A Modern Maximum-Likelihood Theory for High-dimensional Logistic Regression

Pragya Sur[*] Emmanuel J. Candès[*†]

April 2018

Abstract

Every student in statistics or data science learns early on that when the sample size n largely exceeds the number p of variables, fitting a logistic model produces estimates that are approximately unbiased. Every student also learns that there are formulas to predict the variability of these estimates which are used for the purpose of statistical inference; for instance, to produce p-values for testing the significance of regression coefficients. Although these formulas come from large sample asymptotics, we are often told that we are on reasonably safe grounds when n is large in such a way that $n \geq 5p$ or $n \geq 10p$. This paper shows that this is far from the case, and consequently, inferences routinely 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. Then we show that (1) the MLE is biased, (2) the variability of the MLE is far greater than classically predicted, and (3) the commonly used likelihood-ratio test (LRT) is not distributed as a chi-square. The bias of the MLE is extremely problematic as it yields completely wrong predictions for the probability of a case based on observed values of the covariates. We develop a new theory, which asymptotically predicts (1) the bias of the MLE, (2) the variability of the MLE, and (3) the distribution of the LRT. We empirically also demonstrate that these predictions are extremely accurate in finite samples. Further, an appealing feature is that these novel predictions depend on the unknown sequence of regression coefficients only through a single scalar, the overall strength of the signal. This suggests very concrete procedures to adjust inference; we describe one such procedure learning a single parameter from data and producing accurate inference. For space reasons, we do not provide a full mathematical analysis of our results. However, we give a brief overview of the key arguments, which rely on the theory of (generalized) approximate message passing algorithms as well as on leave-one-observation/predictor out approaches.

1 Introduction

1.1 Logistic regression: classical theory and practice

Logistic regression [40, 41, 55] is by and large the most frequently used model to estimate the probability of a binary response from the value of multiple features/predictor variables. It is used all the time 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, every graduate student in statistics or any field that remotely involves data analysis learns about logistic regression, perhaps before any other nonlinear multivariate model. In particular, every student knows how to interpret the excerpt of the computer output from Figure 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:

[*] Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

[†] Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.

```
> 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
```

Figure 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 standard deviation of $\hat\beta_j$ as well as a p-value for testing whether $\beta_j = 0$ or not.

1. Fitting a model via maximum likelihood produces estimates that are *approximately unbiased.*
2. There are formulas to *predict the accuracy or variability* of the maximum-likelihood estimate (MLE) (used in the computer output from Figure 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

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

where $\rho'(t) = e^t/(1 + e^t)$ is the standard sigmoidal function. Then everyone knows [40, 41, 55] that in the limit of large samples in which p is fixed and $n \to \infty$, the MLE $\hat{\boldsymbol{\beta}}$ obeys

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

where $\boldsymbol{I}_{\boldsymbol{\beta}}$ is the $p \times p$ Fisher information matrix evaluated at the true $\boldsymbol{\beta}$. A classical way of understanding (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 approximation (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 essentially what R does in Figure 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 [57], which gives the asymptotic distribution of the likelihood ratio test (LRT):

3. 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 chi-square distribution* with k degrees of freedom in the limit of large samples.

Once more, this approximation is often used in lots of statistical software packages to obtain p-values for testing the significance of individual and/or groups of coefficients.

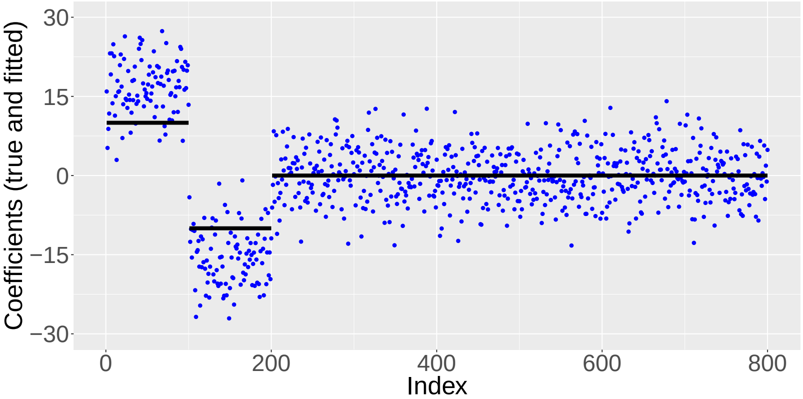


Figure 2: True signal values β_j in black and corresponding ML estimates $\hat{\beta}_j$ (blue points). Observe that estimates of effect magnitudes are seriously biased upward.

1.2 Failures in moderately large dimensions

In modern-day data analysis, new technologies now produce extremely large datasets, often with huge numbers of features on each of a comparatively small number of experimental units. Yet, 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 the modern setting where p is not vanishingly small compared to n?

To address this question, we begin by showing results from an empirical study. Throughout this section, we set $n = 4000$ 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

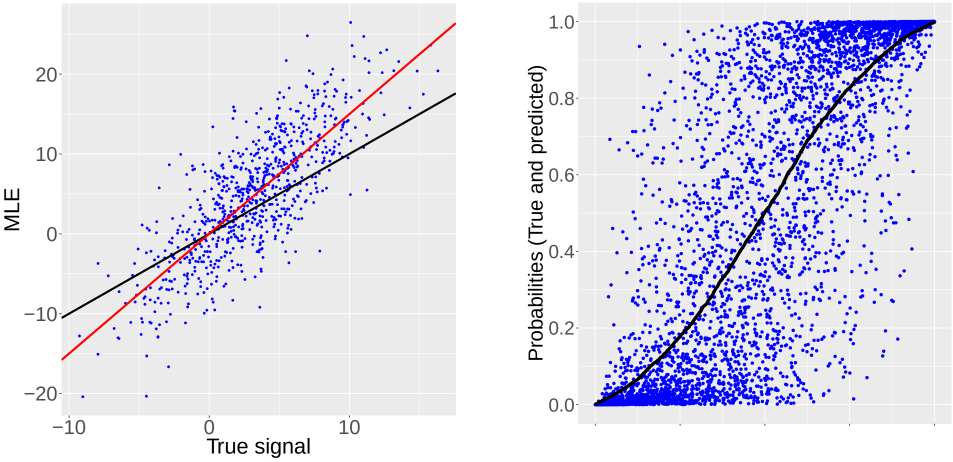
$$\gamma^2 := \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = 5.$$

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 $-4.472 \leq \boldsymbol{X}_i'\boldsymbol{\beta} \leq 4.472$ so that $0.011 \leq \rho'(\boldsymbol{X}_i'\boldsymbol{\beta}) \leq 0.989$.

