arXiv:1705.03604v1 [stat.ME] 10 May 2017

NONUNIFORMITY OF P-VALUES CAN OCCUR EARLY IN DIVERGING DIMENSIONS *

By Yingying Fan, Emre Demirkaya and Jinchi Lv

University of Southern California

Evaluating the joint significance of covariates is of fundamental importance in a wide range of applications. To this end, p-values are frequently employed and produced by algorithms that are powered by classical large-sample asymptotic theory. It is well known that the conventional p-values in Gaussian linear model are valid even when the dimensionality is a non-vanishing fraction of the sample size, but can break down when the design matrix becomes singular in higher dimensions or when the error distribution deviates from Gaussianity. A natural question is when the conventional p-values in generalized linear models become invalid in diverging dimensions. We establish that such a breakdown can occur early in nonlinear models. Our theoretical characterizations are confirmed by simulation studies.

1. Introduction. In many applications it is often desirable to evaluate the significance of covariates in a predictive model for some response of interest. Identifying a set of significant covariates can facilitate domain experts to further probe their causal relationships with the response. Ruling out insignificant covariates can also help reduce the fraction of false discoveries and narrow down the scope of follow-up experimental studies by scientists. These tasks certainly require an accurate measure of feature significance in finite samples. The tool of p-values has provided a powerful framework for such investigations.

As p-values are routinely produced by algorithms, practitioners should perhaps be aware that those p-values are usually based on classical large-sample asymptotic theory. For example, marginal p-values have been employed frequently in large-scale applications when the number of covariates p greatly exceeds the number of observations n. Those p-values are based on marginal regression models linking each individual covariate to the response separately. In these marginal regression models, the ratio of sample

size to model dimensionality is equal to n, which results in justified p-values as sample size increases. Yet due to the correlations among the covariates, we often would like to investigate the joint significance of a covariate in a regression model conditional on all other covariates, which is the main focus of this paper. A natural question is whether conventional joint p-values continue to be valid in the regime of diverging dimensionality p.

It is well known that fitting the linear regression model with $p > n$ using the ordinary least squares can lead to perfect fit giving rise to zero residual vector, which renders the p-values undefined. When $p \leq n$ and the design matrix is nonsingular, the p-values in the linear regression model are well defined and valid thanks to the exact normality of the least-squares estimator when the random error is Gaussian and the design matrix is deterministic. When the error is non-Gaussian, [14] showed that the least-squares estimator can still be asymptotically normal under the assumption of $p = o(n)$, but is generally no longer normal when $p = o(n)$ fails to hold, making the conventional p-values inaccurate in higher dimensions. For the asymptotic properties of M-estimators for robust regression, see, for example, [14, 19, 20] for the case of diverging dimensionality $p = o(n)$ and [15, 6] for the scenario when the dimensionality p grows proportionally to sample size n.

We have seen that the conventional p-values for the least-squares estimator in linear regression model can start behaving wildly and become invalid when the dimensionality p is of the same order as sample size n and the error distribution deviates from Gaussianity. A natural question is whether similar phenomenon holds for the conventional p-values for the maximum likelihood estimator (MLE) in the setting of diverging-dimensional nonlinear models. More specifically, we aim to answer the question of whether $p \sim n$ is still the breakdown point of the conventional p-values when we move away from the regime of linear regression model, where $\sim$ stands for asymptotic order. To simplify the technical presentation, in this paper we adopt the generalized linear model (GLM) as a specific family of nonlinear models [18]. The GLM with a canonical link assumes that the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$ belongs to the canonical exponential family, having the following density function with respect to some fixed measure

$$(1) \quad f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[\frac{y_i \theta_i - b(\theta_i)}{\phi} \right] \right\},$$

where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ design matrix with $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$, $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-dimensional response vector, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional regression coefficient vector, $\{f_0(y; \theta) : \theta \in \mathbb{R}\}$ is a family of distributions in the regular exponential family with disper-

sion parameter $\phi \in (0, \infty)$, and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$. As is common in GLM, the function $b(\theta)$ in (1) is implicitly assumed to be twice continuously differentiable with $b''(\theta)$ always positive. Popularly used GLMs include the linear regression model, logistic regression model, and Poisson regression model for continuous, binary, and count data of responses, respectively.

The key innovation of our paper is the formal justification that the conventional p-values in nonlinear models of GLMs can become invalid in diverging dimensions and such a breakdown can occur *much earlier* than in linear models, which spells out a fundamental difference between linear models and nonlinear models. To begin the journey of p-values in diverging-dimensional GLMs, let us gain some insights into this problem by looking at the specific case of logistic regression. Recently, [7] established an interesting phase transition phenomenon of perfect hyperplane separation for high-dimensional classification with an elegant probabilistic argument. Suppose we are given a random design matrix $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$ and arbitrary binary y_i's that are not all the same. The phase transition of perfect hyperplane separation happens at the point $p/n = 1/2$. With such a separating hyperplane, there exist some $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $t \in \mathbb{R}$ such that $\mathbf{x}_i^T \boldsymbol{\beta}^* > t$ for all cases $y_i = 1$ and $\mathbf{x}_i^T \boldsymbol{\beta}^* < t$ for all controls $y_i = 0$. Let us fit a logistic regression model with an intercept. It is easy to show that multiplying the vector $(-t, (\boldsymbol{\beta}^*)^T)^T$ by a divergence sequence of positive numbers c, we can obtain a sequence of logistic regression fits with the fitted response vector approaching $\mathbf{y} = (y_1, \cdots, y_n)^T$ as $c \to \infty$. As a consequence, the MLE algorithm can return a pretty wild estimate that is close to infinity in topology when the algorithm is set to stop. Clearly, in such a case the p-value of the MLE is no longer justified and meaningful. The results in [7] motivate us to characterize the breakdown point of p-values in nonlinear GLMs with $p \sim n^{\alpha_0}$ in the regime of $\alpha_0 \in [0, 1)$.

It is worth mentioning that our work is different in goals from the limited but growing literature on p-values for high-dimensional nonlinear models, and makes novel contributions to such a problem. The key distinction is that existing work has focused primarily on identifying the scenarios in which conventional p-values or their modifications continue to be valid with some sparsity assumption limiting the growth of intrinsic dimensions. For example, [11] established the oracle property including the asymptotic normality for nonconcave penalized likelihood estimators in the scenario of $p = o(n^{1/5})$, while [9] extended their results to the GLM setting of non-polynomial (NP) dimensionality. In the latter work, the p-values were proved to be valid under the assumption that the intrinsic dimensionality $s = o(n^{1/3})$. More recent work on high-dimensional inference in nonlinear model settings includes

[23, 3] under sparsity assumptions. In addition, two tests were introduced in [13] for high-dimensional GLMs without or with nuisance regression parameters, but the p-values were obtained for testing the global hypothesis for a given set of covariates, which is different from our goal of testing the significance of individual covariates simultaneously.

The rest of the paper is organized as follows. Section 2 provides characterizations of p-values in low dimensions. We establish the nonuniformity of GLM p-values in diverging dimensions in Section 3. Section 4 presents several simulation examples verifying the theoretical phenomenon. We discuss some implications of our results in Section 5. The proofs of all the results are relegated to the Appendix.

2. Characterizations of p-values in low dimensions. To pinpoint the breakdown point of GLM p-values in diverging dimensions, we start with characterizing p-values in low dimensions. In contrast to existing work on the asymptotic distribution of the penalized MLE, our results in this section focus on the asymptotic normality of the unpenalized MLE in diverging-dimensional GLMs, which justifies the validity of conventional p-values. Although Theorem 1 to be established is in the conventional sense, to the best of our knowledge such results are not available in the literature before in terms of the maximum range of the dimensionality p without any sparsity assumption.

