mathsf{prox}_{b\rho}(\tau Z))}\right].$$

Since ρ'' is bounded and the integrand is at most 1, the dominated convergence theorem gives

$$\lim_{b \to 0} G(b) = 0.$$

When $b \to \infty$, $b\rho''(\mathsf{prox}_{b\rho}(\tau z)) \to \infty$ for a fixed z. Again by applying the dominated convergence theorem,

$$\lim_{b \to \infty} G(b) = 1.$$

It follows that $\lim_{b\to 0} G(b) < \kappa < \lim_{b\to\infty} G(b)$ and, therefore, $G(b) = \kappa$ has a unique positive solution.

Remark 3. Finally, we show that the logistic and the probit effective links obey the assumptions of Lemma 5. We work with a fixed $\tau > 0$.

- A direct computation shows that ρ' is log-concave for the logistic model. For the probit, it is well-known that the reciprocal of the hazard function (also known as Mills' ratio) is strictly log-convex [4].
- To check the other condition, recall that the proximal mapping operator satisfies

$$b\rho'(\mathsf{prox}_{b\rho}(\tau z)) + \mathsf{prox}_{b\rho}(\tau z) = \tau z. \tag{115}$$

  For a fixed z, we claim that if $b \to \infty$, $\mathsf{prox}_{b\rho}(\tau z) \to -\infty$. To prove this claim, we start by assuming that this is not true. Then either $\mathsf{prox}_{b\rho}(\tau z)$ is bounded or diverges to ∞. If it is bounded, the LHS above diverges to ∞ while the RHS is fixed, which is a contradiction. Similarly if $\mathsf{prox}_{b\rho}(\tau z)$ diverges to ∞, the left-hand side of (115) diverges to ∞ while the right-hand side is fixed, which cannot be true as well. Further, when $b \to \infty$, we must have $\mathsf{prox}_{b\rho}(\tau z) \to -\infty$, $b\rho'(\mathsf{prox}_{b\rho}(\tau z)) \to \infty$, such that the difference of these two is τz. Observe that for the logistic, $\rho''(x) = \rho'(x)(1 - \rho'(x))$ and for the probit, $\rho''(x) = \rho'(x)(\rho'(x) - x)$ [45]. Hence, combining the asymptotic behavior of $\mathsf{prox}_{b\rho}(\tau z)$ and $b\rho'(\mathsf{prox}_{b\rho}(\tau z))$, we obtain that $b\rho''(\mathsf{prox}_{b\rho}(\tau z))$ diverges to ∞ in both models when $b \to \infty$.

C Proof of Lemma 6

C.1 Proof of Part (i)

Recall from [18, Proposition 6.4] that

$$\kappa \;=\; \mathbb{E}\left[\Psi'\left(\tau Z;\; b(\tau)\right)\right] = 1 - \mathbb{E}\left[\frac{1}{1 + b(\tau)\rho''\big(\mathsf{prox}_{b(\tau)\rho}\left(\tau Z\right)\big)}\right]. \tag{116}$$

If we denote $c := \mathsf{prox}_{b\rho}(0)$, then $b(0)$ is given by the following relation:

$$1 - \kappa = \frac{1}{1 + b(0)\rho''(c)} \quad \Longrightarrow \quad b(0) = \frac{\kappa}{\rho''(c)(1-\kappa)} > 0$$

as $\rho''(c) > 0$ for any given $c > 0$. In addition, since $\rho'(c) > 0$, we have

$$\mathcal{V}(0) = \frac{\Psi(0, b(0))^2}{\kappa} \overset{(a)}{=} \frac{b(0)^2 \rho'(c)^2}{\kappa} > 0,$$

where (a) comes from (20).

C.2 Proof of Part (ii)

We defer the proof of this part to the supplemental materials [47].

D Proof of Part (ii) of Theorem 4

As discussed in Section 5.2.2, it suffices to (1) construct a set $\{\mathcal{B}_i \mid 1 \leq i \leq N\}$ that forms a cover of the cone $\mathcal{A}$ defined in (50), and (2) upper bound $\mathbb{P}\left\{\{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathcal{B}_i \neq \{\boldsymbol{0}\}\right\}$. In what follows, we elaborate on these two steps.

- **Step 1.** Generate $N = \exp\left(2\epsilon^2 p\right)$ i.i.d. points $\boldsymbol{z}^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{p}\boldsymbol{I}_p)$, $1 \leq i \leq N$, and construct a collection of convex cones

$$\mathcal{C}_i := \left\{ \boldsymbol{u} \in \mathbb{R}^p \,\middle|\, \left\langle \boldsymbol{u}, \frac{\boldsymbol{z}^{(i)}}{\|\boldsymbol{z}^{(i)}\|} \right\rangle \geq \epsilon \|\boldsymbol{u}\| \right\}, \qquad 1 \leq i \leq N.$$

In words, $\mathcal{C}_i$ consists of all directions that have nontrivial positive correlation with $\boldsymbol{z}^{(i)}$. With high probability, this collection $\{\mathcal{C}_i \mid 1 \leq i \leq N\}$ forms a cover of $\mathbb{R}^p$, a fact which is an immediate consequence of the following lemma.

Lemma 12. *Consider any given constant $0 < \epsilon < 1$, and let $N = \exp\left(2\epsilon^2 p\right)$. Then there exist some positive universal constants $c_5, C_5 > 0$ such that with probability exceeding $1 - C_5 \exp\left(-c_5 \epsilon^2 p\right)$,*

$$\sum_{i=1}^{N} \boldsymbol{1}_{\left\{\left\langle \boldsymbol{x}, \boldsymbol{z}^{(i)} \right\rangle \geq \epsilon \|\boldsymbol{x}\| \|\boldsymbol{z}^{(i)}\|\right\}} \geq 1$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^p$.

With our family $\{\mathcal{C}_i \mid 1 \leq i \leq N\}$ we can introduce

$$\mathcal{B}_i := \mathcal{C}_i \cap \left\{ \boldsymbol{u} \in \mathbb{R}^n \mid \sum_{j=1}^{n} \max\left\{-u_j, 0\right\} \leq \epsilon\sqrt{n} \left\langle \boldsymbol{u}, \frac{\boldsymbol{z}^{(i)}}{\|\boldsymbol{z}^{(i)}\|} \right\rangle \right\}, \quad 1 \leq i \leq N, \tag{117}$$

which in turn forms a cover of the nonconvex cone $\mathcal{A}$ defined in (50). To justify this, note that for any $\boldsymbol{u} \in \mathcal{A}$, one can find $i \in \{1, \cdots, N\}$ obeying $\boldsymbol{u} \in \mathcal{C}_i$, or equivalently, $\left\langle \boldsymbol{u}, \frac{\boldsymbol{z}^{(i)}}{\|\boldsymbol{z}^{(i)}\|} \right\rangle \geq \epsilon \|\boldsymbol{u}\|$, with high probability. Combined with the membership to $\mathcal{A}$ this gives

$$\sum_{j=1}^{n} \max\left\{-u_j, 0\right\} \leq \epsilon^2 \sqrt{n} \|\boldsymbol{u}\| \leq \epsilon\sqrt{n} \left\langle \boldsymbol{u}, \frac{\boldsymbol{z}^{(i)}}{\|\boldsymbol{z}^{(i)}\|} \right\rangle,$$

indicating that $\boldsymbol{u}$ is contained within some $\mathcal{B}_i$.

- **Step 2.** We now move on to control $\mathbb{P}\left\{\{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathcal{B}_i \neq \{\boldsymbol{0}\}\right\}$. If the statistical dimensions of the two cones obey $\delta\left(\mathcal{B}_i\right) < n - \delta\left(\{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}\right) = n - p$, then an application of [3, Theorem I] gives

$$\begin{aligned}
\mathbb{P}\left\{\{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathcal{B}_i \neq \{\boldsymbol{0}\}\right\} \quad &\leq \quad 4\exp\left\{ -\frac{1}{8} \left(\frac{n - \delta\left(\{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}\right) - \delta\left(\mathcal{B}_i\right)}{\sqrt{n}} \right)^2 \right\} \\
&\leq \quad 4\exp\left\{ -\frac{\left(n - p - \delta(\mathcal{B}_i)\right)^2}{8n} \right\}.
\end{aligned} \tag{118}$$

It then comes down to upper bounding $\delta(\mathcal{B}_i)$, which is the content of the following lemma.

Lemma 13. *Fix $\epsilon > 0$. When n is sufficiently large, the statistical dimension of the convex cone $\mathcal{B}_i$ defined in (117) obeys*

$$\delta(\mathcal{B}_i) \quad \leq \quad \left(\frac{1}{2} + 2\sqrt{2}\epsilon^{\frac{3}{4}} + 10H(2\sqrt{\epsilon})\right) n, \tag{119}$$

where $H(x) := -x\log x - (1-x)\log(1-x)$.

Substitution into (118) gives

$$\begin{aligned}\mathbb{P}\left\{\{\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\} \cap \mathcal{B}_i \neq \{\mathbf{0}\}\right\} \quad &\leq \quad 4\exp\left\{-\frac{\left(\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon})\right)n - p\right)^2}{8n}\right\} \\ &= \quad 4\exp\left\{-\frac{1}{8}\left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon}) - \frac{p}{n}\right)^2 n\right\}. \end{aligned} \tag{120}$$

Finally, we prove Lemmas 12-13 in the next subsections. These are the only remaining parts for the proof of Theorem 4.

D.1 Proof of Lemma 12

To begin with, it is seen that all $\|\boldsymbol{z}^{(i)}\|$ concentrates around 1. Specifically, apply [29, Proposition 1] to get

$$\mathbb{P}\left\{\|\boldsymbol{z}^{(i)}\|^2 > 1 + 2\sqrt{\frac{t}{p}} + \frac{2t}{p}\right\} \leq e^{-t},$$

and set $t = 3\epsilon^2 p$ to reach

$$\mathbb{P}\left\{\|\boldsymbol{z}^{(i)}\|^2 > 1 + 10\epsilon\right\} \quad \leq \quad \mathbb{P}\left\{\|\boldsymbol{z}^{(i)}\|^2 > 1 + 2\sqrt{3}\epsilon + 6\epsilon^2\right\} \quad \leq \quad e^{-3\epsilon^2 p}.$$

Taking the union bound we obtain

$$\mathbb{P}\left\{\exists 1 \leq i \leq N \text{ s.t. } \|\boldsymbol{z}^{(i)}\|^2 > 1 + 10\epsilon\right\} \quad \leq \quad Ne^{-3\epsilon^2 p} = e^{-\epsilon^2 p}. \tag{121}$$

Next, we note that it suffices to prove Lemma 12 for all unit vectors $\boldsymbol{x}$. The following lemma provides a bound on $\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}\rangle$ for any fixed unit vector $\boldsymbol{x} \in \mathbb{R}^p$.