Unbiasedness? Figure 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 zero. 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 is in stark contradiction to 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 has a strong tendency 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 indices $|\hat{\beta}_j| > |\beta_j|$ so that we are over-estimating the magnitudes of the effects.

The strong 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). Figure 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 $\mathbb{E}\,\hat{\beta}_j \approx 1.499\beta_j$.

The strong bias is highly problematic for estimating the probability of our binary response. Suppose we



(a)

(b)

Figure 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 (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 $\hat{f}(\boldsymbol{X}_*) = \rho'(\boldsymbol{X}_*' \hat{\boldsymbol{\beta}})$. Observe the dramatic shrinkage of the estimates toward the end points.

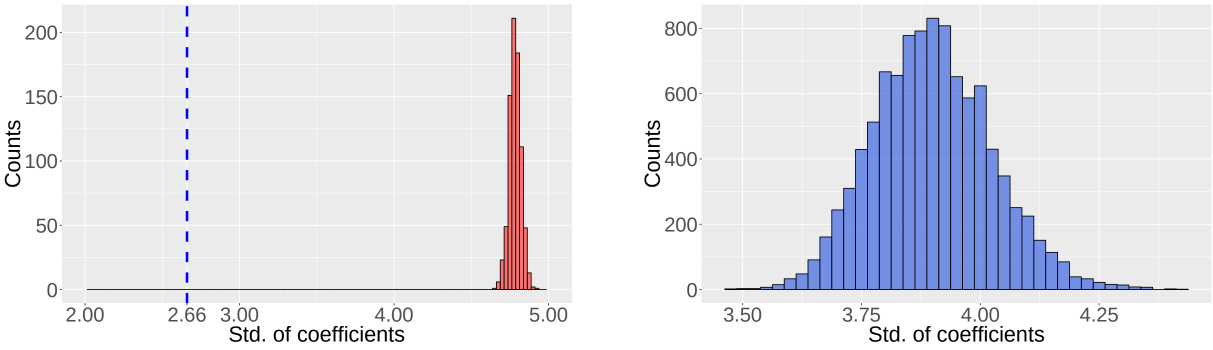
are given a vector of covariates $\boldsymbol{X}_*$ and estimate the regression function $f(\boldsymbol{X}_*) = \mathbb{P}(y = 1 \mid \boldsymbol{X}_*)$ with

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

Then because we tend to over-estimate the magnitudes of the effects, we will also tend to over-estimate or under-estimate the probabilities depending on whether $f(\boldsymbol{X}_*)$ is greater or less than a half. This is illustrated in Figure 3(b). Observe that when $f(\boldsymbol{X}_*) < 1/2$, lots of predictions tend to be close to zero, even when $f(\boldsymbol{X}_*)$ is nowhere near zero. A similar behavior is obtained by symmetry; when $f(\boldsymbol{X}_*) > 1/2$, we see a shrinkage toward the other end point, namely, one. Hence, we see a massive shrinkage towards the extremes and the phenomenon is amplified as the true probability $f(\boldsymbol{X}_*)$ approaches zero or one. Expressed differently, the MLE may predict that an outcome is almost certain (i.e. $\hat{f}$ is close to zero or one) 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_j's are i.i.d. draws from $\mathcal{N}(7, 1)$, and the other half vanish. Figure 4(a) shows standard errors computed via Monte Carlo of 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 prediction for the standard deviation based on the inverse Fisher information can be explicitly calculated in this setting and turns out to be equal to 2.66, see Appendix A. Since the standard deviation values evidently concentrate around 4.75, we see that in higher dimensions, the variance of the MLE is likely to be much larger than that predicted classically. Naturally, using classical predictions would lead to grossly incorrect p-values and confidence statements, a major issue first noticed in [11].

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), see Appendix A. Since practitioners use these estimates all the time, 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 standard error estimate



(a)

(b)

Figure 4: (a) The histogram is the distribution of $\mathrm{STD}(\hat{\beta}_j)$ for each variable j, in which the standard deviation is estimated from $10,000$ samples. The classical predicted standard error value is shown in blue. Classical theory underestimates the variability of the MLE. (b) Standard error 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.

for a single MLE coordinate corresponding to a null variable. The histogram is shown in Figure 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 Figure 4(a).

Distribution of the LRT? By now, the reader should be suspicious that the chi-square 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 [52], the authors established that for a class of logistic regression models, the LRT converges weakly to a *multiple* of a chi-square 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 one as soon as κ is positive. This factor can be computed by solving a nonlinear system of two equations in two unknowns given in (8) below. Furthermore, [52] 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 [52] assume that $\boldsymbol{\beta} = 0$; that is, they apply under the global null where the response does not depend upon the predictor variables, and it is a priori not clear 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.

To investigate what happens when we are not under the global null, consider the same setting as in Figure 4. Figure 5 shows the histogram of the p-values for testing a null coefficient based on the chi-square approximation. Not only are the p-values far from uniform, the enormous mass near zero is extremely 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 very 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 zero. 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.

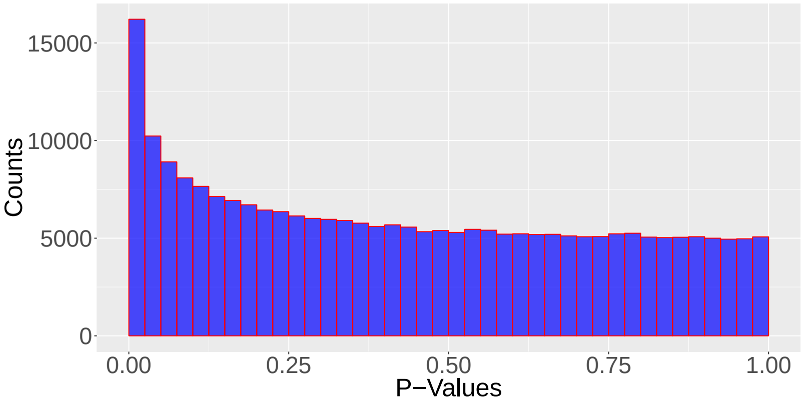


Figure 5: P-values calculated from the χ_1^2 approximation to the LLR. Parameters: $n = 4000, \kappa = 0.2$, with half the co-ordinates of $\boldsymbol{\beta}$ non-zero, generated i.i.d. from $\mathcal{N}(7, 1)$.

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

Table 1: P-value probabilities with standard errors in parentheses. Here, $n = 4000$, $p = 800$, $\boldsymbol{X}$ has i.i.d. Gaussian entries, and half of the entries of $\boldsymbol{\beta}$ are drawn from $\mathcal{N}(7, 1)$.