2.1. *Maximum likelihood estimation and technical conditions.* For the GLM (1), the log-likelihood $\log f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$ of the sample is given, up to an affine transformation, by

$$(2) \qquad \ell_n(\boldsymbol{\beta}) = n^{-1} \left[\mathbf{y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X} \boldsymbol{\beta}) \right],$$

where $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \cdots, b(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T \in \mathbb{R}^n$. Denote by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T \in \mathbb{R}^p$ the MLE which is the maximizer of (2), and

$$(3) \quad \boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \cdots, b'(\theta_n))^T \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}\{b''(\theta_1), \cdots, b''(\theta_n)\}.$$

A well-known fact is that the n-dimensional response vector $\mathbf{y}$ in GLM (1) has mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\phi \boldsymbol{\Sigma}(\boldsymbol{\theta})$. Clearly, the MLE $\widehat{\boldsymbol{\beta}}$ is given by the unique solution to the score equation

$$(4) \qquad \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0}$$

when the design matrix $\mathbf{X}$ is of full column rank p.

We first introduce a deviation probability bound that facilitates our technical analysis. Consider both cases of bounded responses and unbounded responses. In the latter case, assume that there exist some constants $M, v_0 > 0$ such that

$$(5) \qquad \max_{1 \leq i \leq n} E \left\{ \exp \left[\frac{|y_i - b'(\theta_{0,i})|}{M} \right] - 1 - \frac{|y_i - b'(\theta_{0,i})|}{M} \right\} M^2 \leq \frac{v_0}{2}$$

with $(\theta_{0,1}, \cdots, \theta_{0,n})^T = \boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ denotes the true regression coefficient vector in model (1). Then by [9, 10], it holds that for any $\mathbf{a} \in \mathbb{R}^n$,

$$(6) \qquad P\left(\left|\mathbf{a}^T\mathbf{Y} - \mathbf{a}^T\boldsymbol{\mu}(\boldsymbol{\theta}_0)\right| > \|\mathbf{a}\|_2 \varepsilon\right) \leq \varphi(\varepsilon),$$

where $\varphi(\varepsilon) = 2e^{-c_1\varepsilon^2}$ with $c_1 > 0$ some constant, and $\varepsilon \in (0, \infty)$ if the responses are bounded and $\varepsilon \in (0, \|\mathbf{a}\|_2/\|\mathbf{a}\|_\infty]$ if the responses are unbounded.

For nonlinear GLMs, the MLE $\widehat{\boldsymbol{\beta}}$ solves the nonlinear score equation (4) whose solution generally does not admit an explicit form. To address such a challenge, we construct a solution to equation (4) in an asymptotically shrinking neighborhood of $\boldsymbol{\beta}_0$ that meets the MLE $\widehat{\boldsymbol{\beta}}$ thanks to the uniqueness of the solution. Specifically, define a neighborhood of $\boldsymbol{\beta}_0$ as

$$(7) \qquad \mathcal{N}_0 = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty \leq n^{-\gamma}\log n\}$$

for some constant $\gamma \in (0, 1/2]$. Assume that $p = O(n^{\alpha_0})$ for some $\alpha_0 \in (0, \gamma)$ and let $b_n = o\{\min(n^{1/2-\gamma}\sqrt{\log n}, n^{2\gamma-\alpha_0-1/2}/(\log n)^2\}$ be a diverging sequence of positive numbers. We need some basic regularity conditions to establish the asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}$.

CONDITION 1. *The design matrix* $\mathbf{X}$ *satisfies*

$$(8) \qquad \left\|\left[\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}\right]^{-1}\right\|_\infty = O(b_n n^{-1}),$$

$$(9) \qquad \max_{\boldsymbol{\beta}\in\mathcal{N}_0} \max_{j=1}^p \lambda_{\max}\left[\mathbf{X}^T \text{diag}\left\{|\mathbf{x}_j| \circ |\boldsymbol{\mu}''(\mathbf{X}\boldsymbol{\beta})|\right\}\mathbf{X}\right] = O(n)$$

with $\circ$ *denoting the Hadamard product and derivatives understood componentwise. Assume that* $\max_{j=1}^p \|\mathbf{x}_j\|_\infty < c_1^{1/2}\{n/(\log n)\}^{1/2}$ *if the responses are unbounded.*

CONDITION 2. *The eigenvalues of* $n^{-1}\mathbf{A}_n$ *are bounded away from* 0 *and* ∞, $\sum_{i=1}^n (\mathbf{z}_i^T\mathbf{A}_n^{-1}\mathbf{z}_i)^{3/2} = o(1)$, *and* $\max_{i=1}^n E|y_i - b'(\theta_{0,i})|^3 = O(1)$, *where* $\mathbf{A}_n = \mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}$ *and* $(\mathbf{z}_1, \cdots, \mathbf{z}_n)^T = \mathbf{X}$.

Conditions 1 and 2 put some basic restrictions on the design matrix $\mathbf{X}$ and a moment condition on the responses. For the case of linear model, bound (8) becomes $\|(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty = O(b_n/n)$ and bound (9) holds automatically since $b'''(\theta) \equiv 0$. Condition 2 is related to the Lyapunov condition.

2.2. *Conventional p-values in low dimensions.*

THEOREM 1 (Asymptotic normality).   *Assume that Conditions 1–2 and probability bound (6) hold. Then the MLE $\widehat{\boldsymbol{\beta}}$ satisfies that for each $1 \leq j \leq p$,*

$$(10) \qquad (\mathbf{A}_n^{-1})_{jj}^{-1/2}(\widehat{\beta}_j - \beta_{0,j}) \xrightarrow{\mathscr{D}} N(0,\phi),$$

where $\mathbf{A}_n = \mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}$ and $(\mathbf{A}_n^{-1})_{jj}$ denotes the jth diagonal entry.

Theorem 1 establishes the asymptotic normality of the MLE and consequently justifies the validity of the conventional p-values in low dimensions. Note that for simplicity, we present here only the marginal asymptotic normality, and the joint asymptotic normality also holds for the projection of the MLE onto any fixed-dimensional subspace. This result can also be extended to the case of misspecified models; see, for example, [17].

As mentioned in the Introduction, the asymptotic normality was shown in [9] for nonconcave penalized MLE having intrinsic dimensionality $s = o(n^{1/3})$. In contrast, our result in Theorem 1 allows for the scenario of $p = o(n^{1/2})$ with no sparsity assumption in view of our technical conditions. In particular, we see that the conventional p-values in nonlinear GLMs generally remain valid in the regime of slowly diverging dimensionality $p = o(n^{1/2})$.

3. Nonuniformity of GLM p-values in diverging dimensions.   So far we have seen that for nonlinear GLMs, the p-values can be valid when $p = o(n^{1/2})$ as shown in Section 2, and can become meaningless when $p \geq n/2$ as discussed in the Introduction. Apparently, there is a big gap between these two regimes of growth of dimensionality p. To provide some guidance on the practical use of p-values in nonlinear GLMs, it is of crucial importance to characterize their breakdown point. To highlight the main message with simplified technical presentation, hereafter we content ourselves on the specific case of logistic regression model for binary response. We argue that this specific model is sufficient for our purpose because for conventional p-values derived from MLEs in diverging-dimensional GLMs to be valid, it must be at least valid for the specific model of logistic regression. Therefore, the breakdown point for logistic regression is at least the breakdown point for general nonlinear GLMs. This argument is fundamentally

different from that of proving the overall validity of conventional p-values, where one needs to prove the asymptotic normality of MLEs under general GLMs rather than any specific model.