Lemma 14. *Consider any fixed unit vector $\boldsymbol{x} \in \mathbb{R}^p$ and any given constant $0 < \epsilon < 1$, and set $N = \exp\left(2\epsilon^2 p\right)$. There exist positive universal constants $c_5, c_6, C_6 > 0$ such that*

$$\mathbb{P}\left\{\sum_{i=1}^{N} \mathbf{1}_{\left\{\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}\rangle \geq \frac{1}{2}\epsilon\right\}} \leq \exp\left((1 - o(1))\,\frac{7}{4}\epsilon^2 p\right)\right\} \leq \exp\left\{-2\exp\left((1 - o(1))\,\frac{7}{4}\epsilon^2 p\right)\right\}. \tag{122}$$

Recognizing that Lemma 12 is a uniform result, we need to extend Lemma 14 to all $\boldsymbol{x}$ simultaneously, which we achieve via the standard covering argument. Specifically, one can find a set $\mathcal{C} := \left\{\boldsymbol{x}^{(j)} \in \mathbb{R}^p \mid 1 \leq j \leq K\right\}$ of unit vectors with cardinality $K = \left(1 + 2p^2\right)^p$ to form a cover of the unit ball of resolution p^{-2} [51, Lemma 5.2]; that is, for any unit vector $\boldsymbol{x} \in \mathbb{R}^p$, there exists a $\boldsymbol{x}^{(j)} \in \mathcal{C}$ such that

$$\|\boldsymbol{x}^{(j)} - \boldsymbol{x}\| \leq p^{-2}.$$

Apply Lemma 14 and take the union bound to arrive at

$$\sum_{i=1}^{N} \mathbf{1}_{\left\{\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle \geq \frac{1}{2}\epsilon\right\}} \geq \exp\left((1-o(1))\,\frac{7}{4}\epsilon^2 p\right) > 1, \qquad 1 \leq j \leq K \tag{123}$$

with probability exceeding $1-K\exp\left\{-2\exp\left((1-o(1))\,\frac{7}{4}\epsilon^2 p\right)\right\} \geq 1-\exp\left\{-2\left(1-o\left(1\right)\right)\exp\left((1-o(1))\,\frac{7}{4}\epsilon^2 p\right)\right\}$. This guarantees that for each $\boldsymbol{x}^{(j)}$, one can find at least one $\boldsymbol{z}^{(i)}$ obeying

$$\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle \geq \frac{1}{2}\epsilon.$$

This result together with (121) yields that with probability exceeding $1-C\exp\left(-c\epsilon^2 p\right)$, for some universal constants $C, c > 0$.

$$\begin{aligned}
\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}\right\rangle \geq \left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle - \left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}^{(j)}-\boldsymbol{x}\right\rangle \quad & \geq \quad \left\langle \boldsymbol{z}^{(i)}, \boldsymbol{x}^{(j)}\right\rangle - \|\boldsymbol{z}^{(i)}\|\cdot\|\boldsymbol{x}^{(j)}-\boldsymbol{x}\| \\
& \geq \quad \frac{1}{2}\epsilon - \frac{1}{p^2}\|\boldsymbol{z}^{(i)}\| \geq \frac{\frac{1}{2}\epsilon}{\sqrt{1+10\epsilon}}\|\boldsymbol{z}^{(i)}\| - \frac{1}{p^2}\|\boldsymbol{z}^{(i)}\| \\
& \geq \quad \frac{1}{30}\epsilon\|\boldsymbol{z}^{(i)}\|
\end{aligned}$$

holds simultaneously for all unit vectors $\boldsymbol{x} \in \mathbb{R}^p$. Since $\epsilon > 0$ can be an arbitrary constant, this concludes the proof.

Proof of Lemma 14: Without loss of generality, it suffices to consider $\boldsymbol{x} = \boldsymbol{e}_1 = [1, 0, \cdots, 0]^\top$. For any $t > 0$ and any constant $\zeta > 0$, it comes from [2, Theorem A.1.4] that

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{e}_1\right\rangle < \zeta\right\}} > (1+t)\,\Phi\left(\zeta\sqrt{p}\right)\right\} \leq \exp\left(-2t^2\Phi^2\left(\zeta\sqrt{p}\right)N\right).$$

Setting $t = 1-\Phi\left(\zeta\sqrt{p}\right)$ gives

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{e}_1\right\rangle < \zeta\right\}} > \left(2-\Phi\left(\zeta\sqrt{p}\right)\right)\Phi\left(\zeta\sqrt{p}\right)\right\} \leq \exp\left(-2\left(1-\Phi\left(\zeta\sqrt{p}\right)\right)^2\Phi^2\left(\zeta\sqrt{p}\right)N\right).$$

Recall that for any $t > 1$, one has $(t^{-1}-t^{-3})\phi(t) \leq 1-\Phi(t) \leq t^{-1}\phi(t)$ which implies that

$$1-\Phi\left(\zeta\sqrt{p}\right) = \exp\left(-\frac{\left(1+o\left(1\right)\right)\zeta^2 p}{2}\right).$$

Taking $\zeta = \frac{1}{2}\epsilon$, we arrive at

$$\begin{aligned}
\left(2-\Phi\left(\zeta\sqrt{p}\right)\right)\Phi\left(\zeta\sqrt{p}\right) \quad &= \quad 1-\exp\left(-\left(1+o\left(1\right)\right)\zeta^2 p\right) = 1-\exp\left(-\left(1+o\left(1\right)\right)\frac{1}{4}\epsilon^2 p\right), \\
\left(1-\Phi\left(\zeta\sqrt{p}\right)\right)^2\Phi^2\left(\zeta\sqrt{p}\right) \quad &= \quad \exp\left(-\left(1+o\left(1\right)\right)\zeta^2 p\right) = \exp\left(-\left(1+o\left(1\right)\right)\frac{1}{4}\epsilon^2 p\right) \gg \frac{1}{N}.
\end{aligned}$$

This justifies that

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^{N} \mathbf{1}_{\left\{\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{e}_1\right\rangle \geq \frac{1}{2}\epsilon\right\}} \leq N\exp\left(-\left(1+o\left(1\right)\right)\frac{1}{4}\epsilon^2 p\right)\right\} &= \mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{\left\langle \boldsymbol{z}^{(i)}, \boldsymbol{e}_1\right\rangle < \zeta\right\}} > \left(2-\Phi\left(\zeta\sqrt{p}\right)\right)\Phi\left(\zeta\sqrt{p}\right)\right\} \\
&\leq \exp\left\{-2\exp\left(-\left(1+o\left(1\right)\right)\frac{1}{4}\epsilon^2 p\right)N\right\} \\
&= \exp\left\{-2\exp\left(\left(1-o\left(1\right)\right)\frac{7}{4}\epsilon^2 p\right)\right\}
\end{aligned}$$

as claimed. ∎

D.2 Proof of Lemma 13

First of all, recall from the definition (17) that

$$\begin{aligned}\delta(\mathcal{B}_i) &= \mathbb{E}\left[\|\Pi_{\mathcal{B}_i}(\boldsymbol{g})\|^2\right] = \mathbb{E}\left[\|\boldsymbol{g}\|^2 - \min_{\boldsymbol{u}\in\mathcal{B}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right] = n - \mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{B}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right] \\ &\leq n - \mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{D}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right],\end{aligned}$$

where $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)$, and $\mathcal{D}_i$ is a superset of $\mathcal{B}_i$ defined by

$$\mathcal{D}_i := \left\{\boldsymbol{u}\in\mathbb{R}^n \mid \sum\nolimits_{j=1}^n \max\{-u_j,0\} \leq \epsilon\sqrt{n}\|\boldsymbol{u}\|\right\}. \tag{124}$$

Recall from the triangle inequality that

$$\|\boldsymbol{g}-\boldsymbol{u}\| \geq \|\boldsymbol{u}\|-\|\boldsymbol{g}\| > \|\boldsymbol{g}\| = \|\boldsymbol{g}-\boldsymbol{0}\|, \qquad \forall\boldsymbol{u}: \ \|\boldsymbol{u}\|>2\|\boldsymbol{g}\|.$$

Since $\boldsymbol{0}\in\mathcal{D}_i$, this implies that

$$\left\|\arg\min_{\boldsymbol{u}\in\mathcal{D}_i}\|\boldsymbol{g}-\boldsymbol{u}\|\right\| \leq 2\|\boldsymbol{g}\|,$$

revealing that

$$\mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{D}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right] = \mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{D}_i,\|\boldsymbol{u}\|\leq 2\|\boldsymbol{g}\|}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right].$$

In what follows, it suffices to look at the set of $\boldsymbol{u}$'s within $\mathcal{D}_i$ obeying $\|\boldsymbol{u}\|\leq 2\|\boldsymbol{g}\|$, which verify

$$\sum\nolimits_{j=1}^n \max\{-u_j,0\} \leq \epsilon\sqrt{n}\|\boldsymbol{u}\| \leq 2\epsilon\sqrt{n}\|\boldsymbol{g}\|. \tag{125}$$

It is seen that

$$\begin{aligned}\|\boldsymbol{g}-\boldsymbol{u}\|^2 &\geq \sum_{i:g_i<0}(g_i-u_i)^2 = \left\{\sum_{i:g_i<0,u_i\geq 0} + \sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} + \sum_{i:g_i<0,\ u_i\leq-\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|}\right\}(g_i-u_i)^2 \\ &\geq \sum_{i:g_i<0,u_i\geq 0} g_i^2 + \sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0}(g_i-u_i)^2 \\ &\geq \sum_{i:g_i<0,u_i\geq 0} g_i^2 + \sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0}\left(g_i^2-2u_ig_i\right) \\ &\geq \sum_{i:g_i<0,\ u_i>-\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|} g_i^2 - \sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} 2u_ig_i.\end{aligned} \tag{126}$$

1. Regarding the first term of (126), we first recognize that

$$\left\{i \mid u_i \leq -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|\right\} \leq \frac{\sum_{i:\ u_i<0}|u_i|}{\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|} = \frac{\sum_{i=1}^n \max\{-u_i,0\}}{\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|} \leq 2\sqrt{\epsilon}n,$$

where the last inequality follows from the constraint (125). As a consequence,

$$\begin{aligned}\sum_{i:g_i<0,\ u_i>-\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|} g_i^2 &\geq \sum_{i:g_i<0} g_i^2 - \sum_{i:u_i\leq-\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|} g_i^2 \\ &\geq \sum_{i:g_i<0} g_i^2 - \max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2.\end{aligned}$$

2. Next, we turn to the second term of (126), which can be bounded by

$$\begin{aligned}
\sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} u_i g_i &\leq \sqrt{\left(\sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} u_i^2\right)\left(\sum_{i:g_i<0,\ -\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} g_i^2\right)} \\
&\leq \sqrt{\left(\max_{i:-\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|<u_i<0} |u_i|\right)\left(\sum_{i:u_i<0} |u_i|\right)\cdot\|\boldsymbol{g}\|^2} \\
&\leq \sqrt{\sqrt{\frac{\epsilon}{n}}\|\boldsymbol{g}\|\left(\sum_{i:u_i<0}|u_i|\right)\cdot\|\boldsymbol{g}\|^2} \leq \sqrt{2}\epsilon^{\frac{3}{4}}\|\boldsymbol{g}\|^2,
\end{aligned}$$

where the last inequality follows from the constraint (125).

