A modern maximum-likelihood theory for high-dimensional logistic regression

Pragya Sur[a,1,2] and Emmanuel J. Candès[a,b,1,2]

[a]Department of Statistics, Stanford University, Stanford, CA 94305; and [b]Department of Mathematics, Stanford University, Stanford, CA 94305

Contributed by Emmanuel J. Candès, September 19, 2018 (sent for review June 21, 2018; reviewed by Nancy M. Reid and Huibin [Harry] Zhou)

Students in statistics or data science usually learn early on that when the sample size *n* is large relative to the number of variables *p*, fitting a logistic model by the method of maximum likelihood produces estimates that are consistent and that there are well-known formulas that quantify the variability of these estimates which are used for the purpose of statistical inference. We are often told that these calculations are approximately valid if we have 5 to 10 observations per unknown parameter. This paper shows that this is far from the case, and consequently, inferences produced by common software packages are often unreliable. Consider a logistic model with independent features in which *n* and *p* become increasingly large in a fixed ratio. We prove that (*i*) the maximum-likelihood estimate (MLE) is biased, (*ii*) the variability of the MLE is far greater than classically estimated, and (*iii*) the likelihood-ratio test (LRT) is not distributed as a χ^2. The bias of the MLE yields wrong predictions for the probability of a case based on observed values of the covariates. We present a theory, which provides explicit expressions for the asymptotic bias and variance of the MLE and the asymptotic distribution of the LRT. We empirically demonstrate that these results are accurate in finite samples. Our results depend only on a single measure of signal strength, which leads to concrete proposals for obtaining accurate inference in finite samples through the estimate of this measure.

logistic regression | high-dimensional inference | maximum-likelihood estimate | likelihood-ratio test

Logistic regression (1, 2) is one of the most frequently used models to estimate the probability of a binary response from the value of multiple features/predictor variables. It is widely used in the social sciences, the finance industry, the medical sciences, and so on. As an example, a typical application of logistic regression may be to predict the risk of developing a given coronary heart disease from a patient's observed characteristics. Consequently, graduate students in statistics and many fields that involve data analysis learn about logistic regression, perhaps before any other nonlinear multivariate model. In particular, most students know how to interpret the excerpt of the computer output from Fig. 1, which displays regression coefficient estimates, standard errors, and P values for testing the significance of the regression coefficients. In textbooks we learn the following: (*i*) Fitting a model via maximum likelihood produces estimates that are approximately unbiased. (*ii*) There are formulas to estimate the accuracy or variability of the maximum-likelihood estimate (MLE) (used in the computer output from Fig. 1).

These approximations come from asymptotic results. Imagine we have n independent observations $(y_i, \boldsymbol{X}_i)$ where $y_i \in \{0, 1\}$ is the response variable and $\boldsymbol{X}_i \in \mathbb{R}^p$ the vector of predictor variables. The logistic model posits that the probability of a case conditional on the covariates is given by

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{X}_i) = \rho'(\boldsymbol{X}_i'\boldsymbol{\beta}),$$

where $\rho'(t) = e^t/(1 + e^t)$ is the standard sigmoidal function. When p is fixed and $n \to \infty$, the MLE $\hat{\boldsymbol{\beta}}$ obeys

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \boldsymbol{I}_{\boldsymbol{\beta}}^{-1}), \qquad [1]$$

where $\boldsymbol{I}_{\boldsymbol{\beta}}$ is the $p \times p$ Fisher information matrix evaluated at the true $\boldsymbol{\beta}$ (3). A classical way of understanding Eq. **1** is in the case where the pairs $(\boldsymbol{X}_i, y_i)$ are i.i.d. and the covariates $\boldsymbol{X}_i$ are drawn from a distribution obeying mild conditions so that the MLE exists and is unique. Now the limiting result Eq. **1** justifies the first claim of near unbiasedness. Further, software packages then return standard errors by evaluating the inverse Fisher information matrix at the MLE $\hat{\boldsymbol{\beta}}$ [this is what R (4) does in Fig. 1]. In turn, these standard errors are then used for the purpose of statistical inference; for instance, they are used to produce P values for testing the significance of regression coefficients, which researchers use in thousands of scientific studies.

Another well-known result in logistic regression is Wilks' theorem (5), which gives the asymptotic distribution of the likelihood-ratio test (LRT): (*iii*) Consider the likelihood ratio obtained by dropping k variables from the model under study. Then under the null hypothesis that none of the dropped variables belongs to the model, twice the log-likelihood ratio (LLR) converges to a χ^2 distribution with k degrees of freedom in the limit of large samples. Once more, this approximation is often used in many statistical software packages to obtain P values for testing the significance of individual and/or groups of coefficients.

1. Failures in Moderately Large Dimensions

New technologies now produce extremely large datasets, often with huge numbers of features on each of a comparatively small

Significance

Logistic regression is a popular model in statistics and machine learning to fit binary outcomes and assess the statistical significance of explanatory variables. Here, the classical theory of maximum-likelihood (ML) estimation is used by most software packages to produce inference. In the now common setting where the number of explanatory variables is not negligible compared with the sample size, we show that classical theory leads to inferential conclusions that cannot be trusted. We develop a theory that provides expressions for the bias and variance of the ML estimate and characterizes the asymptotic distribution of the likelihood-ratio statistic under some assumptions regarding the distribution of the explanatory variables. This theory can be used to provide valid inference.

Author contributions: P.S. and E.J.C. designed research, performed research, and wrote the paper.

Reviewers: N.M.R., University of Toronto; and H.Z., Yale University.

The authors declare no conflict of interest.

[1]P.S. and E.J.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: candes@stanford.edu or pragya@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810420116/-/DCSupplemental.