Summary. We have hopefully made the case that classical results, which software packages continue to rely upon, are downright erroneous in higher dimensions.

1. Estimates seem systematically biased in the sense that effect magnitudes are overestimated.
2. Estimates are far more variable than classically predicted.
3. Inference measures, e.g. p-values, are unreliable especially at small values.

Given the widespread use of logistic regression in high dimensions, a novel theory explaining how to adjust inference to make it valid is seriously needed.

1.3 Our contribution

Our contribution is to develop a brand new theory, which applies to high-dimensional logistic regression models with independent variables, and is capable of accurately describing all the phenomena we have discussed. Taking them one by one, the theory from this paper predicts:

1. the bias of the MLE;
2. the variability of the MLE;
3. and the distribution of the LRT.

These predictions are, in fact, asymptotically exact in a regime where the sample size and the number of features grow to infinity in a fixed ratio. Moreover, we shall see that our theoretical predictions are extremely accurate in finite sample settings in which p is a fraction of n, e.g. $p = 0.2n$.

A very useful feature of this novel theory is that in our model, all of our predictions 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 3 a method for estimating γ and empirically study the quality of inference based on this estimate.

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 requires a great number of steps. This is why in this expository paper we have decided to provide the reader only with the main ideas. All the details may be found in the separate document [51].

1.4 Prior work

Asymptotic properties of M-estimators in the context of linear regression have been extensively studied in diverging dimensions starting from [34], followed by [45] and [46]. These papers investigated the consistency and asymptotic normality properties of M-estimators in a regime where $p = o(n^\alpha)$, for some $\alpha < 1$. Later on, the regime where p is comparable to n became the subject of a series of remarkable works [5, 22, 26–28]; these works only concern the linear model. The finding in these papers is that although M-estimators remain asymptotically unbiased, they are shown to exhibit a form of 'variance inflation'.

Moving on to more general exponential families, [47] studied the asymptotic behavior of likelihood methods and established that the classical Wilks' theorem holds if $p^{3/2}/n \to 0$ and, moreover, that the classical normal approximation to the MLE holds if $p^2/n \to 0$. Subsequently, [32] quantified the ℓ_2 estimation error of the MLE when $p \log p/n \to 0$. Very recently, the authors from [30] investigated the 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 also exists an extensive body of literature on penalized maximum likelihood estimates/procedures for generalized linear models, see [6, 10, 29, 35, 38, 53, 54], for example. This body of literature often allows p to be larger than n but relies upon strong assumptions about the extreme sparsity of the underlying signal. The setting in these works is, therefore, completely different from ours.

Finite sample behavior of both the MLE and the LRT have been extensively studied in the literature. It has been observed that when the sample size is small, the MLE is found to be biased for the regression coefficients. In this context, a series of works— [1, 13, 15, 17, 31, 42, 50] and the references therein—proposed finite sample corrections to the MLE, which typically hinges on an asymptotic expansion of the MLE up to $O(1/n)$ terms. One can plug in an estimator of the $O(1/n)$ term, which would make the resultant corrected statistic $o(1/n)$ accurate. All of these works are in the low-dimensional setting where the MLE is still asymptotically unbiased. The observed bias was simply attributed to a finite sample effect. Jackknife bias reduction procedures for finite samples have been proposed (see [9] and the references cited therein for other finite sample corrections). Similar corrections for the LRT have been studied, see for instance, [3, 7, 8, 14–16, 21, 39, 43]. It was demonstrated in [52] that such finite sample corrections do not yield valid p-values in the high-dimensional regime we consider. In [37], the author proposed a measure for detecting inadequacy of inference that was based on explicit computation of the third order term in the Taylor expansion of the likelihood function. This term is known to be asymptotically negligible in the low-dimensional setting, and is found to be negligible asymptotically in our high-dimensional regime as well, as will be shown in this paper. Thus, this proposal also falls under the niche of a finite sample correction.

A line of simulation based results exist to guide practitioners regarding how large sample sizes are needed so that such finite sample problems would not arise while using classical inference for logistic regression. The rule of thumb is usually 10 events per variable (EPV) or more as mentioned in [33, 44], while a later study [56] suggested that it could be even less. As we clearly see in this paper, such a rule is not at all valid when the number of features is large. [18] 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. As already mentioned, this paper follows up on the earlier contribution [52] 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. As we will see later in Section 2, this former result may be seen as a special case of the novel Theorem 4, which deals with general signal strengths. As is expected, the arguments are now more complicated than when working under the global null.

2 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 $p/n \to \kappa$. We work with independent observations $\{\boldsymbol{X}_i, y_i\}$ from a logistic model such that $\mathbb{P}(y_i = 1 \,|\, \boldsymbol{X}_i) = \rho'(\boldsymbol{X}_i'\boldsymbol{\beta})$. 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. (This means that the columns of the matrix $\boldsymbol{X}$ of covariates are unit-normed in the limit of large samples.). 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

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

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

2.1 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)

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

A first important remark is that in high dimensions, the MLE does not asymptotically exist if the signal strength γ exceeds a certain functional $g_{\text{MLE}}(\kappa)$ of the dimensionality: i.e. $\gamma > g_{\text{MLE}}(\kappa)$. 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. In [52], the authors proved that if $\gamma = 0$ then $g_{\text{MLE}}(1/2) = 0$ (to be exact, they assumed $\boldsymbol{\beta} = \boldsymbol{0}$). To be more precise, the MLE exists if $\kappa < 1/2$ whereas it does not if $\kappa > 1/2$ [19, 20]. Here, it turns out that a companion paper [12] precisely characterizes the region in which the MLE exists.

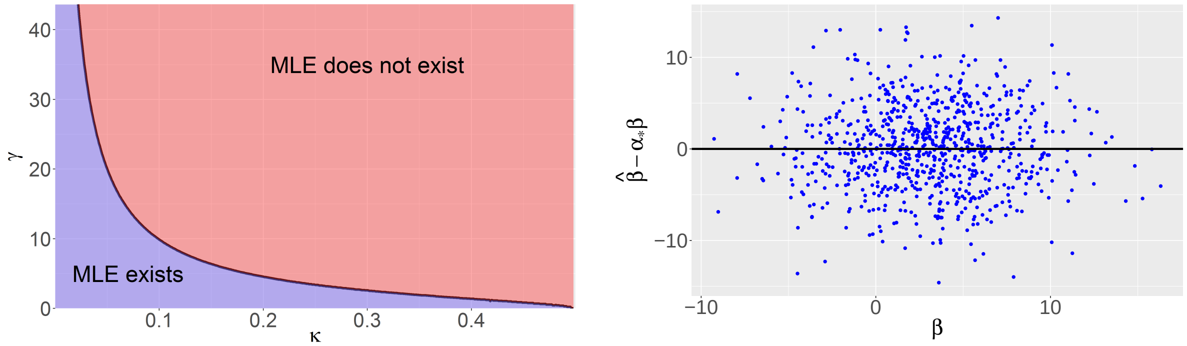
Theorem 1 ([12]). *Let Z be a standard normal variable with density $\varphi(t)$ and V be an independent continuous random variable with density $2\rho'(\gamma t)\varphi(t)$. With $x_+ = \max(x, 0)$, set*