Putting the above results together, we have

$$\|\boldsymbol{g}-\boldsymbol{u}\|^2 \geq \sum_{i:g_i<0} g_i^2 - \max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n} \sum_{i\in S} g_i^2 - 2\sqrt{2}\epsilon^{\frac{3}{4}}\|\boldsymbol{g}\|^2$$

for any $\boldsymbol{u}\in\mathcal{D}_i$ obeying $\|\boldsymbol{u}\|\leq 2\|\boldsymbol{g}\|$, whence

$$\begin{aligned}
\mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{D}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right] &\geq \mathbb{E}\left[\sum_{i:g_i<0} g_i^2 - \max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2 - 2\sqrt{2}\epsilon^{\frac{3}{4}}\|\boldsymbol{g}\|^2\right] \\
&= \left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}}\right)n - \mathbb{E}\left[\max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2\right]. \qquad (127)
\end{aligned}$$

Finally, it follows from [29, Proposition 1] that for any $t>2\sqrt{\epsilon}n$,

$$\mathbb{P}\left\{\sum_{i\in S} g_i^2 \geq 5t\right\} \leq \mathbb{P}\left\{\sum_{i\in S} g_i^2 \geq |S| + 2\sqrt{|S|t} + 2t\right\} \leq e^{-t},$$

which together with the union bound gives

$$\mathbb{P}\left\{\max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2 \geq 5t\right\} \leq \sum_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n} \mathbb{P}\left\{\sum_{i\in S} g_i^2 \geq 5t\right\} \leq \exp\left\{H\left(2\sqrt{\epsilon}\right)n - t\right\}.$$

This gives

$$\begin{aligned}
\mathbb{E}\left[\max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2\right] &= \int_0^\infty \mathbb{P}\left\{\max_{S\subseteq[n]:\ |S|=2\sqrt{\epsilon}n}\sum_{i\in S} g_i^2 \geq t\right\}\mathrm{d}t \\
&\leq 5H\left(2\sqrt{\epsilon}\right)n + \int_{5H\left(2\sqrt{\epsilon}\right)n}^\infty \exp\left\{H\left(2\sqrt{\epsilon}\right)n - \frac{1}{5}t\right\}\mathrm{d}t \\
&< 10H\left(2\sqrt{\epsilon}\right)n,
\end{aligned}$$

for any given $\epsilon>0$ with the proviso that n is sufficiently large. This combined with (127) yields

$$\mathbb{E}\left[\min_{\boldsymbol{u}\in\mathcal{D}_i}\|\boldsymbol{g}-\boldsymbol{u}\|^2\right] \geq \left(\frac{1}{2} - 2\sqrt{2}\epsilon^{\frac{3}{4}} - 10H(2\sqrt{\epsilon})\right)n \qquad (128)$$

as claimed.

E Proof of Lemma 8

Throughout, we shall restrict ourselves on the event $\mathcal{A}_n$ as defined in (84), on which $\tilde{\boldsymbol{G}} \succeq \lambda_{\mathrm{lb}} \boldsymbol{I}$. Recalling the definitions of $\tilde{\boldsymbol{G}}$ and $\boldsymbol{w}$ from (80) and (87), we see that

$$\begin{aligned}\boldsymbol{w}^\top \tilde{\boldsymbol{G}}^{-2} \boldsymbol{w} = &\ \frac{1}{n^2} \boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \left(\frac{1}{n} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \right)^{-2} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1} \\ \leq &\ \frac{\|\boldsymbol{X}_{\cdot 1}^\top\|^2}{n} \left\| \frac{1}{n} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \left(\frac{1}{n} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \right)^{-2} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \right\|. \end{aligned} \tag{129}$$

If we let the singular value decomposition of $\frac{1}{\sqrt{n}} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \tilde{\boldsymbol{X}}$ be $\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$, then a little algebra gives $\boldsymbol{\Sigma} \succeq \sqrt{\lambda_{\mathrm{lb}}} \boldsymbol{I}$ and

$$\frac{1}{n} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \tilde{\boldsymbol{X}} \left(\frac{1}{n} \tilde{\boldsymbol{X}}' \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \right)^{-2} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} = \boldsymbol{U} \boldsymbol{\Sigma}^{-2} \boldsymbol{U}^\top \ \preceq \ \lambda_{\mathrm{lb}}^{-1} \boldsymbol{I}.$$

Substituting this into (129) and using the fact $\|\boldsymbol{X}_{\cdot 1}\|^2 \lesssim n$ with high probability (by Lemma 2), we obtain

$$\boldsymbol{w}^\top \tilde{\boldsymbol{G}}^{-2} \boldsymbol{w} \ \lesssim \ \frac{1}{n \lambda_{\mathrm{lb}}} \|\boldsymbol{X}_{\cdot 1}\|^2 \lesssim 1$$

with probability at least $1 - \exp(-\Omega(n))$.

F Proof of Lemma 9

Throughout this and the subsequent sections, we consider H_n and K_n to be two diverging sequences with the following properties:

$$H_n = o(n^\epsilon), \quad K_n = o(n^\epsilon), \quad n^2 \exp\left(-c_1 H_n^2\right) = o(1), \quad n \exp\left(-c_2 K_n^2\right) = o(1), \tag{130}$$

for any constants $c_i > 0$, $i = 1, 2$ and any $\epsilon > 0$. This lemma is an analogue of [20, Proposition 3.18]. We modify and adapt the proof ideas to establish the result in our setup. Throughout we shall restrict ourselves to the event $\mathcal{A}_n$, on which $\tilde{\boldsymbol{G}} \succeq \lambda_{\mathrm{lb}} \boldsymbol{I}$.

Due to independence between $\boldsymbol{X}_{\cdot 1}$ and $\{\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}, \boldsymbol{H}\}$, one can invoke the Hanson-Wright inequality [44, Theorem 1.1] to yield

$$\begin{aligned}&\mathbb{P}\left(\left| \frac{1}{n} \boldsymbol{X}_{\cdot 1}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{X}_{\cdot 1} - \frac{1}{n} \mathrm{Tr}\left(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \right) \right| > t \ \middle| \ \boldsymbol{H}, \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \right) \\ &\qquad \leq 2 \exp\left(-c \min\left\{ \frac{t^2}{\frac{K^4}{n^2} \|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\|_{\mathrm{F}}^2}, \frac{t}{\frac{K^2}{n} \|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\|} \right\} \right) \\ &\qquad\qquad \leq 2 \exp\left(-c \min\left\{ \frac{t^2}{\frac{K^4}{n} \|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\|^2}, \frac{t}{\frac{K^2}{n} \|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\|} \right\} \right), \end{aligned}$$

where $\|.\|_{\mathrm{F}}$ denotes the Frobenius norm. Choose $t = C^2 \|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\| H_n / \sqrt{n}$ with $C > 0$ a sufficiently large constant, and take H_n to be as in (130). Substitution into the above inequality and unconditioning give

$$\mathbb{P}\left(\left|\frac{1}{n}\boldsymbol{X}_{\cdot 1}^{\top}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{X}_{\cdot 1}-\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\right)\right|>\frac{1}{\sqrt{n}}C^2H_n\|\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\|\right)$$
$$\leq\ 2\exp\left(-c\min\left\{\frac{C^4H_n^2}{K^4},\frac{C^2\sqrt{n}H_n}{K^2}\right\}\right)=C\exp\left(-cH_n^2\right)=o(1),\tag{131}$$

for some universal constants $C,c>0$.

We are left to analyzing $\mathrm{Tr}\big(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\big)$. Recall from the definition (90) of $\boldsymbol{H}$ that

$$\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}=\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}-\frac{1}{n}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}},$$

and, hence,

$$\mathrm{Tr}\left(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\right)=\sum_{i=1}^{n}\left(\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})-\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})^2}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}_i\right).\tag{132}$$

This requires us to analyze $\tilde{\boldsymbol{G}}^{-1}$ carefully. To this end, recall that the matrix $\tilde{\boldsymbol{G}}_{(i)}$ defined in (81) obeys

$$\tilde{\boldsymbol{G}}_{(i)}=\tilde{\boldsymbol{G}}-\frac{1}{n}\rho''(\tilde{\boldsymbol{X}}^{\top}\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^{\top}.$$

Invoking Sherman-Morrison-Woodbury formula (e.g. [26]), we have

$$\tilde{\boldsymbol{G}}^{-1}=\tilde{\boldsymbol{G}}_{(i)}^{-1}-\frac{\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}.\tag{133}$$

It follows that

$$\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}_i=\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i-\frac{\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}(\boldsymbol{X}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i)^2}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i},$$

which implies that

$$\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}_i=\frac{\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}.\tag{134}$$

The relations (132) and (134) taken collectively reveal that

$$\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\boldsymbol{H}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\rho''(\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\beta}})}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}.\tag{135}$$

We shall show that the trace above is close to $\mathrm{Tr}(\boldsymbol{I}-\boldsymbol{H})$ up to some factors. For this purpose we analyze the latter quantity in two different ways. To begin with, observe that

$$\mathrm{Tr}(\boldsymbol{I}-\boldsymbol{H})=\mathrm{Tr}\left(\frac{\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}}{n}\right)=\mathrm{Tr}(\tilde{\boldsymbol{G}}\tilde{\boldsymbol{G}}^{-1})=p-1.\tag{136}$$

On the other hand, it directly follows from the definition of $\boldsymbol{H}$ and (134) that the ith diagonal entry of $\boldsymbol{H}$ is given by

$$H_{i,i}=\frac{1}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^{\top}\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}.$$

Applying this relation, we can compute $\mathrm{Tr}(\boldsymbol{I}-\boldsymbol{H})$ analytically as follows:

$$\mathrm{Tr}(\boldsymbol{I}-\boldsymbol{H}) = \sum_i \frac{\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i}{1 + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i} \tag{137}$$

$$= \sum_i \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\tilde{\alpha} + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i - \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\tilde{\alpha}}{1 + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i}$$

$$= \sum_i \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})\tilde{\alpha} H_{i,i} + \sum_i \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) \left(\frac{1}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i - \tilde{\alpha} \right)}{1 + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i}, \tag{138}$$

where $\tilde{\alpha} := \frac{1}{n}\mathrm{Tr}\left(\tilde{\boldsymbol{G}}^{-1}\right)$.

Observe that the first quantity in the right-hand side above is simply $\tilde{\alpha}\mathrm{Tr}\big(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2}\big)$. For simplicity, denote

$$\eta_i = \frac{1}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i - \tilde{\alpha}. \tag{139}$$

Note that $\tilde{\boldsymbol{G}}_{(i)} \succ \boldsymbol{0}$ on $\mathcal{A}_n$ and that $\rho'' > 0$. Hence the denominator in the second term in (138) is greater than 1 for all i. Comparing (136) and (138), we deduce that

$$\left| \frac{p-1}{n} - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \boldsymbol{H} \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^{1/2} \right) \tilde{\alpha} \right| \;\le\; \sup_i |\eta_i| \cdot \frac{1}{n} \sum_i |\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})| \;\lesssim\; \sup_i |\eta_i| \tag{140}$$

on $\mathcal{A}_n$. It thus suffices to control $\sup_i |\eta_i|$. The above bounds together with Lemma (85) and the proposition below complete the proof.