Published online July 1, 2019.

```
> fit = glm(y ~ X, family = binomial)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.25602    0.43191   0.593  0.55334
X1           7.78102    4.09069   1.902  0.05715 .
X2           9.80854    5.66019   1.733  0.08311 .
X3          -8.14106    5.50490  -1.479  0.13917
X4           0.01953    5.99945   0.003  0.99740
X5          -5.18298    3.88752  -1.333  0.18245
X6           9.48063    4.65335   2.037  0.04161 *

...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 1. Excerpt from an object of class "glm" obtained by fitting a logistic model in R. The coefficient estimates $\hat{\beta}_j$ are obtained by maximum likelihood, and for each variable, R provides an estimate of the SD of $\hat{\beta}_j$ as well as a *P* value for testing whether $\beta_j = 0$ or not.

number of experimental units. However, software packages and practitioners continue to perform calculations as if classical theory applies and, therefore, the main issue is this: Do these approximations hold in high-dimensional settings where p is not vanishingly small compared with n?

To address this question, we begin by showing results from an empirical study. Throughout this section, we set $n = 4,000$ and unless otherwise specified, $p = 800$ (so that the "dimensionality" p/n is equal to 1/5). We work with a matrix of covariates, which has i.i.d. $\mathcal{N}(0, 1/n)$ entries, and different types of regression coefficients scaled in such a way that

$$\gamma^2 := \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = 5.$$

This is a crucial point: We want to make sure that the size of the log-odds ratio $\boldsymbol{X}_i'\boldsymbol{\beta}$ does not increase with n or p, so that $\rho'(\boldsymbol{X}_i'\boldsymbol{\beta})$ is not trivially equal to either 0 or 1. Instead, we want to be in a regime where accurate estimates of $\boldsymbol{\beta}$ translate into a precise evaluation of a nontrivial probability. With our scaling $\gamma = \sqrt{5} \approx 2.236$, about 95% of the observations will be such that $-4.472 \le \boldsymbol{X}_i'\boldsymbol{\beta} \le 4.472$ so that $0.011 \le \rho'(\boldsymbol{X}_i'\boldsymbol{\beta}) \le 0.989$.

Unbiasedness? Fig. 2 plots the true and fitted coefficients in the setting where one-quarter of the regression coefficients have a magnitude equal to 10, and the rest are 0. Half of the nonzero coefficients are positive and the other half are negative. A striking feature is that the black curve does not pass through the center of the blue scatter. This disagrees with what we would expect from classical theory. Clearly, the regression estimates are not close to being unbiased. When the true effect size β_j is positive, we see that the MLE tends to overestimate it. Symmetrically, when β_j is negative, the MLE tends to underestimate the effect sizes in the sense that the fitted values are in the same direction but with magnitudes that are too large. In other words, for most indexes $|\hat{\beta}_j| > |\beta_j|$ so that we are overestimating the magnitudes of the effects.

The bias is not specific to this example as the theory we develop in this paper will make clear. Consider a case where the entries of $\boldsymbol{\beta}$ are drawn i.i.d. from $\mathcal{N}(3, 16)$ (the setup is otherwise unchanged). Fig. 3*A* shows that the pairs $(\beta_j, \hat{\beta}_j)$ are not distributed around a straight line of slope 1; rather, they are distributed around a line with a larger slope. Our theory predicts that the points should be scattered around a line with slope 1.499 shown in red, as if we could think that $\mathbb{E}\hat{\beta}_j \approx 1.499\beta_j$.

This bias is highly problematic for estimating the probability of our binary response. Suppose we are given a vector of covariates $\boldsymbol{X}_*$ and estimate the regression function $f(\boldsymbol{X}_*) = \mathbb{P}(y = 1 \mid \boldsymbol{X}_*)$ with

$$\hat{f}(\boldsymbol{X}_*) = \rho'(\boldsymbol{X}_*'\hat{\boldsymbol{\beta}}).$$

Then because we tend to overestimate the magnitudes of the effects, we will also tend to overestimate or underestimate the probabilities depending on whether $f(\boldsymbol{X}_*)$ is greater or less than a half. This is illustrated in Fig. 3*B*. Observe that when $f(\boldsymbol{X}_*) < 1/2$, many predictions tend to be close to 0, even when $f(\boldsymbol{X}_*)$ is nowhere near 0. A similar behavior is obtained by symmetry; when $f(\boldsymbol{X}_*) > 1/2$, we see a shrinkage toward the other end point, namely, 1. Hence, we see a shrinkage toward the extremes and the phenomenon is amplified as the true probability $f(\boldsymbol{X}_*)$ approaches 0 or 1. Expressed differently, the MLE may predict that an outcome is almost certain (i.e., $\hat{f}$ is close to 0 or 1) when, in fact, the outcome is not at all certain. This behavior is misleading.

Accuracy of Classical Standard Errors? Consider the same matrix $\boldsymbol{X}$ as before and regression coefficients now sampled as follows: Half of the β_js are i.i.d. draws from $\mathcal{N}(7, 1)$, and the other half vanish. Fig. 4*A* shows standard errors computed via Monte Carlo of maximum-likelihood (ML) estimates $\hat{\beta}_j$ corresponding to null coordinates. This is obtained by fixing the signal $\boldsymbol{\beta}$ and resampling the response vector and covariate matrix 10,000 times. Note that for any null coordinate, the classical estimate of SE based on the inverse Fisher information can be explicitly calculated in this setting and turns out to be equal to 2.66 (*SI Appendix*, section A). Since the SE values evidently concentrate

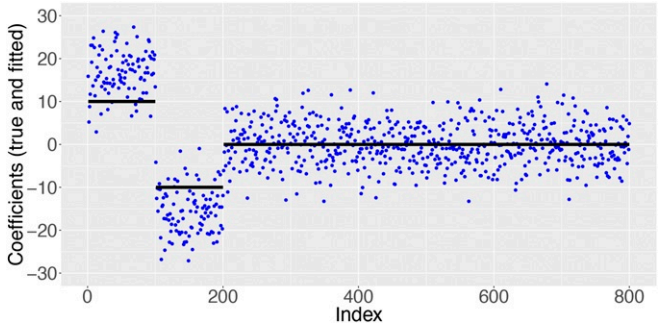

Fig. 2. True signal values β_j in black and corresponding ML estimates $\hat{\beta}_j$ (blue points). Observe that estimates of effect magnitudes are seriously biased upward.

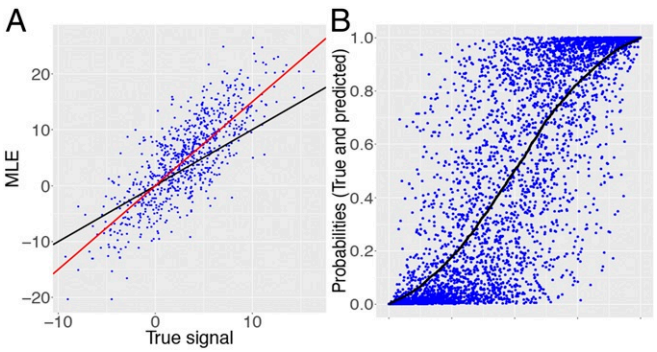

Fig. 3. (*A*) Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$ for i.i.d. $\mathcal{N}(3, 16)$ regression coefficients. The black line has slope 1. Again, we see that the MLE seriously overestimates effect magnitudes. The red line has slope $\alpha^\star \approx 1.499$ predicted by the solution to Eq. **5**. We can see that $\hat{\beta}_j$ seems centered around $\alpha^\star\beta_j$. (*B*) True conditional probability $f(\boldsymbol{X}_*) = \rho'(\boldsymbol{X}_*'\boldsymbol{\beta})$ (black curve) and corresponding estimated probabilities $\hat{f}(\boldsymbol{X}_*) = \rho'(\boldsymbol{X}_*'\hat{\boldsymbol{\beta}})$ (blue points). Observe the dramatic shrinkage of the estimates toward the end points.

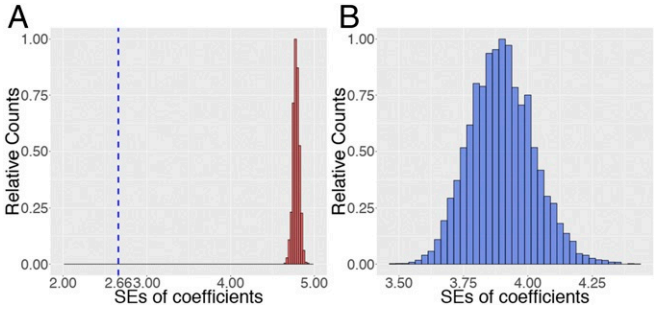

Fig. 4. (*A*) Distribution of SE($\hat{\beta}_j$) for each variable j, in which the SE is estimated from 10, 000 samples. The classical SE value is shown in blue. Classical theory underestimates the variability of the MLE. (*B*) SE estimates computed from R for a single null (for which $\beta_j = 0$) obtained across 10, 000 replicates resampling the response vector and the covariate matrix.

around 4.75, we see that in higher dimensions, the variance of the MLE is likely to be much larger than the classical asymptotic variance. Naturally, using classical results would lead to incorrect P values and confidence statements, a major issue first noted in ref. 6.

The variance estimates obtained from statistical software packages are different from the value 2.66 above because they do not take expectation over the covariates and use the MLE $\hat{\boldsymbol{\beta}}$ in lieu of $\boldsymbol{\beta}$ (plugin estimate) (*SI Appendix*, section A). Since practitioners often use these estimates, it is useful to describe how they behave. To this end, for each of the $10,000$ samples $(\boldsymbol{X}, \boldsymbol{y})$ drawn above, we obtain the R SE estimate for a single MLE coordinate corresponding to a null variable. The histogram is shown in Fig. 4*B*. The behavior for this specific coordinate is typical of that observed for any other null coordinate, and the maximum value for these standard errors remains below 4.5, significantly below the typical values observed via Monte Carlo simulations in Fig. 4*A*.