3.1. *The wild side of nonlinear regime.* For the logistic regression model (1), we have $b(\theta) = \log(1 + e^\theta)$, $\theta \in \mathbb{R}$ and $\phi = 1$. The mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\phi\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of the n-dimensional response vector $\mathbf{y}$ given by (3) now take the familiar form of $\boldsymbol{\mu}(\boldsymbol{\theta}) = \left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, \cdots, \frac{e^{\theta_n}}{1+e^{\theta_n}} \right)^T$ and

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag} \left\{ \frac{e^{\theta_1}}{\left(1 + e^{\theta_1}\right)^2}, \cdots, \frac{e^{\theta_n}}{\left(1 + e^{\theta_n}\right)^2} \right\}$$

with $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$. In many real applications, one would like to interpret the significance of each individual covariate produced by algorithms based on the conventional asymptotic normality of the MLE as established in Theorem 1. As argued at the beginning of this section, in order to justify any significant discoveries, the underlying theory of p-values in diverging-dimensional GLMs should ideally at least ensure that the distributional property in (10) holds for the scenario of true regression coefficient vector $\boldsymbol{\beta}_0 = \mathbf{0}$, that is, under the global null. Otherwise practitioners may simply lose the theoretical backup and the resulting decisions based on the p-values can become ineffective or even misleading. For this reason, we identify the breakdown point of p-values in diverging-dimensional logistic regression model under the global null.

Characterizing the breakdown point of p-values in nonlinear GLMs is highly nontrivial and challenging. First, the nonlinearity generally renders the MLE to take no analytical form, which makes it difficult to analyze its behavior in diverging dimensions. Second, conventional probabilistic arguments for establishing the central limit theorem of MLE only enable us to see one side of the coin, but not exactly at what point the distributional property fails to hold. To address these important challenges, we introduce novel geometric and probabilistic arguments presented later in the proofs of Theorems 2–3 that provide a rather delicate analysis of the MLE. In particular, our arguments unveil that the early breakdown point of p-values in nonlinear GLMs is essentially due to the *nonlinearity* of the mean function $\boldsymbol{\mu}(\cdot)$. This shows that p-values can behave wildly much early on in diverging dimensions when we travel from the linear regime to the nonlinear world as simple as the widely applied logistic regression; see the Introduction for detailed discussions on the p-values in diverging-dimensional linear models.

Before presenting the main results, let us look at the specific case of logistic regression model under the global null. In such a scenario, it holds

that $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{0}$ and thus $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = 4^{-1}I_n$, which results in

$$\mathbf{A}_n = \mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X} = 4^{-1}\mathbf{X}^T\mathbf{X}.$$

In particular, we see that when $n^{-1}\mathbf{X}^T\mathbf{X}$ is close to the identity matrix I_p, the asymptotic standard deviation of the jth component $\widehat{\beta}_j$ of the MLE $\widehat{\boldsymbol{\beta}}$ is close to $2n^{-1/2}$ when the asymptotic theory in (10) holds. As mentioned in the Introduction, when $p \geq n/2$ the MLE can blow up with excessively large variance, a strong evidence against the distributional property in (10). In fact, one can also observe inflated variance of the MLE relative to what is predicted by the asymptotic theory in (10) even when the dimensionality p grows at a slower rate with sample size n. As a consequence, the conventional p-values given by algorithms according to property (10) can be much biased toward zero and thus produce more significant discoveries than the truth. Such a breakdown of conventional p-values is delineated clearly in the simulation examples presented in Section 4.

3.2. *Main results.* We now present the formal results on the invalidity of GLM p-values in diverging dimensions.

THEOREM 2 (Uniform orthonormal design). *Assume that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ consisting of all $n \times p$ orthonormal matrices. Then for the logistic regression model under the global null, the asymptotic normality of the MLE established in (10) fails to hold when $p \sim n^{2/3}$, where $\sim$ stands for asymptotic order.*

THEOREM 3 (Correlated Gaussian design). *Assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$ nonsingular. Then for the logistic regression model under the global null, the same conclusion as in Theorem 2 holds.*

The key ingredients of our new geometric and probabilistic arguments are demonstrated in the proof of Theorem 2 in Section A.2. The assumption that the rescaled random design matrix $n^{-1/2}\mathbf{X}$ has the Haar measure on the Stiefel manifold $V_p(\mathbb{R}^n)$ greatly facilitates our technical analysis. The major theoretical finding is that the nonlinearity of the mean function $\boldsymbol{\mu}(\cdot)$ can be negligible in determining the asymptotic distribution of MLE as given in (10) when the dimensionality p grows at a slower rate than $n^{2/3}$, but such nonlinearity can become dominant and deform the conventional asymptotic normality when p grows at rate $n^{2/3}$ or faster. See the last paragraph of Section A.2 for more detailed in-depth discussions on such an interesting phenomenon.

Theorem 3 further establishes that the invalidity of GLM p-values in high dimensions beyond the scenario of orthonormal design matrices considered in Theorem 2. The breakdown of the conventional p-values can be regardless of the correlation structure of the covariates.

Our theoretical derivations detailed in the Appendix also suggest that the conventional p-values in nonlinear GLMs can generally fail to be valid when $p \sim n^{\alpha_0}$ with α_0 ranging between $1/2$ and $2/3$, which differs significantly from the phenomenon for linear models as discussed in the Introduction. The special feature of logistic regression model that the variance function $b''(\theta)$ takes the maximum value $1/4$ at natural parameter $\theta = 0$ leads to a higher transition point of $p \sim n^{\alpha_0}$ with $\alpha_0 = 2/3$ for the case of global null $\boldsymbol{\beta}_0 = \mathbf{0}$.

4. Numerical studies. We now delineate the journey of p-values of nonlinear GLMs in diverging dimensions as predicted by our major theoretical results in Section 3 with several simulation examples. Indeed, these theoretical results are well supported by the numerical studies.

4.1. *Simulation examples.* Following Theorems 2–3 in Section 3, we consider three examples of the logistic regression model (1). The response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$ has independent components and each y_i has Bernoulli distribution with parameter $e^{\theta_i}/(1+e^{\theta_i})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}_0$. In example 1, we generate the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ such that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ as in Theorem 2, while examples 2 and 3 assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$ as in Theorem 3. In particular, we choose $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{1 \leq j,k \leq p}$ with $\rho = 0, 0.5$, and 0.8 to reflect low, moderate, and high correlation levels among the covariates. Moreover, examples 1 and 2 assume the global null model with $\boldsymbol{\beta}_0 = \mathbf{0}$ following our theoretical results, whereas example 3 allows sparsity $s = \|\boldsymbol{\beta}_0\|_0$ to vary.

To examine the asymptotic results we set the sample size $n = 1000$. In each example, we consider a spectrum of dimensionality p with varying rate of growth with sample size n. As mentioned in the Introduction, the phase transition of perfect hyperplane separation happens at the point $p/n = 1/2$. Recall that Theorems 2–3 establish that the conventional GLM p-values can become invalid when $p \sim n^{2/3}$. We set $p = [n^{\alpha_0}]$ with α_0 in the grid $\{2/3 - 4\delta, \cdots, 2/3 - \delta, 2/3, 2/3 + \delta, \cdots, 2/3 + 4\delta, (\log(n) - \log(2))/\log(n)\}$ for $\delta = 0.05$. For example 3, we pick s signals uniformly at random among all but the first components, where a random half of them are chosen as 3 and the other half are set as -3.