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

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

$$\begin{array}{lcl} \gamma > g_{\text{MLE}}(\kappa) & \Longrightarrow & \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} \to 0, \\ \gamma < g_{\text{MLE}}(\kappa) & \Longrightarrow & \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} \to 1. \end{array}$$

Hence, the curve $\gamma = g_{\text{MLE}}(\kappa)$, or, equivalently, $\kappa = g_{\text{MLE}}^{-1}(\gamma)$ shown in Figure 6 separates the κ–γ 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 Figure 6).



(a)

(b)

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

2.2 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 $(\alpha, \sigma, \lambda)$:

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

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

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

With ρ as in (3), the proximal mapping operator is defined via

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

The system of equations (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 (5) reduces to the following two-dimensional system:

$$\begin{cases} \sigma^2 = \dfrac{1}{\kappa^2}\,\mathbb{E}\left[\left(\lambda\rho'(\mathsf{prox}_{\lambda\rho}(\tau Z))\right)^2\right] \\ 1-\kappa = \mathbb{E}\left[\dfrac{1}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(\tau Z))}\right] \end{cases} \qquad \tau^2 := \kappa\sigma^2, \quad Z \sim \mathcal{N}(0,1). \tag{8}$$

This holds because $Q_1 = 0$. It is not surprising that this system be that from [52] since that work considers $\boldsymbol{\beta} = 0$ and, therefore, $\gamma = 0$.

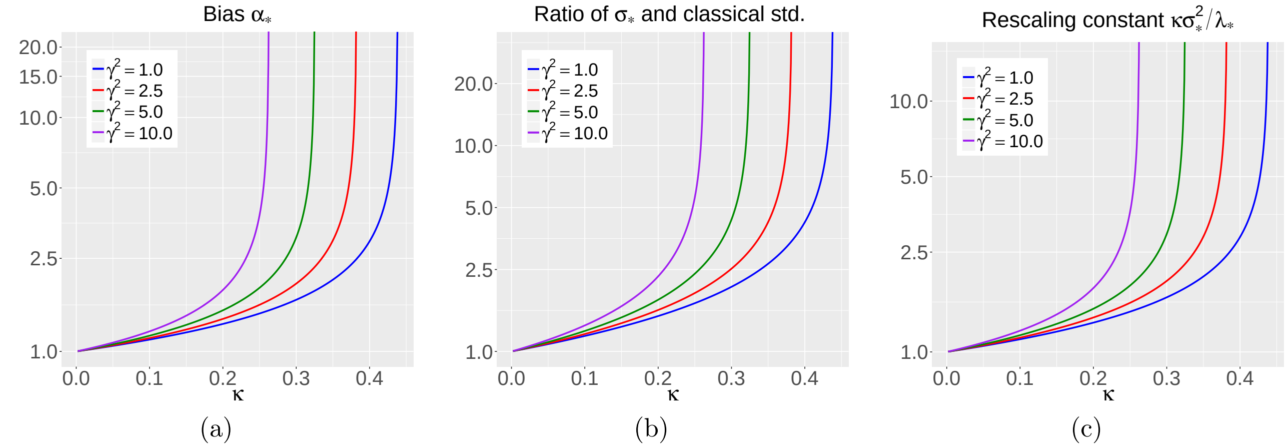


Figure 7: (a) Bias $\alpha_\star$ as a function of κ, for different values of the signal strength γ. Note the logarithmic scale for the y-axis. The curves asymptote at the value of κ for which the MLE ceases to exist. (b) Ratio of the theoretical prediction $\sigma_\star$ and the average standard deviation of the coordinates, as predicted 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.

2.3 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 $\gamma < g_{MLE}(\kappa)$ (the region where the MLE exists asymptotically and shown in Figure 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,*[1] *the marginal distributions of the MLE coordinates obey*

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

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 (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 (5) obeys $\alpha_\star = 1.499$ and Figure 3(a) shows that this is the correct centering.

[1] 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 $\boldsymbol{t}_0, \boldsymbol{t}_1 \in \mathbb{R}^m$, $\|\psi(\boldsymbol{t}_0) - \psi(\boldsymbol{t}_1)\| \le L\left(1 + \|\boldsymbol{t}_0\|^{k-1} + \|\boldsymbol{t}_1\|^{k-1}\right)\|\boldsymbol{t}_0 - \boldsymbol{t}_1\|$.

- 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 (5) obeys $\sigma_\star = 4.744$ and this is what we see in Figure 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 Figure 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 standard deviation $\sigma_\star$ 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 $\alpha_\star$. This dependence is illustrated in Figure 7(a). Further, note that as κ approaches zero, 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 Figure 7(b). This plot shows the theoretical prediction $\sigma_\star$ divided by the average classical standard deviation obtained from $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse of the Fisher information. As κ approaches zero, the ratio goes to 1, indicating that the classical standard deviation 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 either the dimensionality or 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 a half of the coordinates of $\boldsymbol{\beta}$ have constant value 10, and the other half be zero. 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 $\{2000, 4000, 8000\}$. Note that there are several ways of determining $\alpha_\star$ empirically. For instance, the limit (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 indices. 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 (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 (9). We defer the study of the asymptotic variance to the next section.

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)

Table 2: Empirical estimates of the centering and standard deviation of the MLE. Standard errors of these estimates are between parentheses. In this setting, $\kappa = 0.1$ and $\gamma^2 = 5$. Half of the β_j's are equal to ten and the others to zero.

2.4 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 j-th 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*

$$\hat{\beta}_j \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_\star^2). \tag{10}$$

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.

In words, the null MLE coordinates are asymptotically normal with mean zero and variance given by the solution to the system (5). An important remark is this: we have observed that $\sigma_\star$ is an increasing function of γ. Hence, we conclude that for a null variable j, the variance of $\hat{\beta}_j$ is increasingly larger as the magnitude of the other regression coefficients increases.

We return to the finite sample precision of the theoretically predicted asymptotic variance $\sigma_\star$. 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 this same table; 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 (10). 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 Figure 8(a) for some fixed null coordinate. Observe the perfect agreement with a straight line of slope 1.