Distribution of the LRT? By now, the reader should be suspicious that the χ^2 approximation for the distribution of the likelihood-ratio test holds in higher dimensions. Indeed, it does not and this actually is not a new observation. In ref. 7, the authors established that for a class of logistic regression models, the LRT converges weakly to a multiple of a χ^2 variable in an asymptotic regime in which both n and p tend to infinity in such a way that $p/n \to \kappa \in (0, 1/2)$. The multiplicative factor is an increasing function of the limiting aspect ratio κ and exceeds 1 as soon as κ is positive. This factor can be computed by solving a nonlinear system of two equations in two unknowns given in Eq. **8** below. Furthermore, ref. 7 links the distribution of the LRT with the asymptotic variance of the marginals of the MLE, which turns out to be provably higher than that given by the inverse Fisher information. These findings are of course completely in line with the conclusions from the previous paragraphs. The issue is that the results from ref. 7 assume that $\boldsymbol{\beta} = 0$; that is, they apply under the global null where the response does not depend upon the predictors, and it is a priori unclear how the theory would extend beyond this case. Our goal in this paper is to study properties of the MLE and the LRT for high-dimensional logistic regression models under general signal strengths—restricting to the regime where the MLE exists.

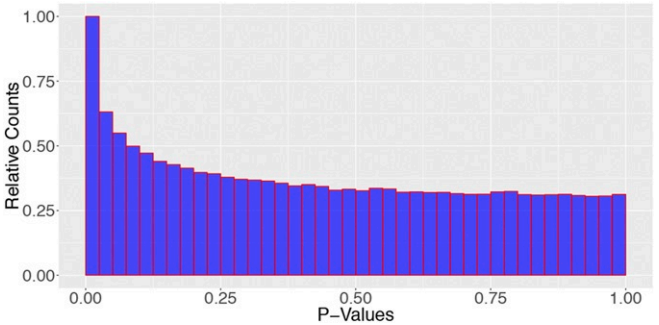

Fig. 5. P values calculated from the χ^2_1 approximation to the LLR. Parameters: $n = 4,000$, $\kappa = 0.2$, with half the coordinates of $\boldsymbol{\beta}$ nonzero, generated i.i.d. from $\mathcal{N}(7, 1)$.

To investigate what happens when we are not under the global null, consider the same setting as in Fig. 4. Fig. 5 shows the distribution of P values for testing a null coefficient based on the χ^2 approximation. Not only are the P values far from uniform, but also the enormous mass near 0 is problematic for multiple-testing applications, where one examines P values at very high levels of significance, e.g., near Bonferroni levels. In such applications, one would be bound to make a large number of false discoveries from using P values produced by software packages. To further demonstrate the large inflation near the small P values, we display in Table 1 estimates of the P-value probabilities in bins near 0. The estimates are much higher than what is expected from a uniform distribution. Clearly, the distribution of the LRT is far from a χ^2_1.

Summary. We have hopefully made the case that classical results, which software packages continue to rely upon, can be inaccurate in higher dimensions. (*i*) Estimates seem systematically biased in the sense that effect magnitudes are overestimated. (*ii*) Estimates are far more variable than classical results. And (*iii*) inference measures, e.g., P values, are unreliable especially at small values. Given the widespread use of logistic regression in high dimensions, a theory explaining how to adjust inference to make it valid is needed.

2. Our Contribution

We develop a theory for high-dimensional logistic regression models with independent variables that is capable of accurately describing all of the phenomena we have discussed. Taking them one by one, the theory from this paper explicitly characterizes (*i*) the bias of the MLE, (*ii*) the variability of the MLE, and (*iii*) the distribution of the LRT, in an asymptotic regime where the sample size and the number of features grow to infinity in a fixed ratio. Moreover, we shall see that our asymptotic results are extremely accurate in finite-sample settings in which p is a fraction of n; e.g., $p = 0.2n$.

A useful feature of this theory is that in our model, all of our results depend on the true coefficients $\boldsymbol{\beta}$ only through the signal strength γ, where $\gamma^2 := \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta})$. This immediately suggests that estimating some high-dimensional parameter is not required to adjust inference. We propose in Section 5 a method for estimating γ and empirically study the quality of inference based on this estimate.

Table 1. *P*-value probabilities with SEs in parentheses

Threshold, %	Classical, %
$\mathbb{P}\{P\,\text{value} \le 5\}$	10.77(0.062)
$\mathbb{P}\{P\,\text{value} \le 1\}$	3.34(0.036)
$\mathbb{P}\{P\,\text{value} \le 0.5\}$	1.98(0.028)
$\mathbb{P}\{P\,\text{value} \le 0.1\}$	0.627(0.016)
$\mathbb{P}\{P\,\text{value} \le 0.05\}$	0.365(0.012)
$\mathbb{P}\{P\,\text{value} \le 0.01\}$	0.136(0.007)

Here, $n = 4,000$, $p = 800$, $\boldsymbol{X}$ has i.i.d. Gaussian entries, and half of the entries of $\boldsymbol{\beta}$ are drawn from $\mathcal{N}(7, 1)$.

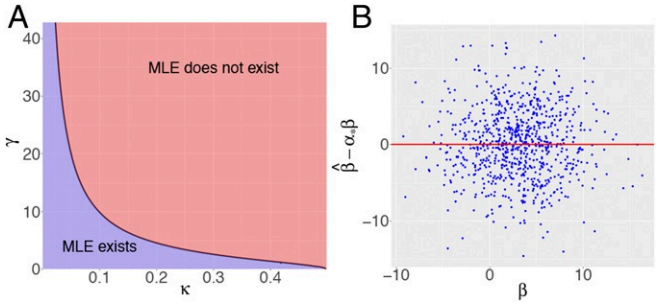

Fig. 6. (*A*) Regions in which the MLE asymptotically exists and is unique and in which it does not. The boundary curve is explicit and given by Eq. **4**. (*B*) In the setting of Fig. 3, scatterplot of the centered MLE $\hat{\beta}_j - \alpha_\star \beta_j$ vs. the true signal β_j.

At the mathematical level, our arguments are very involved. Our strategy is to introduce an approximate message-passing algorithm that tracks the MLE in the limit of a large number of features and samples. In truth, a careful mathematical analysis is delicate and we defer the mathematical details to *SI Appendix*.

3. Prior Work

Asymptotic properties of M estimators in the context of linear regression have been extensively studied in diverging dimensions starting from ref. 8, followed by refs. 9 and 10, in the regime $p = o(n^\alpha)$, for some $\alpha < 1$. Later on, the regime where p is comparable to n became the subject of a series of remarkable works (11–14); these papers concern the distribution of M estimators in linear models. The rigorous results from this literature all assume strongly convex loss functions, a property critically missing in logistic regression. The techniques developed in the work of El Karoui (14) and in ref. 13 play a crucial role in our analysis; the connections are detailed in *SI Appendix*. While this paper was under review, we also learned about extensions to penalized versions of such strongly convex losses (15). Again, this literature is concerned with linear models only and it is natural to wonder what extensions to generalized linear models might look like; see the comments at the end of the talk (16). More general exponential families were studied in refs. 17 and 18; these works were also in setups subsumed under $p = o(n)$. Very recently, ref. 19 investigated classical asymptotic normality of the MLE under the global null and regimes in which it may break down as the dimensionality increases.

In parallel, there exists an extensive body of literature on penalized maximum-likelihood estimates/procedures for generalized linear models; see refs. 20 and 21, for example, and the references cited therein. This body of literature often allows p to be larger than n but relies upon strong sparsity assumptions on the underlying signal. The setting in these works is, therefore, different from ours.