The goal of the simulation examples is to investigate empirically when
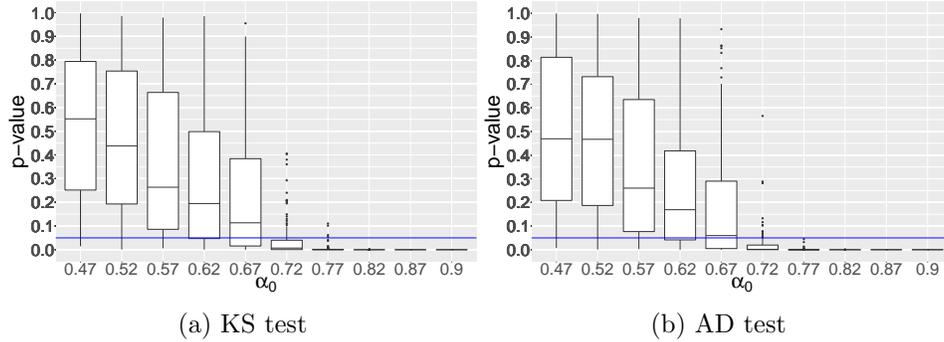
(a) KS test

(b) AD test

Fig 1: Results of KS and AD tests for testing the uniformity of GLM p-values in simulation example 1 for diverging-dimensional logistic regression model with uniform orthonormal design under global null. The vertical axis represents the p-value from the KS and AD tests, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

the conventional GLM p-values could break down in diverging dimensions. When the asymptotic theory for the MLE in (10) holds, the conventional p-values would be valid and distributed uniformly on the interval $[0, 1]$ under the null hypothesis. Note that the first covariate $\mathbf{x}_1$ is a null variable in each simulation example. Thus in each replication, we calculate the conventional p-value for testing the null hypothesis $H_0 : \beta_{0,1} = 0$. To check the validity of these p-values, we further test their uniformity.

For each simulation example, we first calculate the p-values for a total of $1,000$ replications as described above and then test the uniformity of these $1,000$ p-values using, for example, the Kolmogorov–Smirnov (KS) test [16, 22] and the Anderson–Darling (AD) test [1, 2]. We repeat this procedure $1,000$ times to obtain a final set of $1,000$ new p-values from each of these two uniformity tests. Specifically, the KS and AD test statistics for testing the uniformity on $[0, 1]$ are defined as

$$\text{KS} = \sup_{x \in [0,1]} |F_m(x) - x| \ \text{ and } \ \text{AD} = m \int_0^1 \frac{[F_m(x) - x]^2}{x(1 - x)} \, dx,$$

respectively, where $F_m(x) = m^{-1} \sum_{i=1}^m I_{(-\infty, x]}(x_i)$ is the empirical distribution function for a given sample $\{x_i\}_{i=1}^m$.

4.2. *Testing results.* For each simulation example, we apply both KS and AD tests to verify the asymptotic theory for the MLE in (10) by testing the uniformity of conventional p-values at significance level 0.05. As mentioned

(a) KS test for $\rho = 0.5$

(b) AD test for $\rho = 0.5$

(c) KS test for $\rho = 0.8$
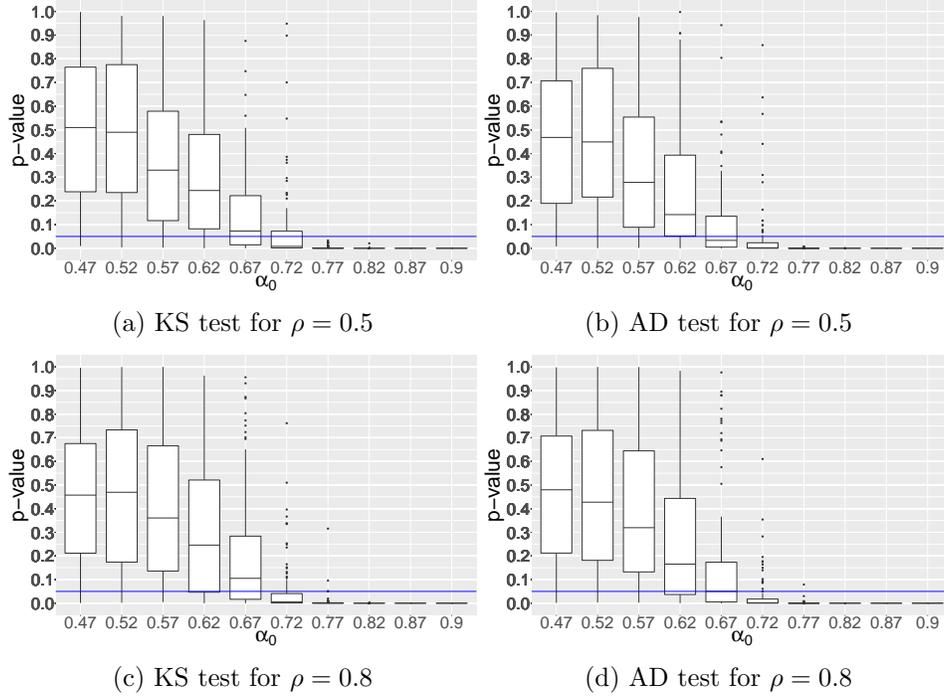
(d) AD test for $\rho = 0.8$

Fig 2: Results of KS and AD tests for testing the uniformity of GLM p-values in simulation example 2 for diverging-dimensional logistic regression model with correlated Gaussian design under global null for varying correlation level ρ. The vertical axis represents the p-value from the KS and AD tests, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

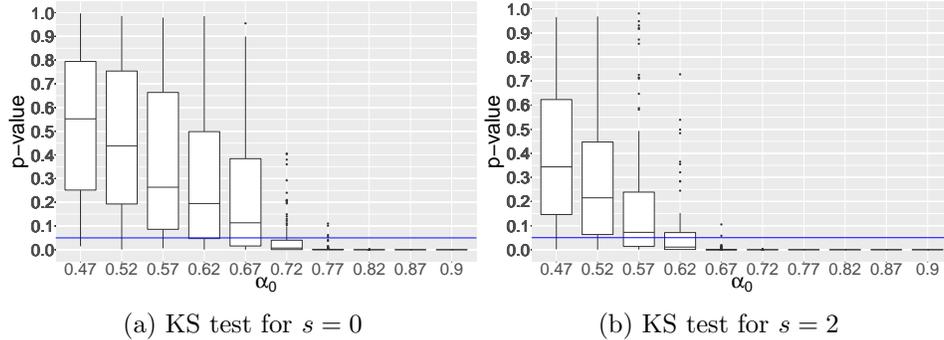(a) KS test for $s = 0$ (b) KS test for $s = 2$

Fig 3: Results of KS test for testing the uniformity of GLM p-values in simulation example 3 for diverging-dimensional logistic regression model with uncorrelated Gaussian design under global null for varying sparsity s. The vertical axis represents the p-value from the KS test, and the horizontal axis stands for the growth rate α_0 of dimensionality $p = [n^{\alpha_0}]$.

in Section 4.1, we end up with two sets of 1000 new p-values from the KS and AD tests. Figures 1–3 depict the boxplots of the p-values obtained from both KS and AD tests for simulation examples 1–3, respectively. In particular, we observe that the numerical results shown in Figures 1–2 for examples 1–2 are in line with our theoretical results established in Theorems 2–3, respectively, for diverging-dimensional logistic regression model under global null that the conventional p-values break down when $p \sim n^{\alpha_0}$ with $\alpha_0 = 2/3$. Figure 3 for example 3 examines the breakdown point of p-values with varying sparsity s. It is interesting to see that the breakdown point shifts even earlier when s increases as suggested in the discussions in Section 3.2. The results from the AD test are similar so we present only the results from the KS test for simplicity.