Proposition 1. *Let η_i be as defined in (139). Then there exist universal constants $C_1, C_2, C_3 > 0$ such that*

$$\mathbb{P}\left(\sup_i |\eta_i| \le \frac{C_1 K_n^2 H_n}{\sqrt{n}} \right) \ge 1 - C_2 n^2 \exp\left(-c_2 H_n^2\right) - C_3 n \exp\left(-c_3 K_n^2\right)$$
$$- \exp\left(-C_4 n \left(1 + o(1)\right)\right) = 1 - o(1),$$

where K_n, H_n are diverging sequences as specified in (130)

Proof of Proposition 1: Fix any index i. Recall that $\tilde{\boldsymbol{\beta}}_{[-i]}$ is the MLE when the 1st predictor and ith observation are removed. Also recall the definition of $\tilde{\boldsymbol{G}}_{[-i]}$ in (83). The proof essentially follows three steps. First, note that $\tilde{\boldsymbol{X}}_i$ and $\tilde{\boldsymbol{G}}_{[-i]}$ are independent. Hence, an application of the Hanson-Wright inequality [44] yields that

$$\mathbb{P}\left(\left| \frac{1}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1} \tilde{\boldsymbol{X}}_i - \frac{1}{n}\mathrm{Tr}\left(\tilde{\boldsymbol{G}}_{[-i]}^{-1}\right) \right| > t \;\middle|\; \tilde{\boldsymbol{G}}_{[-i]} \right) \le 2\exp\left(-c \min\left\{ \frac{t^2}{\frac{K^4}{n^2} \left\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \right\|_{\mathrm{F}}^2}, \frac{t}{\frac{K^2}{n} \left\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \right\|} \right\} \right)$$
$$\le 2\exp\left(-c \min\left\{ \frac{t^2}{\frac{K^4}{n} \left\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \right\|^2}, \frac{t}{\frac{K^2}{n} \left\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \right\|} \right\} \right).$$

We choose $t = C^2 \big\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \big\| H_n / \sqrt{n}$, where $C > 0$ is a sufficiently large constant. Now marginalizing gives

$$\mathbb{P}\left(\left| \frac{1}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1} \tilde{\boldsymbol{X}}_i - \frac{1}{n}\mathrm{Tr}\left(\tilde{\boldsymbol{G}}_{[-i]}^{-1}\right) \right| > C^2 \big\| \tilde{\boldsymbol{G}}_{[-i]}^{-1} \big\| \frac{H_n}{\sqrt{n}} \right) \le 2\exp\left(-c \min\left\{ \frac{C^4 H_n^2}{K^4}, \frac{C^2 \sqrt{n} H_n}{K^2} \right\} \right)$$
$$\le 2\exp\left(-C' H_n^2\right),$$

where $C' > 0$ is a sufficiently large constant. On $\mathcal{A}_n$, the spectral norm $\|\tilde{\boldsymbol{G}}_{(i)}^{-1}\|$ is bounded above by λ_{lb} for all i. Invoking (85) we obtain that there exist universal constants $C_1, C_2, C_3 > 0$ such that

$$\mathbb{P}\left(\sup_i \left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1}\tilde{\boldsymbol{X}}_i - \frac{1}{n}\mathrm{Tr}\left(\tilde{\boldsymbol{G}}_{[-i]}^{-1}\right)\right| > C_1 \frac{H_n}{\sqrt{n}}\right) \leq C_2 n \exp\left(-C_3 H_n^2\right). \tag{141}$$

The next step consists of showing that $\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{[-i]}^{-1}\big)$ (resp. $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1}\tilde{\boldsymbol{X}}_i$) and $\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{(i)}^{-1}\big)$ (resp. $\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i$) are uniformly close across all i. This is established in the following lemma.

Lemma 15. *Let $\tilde{\boldsymbol{G}}_{(i)}$ and $\tilde{\boldsymbol{G}}_{[-i]}$ be defined as in (81) and (83), respectively. Then there exist universal constants $C_1, C_2, C_3, C_4, c_2, c_3 > 0$ such that*

$$\begin{aligned}\mathbb{P}&\left(\sup_i \left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i - \frac{1}{n}\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1}\tilde{\boldsymbol{X}}_i\right| \leq C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\ &\qquad = 1 - C_2 n^2 \exp\left(-c_2 H_n^2\right) - C_3 n \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right) = 1 - o(1),\end{aligned} \tag{142}$$

$$\begin{aligned}\mathbb{P}&\left(\sup_i \left|\frac{1}{n}\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{(i)}^{-1}\big) - \frac{1}{n}\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{[-i]}^{-1}\big)\right| \leq C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\ &\qquad = 1 - C_2 n^2 \exp\left(-c_2 H_n^2\right) - C_3 n \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right) = 1 - o(1),\end{aligned} \tag{143}$$

where K_n, H_n are diverging sequences as defined in (130).

This together with (141) yields that

$$\begin{aligned}\mathbb{P}&\left(\sup_i \left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i - \frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{G}}_{(i)}^{-1})\right| > C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\ &\qquad \leq \; C_2 n^2 \exp\left(-c_2 H_n^2\right) + C_3 n \exp\left(-c_3 K_n^2\right) + \exp\left(-C_4 n\left(1 + o(1)\right)\right).\end{aligned} \tag{144}$$

The final ingredient is to establish that $\frac{1}{n}\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{(i)}^{-1}\big)$ and $\frac{1}{n}\mathrm{Tr}\big(\tilde{\boldsymbol{G}}^{-1}\big)$ are uniformly close across i.

Lemma 16. *Let $\tilde{\boldsymbol{G}}$ and $\tilde{\boldsymbol{G}}_{(i)}$ be as defined in (80) and (81), respectively. Then one has*

$$\mathbb{P}\left(\left|\mathrm{Tr}\big(\tilde{\boldsymbol{G}}_{(i)}^{-1}\big) - \mathrm{Tr}\big(\tilde{\boldsymbol{G}}^{-1}\big)\right| \leq \frac{1}{\lambda_{\mathrm{lb}}}\right) \geq 1 - \exp\left(-\Omega(n)\right). \tag{145}$$

This completes the proof. ■

Proof of Lemma 15: For two invertible matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimensions, the difference of their inverses can be written as

$$\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{B}^{-1}.$$

Applying this identity, we have

$$\tilde{\boldsymbol{G}}_{(i)}^{-1} - \tilde{\boldsymbol{G}}_{[-i]}^{-1} = \tilde{\boldsymbol{G}}_{(i)}^{-1}\left(\tilde{\boldsymbol{G}}_{[-i]} - \tilde{\boldsymbol{G}}_{(i)}\right)\tilde{\boldsymbol{G}}_{[-i]}^{-1}.$$

From the definition of these matrices, it follows directly that

$$\tilde{\boldsymbol{G}}_{[-i]} - \tilde{\boldsymbol{G}}_{(i)} = \frac{1}{n}\sum_{j: j \neq i}\left(\rho''\big(\tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{\beta}}_{[-i]}\big) - \rho''\big(\tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{\beta}}\big)\right)\tilde{\boldsymbol{X}}_j\tilde{\boldsymbol{X}}_j^\top. \tag{146}$$

As ρ''' is bounded, by the mean-value theorem, it suffices to control the differences $\boldsymbol{X}_j^\top \tilde{\boldsymbol{\beta}}_{[-i]} - \tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{\beta}}$ uniformly across all j. This is established in the following lemma, the proof of which is deferred to Appendix H.

Lemma 17. *Let $\hat{\boldsymbol{\beta}}$ be the full model MLE and $\hat{\boldsymbol{\beta}}_{[-i]}$ be the MLE when the ith observation is dropped. Let q_i be as described in Lemma 18 and K_n, H_n be as in (130). Then there exist universal constants $C_1, C_2, C_3, C_4, c_2, c_3 > 0$ such that*

$$\mathbb{P}\left(\sup_{j\neq i}\left|\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]}-\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}\right|\leq C_1\frac{K_n^2H_n}{\sqrt{n}}\right) \\ \geq 1-C_2n\exp\left(-c_2H_n^2\right)-C_3\exp\left(-c_3K_n^2\right)-\exp\left(-C_4n\left(1+o(1)\right)\right)=1-o(1), \quad (147)$$

$$\mathbb{P}\left(\sup_{i}|\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}-\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]})|\leq C_1\frac{K_n^2H_n}{\sqrt{n}}\right) \\ \geq 1-C_2n\exp\left(-c_2H_n^2\right)-C_3\exp\left(-c_3K_n^2\right)-\exp\left(-C_4n\left(1+o(1)\right)\right)=1-o(1). \quad (148)$$

Invoking this lemma, we see that the spectral norm of (146) is bounded above by some constant times

$$\frac{K_n^2H_n}{\sqrt{n}}\Big\|\sum_{j:j\neq i}\tilde{\boldsymbol{X}}_j\tilde{\boldsymbol{X}}_j^\top/n\Big\|$$

with high probability as specified in (147). From Lemma 2, the spectral norm here is bounded by some constant with probability at least $1-c_1\exp(-c_2n)$. These observations together with (85) and the fact that on $\mathcal{A}_n$ the minimum eigenvalues of $\tilde{\boldsymbol{G}}_{(i)}$ and $\tilde{\boldsymbol{G}}_{[-i]}$ are bounded by λ_{lb} yield that

$$\mathbb{P}\left(\|\tilde{\boldsymbol{G}}_{(i)}^{-1}-\tilde{\boldsymbol{G}}_{[-i]}^{-1}\|\leq C_1\frac{K_n^2H_n}{\sqrt{n}}\right)\geq 1-C_2n\exp\left(-c_2H_n^2\right)-C_3\exp\left(-c_3K_n^2\right)-\exp\left(-C_4n\left(1+o(1)\right)\right).$$

This is true for any i. Hence, taking the union bound we obtain

$$\mathbb{P}\left(\sup_i\|\tilde{\boldsymbol{G}}_{(i)}^{-1}-\tilde{\boldsymbol{G}}_{[-i]}^{-1}\|\leq C_1\frac{K_n^2H_n}{\sqrt{n}}\right) \\ \geq 1-C_2n^2\exp\left(-c_2H_n^2\right)-C_3n\exp\left(-c_3K_n^2\right)-\exp\left(-C_4n\left(1+o(1)\right)\right). \quad (149)$$

In order to establish the first result, note that

$$\sup_i\frac{1}{n}\left|\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{[-i]}^{-1}\tilde{\boldsymbol{X}}_i\right|\leq\sup_i\frac{\|\tilde{\boldsymbol{X}}_i\|^2}{n}\sup_i\|\tilde{\boldsymbol{G}}_{(i)}^{-1}-\tilde{\boldsymbol{G}}_{[-i]}^{-1}\|.$$

To obtain the second result, note that

$$\sup_i\left|\frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{G}}_{(i)}^{-1})-\frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{G}}_{[-i]}^{-1})\right|\leq\frac{p-1}{n}\sup_i\|\tilde{\boldsymbol{G}}_{(i)}^{-1}-\tilde{\boldsymbol{G}}_{[-i]}^{-1}\|.$$

Therefore, combining (149) and Lemma 2 gives the desired result. ∎

Proof of Lemma 16: We restrict ourselves to the event $\mathcal{A}_n$ throughout. Recalling (133), one has

$$\mathrm{Tr}(\tilde{\boldsymbol{G}}_{(i)}^{-1})-\mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1})=\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})}{n}\frac{\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-2}\tilde{\boldsymbol{X}}_i}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}.$$