In the low-dimensional setting where the MLE is consistent, finite-sample corrections to the MLE and the LRT have been suggested in a series of works—see, for instance, refs. 22–35. Although these finite-sample approaches aim at correcting the problems described in the preceding sections, that is, the bias of the MLE and nonuniformity of the P values, the corrections are not sufficiently accurate in high dimensions and the methods are often not scalable to high-dimensional data; see *SI Appendix*, section A for some simulations in this direction. A line of simulation-based results exists to guide practitioners about the sample size required to avoid finite-sample problems (36, 37). The rule of thumb is usually 10 events per variable (EPV) or more but we shall later clearly see that such a rule is not valid when the number of features is large. Ref. 38 contested the previously established 10 EPV rule. To the best of our knowledge, logistic regression in the regime where p is comparable to n has been quite sparsely studied. This paper follows up on the earlier contribution (7) of the authors, which characterized the LLR distribution in the case where there is no signal (global null). This earlier reference derived the asymptotic distribution of the LLR as a function of the limiting ratio p/n. This former result may be seen as a special case of *Theorem 4*, which deals with general signal strengths. As is expected, the arguments are now much more complicated than when working under the global null.

4. Main Results

Setting. We describe the asymptotic properties of the MLE and the LRT in a high-dimensional regime, where n and p both go to infinity in such a way that $p/n \to \kappa > 0$. We work with independent observations $\{\boldsymbol{X}_i, y_i\}$ from a logistic model such that $\mathbb{P}(y_i = 1 \mid \boldsymbol{X}_i) = \rho'(\boldsymbol{X}_i'\boldsymbol{\beta})$. We assume here that $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the p-dimensional identity matrix. The exact scaling of $\boldsymbol{X}_i$ is not important. As noted before, the important scaling is the signal strength $\boldsymbol{X}_i'\boldsymbol{\beta}$ and we assume that the p regression coefficients (recall that p increases with n) are scaled in such a way that

$$\lim_{n\to\infty} \operatorname{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \gamma^2, \qquad [2]$$

where γ is fixed. It is useful to think of the parameter γ as the signal strength. Another way to express Eq. **2** is to say that $\lim_{n\to\infty} \|\boldsymbol{\beta}\|^2/n = \gamma^2$.

4.a. When Does the MLE Exist? The MLE $\hat{\boldsymbol{\beta}}$ is the minimizer of the negative log-likelihood ℓ defined via (observe that the sigmoid is the first derivative of ρ)

$$\ell(\boldsymbol{b}) = \sum_{i=1}^{n} \{\rho(\boldsymbol{X}_i'\boldsymbol{b}) - y_i\,(\boldsymbol{X}_i'\boldsymbol{b})\}, \qquad \rho(t) = \log(1 + e^t). \qquad [3]$$

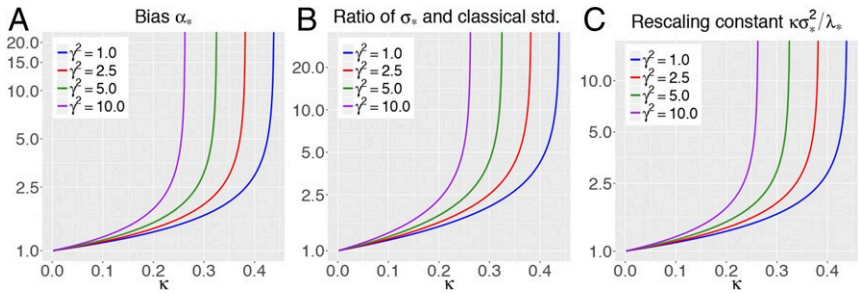

Fig. 7. (*A*) Bias $\alpha_\star$ as a function of κ, for different values of the signal strength γ. Note the logarithmic scale for the *y* axis. The curves asymptote at the value of κ for which the MLE ceases to exist. (*B*) Ratio of the theoretical SD $\sigma_\star$ and the average SD of the coordinates, as obtained from classical theory; i.e., computed using the inverse of the Fisher information. (*C*) Functional dependence of the rescaling constant $\kappa\sigma_\star^2/\lambda_\star$ on the parameters κ and γ.

Table 2. Empirical estimates of the centering and SD of the MLE

Parameter	$p = 200$	$p = 400$	$p = 800$
$\alpha_\star = 1.1678$	1.1703(0.0002)	1.1687(0.0002)	1.1681(0.0001)
$\sigma_\star = 3.3466$	3.3567(0.0011)	3.3519(0.0008)	3.3489(0.0006)

SEs of these estimates are within parentheses. In this setting, $\kappa = 0.1$ and $\gamma^2 = 5$. Half of the β_js are equal to 10 and the others to 0.

A first important remark is that in high dimensions, the MLE does not asymptotically exist if the signal strength γ exceeds a certain functional $g_{\mathrm{MLE}}(\kappa)$ of the dimensionality: i.e., $\gamma > g_{\mathrm{MLE}}(\kappa)$. This happens because in such cases, there is a perfect separating hyperplane—separating the cases from the controls if you will—sending the MLE to infinity. It turns out that a companion paper (39) precisely characterizes the region in which the MLE exists.

Theorem 1 (39). *Let Z be a standard normal variable with density $\varphi(t)$ and V be an independent continuous random variable with density $2\rho'(\gamma t)\varphi(t)$. With $x_+ = \max(x, 0)$, set*

$$g_{\mathrm{MLE}}^{-1}(\gamma) = \min_{t\in\mathbb{R}} \left\{\mathbb{E}(Z - tV)_+^2\right\}, \qquad [4]$$

which is a decreasing function of γ. Then in the setting described above,

$$\gamma > g_{\mathrm{MLE}}(\kappa) \Longrightarrow \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 0,$$
$$\gamma < g_{\mathrm{MLE}}(\kappa) \Longrightarrow \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 1.$$

Hence, the curve $\gamma = g_{\mathrm{MLE}}(\kappa)$, or, equivalently, $\kappa = g_{\mathrm{MLE}}^{-1}(\gamma)$ shown in Fig. 6 separates the $\kappa - \gamma$ plane into two regions: One in which the MLE asymptotically exists and one in which it does not. Clearly, we are interested in this paper in the former region (the purple region in Fig. 6*A*).

4.b. A System of Nonlinear Equations. As we shall soon see, the asymptotic behavior of both the MLE and the LRT is characterized by a system of equations in three variables $(\alpha, \sigma, \lambda)$,

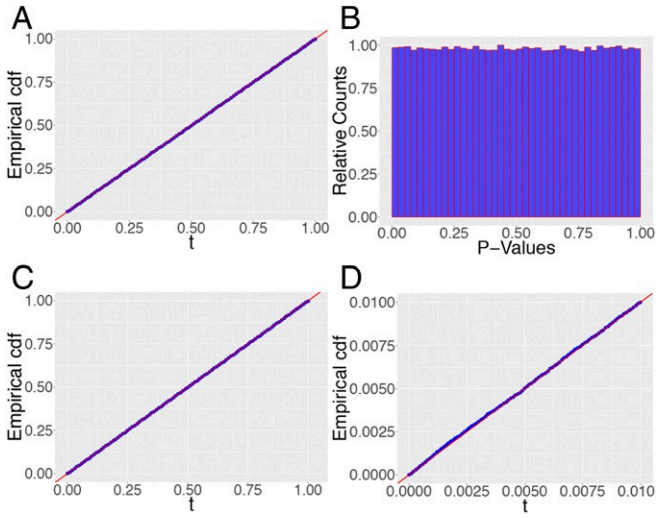

Fig. 8. The setting is that from Table 2 with $n = 4,000$. (*A*) Empirical cdf of $\Phi(\hat{\beta}_j/\sigma_\star)$ for a null variable ($\beta_j = 0$). (*B*) *P* values given by the LLR approximation Eq. **11** for this same null variable. (*C*) Empirical distribution of the *P* values from *B*. (*D*) Same as *C* but showing accuracy in the lower tail (check the range of the horizontal axis). All these plots are based on 500,000 replicates.

Table 3. *P*-value probabilities estimated over 500,000 replicates with standard errors in parentheses

Threshold, %	$p = 400$, %	$p = 800$, %
$\mathbb{P}\{P\text{ value} \le 5\}$	5.03(0.031)	5.01(0.03)
$\mathbb{P}\{P\text{ value} \le 1\}$	1.002(0.014)	1.005(0.014)
$\mathbb{P}\{P\text{ value} \le 0.5\}$	0.503(0.01)	0.49(0.0099)
$\mathbb{P}\{P\text{ value} \le 0.1\}$	0.109(0.004)	0.096(0.0044)
$\mathbb{P}\{P\text{ value} \le 0.05\}$	0.052(0.003)	0.047(0.0031)
$\mathbb{P}\{P\text{ value} \le 0.01\}$	0.008(0.0013)	0.008(0.0013)

Here, $\kappa = 0.1$ and the setting is otherwise the same as in Table 2.

$$\begin{cases} \sigma^2 = \frac{1}{\kappa^2}\mathbb{E}\left[2\rho'(Q_1)\left(\lambda\rho'(\mathrm{prox}_{\lambda\rho}(Q_2))\right)^2\right] \\ 0 = \mathbb{E}\left[\rho'(Q_1)Q_1\lambda\rho'(\mathrm{prox}_{\lambda\rho}(Q_2))\right] \\ 1-\kappa = \mathbb{E}\left[\frac{2\rho'(Q_1)}{1+\lambda\rho''(\mathrm{prox}_{\lambda\rho}(Q_2))}\right] \end{cases}, \qquad [5]$$

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

$$\mathbf{\Sigma}(\alpha, \sigma) = \begin{bmatrix} \gamma^2 & -\alpha\gamma^2 \\ -\alpha\gamma^2 & \alpha^2\gamma^2 + \kappa\sigma^2 \end{bmatrix}. \qquad [6]$$

With ρ as in Eq. **3**, the proximal mapping operator is defined via

$$\mathrm{prox}_{\lambda\rho}(z) = \arg\min_{t\in\mathbb{R}} \left\{\lambda\rho(t) + \frac{1}{2}(t-z)^2\right\}. \qquad [7]$$

The system of equations in Eq. **5** is parameterized by the pair (κ, γ) of dimensionality and signal strength parameters. It turns out that the system admits a unique solution if and only if (κ, γ) is in the region where the MLE asymptotically exists!

It is instructive to note that in the case where the signal strength vanishes, $\gamma = 0$, the system of equations in Eq. **5** reduces to the 2-dimensional system

$$\begin{cases} \sigma^2 = \frac{1}{\kappa^2}\mathbb{E}\left[\left(\lambda\rho'(\mathrm{prox}_{\lambda\rho}(\tau Z))\right)^2\right] \\ 1-\kappa = \mathbb{E}\left[\frac{1}{1+\lambda\rho''(\mathrm{prox}_{\lambda\rho}(\tau Z))}\right] \end{cases}, \qquad [8]$$

where $\tau^2 := \kappa\sigma^2$ and $Z \sim \mathcal{N}(0, 1)$. This holds because $Q_1 = 0$. It is not surprising that this system is that from ref. 7 since that work considers $\boldsymbol{\beta} = 0$ and, therefore, $\gamma = 0$.

We remark that similar equations have been obtained for M estimators in linear models; see, for instance, ref. 11; *SI Appendix*, Eqs. S1 and S2; and refs. 13–15.