5. Discussions. In this paper we have provided characterizations of p-values in nonlinear GLMs with diverging dimensionality. The major findings are that the conventional p-values can remain valid when $p = o(n^{1/2})$, but can become invalid much earlier in nonlinear models of GLMs than in linear models, where the latter case can allow for $p = o(n)$. In particular, our theoretical results pinpoint the breakdown point of $p \sim n^{2/3}$ for p-values in diverging-dimensional logistic regression model under global null with uniform orthonormal design and correlated Gaussian design, as evidenced in the numerical results. It would be interesting to investigate such a phenomenon for more general class of random design matrices.

The problem of identifying the breakdown point of p-values becomes even more complicated and challenging when we move away from the setting of global null. Our technical analysis suggests that the breakdown point $p \sim n^{\alpha_0}$ can shift even earlier with α_0 ranging between $1/2$ and $2/3$. But the exact breakdown point can depend upon the number of signals s, the signal magnitude, and the correlation structure among the covariates in a rather complicated fashion. Thus more delicate mathematical analysis is needed to obtain the exact relationship. We leave such a problem for future investigation. Moving beyond the GLM setting will further complicate the theoretical analysis.

As we routinely produce p-values using algorithms, the phenomenon of nonuniformity of p-values occurring early in diverging dimensions unveiled in the paper poses useful cautions to researchers and practitioners when making decisions in real applications using results from p-value based methods. For instance, when testing the joint significance of covariates in diverging-dimensional nonlinear models, the effective sample size requirement should be checked before interpreting the testing results. Indeed, statistical inference in general high-dimensional nonlinear models is particularly challenging since obtaining accurate p-values is generally uneasy. One possible route is to bypass the use of p-values in certain tasks including the false discovery rate (FDR) control; see, for example, [5, 8] for some initial efforts made along this line.

APPENDIX A: PROOFS OF MAIN RESULTS

We provide the detailed proofs of Theorems 1–3 in this Appendix.

A.1. Proof of Theorem 1. To ease the presentation, we split the proof into two parts, where the first part locates the MLE $\widehat{\boldsymbol{\beta}}$ in an asymptotically shrinking neighborhood $\mathcal{N}_0$ of the true regression coefficient vector $\boldsymbol{\beta}_0$ with significant probability and the second part further establishes its asymptotic normality.

Part 1: Existence of a solution to score equation (4) in $\mathcal{N}_0$ under Condition 1 and probability bound (6). For simplicity, assume that the design matrix $\mathbf{X}$ is rescaled columnwise such that $\|\mathbf{x}_j\|_2 = \sqrt{n}$ for each $1 \leq j \leq p$. Consider an event

(11) $$\mathcal{E} = \left\{ \|\boldsymbol{\xi}\|_\infty \leq c_1^{-1/2} \sqrt{n \log n} \right\},$$

where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)]$. Note that for unbounded responses, the assumption of $\max_{j=1}^p \|\mathbf{x}_j\|_\infty < c_1^{1/2}\{n/(\log n)\}^{1/2}$ in Condition

1 entails that $c_1^{-1/2}\sqrt{\log n} < \min_{j=1}^p\{\|\mathbf{x}_j\|_2/\|\mathbf{x}_j\|_\infty\}$. Thus by $\|\mathbf{x}_j\|_2 = \sqrt{n}$, probability bound (6), and Bonferroni's inequality, we deduce

$$(12) \qquad P\left(\mathcal{E}\right) \geq 1 - \sum_{j=1}^p P\left(|\xi_j| > c_1^{-1/2}\sqrt{n\log n}\right)$$

$$\geq 1 - 2pn^{-1} = 1 - O\{n^{-(1-\alpha_0)}\},$$

since $p = O(n^{\alpha_0})$ for some $\alpha_0 \in (0,\gamma)$ with $\gamma \in (0, 1/2]$ by assumption. Hereafter we condition on the event $\mathcal{E}$ defined in (11) which holds with significant probability.

We will show that for sufficiently large n, the score equation (4) has a solution in the neighborhood $\mathcal{N}_0$ which is a hypercube. Define two vector-valued functions

$$\boldsymbol{\gamma}(\boldsymbol{\beta}) = (\gamma_1(\boldsymbol{\beta}), \cdots, \gamma_p(\boldsymbol{\beta}))^T = \mathbf{X}^T\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})$$

and

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}) - \boldsymbol{\gamma}(\boldsymbol{\beta}_0) - \boldsymbol{\xi}, \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Then equation (4) is equivalent to $\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbf{0}$. We need to show that the latter has a solution inside the hypercube $\mathcal{N}_0$. To this end, applying a second order Taylor expansion of $\boldsymbol{\gamma}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ with the Lagrange remainder term componentwise leads to

$$(13) \qquad \boldsymbol{\gamma}(\boldsymbol{\beta}) = \boldsymbol{\gamma}(\boldsymbol{\beta}_0) + \mathbf{X}^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{r},$$

where $\mathbf{r} = (r_1, \cdots, r_p)^T$ and for each $1 \leq j \leq p$,

$$r_j = \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)^T \nabla^2\gamma_j(\boldsymbol{\beta}_j)\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right)$$

with $\boldsymbol{\beta}_j$ some p-dimensional vector lying on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. It follows from (9) in Condition 1 that

$$(14) \quad \|\mathbf{r}\|_\infty \leq \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{j=1}^p \frac{1}{2}\lambda_{\max}\left[\mathbf{X}^T\text{diag}\left\{|\mathbf{x}_j| \circ \left|\boldsymbol{\mu}''\left(\mathbf{X}\boldsymbol{\delta}\right)\right|\right\}\mathbf{X}\right]\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$$

$$= O\left\{pn^{1-2\gamma}(\log n)^2\right\}.$$

Let us define another vector-valued function

$$(15) \qquad \overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) \equiv \left[\mathbf{X}^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}\right]^{-1}\boldsymbol{\Psi}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}_0 + \mathbf{u},$$

where $\mathbf{u} = -[\mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}]^{-1}(\boldsymbol{\xi} - \mathbf{r})$. It follows from (11), (14), and (8) in Condition 1 that for any $\boldsymbol{\beta} \in \mathcal{N}_0$,

$$(16) \qquad \|\mathbf{u}\|_\infty \leq \left\| \left[\mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X} \right]^{-1} \right\|_\infty (\|\boldsymbol{\xi}\|_\infty + \|\mathbf{r}\|_\infty)$$
$$= O\left[b_n n^{-1/2} \sqrt{\log n} + b_n p n^{-2\gamma} (\log n)^2 \right].$$

By the assumptions of $p = O(n^{\alpha_0})$ with constant $\alpha_0 \in (0, \gamma)$ and $b_n = o\{\min(n^{1/2-\gamma}\sqrt{\log n}, n^{2\gamma-\alpha_0-1/2}/(\log n)^2\}$, we have

$$\|\mathbf{u}\|_\infty = o(n^{-\gamma} \log n).$$

Thus in light of (15), it holds for large enough n that when $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)_j = n^{-\gamma}\sqrt{\log n}$,

$$(17) \qquad \overline{\Psi}_j(\boldsymbol{\beta}) \geq n^{-\gamma} \sqrt{\log n} - \|\mathbf{u}\|_\infty \geq 0,$$

and when $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)_j = -n^{-\gamma}\sqrt{\log n}$,

$$(18) \qquad \overline{\Psi}_j(\boldsymbol{\beta}) \leq -n^{-\gamma} \sqrt{\log n} + \|\mathbf{u}\|_\infty \leq 0,$$

where $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = (\overline{\Psi}_1(\boldsymbol{\beta}), \cdots, \overline{\Psi}_p(\boldsymbol{\beta}))^T$.