In addition, on $\mathcal{A}_n$ we have

$$\frac{1}{\lambda_{\mathrm{lb}}}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i-\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-2}\tilde{\boldsymbol{X}}_i=\frac{1}{\lambda_{\mathrm{lb}}}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\left(\tilde{\boldsymbol{G}}_{(i)}-\lambda_{\mathrm{lb}}\boldsymbol{I}\right)\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i\geq 0.$$

Combining these results and recognizing that $\rho''>0$, we get

$$\left|\mathrm{Tr}(\tilde{\boldsymbol{G}}_{(i)}^{-1})-\mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1})\right|\leq\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})}{n}\frac{\frac{1}{\lambda_{\mathrm{lb}}}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}\leq\frac{1}{\lambda_{\mathrm{lb}}} \quad (150)$$

as claimed. ∎

G Proof of Lemma 11

Again, we restrict ourselves to the event $\mathcal{A}_n$ on which $\tilde{\boldsymbol{G}} \succeq \lambda_{\mathrm{lb}} \boldsymbol{I}$. Note that

$$\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w} = \frac{1}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \boldsymbol{X}_{\cdot 1}.$$

Note that $\{\tilde{\boldsymbol{G}}, \tilde{\boldsymbol{X}}\}$ and $\boldsymbol{X}_{\cdot 1}$ are independent. Conditional on $\tilde{\boldsymbol{X}}$, the left-hand side is Gaussian with mean zero and variance $\frac{1}{n^2} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^2 \tilde{\boldsymbol{X}} \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}_i$. The variance is bounded above by

$$\begin{aligned}
\sigma_X^2 := \frac{1}{n^2} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}}^2 \tilde{\boldsymbol{X}} \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}_i \;&\leq\; \sup_i \left| \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) \right| \cdot \frac{1}{n^2} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{D}_{\tilde{\boldsymbol{\beta}}} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}_i \\
&= \frac{1}{n} \sup_i \left| \rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}}) \right| \cdot \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \tilde{\boldsymbol{X}}_i \lesssim \frac{1}{n} \|\tilde{\boldsymbol{X}}_i\|^2
\end{aligned} \tag{151}$$

In turn, Lemma 2 asserts that $n^{-1} \|\tilde{\boldsymbol{X}}_i\|^2$ is bounded by a constant with high probability. As a result, applying Gaussian concentration results [49, Theorem 2.1.12] gives

$$|\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}| \lesssim H_n$$

with probability exceeding $1 - C \exp\left(-cH_n^2\right)$, where $C, c > 0$ are universal constants.

In addition, $\sup_i |X_{i1}| \lesssim H_n$ holds with probability exceeding $1 - C \exp\left(-cH_n^2\right)$. Putting the above results together, applying the triangle inequality $|X_{i1} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}| \leq |X_{i1}| + |\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}|$, and taking the union bound, we obtain

$$\mathbb{P}\big(\sup_{1 \leq i \leq n} |X_{i1} - \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}^{-1} \boldsymbol{w}| \lesssim H_n \big) \geq 1 - Cn \exp\left(-cH_n^2\right) = 1 - o(1).$$

H Proof of Lemma 17

The goal of this section is to prove Lemma 17, which relates the full-model MLE $\hat{\boldsymbol{\beta}}$ and the MLE $\hat{\boldsymbol{\beta}}_{[-i]}$. To this end, we establish the key lemma below.

Lemma 18. *Suppose $\hat{\boldsymbol{\beta}}_{[-i]}$ denote the MLE when the ith observation is dropped. Further let $\boldsymbol{G}_{[-i]}$ be as in (82), and define q_i and $\hat{\boldsymbol{b}}$ as follows:*

$$\begin{aligned}
q_i &= \frac{1}{n} \boldsymbol{X}_i^\top \boldsymbol{G}_{[-i]}^{-1} \boldsymbol{X}_i; \\
\hat{\boldsymbol{b}} &= \hat{\boldsymbol{\beta}}_{[-i]} - \frac{1}{n} \boldsymbol{G}_{[-i]}^{-1} \boldsymbol{X}_i \left(\rho' \left(\mathsf{prox}_{q_i \rho} \left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]} \right) \right) \right).
\end{aligned} \tag{152}$$

Suppose K_n, H_n are diverging sequences as in (130). Then there exist universal constants $C_1, C_2, C_3 > 0$ such that

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{b}}\| \leq C_1 \frac{K_n^2 H_n}{n} \right) \geq 1 - C_2 n \exp(-c_2 H_n^2) - C_3 \exp(-c_3 K_n^2) - \exp(-C_4 n(1 + o(1))); \tag{153}$$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{j \neq i} \left| \boldsymbol{X}_j^\top \hat{\boldsymbol{\beta}}_{[-i]} - \boldsymbol{X}_j^\top \hat{\boldsymbol{b}} \right| \leq C_1 \frac{K_n H_n}{\sqrt{n}} \right) \\
&\qquad\qquad \geq 1 - C_2 n \exp\left(-c_2 H_n^2\right) - C_3 \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n \left(1 + o(1)\right)\right).
\end{aligned} \tag{154}$$

The proof ideas are inspired by the leave-one-observation-out approach of [20]. We however emphasize once more that the adaptation of these ideas to our setup is not straightforward and crucially hinges on Theorem 4, Lemma 7 and properties of the effective link function.

Proof of Lemma 18: Invoking techniques similar to that for establishing Lemma 7, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} \rho''(\gamma_i^*)\boldsymbol{X}_i\boldsymbol{X}_i^\top \succeq \lambda_{\mathrm{lb}}\boldsymbol{I} \tag{155}$$

with probability at least $1-\exp(\Omega(n))$, where γ_i^* is between $\boldsymbol{X}_i^\top\hat{\boldsymbol{b}}$ and $\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}$. Denote by $\mathcal{B}_n$ the event where (155) holds. Throughout this proof, we work on the event $\mathcal{C}_n := \mathcal{A}_n \cap \mathcal{B}_n$, which has probability $1-\exp(-\Omega(n))$. As in (105) then,

$$\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{b}}\| \le \frac{1}{n\lambda_{\mathrm{lb}}}\big\|\nabla\ell(\hat{\boldsymbol{b}})\big\|. \tag{156}$$

Next, we simplify (156). To this end, recall the defining relation of the proximal operator

$$b\rho'(\mathsf{prox}_{b\rho}(z)) + \mathsf{prox}_{b\rho}(z) = z,$$

which together with the definitions of $\hat{\boldsymbol{b}}$ and q_i gives

$$\boldsymbol{X}_i^\top\hat{\boldsymbol{b}} = \mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\right). \tag{157}$$

Now, let $\ell_{[-i]}$ denote the negative log-likelihood function when the ith observation is dropped, and hence $\nabla\ell_{[-i]}\big(\hat{\boldsymbol{\beta}}_{[-i]}\big) = \boldsymbol{0}$. Expressing $\nabla\ell(\hat{\boldsymbol{b}})$ as $\nabla\ell(\hat{\boldsymbol{b}}) - \nabla\ell_{[-i]}\big(\hat{\boldsymbol{\beta}}_{[-i]}\big)$, applying the mean value theorem, and using the analysis similar to that in [20, Proposition 3.4], we obtain

$$\frac{1}{n}\nabla\ell(\hat{\boldsymbol{b}}) = \frac{1}{n}\sum_{j:j\neq i}\left[\rho''(\gamma_j^*) - \rho''(\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]})\right]\boldsymbol{X}_j\boldsymbol{X}_j^\top\left(\hat{\boldsymbol{b}}-\hat{\boldsymbol{\beta}}_{[-i]}\right), \tag{158}$$

where γ_j^* is between $\boldsymbol{X}_j^\top\hat{\boldsymbol{b}}$ and $\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]}$. Combining (156) and (158) leads to the upper bound

$$\|\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{b}}\| \le \frac{1}{\lambda_{\mathrm{lb}}}\left\|\frac{1}{n}\sum_{j:j\neq i}\boldsymbol{X}_j\boldsymbol{X}_j^\top\right\|\cdot\sup_{j\neq i}\left|\rho''(\gamma_j^*) - \rho''\big(\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]}\big)\right|\cdot\left\|\frac{1}{n}\boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\right\|\cdot\left|\rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]})\right)\right|. \tag{159}$$

We need to control each term in the right-hand side. To start with, the first term is bounded by a universal constant with probability $1-\exp(-\Omega(n))$ (Lemma 2). For the second term, since γ_j^* is between $\boldsymbol{X}_j^\top\hat{\boldsymbol{b}}$ and $\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]}$ and $\|\rho'''\|_\infty < \infty$, we get

$$\begin{aligned}
\sup_{j\neq i}\left|\rho''(\gamma_j^*) - \rho''(\boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]})\right| &\le \|\rho'''\|_\infty\|\boldsymbol{X}_j^\top\hat{\boldsymbol{b}} - \boldsymbol{X}_j^\top\hat{\boldsymbol{\beta}}_{[-i]}\| & (160)\\
&\le \|\rho'''\|_\infty\left|\frac{1}{n}\boldsymbol{X}_j^\top\boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\rho'\Big(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]})\Big)\right| & (161)\\
&\le \|\rho'''\|_\infty\frac{1}{n}\sup_{j\neq i}\left|\boldsymbol{X}_j^\top\boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\right|\cdot\left|\rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]})\right)\right|. & (162)
\end{aligned}$$

Given that $\{\boldsymbol{X}_j, \boldsymbol{G}_{[-i]}\}$ and $\boldsymbol{X}_i$ are independent for all $j\neq i$, conditional on $\{\boldsymbol{X}_j, \boldsymbol{G}_{[-i]}\}$ one has

$$\boldsymbol{X}_j^\top\boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i \sim \mathcal{N}\left(0, \boldsymbol{X}_j^\top\boldsymbol{G}_{[-i]}^{-2}\boldsymbol{X}_j\right).$$

In addition, the variance satisfies

$$|\boldsymbol{X}_j^\top\boldsymbol{G}_{[-i]}^{-2}\boldsymbol{X}_j| \le \frac{\|\boldsymbol{X}_j\|^2}{\lambda_{\mathrm{lb}}^2} \lesssim n \tag{163}$$

with probability at least $1-\exp(-\Omega(n))$. Applying standard Gaussian concentration results [49, Theorem 2.1.12], we obtain

$$\mathbb{P}\left(\frac{1}{\sqrt{p}}\left|\boldsymbol{X}_j^\top \boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\right| \geq C_1 H_n\right) \leq C_2 \exp\left(-c_2 H_n^2\right) + \exp\left(-C_3 n\left(1+o(1)\right)\right). \tag{164}$$

By the union bound

$$\mathbb{P}\left(\frac{1}{\sqrt{p}}\sup_{j\neq i}\left|\boldsymbol{X}_j^\top \boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\right| \leq C_1 H_n\right) \geq 1 - nC_2 \exp\left(-c_2 H_n^2\right) - \exp\left(-C_3 n\left(1+o(1)\right)\right). \tag{165}$$

Consequently,

$$\sup_{j\neq i}\left|\rho''(\gamma_j^*) - \rho''(\boldsymbol{X}_j^\top \hat{\boldsymbol{\beta}}_{[-i]})\right| \lesssim \sup_{j\neq i}\|\boldsymbol{X}_j^\top \hat{\boldsymbol{b}} - \boldsymbol{X}_j^\top \hat{\boldsymbol{\beta}}_{[-i]}\| \lesssim \frac{1}{\sqrt{n}} H_n \left|\rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})\right)\right|. \tag{166}$$

In addition, the third term in the right-hand side of (159) can be upper bounded as well since

$$\frac{1}{n}\|\boldsymbol{G}_{[-i]}^{-1}\boldsymbol{X}_i\| = \frac{1}{n}\sqrt{|\boldsymbol{X}_i^\top \boldsymbol{G}_{[-i]}^{-2}\boldsymbol{X}_i|} \lesssim \frac{1}{\sqrt{n}} \tag{167}$$

with high probability.