2.5 The distribution of the LRT

We finally turn our attention to 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 chi-square under the null,*

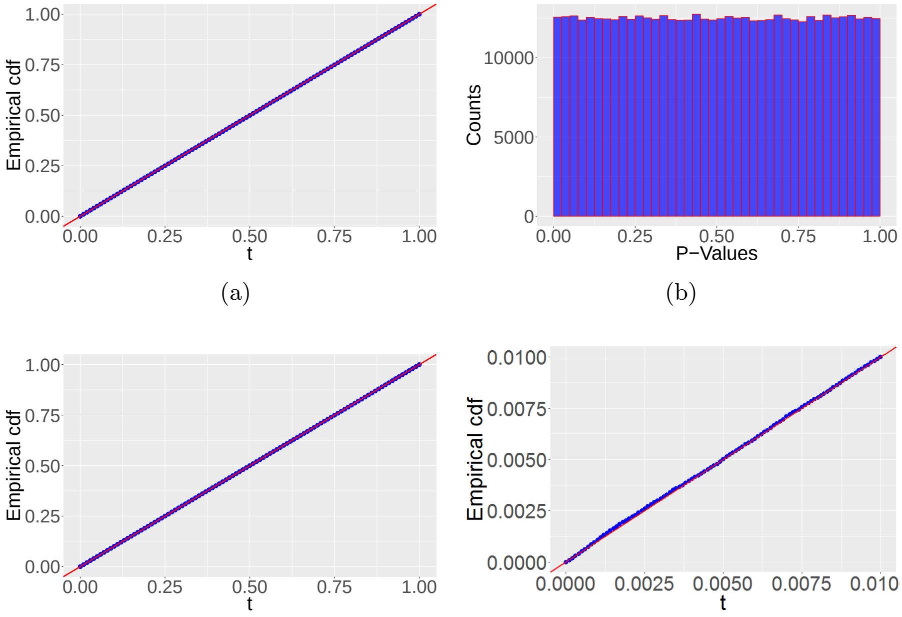
$$2\Lambda_j \xrightarrow{\mathrm{d}} \frac{\kappa\,\sigma_\star^2}{\lambda_\star}\,\chi_1^2. \tag{11}$$

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

This theorem explicitly states that the LRT does not follow a χ_1^2 distribution as soon as $\kappa > 0$ since the multiplicative factor is then larger than one, as demonstrated in Figure 7(c). In other words, the LRT is stochastically quite larger than a χ_1^2, explaining the large spike near zero in Figure 5. Also, Figure 7(c) suggests that as $\kappa \to 0$, the classical result is recovered.

Theorem 4 extends to arbitrary signal strengths the earlier result from [52], which described the distribution of the LRT under the global null ($\beta_j = 0$ for all j). One can quickly verify that when $\gamma = 0$, the multiplicative factor in (11) is that given in [52], which easily follows from the fact that in this case, the system (5) reduces to (8). Furthermore, if the signal is sparse in the sense that $o(n)$ coefficients have non-zero values, $\gamma^2 = 0$, which immediately implies that the asymptotic distribution for the LLR from [52] still holds in such cases.

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



(a)

(b)

(c)

(d)

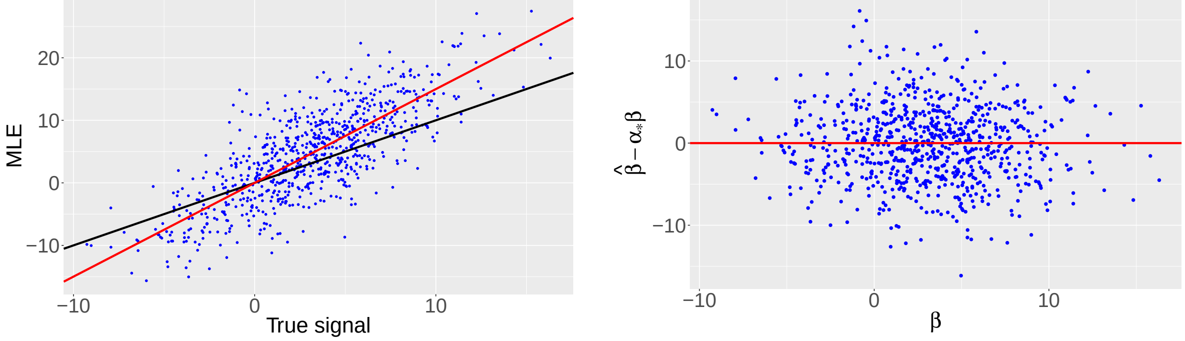
Figure 8: The setting is that from Table 2 with $n = 4000$. (a) Empirical cdf of $\Phi(\hat{\beta}_j/\sigma_\star)$ for a null variable ($\beta_j = 0$). (b) P-values given by the LLR approximation (11) for this same null variable. (c) Empirical distribution of the p-values from (b). (d) Same as (c) but showing accuracy in the lower tail (check the range of the horizontal axis). All these plots are based on 500,000 replicates.

From a multiple testing perspective, it is essential to understand the accuracy of the rescaled chi-square approximation in the tails of the distribution. We plot the empirical cdf of the p-values, zooming in the tail, in Figure 8(d). We find that the rescaled chi-squared approximation works extremely well even in the tails of the distribution.

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

Table 3: P-value probabilities estimated over $500,000$ replicates with standard errors in parentheses. Here, $\kappa = 0.1$ and the setting is otherwise the same as in Table 2.

To obtain a more refined idea of the quality of approximation, we zoom in the smaller bins close to zero and provide estimates of the p-value probabilities in Table 3 for $n = 4000$ and $n = 8000$. The tail approximation is accurate, modulo a slight deviation in the bin for $\mathbb{P}\{\text{p-value}\} \leq 0.1$ for the smaller sample size. For $n = 8000$, 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.



(a)

(b)

Figure 9: Simulation for a non-Gaussian design. The j-th 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 Figure 3. (a) Analogue of Figure 3(a). Red line has slope $\alpha_\star \approx 1.499$.(b) Analogue of Figure 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.