By the continuity of the vector-valued function $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta})$, (17), and (18), Miranda's existence theorem [24] ensures that equation $\overline{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = \mathbf{0}$ has a solution $\widehat{\boldsymbol{\beta}}$ in $\mathcal{N}_0$. Clearly, $\widehat{\boldsymbol{\beta}}$ also solves equation $\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbf{0}$ in view of (15). Therefore, we have shown that score equation (4) indeed has a solution $\widehat{\boldsymbol{\beta}}$ in $\mathcal{N}_0$. The strict concavity of the log-likelihood function (2) by assumptions for model (1) entails that $\widehat{\boldsymbol{\beta}}$ is the MLE.

Part 2: Conventional asymptotic normality of the MLE $\widehat{\boldsymbol{\beta}}$. Fix any $1 \leq j \leq p$. In light of (15), we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{A}_n^{-1}(\boldsymbol{\xi} - \mathbf{r})$, which results in

$$(19) \qquad (\mathbf{A}_n^{-1})_{jj}^{-1/2}(\widehat{\beta}_j - \beta_{0,j}) = (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T \mathbf{A}_n^{-1}\boldsymbol{\xi} - (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T \mathbf{A}_n^{-1}\mathbf{r}$$

with $\mathbf{e}_j \in \mathbb{R}^p$ having one for the jth component and zero otherwise. Note that since the smallest and largest eigenvalues of $n^{-1}\mathbf{A}_n$ are bounded away from 0 and ∞ by Condition 2, it is easy to show that $(\mathbf{A}_n^{-1})_{jj}^{-1/2}$ is of exact order $n^{1/2}$. In view of (16), it holds on the event $\mathcal{E}$ defined in (11) that

$$\begin{aligned} \left\| \mathbf{A}_n^{-1}\mathbf{r} \right\|_\infty &\leq \left\| \left[\mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X} \right]^{-1} \right\|_\infty \|\mathbf{r}\|_\infty \\ &= O\left[b_n p n^{-2\gamma}(\log n)^2 \right] = o(n^{-1/2}), \end{aligned}$$

since $b_n = o\{n^{2\gamma - \alpha_0 - 1/2}/(\log n)^2\}$ by assumption. This leads to

$$(20) \qquad (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{r} = O(n^{1/2}) \cdot o_P(n^{-1/2}) = o_P(1).$$

It remains to consider the term $(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i$, where $\eta_i = (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{z}_i[y_i - b'(\theta_{0,i})]$. Clearly, the n random variables η_i's are independent with mean 0 and

$$\sum_{i=1}^n \text{var}\eta_i = (\mathbf{A}_n^{-1})_{jj}^{-1}\mathbf{e}_j^T\mathbf{A}_n^{-1}(\phi\mathbf{A}_n)\mathbf{A}_n^{-1}\mathbf{e}_j = \phi.$$

It follows from Condition 2 and the Cauchy–Schwarz inequality that

$$\begin{aligned}
\sum_{i=1}^n E\,|\eta_i|^3 &= \sum_{i=1}^n \left|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 E\,\left|y_i - b'\,(\theta_{0,i})\right|^3 \\
&= O(1)\sum_{i=1}^n \left|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\mathbf{z}_i\right|^3 \\
&\leq O(1)\sum_{i=1}^n \left\|(\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1/2}\right\|_2^3 \left\|\mathbf{A}_n^{-1/2}\mathbf{z}_i\right\|_2^3 \\
&= O(1)\sum_{i=1}^n \left(\mathbf{z}_i^T\mathbf{A}_n^{-1}\mathbf{z}_i\right)^{3/2} = o(1).
\end{aligned}$$

Thus an application of Lyapunov's theorem yields

$$(21) \qquad (\mathbf{A}_n^{-1})_{jj}^{-1/2}\mathbf{e}_j^T\mathbf{A}_n^{-1}\boldsymbol{\xi} = \sum_{i=1}^n \eta_i \xrightarrow{\mathscr{D}} N(0,\phi).$$

By Slutsky's lemma, we see from (19)–(21) that

$$(\mathbf{A}_n^{-1})_{jj}^{-1/2}(\widehat{\beta}_j - \beta_{0,j}) \xrightarrow{\mathscr{D}} N(0,\phi),$$

showing the asymptotic normality of each component $\widehat{\beta}_j$ of the MLE $\widehat{\boldsymbol{\beta}}$. This completes the proof of Theorem 1.

A.2. Proof of Theorem 2. To prove the conclusion in Theorem 2, we use the proof by contradiction. Let us make an assumption (A) that the asymptotic normality (10) in Theorem 1 which has been proved to hold when $p = o(n^{1/2})$ continues to hold when $p \sim n^{\alpha_0}$ for some constant $1/2 < \alpha_0 \leq 1$, where $\sim$ stands for asymptotic order. As shown in Section 3.1, in the case

of logistic regression under global null (that is, $\boldsymbol{\beta}_0 = \mathbf{0}$) with deterministic rescaled orthonormal design matrix $\mathbf{X}$ (in the sense of $n^{-1}\mathbf{X}^T\mathbf{X} = I_p$) the limiting distribution in (10) by assumption (A) becomes

$$(22) \qquad 2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1),$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ is the MLE.

Let us now assume that the rescaled random design matrix $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ which can be thought of as the space of all $n \times p$ orthonormal matrices. Then it follows from (22) that

$$(23) \qquad 2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1) \ \text{ conditional on } \mathbf{X}.$$

Based on the limiting distribution in (23), we can make two observations. First, it holds that

$$(24) \qquad 2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1)$$

unconditional on the design matrix $\mathbf{X}$. Second, $\widehat{\beta}_j$ is asymptotically independent of the design matrix $\mathbf{X}$, and so is the MLE $\widehat{\boldsymbol{\beta}}$.

Since the distribution of $n^{-1/2}\mathbf{X}$ is assumed to be the Haar measure on the Stiefel manifold $V_p(\mathbb{R}^n)$, we have

$$(25) \qquad n^{-1/2}\mathbf{X}\mathbf{Q} \overset{d}{=} n^{-1/2}\mathbf{X},$$

where $\mathbf{Q}$ is any fixed $p \times p$ orthogonal matrix and $\overset{d}{=}$ stands for equal in distribution. Recall that the MLE $\widehat{\boldsymbol{\beta}}$ solves the score equation (4), which is in turn equivalent to equation

$$(26) \qquad \mathbf{Q}^T\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0}$$

since $\mathbf{Q}$ is orthogonal. We now use the fact that the model is under global null which entails that the response vector $\mathbf{y}$ is independent of the design matrix $\mathbf{X}$. Combining this fact with (25)–(26) yields

$$(27) \qquad \mathbf{Q}^T\widehat{\boldsymbol{\beta}} \overset{d}{=} \widehat{\boldsymbol{\beta}}$$

by noting that $\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\mathbf{Q})(\mathbf{Q}^T\boldsymbol{\beta})$. Since the distributional identity (27) holds for any fixed $p \times p$ orthogonal matrix $\mathbf{Q}$, we conclude that the MLE $\widehat{\boldsymbol{\beta}}$ has a spherical distribution on $\mathbb{R}^p$. It is a well-known fact that all the

marginal characteristic functions of a spherical distribution have the same generator. Such a fact along with (24) entails that

(28) $2^{-1}n^{1/2}\widehat{\boldsymbol{\beta}}$ is asymptotically close to $N(\mathbf{0}, I_p)$.