4.c. The Average Behavior of the MLE. Our first main result characterizes the "average" behavior of the MLE.

Theorem 2. *Assume the dimensionality and signal strength parameters κ and γ are such that $\gamma < g_{\mathrm{MLE}}(\kappa)$ (the region where the MLE exists asymptotically and is shown in Fig. 6). Assume the logistic model described above where the empirical distribution of $\{\beta_j\}$ converges weakly to a distribution Π with finite second moment. Suppose further that the second moment converges in the sense that as $n \to \infty$, $\mathrm{Ave}_j(\beta_j^2) \to \mathbb{E}\beta^2$, $\beta \sim \Pi$. Then for any pseudo-Lipschitz function ψ of order 2,[†] the marginal distributions of the MLE coordinates obey*

[†]A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is said to be pseudo-Lipschitz of order k if there exists a constant $L > 0$ such that for all $\boldsymbol{t}_0, \boldsymbol{t}_1 \in \mathbb{R}^m$, $\|\psi(\boldsymbol{t}_0) - \psi(\boldsymbol{t}_1)\| \le L\left(1 + \|\boldsymbol{t}_0\|^{k-1} + \|\boldsymbol{t}_1\|^{k-1}\right)\|\boldsymbol{t}_0 - \boldsymbol{t}_1\|$.

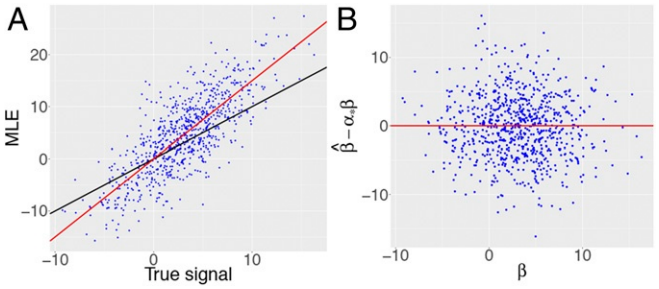

Fig. 9. Simulation for a non-Gaussian design. The jth feature takes values in $\{0, 1, 2\}$ with probabilities p_j^2, $2p_j(1-p_j)$, $(1-p_j)^2$; here, $p_j \in [0.25, 0.75]$ and $p_j \neq p_k$ for $j \neq k$. Features are then centered and rescaled to have unit variance. The setting is otherwise the same as for Fig. 3. (*A*) Analogue of Fig. 3*A*. Red line has slope $\alpha_\star \approx 1.499$. (*B*) Analogue of Fig. 6*B*. Observe the same behavior as earlier: The theory predicts correctly the bias and the decorrelation between the bias-adjusted residuals and the true effect sizes.

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{\beta}_j-\alpha_\star\beta_j,\beta_j)\xrightarrow{\text{a.s.}}\mathbb{E}[\psi(\sigma_\star Z,\beta)],\quad Z\sim\mathcal{N}(0,1),\quad \textbf{[9]}$$

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

Among the many consequences of this result, we give three:

- This result quantifies the exact bias of the MLE in some statistical sense. This can be seen by taking $\psi(t,u)=t$ in Eq. **9**, which leads to

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j-\alpha_\star\beta_j)\xrightarrow{\text{a.s.}}0$$

and says that $\hat{\beta}_j$ is centered about $\alpha_\star\,\beta_j$. This can be seen from the empirical results from the previous sections as well. When $\kappa=0.2$ and $\gamma=\sqrt{5}$, the solution to Eq. **5** obeys $\alpha_\star=1.499$ and Fig. 3*A* shows that this is the correct centering.

- Second, our result also provides the asymptotic variance of the MLE marginals after they are properly centered. This can be seen by taking $\psi(t,u)=t^2$, which leads to

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j-\alpha_\star\beta_j)^2\xrightarrow{\text{a.s.}}\sigma_\star^2.$$

As before, this can also be seen from the empirical results from the previous section. When $\kappa=0.2$ and $\gamma=\sqrt{5}$, the solution to Eq. **5** obeys $\sigma_\star=4.744$ and this is what we see in Fig. 4.

- Third, our result establishes that upon centering the MLE around $\alpha_\star\boldsymbol{\beta}$, it becomes decorrelated from the signal $\boldsymbol{\beta}$. This can be seen by taking $\psi(t,u)=tu$, which leads to

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j-\alpha_\star\beta_j)\,\beta_j\xrightarrow{\text{a.s.}}0.$$

This can be seen from our earlier empirical results in Fig. 6*B*. The scatter directly shows the decorrelated structure and the x axis passes right through the center, corroborating our theoretical finding.

It is of course interesting to study how the bias $\alpha_\star$ and the SD $\sigma_\star$ depend on the dimensionality κ and the signal strength γ. We numerically observe that the larger the dimensionality and/or the larger the signal strength, the larger the bias $\alpha_\star$ will be. This dependence is illustrated in Fig. 7*A*. Further, note that as κ approaches 0, the bias $\alpha_\star\to 1$, indicating that the MLE is asymptotically unbiased if $p=o(n)$. The same behavior applies to $\sigma_\star$; that is, $\sigma_\star$ increases in either κ or γ as shown in Fig. 7*B*. This plot shows the theoretical prediction $\sigma_\star$ divided by the average classical SD obtained from $\boldsymbol{I}_{\boldsymbol{\beta}}^{-1}$, the inverse of the Fisher information. As κ approaches 0, the ratio goes to 1, indicating that the classical SD value is valid for $p=o(n)$; this is true across all values of γ. As κ increases, the ratio deviates increasingly from 1 and we observe higher and higher variance inflation. In summary, the MLE increasingly deviates from what is classically expected as the dimensionality, the signal strength, or both increase.

Theorem 2 is an asymptotic result, and we study how fast the asymptotic kicks in as we increase the sample size n. To this end, we set $\kappa=0.1$ and let half of the coordinates of $\boldsymbol{\beta}$ have constant value 10 and the other half be 0. Note that in this example, $\gamma^2=5$ as before. Our goal is to empirically determine the parameters $\alpha_\star$ and $\sigma_\star$ from $68,000$ runs, for each n taking values in $\{2,000, 4,000, 8,000\}$. Note that there are several ways of determining $\alpha_\star$ empirically. For instance, the limit Eq. **9** directly suggests taking the ratio $\sum_j\hat{\beta}_j/\sum_j\beta_j$. An alternative is to consider taking the ratio when restricting the summation to nonzero indexes. Empirically, we find there is not much difference between these two choices and choose the latter option, denoting it as $\hat{\alpha}$. With $\kappa=0.1, \gamma=\sqrt{5}$, the solution to Eq. **5** is equal to $\alpha_\star=1.1678, \sigma_\star=3.3466, \lambda_\star=0.9605$. Table 2 shows that $\hat{\alpha}$ is very slightly larger than $\alpha_\star$ in finite samples. However, observe that as the sample size increases, $\hat{\alpha}$ approaches $\alpha_\star$, confirming the result from Eq. **9**.

4.d. The Distribution of the Null MLE Coordinates. Whereas *Theorem 2* describes the average or bulk behavior of the MLE across all of its entries, our next result provides the explicit distribution of $\hat{\beta}_j$ whenever $\beta_j=0$, i.e., whenever the jth variable is independent from the response y.

Theorem 3. *Let j be any variable such that $\beta_j=0$. Then in the setting of Theorem 2, the MLE obeys*

$$\hat{\beta}_j\xrightarrow{\text{d}}\mathcal{N}(0,\sigma_\star^2).\qquad\textbf{[10]}$$

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

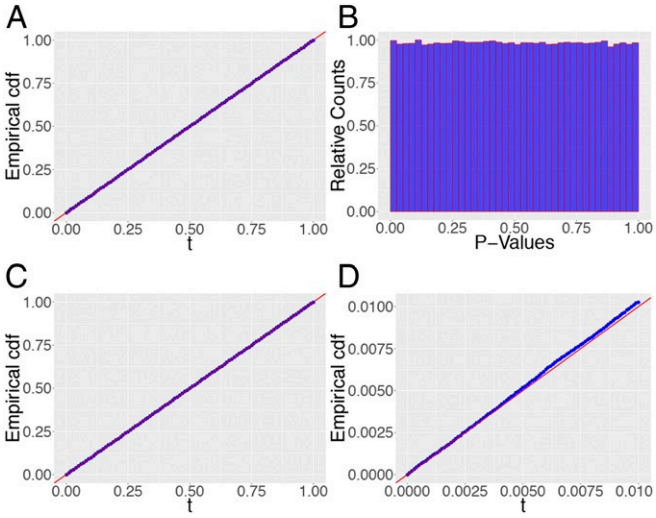

Fig. 10. The features are multinomial as in Fig. 9 and the setting is otherwise the same as in Fig. 8. (*A*) Empirical cdf of $\Phi(\hat{\beta}_j/\sigma_\star)$ for a null variable ($\beta_j=0$). (*B*) *P* values given by the LLR approximation Eq. **11** for this same null variable. (*C*) Empirical distribution of the *P* values from *B*. (*D*) Same as *C* but displaying accuracy in the extreme. These results are based on 500, 000 replicates.

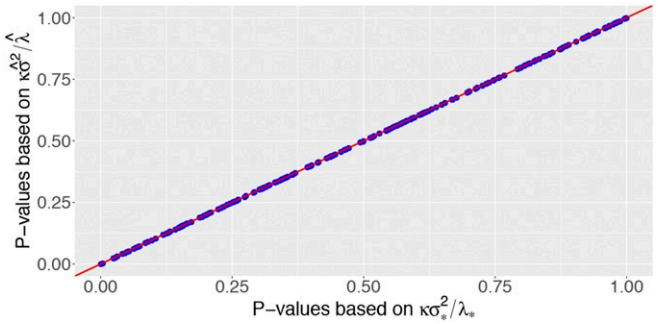

Fig. 11. Null P values obtained using the $(\kappa\hat{\sigma}^2/\hat{\lambda})\,\chi_1^2$ approximation plotted against those obtained using $(\kappa\sigma_\star^2/\lambda_\star)\,\chi_1^2$. Observe the perfect agreement with the red diagonal.

In words, the null MLE coordinates are asymptotically normal with mean 0 and variance given by the solution to the system Eq. **5**. An important remark is this: We have observed that $\sigma_\star$ is an increasing function of γ. Hence, the distribution of a null MLE coordinate depends on the magnitude of the remaining coordinates of the signal: The variance increases as the other coefficient magnitudes increase.