2.6 Other scalings

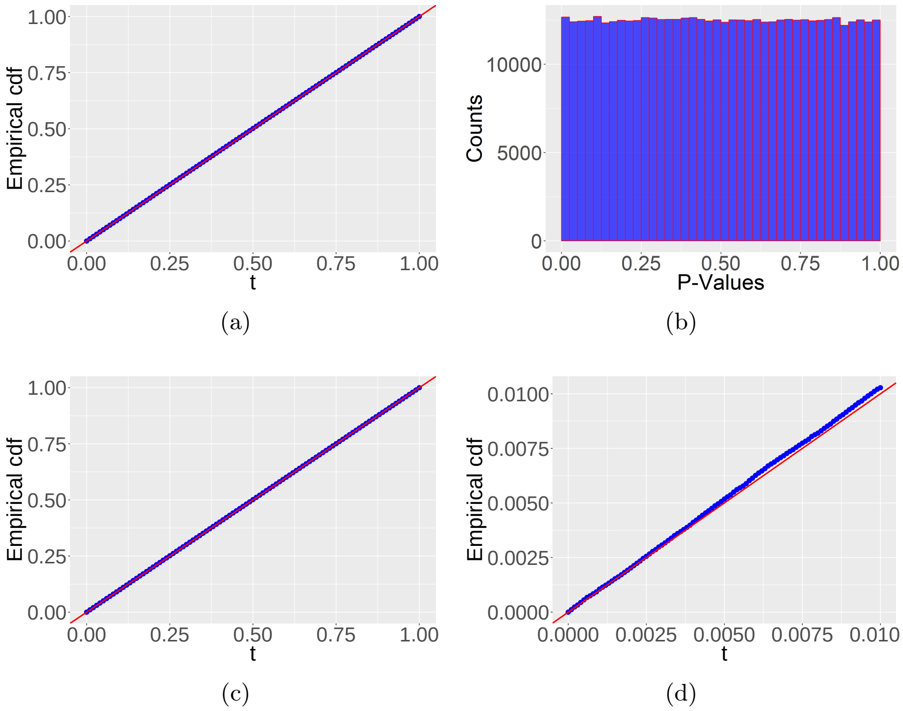
Throughout this section, we worked under the assumption that $\lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \gamma^2$, which does not depend on n, and we explained that this is the only scaling that makes sense to avoid a trivial problem. We set the variables to have variance $1/n$ 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 Theorem 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 asymptotic standard deviation of $\hat{\boldsymbol{\beta}}$ is equal to $\sigma_\star/\sqrt{nv}$, where $\sigma_\star$ is just as in Theorems 2 and 3. On the other hand, the result for the LRT, namely, Theorem 4 is scale invariant; no such trivial adjustment is necessary.

2.7 Non-Gaussian covariates

Our model assumes that the features are Gaussian, yet, 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 j-th 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 zero mean and unit variance. Keeping everything else as in the setting of Figure 3, we study the bias of the MLE in Figure 9(a). As for Gaussian designs, the MLE seriously over-estimates 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 Figure 9(b).

The bulk distribution of a null coordinate suggested by Theorem 3 and the LRT distribution from Theorem 4 are displayed in Figure 10. Other than the design, the setting is the same as for Figure 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



(a)

(b)

(c)

(d)

Figure 10: The features are multinomial as in Figure 9 and the setting is otherwise the same as for Figure 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 (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.

findings indicate that our theory is expected to apply to a far broader class of features.

3 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 the second author presented the new high-dimensional ML theory from this paper at the Mathematisches Forshunginstitut Oberwolfach on March 12, 2018.

3.1 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 2 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 shall 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 $\kappa \leq \kappa_1 \leq \kappa_2 \leq ... \leq \kappa_K \leq 1/2$. 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 programming 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 $\hat{\pi}(\kappa_j) \geq 0.5$. 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 Figure 6, choosing a sufficiently fine grid $\{\kappa_j\}$ would make the linear interpolation step sufficiently precise.)

3.2 Empirical performance of adjusted inference

We demonstrate the accuracy of ProbeFrontier via some empirical results. We begin by generating 4000 i.i.d. observations $(y_i, \boldsymbol{X}_i)$ using the same setup as in Figure 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 (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 (11) equal to 1.1663. Recall from Section 2 that the theoretical values are $(\alpha_\star, \tau_\star, \lambda_\star) = (1.1678, 3.3466, 0.9605)$ and $\kappa\sigma_\star^2/\lambda_\star = 1.1660$. Next, we compute the LLR statistic for each null and p-values from the approximation (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 Figure 11(a) (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 $n = 4000$, $p = 400$ (so that $\kappa = 0.1$), and half the coefficients have a nonvanishing magnitude scaled in such a way that the signal strength is γ. Figure 12(a) displays $\hat{\gamma}$ versus γ in blue, and we notice that ProbeFrontier works very

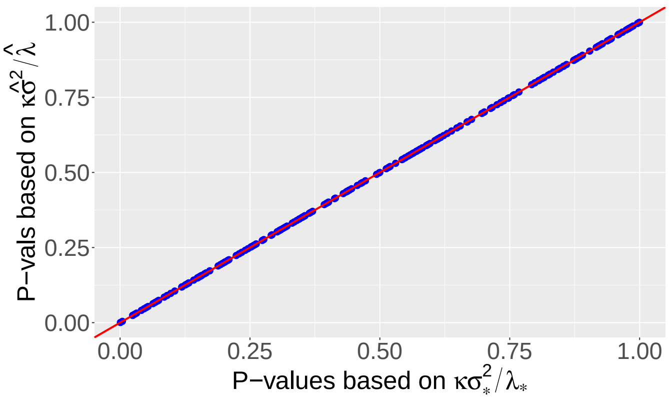
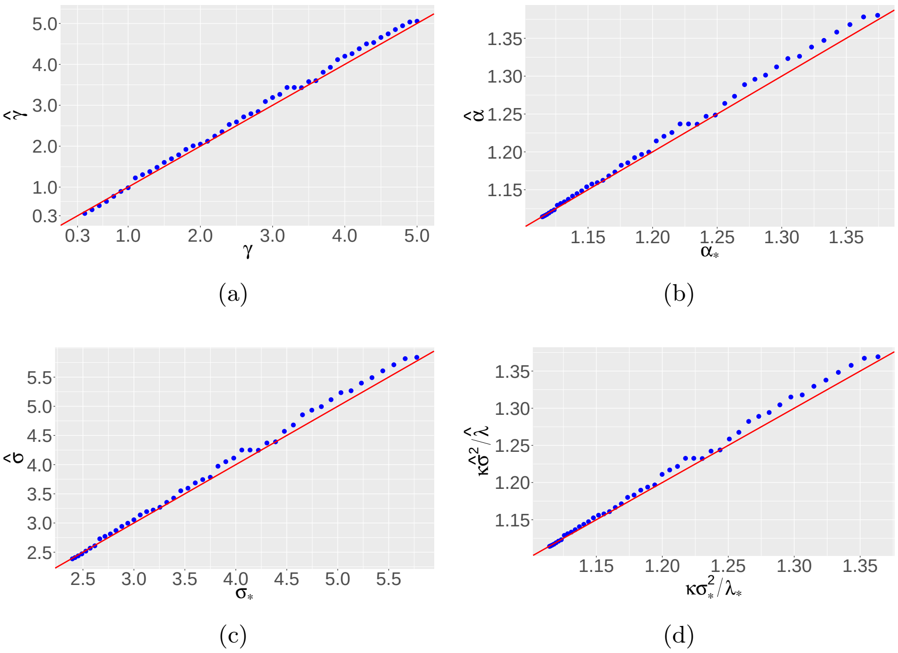


Figure 11: (a) Null p-values obtained using the $(\kappa\hat{\sigma}^2/\hat{\lambda})\,\chi^2_1$ approximation plotted against those obtained using $(\kappa\sigma^2_\star/\lambda_\star)\,\chi^2_1$. Observe the perfect agreement (the diagonal is in red).