To simplify the exposition, let us now make the asymptotic limit exact and assume that

(29) $\widehat{\boldsymbol{\beta}} \sim N(\mathbf{0}, 4n^{-1}I_p)$ and is independent of $\mathbf{X}$.

The remaining analysis focuses on the score equation (4) which is solved exactly by the MLE $\widehat{\boldsymbol{\beta}}$, that is,

$$(30) \qquad \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}})] = \mathbf{0},$$

which leads to

$$(31) \qquad \boldsymbol{\xi} \equiv n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] = n^{-1/2}\mathbf{X}^T[\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}(\mathbf{0})] \equiv \boldsymbol{\eta}.$$

Let us first consider the random variable $\boldsymbol{\xi}$ defined in (31). Note that $2[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})]$ has independent and identically distributed (i.i.d.) components each taking value 1 or -1 with equal probability $1/2$, and is independent of $\mathbf{X}$. Thus since $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$, it is easy to see that

$$(32) \qquad \boldsymbol{\xi} = n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] \overset{d}{=} 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all components being one. Using similar arguments as before, we can show that $\boldsymbol{\xi}$ has a spherical distribution on $\mathbb{R}^p$. Thus the joint distribution of $\boldsymbol{\xi}$ is determined completely by the marginal distribution of $\boldsymbol{\xi}$. For each $1 \leq j \leq p$, denote by ξ_j the jth component of $\boldsymbol{\xi} = 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1}$ using the distributional representation in (32). Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ with each $\mathbf{x}_j \in \mathbb{R}^n$. Then we have

$$(33) \qquad \xi_j = 2^{-1}n^{-1/2}\mathbf{x}_j^T\mathbf{1} \overset{d}{=} 2^{-1}(n^{1/2}/\|\widetilde{\mathbf{x}}_j\|_2)n^{-1/2}\widetilde{\mathbf{x}}_j^T\mathbf{1},$$

where $\widetilde{\mathbf{x}}_j \sim N(\mathbf{0}, 4^{-1}I_n)$. It follows from (33) and the concentration phenomenon of Gaussian measures that each ξ_j is asymptotically close to $N(0, 4^{-1})$ and thus consequently $\boldsymbol{\xi}$ is asymptotically close to $N(\mathbf{0}, 4^{-1}I_p)$. *A key fact (i) for the finite-sample distribution of $\boldsymbol{\xi}$ is that the standard deviation of each component ξ_j converges to $1/2$ at rate $O_P(n^{-1/2})$ that does not depend upon the dimensionality p at all.*

We now turn our attention to the second term $\boldsymbol{\eta}$ defined in (31). In view of (29) and the fact that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$, we can show that with significant probability,

$$(34) \qquad \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_\infty \leq o(1)$$

for $p \sim n^{\alpha_0}$ with $\alpha_0 < 1$. The uniform bound in (34) enables us to apply the mean value theorem for the vector-valued function $\boldsymbol{\eta}$ around $\boldsymbol{\beta}_0 = \mathbf{0}$, which results in

$$(35) \qquad \boldsymbol{\eta} = n^{-1/2}\mathbf{X}^T[\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}(\mathbf{0})] = 4^{-1}n^{-1/2}\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{r}$$
$$= 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}} + \mathbf{r}$$

since $n^{-1/2}\mathbf{X}$ is assumed to be orthonormal, where

$$(36) \qquad \mathbf{r} = n^{-1/2}\mathbf{X}^T\left\{\int_0^1 \left[\boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}}) - 4^{-1}I_n\right] dt\right\}\mathbf{X}\widehat{\boldsymbol{\beta}}.$$

Here, the remainder term $\mathbf{r} = (r_1, \cdots, r_p)^T \in \mathbb{R}^p$ is stochastic and each component r_j is generally of order $O_P\{p^{1/2}n^{-1/2}\}$ in light of (29) when the true model may deviate from the global null case of $\boldsymbol{\beta}_0 = \mathbf{0}$.

Since our focus in this theorem is the logistic regression model under the global null, we can in fact claim that each component r_j is generally of order $O_P\{pn^{-1}\}$, which is a better rate of convergence than the one mentioned above thanks to the assumption of $\boldsymbol{\beta}_0 = \mathbf{0}$. To prove this claim, note that the variance function $b''(\theta)$ is symmetric in $\theta \in \mathbb{R}$ and takes the maximum value $1/4$ at $\theta = 0$. Thus in view of (34), we can show that with significant probability,

$$(37) \quad 4^{-1}I_n - \boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}}) \geq c\,\mathrm{diag}\{(t\mathbf{X}\widehat{\boldsymbol{\beta}}) \circ (t\mathbf{X}\widehat{\boldsymbol{\beta}})\} = ct^2\mathrm{diag}\{(\mathbf{X}\widehat{\boldsymbol{\beta}}) \circ (\mathbf{X}\widehat{\boldsymbol{\beta}})\}$$

for all $t \in [0, 1]$, where $c > 0$ is some constant and $\geq$ stands for the inequality for positive semidefinite matrices. Moreover, it follows from (29) and the fact that $n^{-1/2}\mathbf{X}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^n)$ that with significant probability, all the n components of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ are concentrated in the order of $p^{1/2}n^{-1/2}$. This result along with (37) and the fact that $n^{-1}\mathbf{X}^T\mathbf{X} = I_p$ entails that with significant probability,

$$(38) \qquad n^{-1/2}\mathbf{X}^T\left\{\int_0^1 \left[4^{-1}I_n - \boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}})\right] dt\right\}\mathbf{X}$$
$$\geq n^{-1/2}\mathbf{X}^T\left\{\int_0^1 c_*t^2pn^{-1}dt\right\}\mathbf{X}$$
$$= 3^{-1}c_*pn^{-3/2}\mathbf{X}^T\mathbf{X} = 3^{-1}c_*pn^{-1/2}I_p,$$

where $c_* > 0$ is some constant. Thus combining (36), (38), and (29) proves the above claim.