It remains to bound $\left|\rho'\left(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})\right)\right|$. To do this, we begin by considering $\rho'(\mathsf{prox}_{c\rho}(Z))$ for any constant $c>0$ (rather than a random variable q_i). Recall that for any constant $c>0$ and any $Z\sim\mathcal{N}(0,\sigma^2)$ with finite variance, the random variable $\rho'(\mathsf{prox}_{c\rho}(Z))$ is sub-Gaussian. Conditional on $\hat{\boldsymbol{\beta}}_{[-i]}$, one has $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]} \sim \mathcal{N}\big(0, \|\hat{\boldsymbol{\beta}}_{[-i]}\|^2\big)$. This yields

$$\begin{aligned}\mathbb{P}\left(\rho'\left(\mathsf{prox}_{c\rho}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})\right) \geq C_1 K_n\right) &\leq C_2\,\mathbb{E}\left[\exp\left(-\frac{C_3^2 K_n^2}{\|\hat{\boldsymbol{\beta}}_{[-i]}\|^2}\right)\right] \\ &\leq C_2 \exp\left(-C_3 K_n^2\right) + C_4 \exp\left(-C_5 n\right)\end{aligned} \tag{168}$$

for some constants $C_1, C_2, C_3, C_4, C_5 > 0$ since$\|\hat{\boldsymbol{\beta}}_{[-i]}\|$ is bounded with high probability (see Theorem 4).

Note that $\frac{\partial \mathsf{prox}_{b\rho}(z)}{\partial b} \leq 0$ by [18, Proposition 6.3]. Hence, in order to move over from the above concentration result established for a fixed constant c to the random variables q_i, it suffices to establish a uniform lower bound for q_i with high probability. Observe that for each i,

$$q_i \geq \frac{\|\boldsymbol{X}_i\|^2}{n}\frac{1}{\|\boldsymbol{G}_{[-i]}\|} \geq C^*$$

with probability $1-\exp(-\Omega(n))$, where C^* is some universal constant. On this event, one has

$$\rho'\left(\mathsf{prox}_{q_i\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right) \leq \rho'\left(\mathsf{prox}_{C^*\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right).$$

This taken collectively with (168) yields

$$\begin{aligned}\mathbb{P}\left(\rho'(\mathsf{prox}_{q_i\rho}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})) \leq C_1 K_n\right) &\geq \ \mathbb{P}\left(\rho'(\mathsf{prox}_{C^*\rho}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})) \leq C_1 K_n\right) && (169)\\ &\geq \ 1 - C_2 \exp\left(-C_3 K_n^2\right) - C_4 \exp\left(-C_5 n\right). && (170)\end{aligned}$$

This controls the last term.

To summarize, if $\{K_n\}$ and $\{H_n\}$ are diverging sequences satisfying the assumptions in (130), combining (159) and the bounds for each term in the right-hand side finally gives (153). On the other hand, combining (165) and (170) yields (154). ■

With the help of Lemma 18 we are ready to prove Lemma 17. Indeed, observe that

$$\left|\boldsymbol{X}_j^\top(\hat{\boldsymbol{\beta}}_{[-i]} - \hat{\boldsymbol{\beta}})\right| \leq \left|\boldsymbol{X}_j^\top(\hat{\boldsymbol{b}} - \hat{\boldsymbol{\beta}})\right| + \left|\boldsymbol{X}_j^\top(\hat{\boldsymbol{\beta}}_{[-i]} - \hat{\boldsymbol{b}})\right|,$$

and hence by combining Lemma 2 and Lemma 18, we establish the first claim (147). The second claim (148) follows directly from Lemmas 2, 18 and (157).

I Proof of Theorem 7(b)

This section proves that the random sequence $\tilde{\alpha} = \mathrm{Tr}(\tilde{\boldsymbol{G}}^{-1})/n$ converges in probability to the constant b_* defined by the system of equations (23) and (24). To begin with, we claim that $\tilde{\alpha}$ is close to a set of auxiliary random variables $\{\tilde{q}_i\}$ defined below.

Lemma 19. *Define $\tilde{q}_i$ to be*

$$\tilde{q}_i = \frac{1}{n}\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{[-i]}^{-1}\tilde{\boldsymbol{X}}_i,$$

where $\tilde{\boldsymbol{G}}_{[-i]}$ is defined in (83). Then there exist universal constants $C_1, C_2, C_3, C_4, c_2, c_3 > 0$ such that

$$\mathbb{P}\left(\sup_i |\tilde{q}_i - \tilde{\alpha}| \leq C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\ \geq 1 - C_2 n^2 \exp\left(c_2 H_n^2\right) - C_3 n \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right) = 1 - o(1),$$

where K_n, H_n are as in (130).

Proof: This result follows directly from Proposition 1 and equation (142).

■

A consequence is that $\mathsf{prox}_{\tilde{q}_i \rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)$ becomes close to $\mathsf{prox}_{\tilde{\alpha}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)$.

Lemma 20. *Let $\tilde{q}_i$ and $\tilde{\alpha}$ be as defined earlier. Then one has*

$$\mathbb{P}\left(\sup_i \left|\mathsf{prox}_{\tilde{q}_i \rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right) - \mathsf{prox}_{\tilde{\alpha}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right| \leq C_1 \frac{K_n^3 H_n}{\sqrt{n}}\right) \\ \geq 1 - C_2 n^2 \exp\left(-c_2 H_n^2\right) - C_3 n \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right) = 1 - o(1), \quad (171)$$

where K_n, H_n are as in (130).

The key idea behind studying $\mathsf{prox}_{\tilde{\alpha}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)$ is that it is connected to a random function $\delta_n(\cdot)$ defined below, which happens to be closely related to the equation (24). In fact, we will show that $\delta_n(\tilde{\alpha})$ converges in probability to 0; the proof relies on the connection between $\mathsf{prox}_{\tilde{\alpha}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)$ and the auxiliary quantity $\mathsf{prox}_{\tilde{q}_i \rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)$. The formal results is this:

Proposition 2. *For any index i, let $\hat{\boldsymbol{\beta}}_{[-i]}$ be the MLE obtained on dropping the ith observation. Define $\delta_n(x)$ to be the random function*

$$\delta_n(x) := \frac{p}{n} - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right)}. \quad (172)$$

Then one has $\delta_n(\tilde{\alpha}) \stackrel{\mathbb{P}}{\rightarrow} 0$.

Furthermore, the random function $\delta_n(x)$ converges to a deterministic function $\Delta(x)$ defined by

$$\Delta(x) = \kappa - 1 + \mathbb{E}_Z\left[\frac{1}{1 + x\rho''(\mathsf{prox}_{x\rho}(\tau_* Z))}\right], \tag{173}$$

where $Z \sim \mathcal{N}(0,1)$, and τ_* is such that (τ_*, b_*) is the unique solution to (23) and (24).

Proposition 3. *With $\Delta(x)$ as in (173), $\Delta(\tilde{\alpha}) \xrightarrow{\mathbb{P}} 0$.*

In fact, one can easily verify that

$$\Delta(x) = \kappa - \mathbb{E}\big[\Psi'\left(\tau_* Z;\, x\right)\big], \tag{174}$$

and hence by Lemma 5, the solution to $\Delta(x) = 0$ is exactly b_*. As a result, putting the above claims together, we show that $\tilde{\alpha}$ converges in probability to b_*.

It remains to formally prove the preceding lemmas and propositions, which is the goal of the rest of this section.

Proof of Lemma 20: By [18, Proposition 6.3], one has

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

which yields

$$\begin{aligned}
&\sup_i \left| \mathsf{prox}_{\tilde{q}_i\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right) - \mathsf{prox}_{\tilde{\alpha}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right| \\
&\quad = \sup_i \left[\left| \left.\frac{\rho'(x)}{1 + q_{\tilde{\alpha},i}\rho''(x)}\right|_{x = \mathsf{prox}_{q_{\tilde{\alpha},i}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)}\right| \cdot |\tilde{q}_i - \tilde{\alpha}|\right] \\
&\quad \leq \sup_i \left|\rho'\left(\mathsf{prox}_{q_{\tilde{\alpha},i}}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]})\right)\right| \cdot \sup_i |\tilde{q}_i - \tilde{\alpha}|,
\end{aligned} \tag{175}$$

where $q_{\tilde{\alpha},i}$ is between $\tilde{q}_i$ and $\tilde{\alpha}$. Here, the last inequality holds since $q_{\tilde{\alpha},i}, \rho'' \geq 0$.

In addition, just as in the proof of Lemma 18, one can show that q_i is bounded below by some constant $C^* > 0$ with probability $1 - \exp(-\Omega(n))$. Since $q_{\tilde{\alpha},i} \geq \min\{\tilde{q}_i, \tilde{\alpha}\}$, on the event $\sup_i |\tilde{q}_i - \tilde{\alpha}| \leq C_1 K_n^2 H_n/\sqrt{n}$, which happens with high probability (Lemma 19), $q_{\tilde{\alpha},i} \geq C_\alpha$ for some universal constant $C_\alpha > 0$. Hence, by an argument similar to that establishing (170), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_i \left|\rho'\left(\mathsf{prox}_{q_{\tilde{\alpha},i}}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]}\right)\right)\right| \geq C_1 K_n\right) \\
&\qquad\qquad \leq C_2 n^2 \exp\left(-c_2 H_n^2\right) + C_3 n \exp\left(-c_3 K_n^2\right) + \exp\left(-C_4 n\left(1 + o(1)\right)\right).
\end{aligned}$$

This together with (175) and Lemma 19 concludes the proof. ■

Proof of Proposition 2: To begin with, recall from (136) and (137) that on $\mathcal{A}_n$,

$$\frac{p-1}{n} = \sum_{i=1}^n \frac{\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i}{1 + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i} = 1 - \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + \frac{\rho''(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{\beta}})}{n} \tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{G}}_{(i)}^{-1} \tilde{\boldsymbol{X}}_i}. \tag{176}$$

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

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\frac{\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})}{n}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i}-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\tilde{\alpha}}\right| \\
&\quad\leq \frac{1}{n}\sum_{i=1}^{n}\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i-\tilde{\alpha}\right| \;\leq\; \|\rho''\|_\infty \sup_i\left|\frac{1}{n}\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{G}}_{(i)}^{-1}\tilde{\boldsymbol{X}}_i-\tilde{\alpha}\right| \\
&\quad= \|\rho''\|_\infty \sup_i|\eta_i| \leq C_1\frac{K_n^2H_n}{\sqrt{n}},
\end{aligned}$$

with high probability (Proposition 1). This combined with (176) yields

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{p-1}{n}-1+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\tilde{\alpha}}\right| \geq C_1\frac{K_n^2H_n}{\sqrt{n}}\right) \\
&\qquad\qquad\leq C_2n^2\exp\left(-c_2H_n^2\right)+C_3n\exp\left(-c_3K_n^2\right)+\exp\left(-C_4n\left(1+o(1)\right)\right).
\end{aligned}$$