We return to the finite-sample precision of the asymptotic variance $\sigma_\star^2$. As an empirical estimate, we use $\hat{\beta}_j^2$ averaged over the null coordinates $\{j : \beta_j = 0\}$ since it is approximately unbiased for $\sigma_\star^2$. We work in the setting of Table 2 in which $\sigma_\star = 3.3466$, averaging our $68,000$ estimates. The results are given in Table 2; we observe that $\hat{\sigma}$ is very slightly larger than $\sigma_\star$. However, it progressively gets closer to $\sigma_\star$ as the sample size n increases.

Next, we study the accuracy of the asymptotic convergence results in Eq. **4**. In the setting of Table 2, we fit $500,000$ independent logistic regression models and plot the empirical cumulative distribution function of $\Phi(\hat{\boldsymbol{\beta}}_j/\sigma_\star)$ in Fig. 8*A* for some fixed null coordinate. Observe the perfect agreement with a straight line of slope 1.

4.e. The Distribution of the LRT. We finally consider the distribution of the likelihood-ratio statistic for testing $\beta_j = 0$.

Theorem 4. *Consider the LLR $\Lambda_j = \min_{\boldsymbol{b}\,:\,b_j=0} \ell(\boldsymbol{b}) - \min_{\boldsymbol{b}} \ell(\boldsymbol{b})$ for testing $\beta_j = 0$. In the setting of Theorem 2, twice the LLR is asymptotically distributed as a multiple of a χ^2 under the null,*

$$2\Lambda_j \xrightarrow{\mathrm{d}} \frac{\kappa\,\sigma_\star^2}{\lambda_\star}\,\chi_1^2. \qquad [11]$$

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

Theorem 4 explicitly states that the LLR does not follow a χ_1^2 distribution as soon as $\kappa > 0$ since the multiplicative factor is then larger than 1, as demonstrated in Fig. 7*C*. In other words, the LLR is stochastically much larger than a χ_1^2, explaining the large spike near 0 in Fig. 5. Also, Fig. 7*C* suggests that as $\kappa \to 0$, the classical result is recovered.[‡] We refer the readers to *SI Appendix*, section G for an empirical comparison of the P values based on Eq. **11** and classical P values in settings with small κ and moderate n, p.

Theorem 4 extends to arbitrary signal strengths the earlier result from ref. 7, which described the distribution of the LLR under the global null ($\beta_j = 0$ for all j). One can quickly verify that when $\gamma = 0$, the multiplicative factor in Eq. **11** is that given in ref. 7, which easily follows from the fact that in this case, Eq. **5** reduces to Eq. **8**. Furthermore, if the signal is sparse in the sense that $o(n)$ coefficients have nonzero values, then $\gamma^2 = 0$, which immediately implies that the asymptotic distribution for the LLR from ref. 7 still holds in such cases.

To investigate the quality of the accuracy of Eq. **11** in finite samples, we work on the P-value scale. We select a null coefficient and compute P values based on Eq. **11**. The histogram for the P values across $500,000$ runs is shown in Fig. 8*B* and the empirical cumulative distribution function (cdf) in Fig. 8*C*. In stark contrast to Fig. 4, we observe that the P values are uniform over the bulk of the distribution.

From a multiple-testing perspective, it is essential to understand the accuracy of the rescaled χ^2 approximation in the tails of the distribution. We plot the empirical cdf of the P values, zooming in on the tail, in Fig. 8*D*. We find that the rescaled χ^2 approximation works well even in the tails of the distribution. To obtain a more refined idea of the quality of approximation, we zoom in on the smaller bins close to 0 and provide estimates of the P-value probabilities in Table 3 for $n = 4,000$ and $n = 8,000$. The tail approximation is accurate, modulo a slight deviation in the bin for $\mathbb{P}\{P\text{-value}\} \le 0.1$ for the smaller sample size. For $n = 8,000$, however, this deviation vanishes and we find perfect coverage of the true values. It seems that our approximation is extremely precise even in the tails.

4.f. Other Scalings. Throughout this section, we worked under the assumption that $\lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \gamma^2$, which does not depend on n, and we explained that this is the only scaling that makes sense to avoid a trivial problem. We set the variables to have variance $1/n$ but this is of course somewhat arbitrary. For example, we could choose them to have variance v as in $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, v\boldsymbol{I}_p)$. This means that $\boldsymbol{X}_i = \sqrt{vn}\boldsymbol{Z}_i$, where $\boldsymbol{Z}_i$ is as before. This gives $\boldsymbol{X}_i'\boldsymbol{\beta} = \boldsymbol{Z}_i'\boldsymbol{b}$, where $\boldsymbol{\beta} = \boldsymbol{b}/\sqrt{nv}$. The conclusions from *Theorems 2* and *3* then hold for the model with predictors $\boldsymbol{Z}_i$ and regression coefficient sequence $\boldsymbol{b}$. Consequently, by simple rescaling, we can pass the properties of the MLE in this model to those of the MLE in the model with predictors $\boldsymbol{X}_i$ and coefficients $\boldsymbol{\beta}$. For instance, the SE of $\hat{\boldsymbol{\beta}}$ is equal to $\sigma_\star/\sqrt{nv}$, where $\sigma_\star$ is just as in *Theorems 2* and *3*. On the

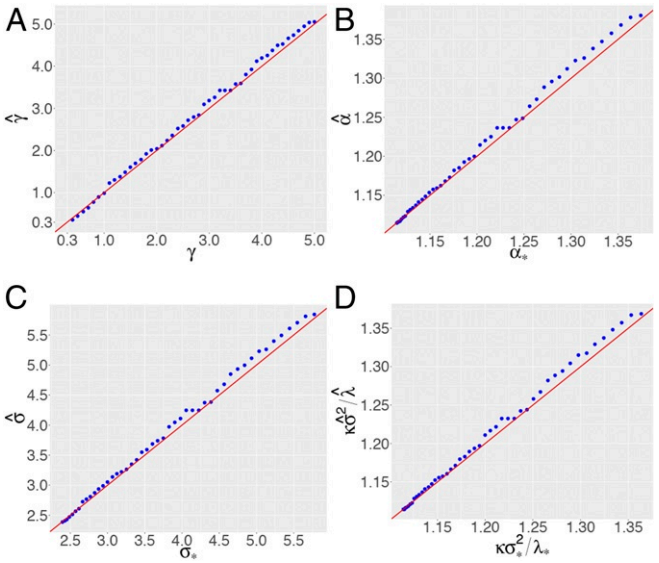

Fig. 12. (*A–D*) ProbeFrontier estimates of signal strength $\hat{\gamma}$ (*A*), bias $\hat{\alpha}$ (*B*), SD $\hat{\sigma}$ (*C*), and LRT factor $\kappa\hat{\sigma}^2/\hat{\lambda}$ (*D*) in Eq. **11**, plotted against the theoretical values.

[‡]For the analytically motivated reader, we remark that unlike classical theory, here the asymptotic distribution of the LRT, i.e., *Theorem 4*, does not follow directly from the asymptotic normality result in *Theorem 3*. It requires additional probabilistic analysis.

Table 4. Parameter estimates in the setting of Table 2: Averages over 6,000 replicates with SEs within parentheses

Parameters	True	Estimates
γ	2.2361	2.2771(0.0012)
α	1.1678	1.1698(0.0001)
σ	3.3466	3.3751(0.0008)
$\kappa\sigma^2/\lambda$	1.166	1.1680(0.0001)

other hand, the result for the LRT, namely *Theorem 4*, is scale invariant.

4.g. Non-Gaussian Covariates. Our model assumes that the features are Gaussian. However, we expect that the same results hold under other distributions with the proviso that they have sufficiently light tails. In this section, we empirically study the applicability of our results for certain non-Gaussian features.

In genetic studies, we often wish to understand how a binary response/phenotype depends on single-nucleotide polymorphisms (SNPs), which typically take on values in $\{0, 1, 2\}$. When the jth SNP is in Hardy–Weinberg equilibrium, the chance of observing 0, 1, and 2 is respectively p_j^2, $2p_j(1-p_j)$, and $(1-p_j)^2$, where p_j is between 0 and 1. Below we generate independent features with marginal distributions as above for parameters p_j varying in $[0.25, 0.75]$. We then center and normalize each column of the feature matrix $\boldsymbol{X}$ to have 0 mean and unit variance. Keeping everything else as in the setting of Fig. 3, we study the bias of the MLE in Fig. 9*A*. As for Gaussian designs, the MLE seriously overestimates effect magnitudes and our theoretical prediction $\alpha_\star$ accurately corrects for the bias. We also see that the bias-adjusted residuals $\hat{\boldsymbol{\beta}} - \alpha_\star\boldsymbol{\beta}$ are uncorrelated with the effect sizes $\boldsymbol{\beta}$, as shown in Fig. 9*B*.

The bulk distribution of a null coordinate suggested by *Theorem 3* and the LRT distribution from *Theorem 4* are displayed in Fig. 10. Other than the design, the setting is the same as in Fig. 8. The theoretical predictions are once again accurate. Furthermore, upon examining the tails of the *P*-value distribution, we once more observe a close agreement with our theoretical predictions. All in all, these findings indicate that our theory is expected to apply to a far broader class of features. We conduct additional experiments based on real-data design matrices in *SI Appendix*, section F.