(a)

(b)

(c)

(d)

Figure 12: ProbeFrontier estimates of signal strength $\hat{\gamma}$, bias $\hat{\alpha}$, std. dev. $\hat{\sigma}$, and LRT factor $\kappa\hat{\sigma}^2/\hat{\lambda}$ in (11), plotted against the theoretical values.

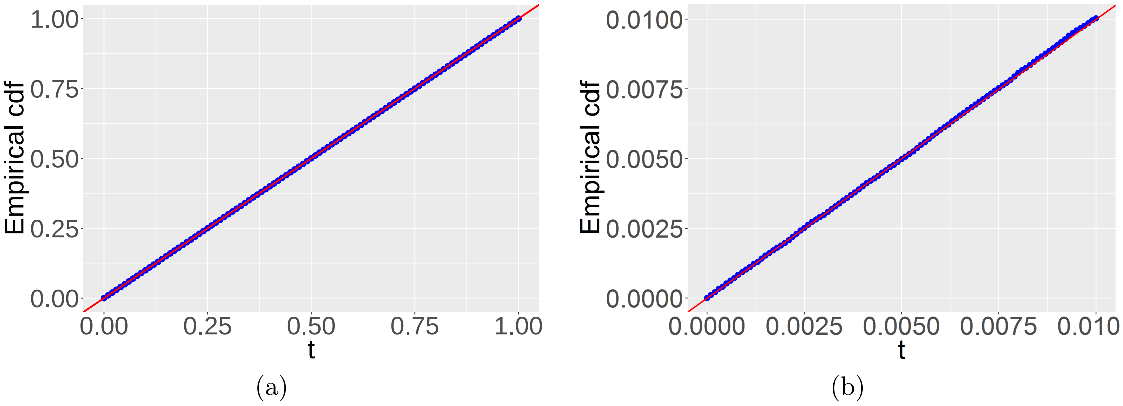


Figure 13: (a) Empirical distribution of the p-values based on the LLR approximation (11), obtained using the estimated factor $\kappa\hat{\sigma}^2/\hat{\lambda}$. (b) Same as (a), but showing the tail of the empirical cdf. The calculations are based on 500,000 replicates.

well. We observe that the blue points fluctuate very 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 (5) immediately yields an estimate for the bias $\alpha_\star$, standard deviation $\sigma_\star$ and the rescaling factor in (11). We study the accuracy of these estimates in Figure 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.

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

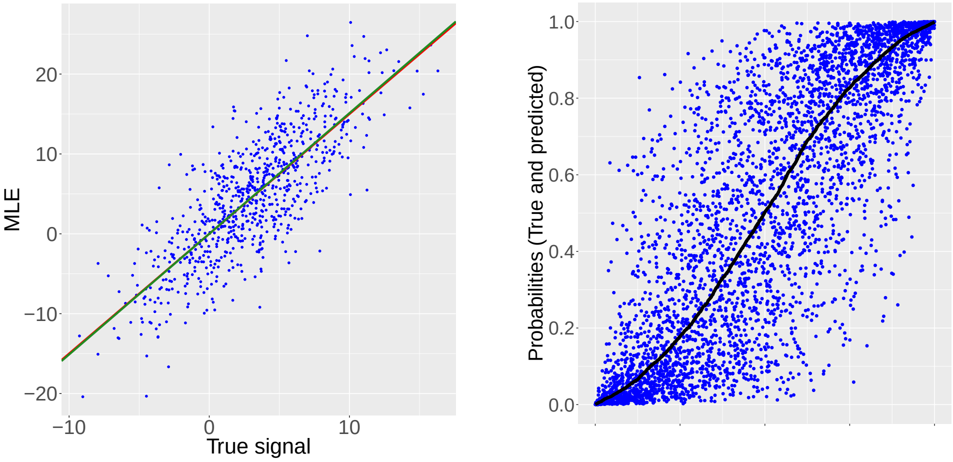
Table 4: Parameter estimates in the setting of Table 2. We report an average over 6000 replicates with the std. dev. between parentheses.

.

Finally, we focus on the estimation accuracy for a particular (κ, γ) pair across several replicates. In the setting of Figure 8, we generate 6000 samples and obtain estimates of bias $(\hat{\alpha})$, std. dev. $(\hat{\sigma})$, and rescaling factor for the LRT $(\kappa\hat{\sigma}^2/\hat{\lambda})$. The average 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 (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 Figure 13. We observe the perfect agreement with the uniform distribution, establishing the practical applicability of our theory and methods.

3.3 De-biasing 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 Figure 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 Figure 14(a). Clearly, the estimate of bias is extremely precise and coefficient estimates $\hat{\beta}_j/\hat{\alpha}$ appear nearly unbiased.



(a) (b)

Figure 14: (a) Scatterplot of the pairs $(\beta_j, \hat{\beta}_j)$ for the dataset from Figure 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 *de-biased MLE* (blue point). Observe that the black curve now passes through the center of the blue point cloud. Our predictions are fairly unbiased.

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

4 Main Mathematical Ideas

As we mentioned earlier, we do not provide detailed proofs in this paper. The reader will find them in [51] and the first author's Ph. D. thesis. However, in this section we give some of the key mathematical ideas and explain some of the main steps in the arguments, relying as much as possible on published results from [52].