We make two important observations about the remainder term $\mathbf{r}$ in (35). First, $\mathbf{r}$ has a spherical distribution on $\mathbb{R}^p$. This is because by (35) and (31) it holds that

$$\mathbf{r} = \boldsymbol{\eta} - 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}} = \boldsymbol{\xi} - 4^{-1}n^{1/2}\widehat{\boldsymbol{\beta}},$$

which has a spherical distribution on $\mathbb{R}^p$. Thus the joint distribution of $\mathbf{r}$ is determined completely by the marginal distribution of $\mathbf{r}$. Second, for the nonlinear setting of logistic regression model, the appearance of the remainder term $\mathbf{r}$ in (35) is due solely to *the nonlinearity of the mean function* $\boldsymbol{\mu}(\cdot)$, and we have shown that each component r_j can indeed achieve the worst-case order pn^{-1} in probability. For each $1 \leq j \leq p$, denote by η_j the jth component of $\boldsymbol{\eta}$. Then in view of (29) and (35), *a key fact (ii) for the finite-sample distribution of $\boldsymbol{\eta}$ is that the standard deviation of each component η_j converges to $1/2$ at rate $O_P\{pn^{-1}\}$ that generally does depend upon the dimensionality p.*

Finally, we are ready to compare the two random variables $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ on the two sides of equation (31). Since equation (31) is a distributional identity in $\mathbb{R}^p$, naturally the square root of the sum of varξ_j's and the square root of the sum of varη_j's are expected to converge to the common value $2^{-1}p^{1/2}$ at rates that are asymptotically negligible. However, the former has rate $p^{1/2}O_P(n^{-1/2}) = O_P\{p^{1/2}n^{-1/2}\}$, whereas the latter has rate $p^{1/2}O_P\{pn^{-1}\} = O_P\{p^{3/2}n^{-1}\}$. A key consequence is that when $p \sim n^{\alpha_0}$ for some constant $2/3 \leq \alpha_0 < 1$, there is a profound difference between the two asymptotic rates in that the former rate is $O_P\{n^{-(1-\alpha_0)/2}\} = o_P(1)$, while the latter rate becomes $O_P\{n^{3\alpha_0/2-1}\}$ which is now asymptotically diverging or nonvanishing. Such an intrinsic asymptotic difference is, however, prohibited by the distributional identity (31) in $\mathbb{R}^p$, which results in a contradiction. Therefore, we have now argued that assumption (A) we started with for $2/3 \leq \alpha_0 < 1$ must be false, that is, the asymptotic normality (10) which has been proved to hold when $p = o(n^{1/2})$ generally would not continue to hold when $p \sim n^{\alpha_0}$ with constant $2/3 \leq \alpha_0 \leq 1$. In other words, we have proved the invalidity of the conventional GLM p-values in this regime of diverging dimensionality, which concludes the proof of Theorem 2.

A.3. Proof of Theorem 3. By assumption, $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$ nonsingular. Let us first make a useful observation. For the general case of nonsingular covariance matrix $\boldsymbol{\Sigma}$, we can introduce a change of variable by letting $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$ and correspondingly $\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}$. Clearly, $\widetilde{\mathbf{X}} \sim N(\mathbf{0}, I_n \otimes I_p)$ and the MLE for the transformed parameter

vector $\widetilde{\boldsymbol{\beta}}$ is exactly $\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ denotes the MLE under the original design matrix $\mathbf{X}$. Thus to show the breakdown point of the conventional asymptotic normality of the MLE, it suffices to focus on the specific case of $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$.

Hereafter we assume that $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes I_p)$ with $p = o(n)$. The rest of the arguments are similar to those in the proof of Theorem 2 in Section A.2 except for some modifications needed for the case of Gaussian design. Specifically, for the case of logistic regression model under global null (that is, $\boldsymbol{\beta}_0 = \mathbf{0}$), the limiting distribution in (10) becomes

$$(39) \qquad 2^{-1}n^{1/2}\widehat{\beta}_j \xrightarrow{\mathscr{D}} N(0,1),$$

since $n^{-1}\mathbf{X}^T\mathbf{X} \to I_p$ almost surely in spectrum and thus $4^{-1}n(\mathbf{A}_n^{-1})_{jj} \to 1$ in probability as $n \to \infty$. Here, we have used a claim that both the largest and smallest eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$ converge to 1 almost surely as $n \to \infty$ for the case of $p = o(n)$, which can be shown by using the classical results from random matrix theory (RMT) [12, 21, 4].

Note that since $\mathbf{X} \sim N(0, I_n \otimes I_p)$, it holds that

$$(40) \qquad n^{-1/2}\mathbf{X}\mathbf{Q} \overset{d}{=} n^{-1/2}\mathbf{X},$$

where $\mathbf{Q}$ is any fixed $p \times p$ orthogonal matrix and $\overset{d}{=}$ stands for equal in distribution. By $\mathbf{X} \sim N(0, I_n \otimes I_p)$, it is also easy to see that

$$(41) \qquad \boldsymbol{\xi} = n^{-1/2}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})] \overset{d}{=} 2^{-1}n^{-1/2}\mathbf{X}^T\mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all components being one. In view of (29) and the assumption of $\mathbf{X} \sim N(0, I_n \otimes I_p)$, we can show that with significant probability,

$$(42) \qquad \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_\infty \leq o(1)$$

for $p \sim n^{\alpha_0}$ with constant $\alpha_0 < 1$. It holds further that with significant probability, all the n components of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ are concentrated in the order of $p^{1/2}n^{-1/2}$. This result along with (37) and the fact that $n^{-1}\mathbf{X}^T\mathbf{X} \to I_p$ almost surely in spectrum entails that with asymptotic probability one,

$$(43) \qquad n^{-1/2}\mathbf{X}^T \left\{ \int_0^1 \left[4^{-1}I_n - \boldsymbol{\Sigma}(t\mathbf{X}\widehat{\boldsymbol{\beta}}) \right] dt \right\} \mathbf{X}$$
$$\geq n^{-1/2}\mathbf{X}^T \left\{ \int_0^1 c_* t^2 p n^{-1} dt \right\} \mathbf{X}$$
$$= 3^{-1}c_* p n^{-3/2}\mathbf{X}^T\mathbf{X} \to 3^{-1}c_* p n^{-1/2} I_p,$$

where $c_* > 0$ is some constant. This completes the proof of Theorem 3.

REFERENCES

[1] ANDERSON, T. W. and DARLING, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* **23** 193–212.

[2] ANDERSON, T. W. and DARLING, D. A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association* **49** 765–769.

[3] ATHEY, S., IMBENS, G. W. and WAGER, S. (2016). Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*.

[4] BAI, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sin.* **9** 611–677.

[5] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085.

[6] BEAN, D., BICKEL, P. J., KAROUI, N. E. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences of the United States of America* **110** 14563–14568.

[7] CANDÈS, E. J. (2016). Private communication.

[8] CANDÈS, E. J., FAN, Y., JANSON, L. and LV, J. (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *Manuscript*.

[9] FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57** 5467–5484.

[10] FAN, Y. and LV, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association* **108** 1044–1061.

[11] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32** 928–961.

[12] GEMAN, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8** 252–261.

[13] GUO, B. and CHEN, S. X. (2016). Tests for high dimensional generalized linear models. *J. R. Statist. Soc. B* **78** 1079–1102.

[14] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1** 799–821.

[15] KAROUI, N. E., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences of the United States of America* **110** 14557–14562.

[16] KOLMOGOROV, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.* **4** 83–91.

[17] LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B* **76** 141–167.

[18] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

[19] PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *The Annals of Statistics* **12** 1298–1309.

[20] PORTNOY, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal Approximation. *The Annals of Statistics* **13** 1403–1417.

[21] SILVERSTEIN, J. W. (1985). The smallest eigenvalue of a large dimensional Wishart matrix. *Ann. Probab.* **13** 1364–1368.

[22] SMIRNOV, N. (1948). Table for estimating the goodness of fit of empirical distribu-

tions. *Annals of Mathematical Statistics* **19** 279–281.

[23] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.

[24] VRAHATIS, M. N. (1989). A short proof and a generalization of Miranda's existence theorem. *Proceedings of the American Mathematical Society* **107** 701–703.

DATA SCIENCES AND OPERATIONS DEPARTMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CA 90089
USA
E-MAIL: fanyingy@marshall.usc.edu
        jinchilv@marshall.usc.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CA 90089
USA
E-MAIL: demirkay@usc.edu