That said, we caution readers against overinterpreting our results. For linear models, for example, it is known that the distribution of M estimators in high dimensions can be significantly different if the covariates follow a general elliptical distribution, rather than a normal distribution (11, 14).

5. Adjusting Inference by Estimating the Signal Strength

All of our asymptotic results, namely, the average behavior of the MLE, the asymptotic distribution of a null coordinate, and the LLR, depend on the unknown signal strength γ. In this section, we describe a simple procedure for estimating this single parameter from an idea proposed by Boaz Nadler and Rina Barber after E.J.C. presented the results from this paper at the Mathematisches Forshunginstitut Oberwolfach on March 12, 2018.

5.a. ProbeFrontier: Estimating γ by Probing the MLE Frontier. We estimate the signal strength by actually using the predictions from our theory, namely, the fact that we have asymptotically characterized in Section 4.a the region where the MLE exists. We know from *Theorem 1* that for each γ, there is a maximum dimensionality $g_{\mathrm{MLE}}^{-1}(\gamma)$ at which the MLE ceases to exist. We propose an estimate $\hat{\kappa}$ of $g_{\mathrm{MLE}}^{-1}(\gamma)$ and set $\hat{\gamma} = g_{\mathrm{MLE}}(\hat{\kappa})$. Below, we refer to this as the ProbeFrontier method.

Given a data sample $(y_i, \boldsymbol{X}_i)$, we begin by choosing a fine grid of values $\kappa \le \kappa_1 \le \kappa_2 \le \ldots \le \kappa_K \le 1/2$. For each κ_j, we execute the following procedure:

Subsample. Sample $n_j = p/\kappa_j$ observations from the data without replacement, rounding to the nearest integer. Ignoring the rounding, the dimensionality of this subsample is $p/n_j = \kappa_j$.

Check whether MLE exists. For the subsample, check whether the MLE exists or not. This is done by solving a linear programing feasibility problem; if there exists a vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{X}_i'\boldsymbol{b}$ is positive when $y_i = 1$ and negative otherwise, then perfect separation between cases and controls occurs and the MLE does not exist. Conversely, if the linear program is infeasible, then the MLE exists.

Repeat. Repeat the two previous steps B times and compute the proportion of times $\hat{\pi}(\kappa_j)$ the MLE does not exist.

We next find (κ_{j-1}, κ_j), such that κ_j is the smallest value in $\mathcal{K}$ for which $\hat{\pi}(\kappa_j) \ge 0.5$. By linear interpolation between κ_{j-1} and κ_j, we obtain $\hat{\kappa}$ for which the proportion of times the MLE does not exist would be 0.5. We set $\hat{\gamma} = g_{\mathrm{MLE}}(\hat{\kappa})$. (Since the "phase-transition" boundary for the existence of the MLE is a smooth function of κ, as is clear from Fig. 6, choosing a sufficiently fine grid $\{\kappa_j\}$ would make the linear interpolation step sufficiently precise.)

5.b. Empirical Performance of Adjusted Inference. We demonstrate the accuracy of ProbeFrontier via some empirical results. We begin by generating 4,000 i.i.d. observations $(y_i, \boldsymbol{X}_i)$ using the same setup as in Fig. 8 ($\kappa = 0.1$ and half of the regression coefficients are null). We work with a sequence $\{\kappa_j\}$ of points spaced apart by 10^{-3} and obtain $\hat{\gamma}$ via the procedure described above, drawing 50 subsamples. Solving the system Eq. **5** using $\kappa = 0.1$ and $\hat{\gamma}$ yields estimates for the theoretical predictions $(\alpha_\star, \sigma_\star, \lambda_\star)$ equal to $(\hat{\alpha}, \hat{\sigma}, \hat{\lambda}) = (1.1681, 3.3513, 0.9629)$. In turn, this yields an estimate for the multiplicative factor $\kappa\sigma_\star^2/\lambda_\star$ in Eq. **11** equal to 1.1663. Recall from Section 4 that the theoretical values are $(\alpha_\star, \sigma_\star, \lambda_\star) = (1.1678, 3.3466, 0.9605)$ and $\kappa\sigma_\star^2/\lambda_\star = 1.1660$. Next, we compute the LLR statistic for each null and *P* values from the approximation Eq. **11** in two ways: First, by using the theoretically predicted values, and second, by using our estimates. A scatterplot of these two sets of *P* values is shown in Fig. 11 (blue). We observe impeccable agreement.

Next, we study the accuracy of $\hat{\gamma}$ across different choices for γ, ranging from 0.3 to 5. We begin by selecting a fine grid of γ values and for each, we generate observations $(y_i, \boldsymbol{X}_i)$ with $n = 4,000$, $p = 400$ (so that $\kappa = 0.1$), and half the coefficients have a nonvanishing magnitude scaled in such a way that the signal strength is γ. Fig. 12*A* displays $\hat{\gamma}$ vs. γ in blue, and we note that ProbeFrontier works very well. We observe that the blue points fluctuate very

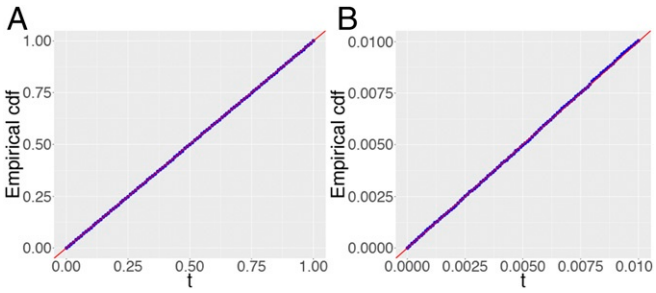

Fig. 13. (*A*) Empirical distribution of the *P* values based on the LLR approximation Eq. **11**, obtained using the estimated factor $\kappa\hat{\sigma}^2/\hat{\lambda}$. (*B*) Same as *A*, but showing the tail of the empirical cdf. The calculations are based on 500,000 replicates.

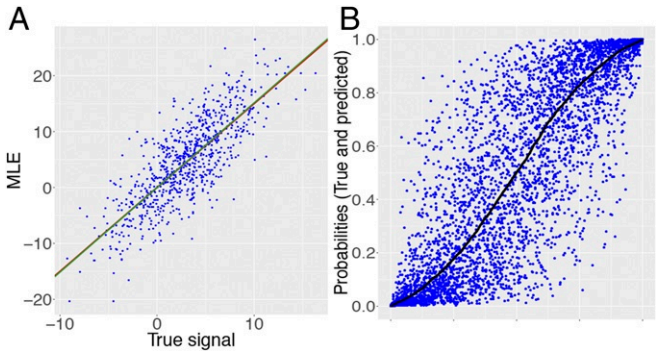

Fig. 14. (*A*) Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$ for the dataset from Fig. 3. Here, $\alpha_\star = 1.499$ (red line) and our ProbeFrontier estimate is $\hat{\alpha} = 1.511$ (green line). The estimate is so close that the green line masks the red. (*B*) True conditional probabilities $\rho'(\boldsymbol{X}_i'\boldsymbol{\beta})$ (black curve) and corresponding estimated probabilities $\rho'(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}/\hat{\alpha})$ computed from the debiased MLE (blue points). Observe that the black curve now passes through the center of the blue point cloud. Our predictions are fairly unbiased.

mildly above the diagonal for larger values of the signal strength but remain extremely close to the diagonal throughout. This confirms that ProbeFrontier estimates the signal strength γ with reasonably high precision. Having obtained an accurate estimate for γ, plugging it into Eq. **5** immediately yields an estimate for the bias $\alpha_\star$, SD$\sigma_\star$, and the rescaling factor in Eq. **11**. We study the accuracy of these estimates in Fig. 12 *B–D*. We observe a similar behavior in all these cases, with the procedure yielding extremely precise estimates for smaller values and reasonably accurate estimates for higher values.

Finally, we focus on the estimation accuracy for a particular (κ, γ) pair across several replicates. In the setting of Fig. 8, we generate 6,000 samples and obtain estimates of bias $(\hat{\alpha})$, SD $(\hat{\sigma})$, and rescaling factor for the LRT $(\kappa\hat{\sigma}^2/\hat{\lambda})$. The averages of these estimates are reported in Table 4. Our estimates always recover the true values up to the first digit. It is instructive to study the precision of the procedure on the *P*-value scale. To this end, we compute *P* values from Eq. **11**, using the estimated multiplicative factor $\kappa\hat{\sigma}^2/\hat{\lambda}$. The empirical cdf of the *P* values both in the bulk and in the extreme tails is shown in Fig. 13. We observe perfect agreement with the uniform distribution, establishing the practical applicability of our theory and methods.

5.c. Debiasing the MLE and Its Predictions. We have seen that maximum likelihood produces biased coefficient estimates and predictions. The question is, how precisely can our proposed theory and methods correct this? Recall the example from Fig. 3, where the theoretical prediction for the bias is $\alpha_\star = 1.499$. For this dataset, ProbeFrontier yields $\hat{\alpha} = 1.511$, shown as the green line in Fig. 14*A*. Clearly, the estimate of bias is extremely precise and coefficient estimates $\hat{\beta}_j/\hat{\alpha}$ appear nearly unbiased.