The above bound concerns $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\tilde{\alpha}}$, and it remains to relate it to $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''\left(\mathsf{prox}_{\tilde{\alpha}\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\right)\tilde{\alpha}}$. To this end, we first get from the uniform boundedness of ρ''' and Lemma 17 that

$$\begin{aligned}
&\mathbb{P}\left(\sup_i\left|\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})-\rho''\left(\mathsf{prox}_{\tilde{q}_i\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]})\right)\right| \geq C_1\frac{K_n^2H_n}{\sqrt{n}}\right) \\
&\qquad\qquad\leq C_2n\exp(-c_2H_n^2)+C_3\exp(-c_3K_n^2)+\exp(-C_4n(1+o(1))). \qquad (177)
\end{aligned}$$

Note that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})\tilde{\alpha}}-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''(\mathsf{prox}_{\tilde{\alpha}\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]}))\tilde{\alpha}}\right| \\
&\quad\leq |\tilde{\alpha}|\sup_i\left|\rho''(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}})-\rho''\left(\mathsf{prox}_{\tilde{\alpha}\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]})\right)\right| \\
&\quad\leq |\tilde{\alpha}|\sup_i\left\{\left|\rho''\left(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}\right)-\rho''\left(\mathsf{prox}_{\tilde{q}_i\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]})\right)\right|+\left|\rho''\left(\mathsf{prox}_{\tilde{q}_i\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]})\right)-\rho''\left(\mathsf{prox}_{\tilde{\alpha}\rho}(\tilde{\boldsymbol{X}}_i^\top\tilde{\boldsymbol{\beta}}_{[-i]})\right)\right|\right\}.
\end{aligned}$$

By the bound (177), an application of Lemma 20, and the fact that $\tilde{\alpha} \leq p/(n\lambda_{\mathrm{lb}})$ (on $\mathcal{A}_n$), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{p}{n}-1+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\rho''\left(\mathsf{prox}_{\tilde{\alpha}\rho}(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]})\right)\tilde{\alpha}}\right| \geq C_1\frac{K_n^3H_n}{\sqrt{n}}\right) \\
&\qquad\qquad\leq C_2n^2\exp\big(-c_2H_n^2\big)+C_3n\exp\big(-c_3K_n^2\big)+\exp\big(-C_4n(1+o(1))\big).
\end{aligned}$$

This establishes that $\delta_n(\tilde{\alpha}) \xrightarrow{\mathbb{P}} 0$. ■

Proof of Proposition 3: Note that since $0 < \alpha \leq p/(n\lambda_{\mathrm{lb}}) := B$ on $\mathcal{A}_n$, it suffices to show that

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

We do this by following three steps. Below, $M > 0$ is some sufficiently large constant.

1. First we truncate the random function $\delta_n(x)$ and define

$$\tilde{\delta}_n(x) = \frac{p}{n}-1+\sum_{i=1}^{n}\frac{1}{1+x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\leq M\}}\right)\right)}.$$

The first step is to show that $\sup_{x\in[0,B]}\left|\tilde{\delta}_n(x)-\delta_n(x)\right| \xrightarrow{\mathbb{P}} 0$. We stress that this truncation does not arise in [20], and we keep track of the truncation throughout the rest of the proof.

2. Show that $\sup_{x\in[0,B]}\left|\tilde{\delta}_n(x)-\mathbb{E}\left[\tilde{\delta}_n(x)\right]\right| \xrightarrow{\mathbb{P}} 0$.

3. Show that $\sup_{x\in[0,B]}\left|\mathbb{E}\left[\tilde{\delta}_n(x)\right]-\Delta(x)\right| \xrightarrow{\mathbb{P}} 0$.

To argue about the first step, observe that $|\frac{1}{1+y}-\frac{1}{1+z}|\le|y-z|$ for any $y,z>0$ and that $|\frac{\partial \mathsf{prox}_{c\rho}(x)}{\partial x}|\le 1$ [18, Proposition 6.3]. Then

$$|\delta_n(x)-\tilde{\delta}_n(x)|\le|x|\cdot\|\rho'''\|_\infty\cdot\sup_i\left|\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}-\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right|.$$

For a sufficiently large constant $M>0$, $\mathbb{P}(\|\hat{\boldsymbol{\beta}}_{[-i]}\|\ge M)\le\exp(-\Omega(n))$ by Theorem 4. Hence, for any $\epsilon>0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in[0,B]}|\delta_n(x)-\tilde{\delta}_n(x)|\ge\epsilon\right)&\le\mathbb{P}\left(\sup_i\left|\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\ge M\}}\right|\ge\frac{\epsilon}{B\|\rho'''\|_\infty}\right)\\&\le\sum_{i=1}^n\mathbb{P}\left(\|\hat{\boldsymbol{\beta}}_{[-i]}\|\ge M\right)\;=\;o(1),\end{aligned}\tag{178}$$

establishing Step 1.

To argue about Step 2, note that for any x and z,

$$\begin{aligned}\left|\tilde{\delta}_n(x)-\tilde{\delta}_n(z)\right|&\le\frac{1}{n}\sum_{i=1}^n\left|\frac{1}{1+x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)}-\frac{1}{1+z\rho''\left(\mathsf{prox}_{z\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)}\right|\\&\le\frac{1}{n}\sum_{i=1}^n\left|x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)-z\rho''\left(\mathsf{prox}_{z\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)\right|\\&\le\frac{1}{n}\sum_{i=1}^n\left(\|\rho''\|_\infty|x-z|+|z|\cdot\|\rho'''\|_\infty\left|\frac{\rho'(x)}{1+\tilde{z}\rho''(x)}\Big|_{x=\mathsf{prox}_{\tilde{z}\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)}\right||x-z|\right)\\&\le|x-z|\left(\|\rho''\|_\infty+|z|\cdot\|\rho'''\|_\infty\frac{1}{n}\sum_{i=1}^n\left|\rho'\left(\mathsf{prox}_{\tilde{z}\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)\right|\right),\end{aligned}$$

where $\tilde{z}\in(x,z)$. Setting

$$Y_n:=\|\rho''\|_\infty+B\|\rho'''\|_\infty\frac{1}{n}\sum\nolimits_{i=1}^n\left|\rho'\left(\mathsf{prox}_{\tilde{z}\rho}\left(\boldsymbol{X}_i^\top\hat{\boldsymbol{\beta}}_{[-i]}\mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\|\le M\}}\right)\right)\right|,$$

then for any $\epsilon,\eta>0$ we have

$$\mathbb{P}\left(\sup_{x,z\in(0,B],|x-z|\le\eta}|\tilde{\delta}_n(x)-\tilde{\delta}_n(z)|\ge\epsilon\right)\;\le\;\mathbb{P}\left(Y_n\ge\frac{\epsilon}{\eta}\right)\;\le\;\frac{\eta}{\epsilon}\,\mathbb{E}[Y_n]\;\le\;\eta C_1(\epsilon),\tag{179}$$

where $C_1(\epsilon)$ is some function independent of n. The inequality (179) is an analogue of [20, Lemma 3.24]. We remark that the truncation is particularly important here in guaranteeing that $\mathbb{E}[Y_n]<\infty$.

Set

$$G_n(x):=\mathbb{E}\left[\tilde{\delta}_n(x)\right],$$

and observe that

$$\left|G_n(x) - G_n(z)\right| \le |x-z| \left(\|\rho''\|_\infty + |z| \|\rho'''\|_\infty \, \mathbb{E}\left[\left| \rho'\left(\mathsf{prox}_{\tilde{z}\rho}\left(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\| \le M\}} \right)\right) \right| \right] \right).$$

A similarly inequality applies to $\Delta(x)$ in which $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\| \le M\}}$ is replaced by $\tau_* Z$. In either case,

$$\sup_{x,z \in (0,B], |x-z| \le \eta} |G_n(x) - G_n(z)| \le C_2 \eta \qquad \text{and} \qquad \sup_{x,z \in (0,B], |x-z| \le \eta} |\Delta(x) - \Delta(z)| \le C_3 \eta \tag{180}$$

for any $\eta > 0$.

For any $\epsilon' > 0$, set $K = \max\{C_1(\epsilon'), C_2\}$. Next, divide $[0, B]$ into finitely many segments

$$[0, x_1),\ [x_1, x_2),\ \ldots,\ [x_{K-1}, x_K := B]$$

such that the length of each segment is η/K for any $\eta > 0$. Then for every $x \in [0, B]$, there exists l such that $|x - x_l| \le \eta/K$. As a result, for any $x \in [0, B]$,

$$\sup_{x \in (0,B]} \left| \tilde{\delta}_n(x) - G_n(x) \right| \ \le\ \eta + \sup_{x, x_l \in (0,B], |x - x_l| \le \eta/K} \left| \tilde{\delta}_n(x) - \tilde{\delta}_n(x_l) \right| + \sup_{1 \le l \le k} \left| \tilde{\delta}_n(x_l) - G_n(x_l) \right|.$$

Now fix $\delta > 0, \epsilon > 0$. Applying the above inequality gives

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in (0,B]} |\tilde{\delta}_n(x) - G_n(x)| \ge \delta \right) &\le \mathbb{P}\left(\sup_{x, x_l \in (0,B], |x - x_l| \le \eta/K} |\tilde{\delta}_n(x) - \tilde{\delta}_n(x_l)| \ge \frac{\delta - \eta}{2} \right) \\ &\quad + \mathbb{P}\left(\sup_{1 \le l \le k} |\tilde{\delta}_n(x_l) - G_n(x_l)| \ge \frac{\delta - \eta}{2} \right) \\ &\le \frac{\eta}{K} C_1\left(\frac{\delta - \eta}{2} \right) + \mathbb{P}\left(\sup_{1 \le l \le k} |\tilde{\delta}_n(x_l) - G_n(x_l)| \ge \frac{\delta - \eta}{2} \right). \end{aligned} \tag{181}$$

Choose $\eta < \min\{\epsilon/2, \delta\}, K = \max\{C_1(\frac{\delta - \eta}{2}), C_2\}$. Then the first term in the right-hand side is at most $\epsilon/2$. Furthermore, suppose one can establish that for any fixed x,

$$|\tilde{\delta}_n(x) - G_n(x)| \ \overset{\mathbb{P}}{\to}\ 0. \tag{182}$$

Since the second term in the right-hand side is a supremum over finitely many points, there exists an integer N such that for all $n \ge N$, the second term is less than or equal to $\epsilon/2$. Hence, for all $n \ge N$, right-hand side is at most ϵ, which proves Step 2. Thus, it remains to prove (182) for any fixed $x \ge 0$. We do this after the analysis of Step 3.