Further, we can also use our estimate of bias to refine the predictions since we can estimate the regression function by $\rho'(\boldsymbol{X}'\hat{\boldsymbol{\beta}}/\hat{\alpha})$. Fig. 14*B* shows our predictions on the same dataset. In stark contrast to Fig. 3*B*, the predictions are now centered around the regression function and the massive shrinkage toward the extremes has disappeared. The predictions constructed from the debiased MLE are more accurate and no longer falsely predict almost certain outcomes. Rather, we obtain fairly nontrivial chances of being classified in either of the two response categories—as it should be.

6. Broader Implications and Future Directions

This paper shows that in high dimensions, classical ML theory has limitations; e.g., classical theory predicts that the MLE is approximately unbiased when in reality it overestimates effect magnitudes. Since the purpose of logistic modeling is to estimate the risk of a specific disease given a patient's observed characteristics, for example, the bias of the MLE is problematic. A consequence of the bias is that the MLE pushes the predicted chance of being sick toward 0 or 1. This, along with the fact that *P* values computed from classical approximations are misleading in high dimensions, clearly make the case that routinely used statistical tools fail to provide meaningful inferences from both an estimation and a testing perspective.

We have developed a theory which gives the asymptotic distribution of the MLE and the LRT in a model with independent covariates. As seen in Section 4.g, our results likely hold for a broader range of feature distributions (i.e., other than Gaussian) and it would be important to establish this rigorously. Further, we have also shown how to adjust inference by plugging in an estimate of signal strength in our theoretical predictions.

We conclude with a few directions for future work: It would be of interest to develop corresponding results in the case where the predictors are correlated and to extend the results from this paper to other generalized linear models. Further, it is crucial to understand the robustness of our proposed *P* values to model misspecifications. We provide some preliminary simulations in *SI Appendix*, section E. Finally, covariates following general elliptical distributions can be challenging, as shown in refs. 11 and 14 in the context of linear models. Hence, caution is in order as to the broader applicability of our theory.

ACKNOWLEDGMENTS. P.S. was partially supported by a Ric Weiland Graduate Fellowship. E.J.C. was partially supported by the Office of Naval Research under Grant N00014-16-1-2712, by the National Science Foundation via DMS 1712800, by the Math + X Award from the Simons Foundation, and by a generous gift from TwoSigma. P.S. thanks Andrea Montanari and Subhabrata Sen for helpful discussions about approximate message passing. We thank Małgorzata Bogdan, Nikolaos Ignatiadis, Asaf Weinstein, Lucas Janson, Stephen Bates, and Chiara Sabatti for useful comments about an early version of this paper.

1. D. R. Cox, The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B (Methodol.)* **20**, 215–232 (1958).
2. D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression* (John Wiley & Sons, 2013), vol. 398.
3. A. W. Van der Vaart, *Asymptotic Statistics* (Cambridge University Press), vol. 3.
4. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).
5. S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).
6. E. Candes, Y. Fan, L. Janson, J. Lv, Panning for gold: 'Model-x' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **80**, 551–577 (2018).
7. P. Sur, Y. Chen, E. J. Candès, The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Relat. Fields*, 10.1007/s00440-018-00896-9 (23 January 2019).
8. P. J. Huber, Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1**, 799–821 (1973).
9. S. Portnoy, Asymptotic behavior of M-estimators of *p* regression parameters when p^2/n is large. I. Consistency. *Ann. Stat.* **12**, 1298–1309 (1984) .
10. S. Portnoy, Asymptotic behavior of M-estimators of *p* regression parameters when p^2/n is large; II. Normal approximation. *Ann. Stat.* **13**, 1403–1417 (1985).
11. N. El Karoui, D. Bean, P. J. Bickel, C. Lim, B. Yu, On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. U.S.A* **110**, 14557–14562 (2013).
12. D. Bean, P. J. Bickel, N. El Karoui, B. Yu, Optimal m-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14563–14568 (2013).
13. D. Donoho, A. Montanari, High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields* **166**, 935–969 (2013).
14. N. El Karoui, On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* **170**, 95–175 (2017).
15. J. Bradic, Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electron. J. Stat.* **10**, 3894–3944 (2016).

16. N. El Karoui, Random matrices and high-dimensional M-estimation: Applications to robust regression, penalized robust regression and GLMs. https://video.seas.harvard.edu/media/14_03_31+Noureddine+El+Karoui–AM.mp4/1_e5szurf0/13151391 (31 March 2014).
17. S. Portnoy, Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Stat.* **16**, 356–366 (1988).
18. X. He, Q. M. Shao, On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120–135 (2000).
19. Y. Fan, E. Demirkaya, J. Lv, Nonuniformity of p-values can occur early in diverging dimensions. arXiv:1705.03604 (10 May 2017).
20. F. Bunea, Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization. *Electron J. Stat.* **2**, 1153–1194 (2008).
21. S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* **42**, 1166–1202 (2014).
22. D. Firth, Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
23. J. Anderson, S. Richardson, Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics* **21**, 71–78 (1979).
24. G. McLachlan, A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics* **22**, 621–627 (1980).
25. R. L. Schaefer, Bias correction in maximum likelihood logistic regression. *Stat. Med.* **2**, 71–78 (1983).
26. J. B. Copas, Binary regression models for contaminated data. *J. R. Stat. Soc. Ser. B (Methodol.)* **50**, 225–253 (1988).
27. G. M. Cordeiro, P. McCullagh, Bias correction in generalized linear models. *J. R. Stat. Soc. Ser. B (Methodol.)* **53**, 629–643 (1991).
28. S. Bull, W. Hauck, C. Greenwood, Two-step jackknife bias reduction for logistic regression MLEs. *Commun. Stat. Simul. Comput.* **23**, 59–88 (1994).
29. D. E. Jennings, Judging inference adequacy in logistic regression. *J. Am. Stat. Assoc.* **81**, 471–476 (1986).
30. M. S. Bartlett, Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **160**, 268–282 (1937).
31. G. Box, A general distribution theory for a class of likelihood criteria. *Biometrika* **36**, 317–346 (1949).
32. D. Lawley, A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **43**, 295–303 (1956).
33. G. M. Cordeiro, Improved likelihood ratio statistics for generalized linear models. *J. R. Stat. Soc. Ser. B (Methodol.)* **45**, 404–413 (1983).
34. P. J. Bickel, J. Ghosh, A decomposition for the likelihood ratio statistic and the Bartlett correction–A Bayesian argument. *Ann. Stat.* **18**, 1070–1090 (1990).
35. L. H Moulton, L. A. Weissfeld, R. T. S. Laurent, Bartlett correction factors in logistic regression models. *Comput. Stat. Data Anal.* **15**, 1–11 (1993).
36. P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, A. R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).
37. E. Vittinghoff, C. E. McCulloch, Relaxing the rule of ten events per variable in logistic and Cox regression. *Am. J. Epidemiol.* **165**, 710–718 (2007).
38. D. S. Courvoisier, C Combescure, T. Agoritsas, A. Gayet-Ageron, T. V. Perneger, Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *J. Clin. Epidemiol.* **64**, 993–1000 (2011).
39. E. J. Candès, P. Sur, The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. arXiv:1804.09753 (25 April 2018).