We argue about Step 3 in a similar fashion and letting $K = \max\{C_2, C_3\}$, divide $(0, B]$ into segments of length η/K. For every $x \in (0, B]$ there exists x_l such that $|x - x_l| \le \eta/K$. Then

$$\begin{aligned} |G_n(x) - \Delta(x)| &\le\ |G_n(x) - G_n(x_l)| + |G_n(x_l) - \Delta(x_l)| + |\Delta(x_l) - \Delta(x)| \\ &\le\ 2\eta + |G_n(x_l) - \Delta(x_l)|, \end{aligned}$$

$$\Longrightarrow \qquad \sup_{[0,B]} |G_n(x) - \Delta(x)| \le 2\eta + \sup_{1 \le l \le k} |G_n(x_l) - \Delta(x_l)|.$$

Hence, it suffices to show that for any fixed x, $|G_n(x) - \Delta(x)| \to 0$. To this end, observe that for any fixed x,

$$\begin{aligned} &|G_n(x) - \Delta(x)| \\ &\quad \le \left| \frac{p}{n} - \kappa \right| + \left| \mathbb{E}_{\boldsymbol{X}} \left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-1]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\}} \right)\right)} \right] - \mathbb{E}_Z \left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}(\tau_* Z) \right)} \right] \right|. \end{aligned}$$

Additionally,

$$\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-1]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\}} = \|\hat{\boldsymbol{\beta}}_{[-1]}\| \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\}} \tilde{Z},$$

where $\tilde{Z} \sim \mathcal{N}(0,1)$. Since $\|\hat{\boldsymbol{\beta}}_{[-1]}\| \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\}} \stackrel{\mathbb{P}}{\to} \tau_*$, by Slutsky's theorem, $\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-1]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\}}$ converges weakly to $\tau_* \tilde{Z}$. Since $t \mapsto 1/(1 + x\rho\text{"}(\mathsf{prox}_{x\rho}(t)))$ is bounded, one directly gets

$$G_n(x) - \Delta(x) \stackrel{\mathbb{P}}{\to} 0$$

for every x.

Finally, we establish (182). To this end, we will prove instead that for any given x,

$$\delta_n(x) - G_n(x) \stackrel{\mathrm{L}_2}{\to} 0.$$

Define

$$M_i := \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_{[-i]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-i]}\| \le M\}} \qquad \text{and} \qquad f(M_i) := \frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}(M_i)\right)} - \mathbb{E}\left[\frac{1}{1 + x\rho''\left(\mathsf{prox}_{x\rho}(M_i)\right)}\right].$$

Then $\delta_n(x) - G_n(x) = \frac{1}{n}\sum_{i=1}^n f(M_i)$. Hence, for any $x \in [0, B]$,

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

The first term in the right-hand side is at most $1/n$. Hence, it suffices to show that $\mathbb{E}\left[f(M_1)f(M_2)\right] \to 0$. Let $\hat{\boldsymbol{\beta}}_{[-12]}$ be the MLE when the 1st and 2nd observations are dropped. Define

$$\boldsymbol{G}_{[-12]} := \frac{1}{n}\sum_{j \ne 1,2} \rho''\left(\boldsymbol{X}_j^\top \hat{\boldsymbol{\beta}}_{[-12]}\right) \boldsymbol{X}_j \boldsymbol{X}_j^\top, \qquad q_2 := \frac{1}{n}\boldsymbol{X}_2^\top \boldsymbol{G}_{[-12]}^{-1} \boldsymbol{X}_2,$$

$$\hat{\boldsymbol{b}}_{[-1]} := \hat{\boldsymbol{\beta}}_{[-12]} + \frac{1}{n}\boldsymbol{G}_{[-12]}^{-1} \boldsymbol{X}_2 \left(-\rho'\left(\mathsf{prox}_{q_2\rho}\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]}\right)\right)\right).$$

By an application of Lemma 18,

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}_{[-1]} - \hat{\boldsymbol{b}}_{[-1]}\right\| \ge C_1 \frac{K_n^2 H_n}{n}\right) \le C_2 n \exp\left(-c_2 H_n^2\right) - C_3 \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n(1 + o(1))\right). \tag{183}$$

Also, by the triangle inequality,

$$\left|\boldsymbol{X}_1^\top(\hat{\boldsymbol{\beta}}_{[-1]} - \hat{\boldsymbol{\beta}}_{[-12]})\right| \le \left|\boldsymbol{X}_1^\top(\hat{\boldsymbol{\beta}}_{[-1]} - \hat{\boldsymbol{b}}_{[-1]})\right| + \frac{1}{n}\left|\boldsymbol{X}_1^\top \boldsymbol{G}_{[-12]}^{-1} \boldsymbol{X}_2 \left(-\rho'\left(\mathsf{prox}_{q_2\rho}\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]}\right)\right)\right)\right|.$$

Invoking Lemma 2, (183), and an argument similar to that leading to (170) and (164), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{X}_1^\top\left(\hat{\boldsymbol{\beta}}_{[-1]} - \hat{\boldsymbol{\beta}}_{[-12]}\right)\right| \ge C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\
&\qquad\qquad \le C_2 n \exp\left(-c_2 H_n^2\right) - C_3 \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right).
\end{aligned}$$

The event $\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M\} \cap \{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \le M\}$ occurs with probability at least $1 - C\exp(-cn)$. Hence, one obtains

$$\begin{aligned}
&\mathbb{P}\left(\left|\boldsymbol{X}_1^\top\left(\hat{\boldsymbol{\beta}}_{[-1]} \mathbf{1}_{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \le M} - \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \le M}\right)\right| \le C_1 \frac{K_n^2 H_n}{\sqrt{n}}\right) \\
&\qquad\qquad \ge C_2 n \exp\left(-c_2 H_n^2\right) - C_3 \exp\left(-c_3 K_n^2\right) - \exp\left(-C_4 n\left(1 + o(1)\right)\right).
\end{aligned} \tag{184}$$

A similar statement continues to hold with $\boldsymbol{X}_1$ replaced by $\boldsymbol{X}_2$ and $\hat{\boldsymbol{\beta}}_{[-1]}$ replaced by $\hat{\boldsymbol{\beta}}_{[-2]}$. Some simple computation yields that $\|f'\|_\infty$ is bounded by some constant times $|x|$. By the mean value theorem and the fact that $\|f\|_\infty \leq 1$,

$$\begin{aligned}
&\left|f(M_1)f(M_2) - f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right) f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right| \\
&\quad \leq \|f\|_\infty \left\{\left|f(M_1) - f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right| \right. \\
&\qquad \left. + \left|f(M_2) - f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right|\right\} \\
&\quad \leq C|x| \cdot \left|\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-1]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-1]}\| \leq M\}} - \boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right| \\
&\qquad + |x| \cdot \left|\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-2]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-2]}\| \leq M\}} - \boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right|.
\end{aligned}$$

Consequently,

$$f(M_1)f(M_2) - f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right) f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right) \stackrel{\mathbb{P}}{\to} 0.$$

As $\|f\|_\infty \leq 1$, this implies convergence in L_1. Thus, it simply suffices to show that

$$\mathbb{E}\left[f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right) f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right] \to 0.$$

Denote the design matrix on dropping the first and second row as $\boldsymbol{X}_{[-12]}$. Note that conditional on $\boldsymbol{X}_{[-12]}$, $\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}$ and $\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{}\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}$ are independent and have distribution

$$\mathcal{N}\left(0, \|\hat{\boldsymbol{\beta}}_{[-12]}\|^2 \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right).$$

Using this and by arguments similar to [20, Lemma 3.23], one can show that

$$\begin{aligned}
&\mathbb{E}\left[e^{i\left(t\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}} + w\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)}\right] \\
&\qquad - \mathbb{E}\left[e^{it\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}}\right] \mathbb{E}\left[e^{iw\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}}\right] \to 0. \qquad (185)
\end{aligned}$$

On repeated application of the multivariate inversion theorem for obtaining densities from characteristic functions, we get that

$$\begin{aligned}
&\mathbb{E}\left[f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right) f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right] \\
&\qquad - \mathbb{E}\left[f\left(\boldsymbol{X}_1^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right] \mathbb{E}\left[f\left(\boldsymbol{X}_2^\top \hat{\boldsymbol{\beta}}_{[-12]} \mathbf{1}_{\{\|\hat{\boldsymbol{\beta}}_{[-12]}\| \leq M\}}\right)\right] \to 0.
\end{aligned}$$

Since f is centered, this completes the proof. ■

References

[1] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[2] Noga Alon and Joel H Spencer. *The probabilistic method (3rd edition)*. John Wiley & Sons, 2008.

[3] 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.

[4] Árpád Baricz. Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370, 2008.

[5] 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.

[6] 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.

[7] Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[8] 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.

[9] Stéphane Boucheron and Pascal Massart. A high-dimensional Wilks phenomenon. *Probability theory and related fields*, 150(3-4):405–433, 2011.

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

[11] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

[12] Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

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

[14] Gauss M Cordeiro and Francisco Cribari-Neto. *An introduction to Bartlett correction and bias reduction.* Springer, 2014.

[15] 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.

[16] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[17] Francisco Cribari-Neto and Gauss M Cordeiro. On bartlett and bartlett-type corrections francisco cribari-neto. *Econometric reviews*, 15(4):339–367, 1996.

[18] 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.

[19] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[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] Jianqing Fan and Jiancheng Jiang. Nonparametric inference with generalized likelihood ratio tests. *Test*, 16(3):409–444, 2007.

[23] Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

[24] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of statistics*, pages 153–193, 2001.

[25] Yingying Fan, Emre Demirkaya, and Jinchi Lv. Nonuniformity of p-values can occur early in diverging dimensions. *https://arxiv.org/abs/1705.03604*, May 2017.

[26] William W Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

[27] David L. Hanson and Farroll T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[28] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

[29] 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.

[30] Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821, 1973.

[31] Peter J Huber. *Robust statistics*. Springer, 2011.

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

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

[34] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[35] Hua Liang, Pang Du, et al. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics*, 6:1838–1846, 2012.

[36] Enno Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, pages 382–400, 1989.

[37] Peter McCullagh and James A Nelder. Generalized linear models. *Monograph on Statistics and Applied Probability*, 1989.

[38] 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.

[39] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[40] Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. i. consistency. *The Annals of Statistics*, pages 1298–1309, 1984.

[41] Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985.

[42] Stephen Portnoy. Asymptotic behavior of the empiric distribution of m-estimated residuals from a regression model with many parameters. *The Annals of Statistics*, pages 1152–1170, 1986.

[43] Stephen Portnoy et al. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16(1):356–366, 1988.

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

[45] Michael R Sampford. Some inequalities on mill's ratio and related functions. *The Annals of Mathematical Statistics*, 24(1):130–132, 1953.

[46] Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *arXiv preprint arXiv:1205.0498*, 2012.

[47] Pragya Sur, Yuxin Chen, and Emmanuel Candès. Supplemental materials for "the likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square". *http://statweb.stanford.edu/~candes/papers/supplement_LRT.pdf*, 2017.

[48] Cheng Yong Tang and Chenlei Leng. Penalized high-dimensional empirical likelihood. *Biometrika*, pages 905–919, 2010.

[49] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

[50] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

[52] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[53] Ting Yan, Yuanzhuang Li, Jinfeng Xu, Yaning Yang, and Ji Zhu. High-dimensional Wilks phenomena in some exponential random graph models. *arXiv preprint arXiv:1201.0058*, 2012.