4.1 The bulk distribution of the MLE

To analyze the MLE, we introduce an approximate message passing (AMP) algorithm that tracks the MLE in the limit of large n and p. Our purpose is a little different from the work in [48] which, in the context of generalized linear models, proposed AMP algorithms for Bayesian posterior inference, and whose properties have later been studied in [36] and [2]. To the best of our knowledge, an AMP algorithm for tracking the MLE from a logistic model has not yet been proposed in the literature. Our starting point is to write down a sequence of AMP iterates $\{\boldsymbol{S}^t, \hat{\boldsymbol{\beta}}^t\}_{t\geq 0}$, with $\boldsymbol{S}^t \in \mathbb{R}^n, \hat{\boldsymbol{\beta}}^t \in \mathbb{R}^p$, using the following scheme: starting with an initial guess $\boldsymbol{\beta}^0$, set $\boldsymbol{S}^0 = \boldsymbol{X}\boldsymbol{\beta}^0$ and for $t = 1, 2, \ldots$, inductively define

$$\begin{aligned} \hat{\boldsymbol{\beta}}^t &= \hat{\boldsymbol{\beta}}^{t-1} + \kappa^{-1}\boldsymbol{X}'\Psi_{t-1}(\boldsymbol{y}, \boldsymbol{S}^{t-1}) \\ \boldsymbol{S}^t &= \boldsymbol{X}\hat{\boldsymbol{\beta}}^t - \Psi_{t-1}(\boldsymbol{y}, \boldsymbol{S}^{t-1}) \end{aligned} \tag{12}$$

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

$$\Psi_t(y,s) = \lambda_t r_t, \qquad r_t = y - \rho'(\mathsf{prox}_{\lambda_t \rho}(\lambda_t y + s)). \tag{13}$$

Observe that the evolution (12) depends on a sequence of parameters $\{\lambda_t\}$ whose dynamics we describe next.

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

$$\mathbb{E}\left[\frac{2\rho'(Q_1^t)}{1 + \lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2^t))}\right] = 1 - \kappa \tag{14}$$

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

$$\begin{aligned} \alpha_{t+1} &= \alpha_t + \frac{1}{\kappa\gamma^2}\, \mathbb{E}\left[2\rho'(Q_1^t) Q_1^t \lambda_t \rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right] \\ \sigma_{t+1}^2 &= \frac{1}{\kappa^2}\, \mathbb{E}\left[2\rho'(Q_1^t)\left(\lambda_t \rho'(\mathsf{prox}_{\lambda_t\rho}(Q_2^t))\right)^2\right]. \end{aligned} \tag{15}$$

To repeat, we run the AMP iterations (12) using the scalar variables $\{\lambda_t\}$ calculated via the *variance map* updates (14)–(15).

In the regime where the MLE exists (see Figure 6), the variance map updates (14)–(15) converge (as $t \to \infty$) to a unique fixed point $(\alpha_\star, \sigma_\star, \lambda_\star)$. Note that by definition, $(\alpha_\star, \sigma_\star, \lambda_\star)$ is the solution to our system (5) in three unknowns. From now on, we use $\alpha_0 = \alpha_\star$, $\sigma_0 = \sigma_\star$ so that the sequence $\{\alpha_t, \sigma_t, \lambda_t\}$ is stationary; i. e. for all $t \geq 0$,

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

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

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

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

The crucial point is that we can study the properties of the MLE by studying the properties of the AMP iterates with the proviso that they converge. It turns out that the study of the sequence $\{\boldsymbol{S}^t, \hat{\boldsymbol{\beta}}^t\}$ is amenable to a rigorous analysis because several transformations reduce the above algorithm to a generalized AMP algorithm [36], which in turn yields a characterization of the limiting variance of the AMP iterates: for any function ψ as in Theorem 2, we have as $n \to \infty$,

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_j^t - \alpha_\star \beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi(\sigma_\star Z, \beta)\right], \tag{16}$$

where β is drawn from the distribution Π (see Theorem 2) independently of $Z \sim \mathcal{N}(0,1)$, and $\sigma_\star$ is as above. To summarize, the asymptotic behavior of the AMP iterates $\hat{\boldsymbol{\beta}}^t$ can be characterized through a standard Gaussian variable, the distribution Π and the scalar quantity $\sigma_\star$ determined by the iteration (14)–(15). The description of our AMP algorithm and large sample properties of the iterates are understood only when we understand the behavior of the scalar sequences $\{\alpha_t, \sigma_t, \lambda_t\}_{t\geq 0}$, which are known as the state evolution sequence in the literature; this formalism was introduced in [4, 23–25]. From here on, an analysis similar to that in [52] establishes that in the limit of large iteration counts, the AMP iterates converge to the MLE, that is,

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

which is the content of Theorem 2.

4.2 The distribution of a null coordinate

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

Now consider Theorem 2 with $\psi(t,u) = t^2 1(u=0)$. Strictly speaking, this is a discontinuous function which is not pseudo-Lipschitz. However, we can work with a smooth approximation ψ_a, instead, obtained using standard techniques for smoothing an indicator function, such that the error $\|\psi - \psi_a\|_2$ is arbitrarily small. For simplicity, we skip the technical details underlying this approximation, and motivate the subsequent arguments using ψ directly. Theorem 2 then yields

$$\frac{1}{p}\sum_{j\in[p]:\beta_j=0}\hat{\beta}_j^2 \xrightarrow{\text{a.s.}} \sigma_\star^2 \mathbb{P}_\Pi\left[\beta=0\right] \quad\Longrightarrow\quad \frac{1}{|j\in[p]:\beta_j=0|}\sum_{j\in[p]:\beta_j=0}\hat{\beta}_j^2 \xrightarrow{\text{a.s.}} \sigma_\star^2. \tag{17}$$

Without loss of generality, assume that the first k coordinates of $\boldsymbol{\beta}$ vanish, and that $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = \left(\mathbf{0}_{[k]}, \boldsymbol{\beta}_{-[k]}\right)$ and similarly for $\hat{\boldsymbol{\beta}}$. From the rotational distributional invariance of the $\boldsymbol{X}_i$'s, it can be shown that for any fixed orthogonal matrix $\boldsymbol{U}\in\mathbb{R}^{k\times k}$, $\hat{\boldsymbol{\beta}} \overset{\text{d}}{=} \left(\boldsymbol{U}\hat{\boldsymbol{\beta}}_{[k]}, \hat{\boldsymbol{\beta}}_{-[k]}\right)$. Consequently, $\hat{\boldsymbol{\beta}}_{[k]}/\|\hat{\boldsymbol{\beta}}_{[k]}\|$ is uniformly distributed on the unit sphere $\mathbb{S}^{k-1}$ and is independent of $\|\hat{\boldsymbol{\beta}}_{[k]}\|$. Thus, any null coordinate $\hat{\beta}_j$ has the same distribution as $\|\hat{\boldsymbol{\beta}}_{[k]}\|\, Z_j/\|\boldsymbol{Z}\|$, where $\boldsymbol{Z}\sim\mathcal{N}(0,\boldsymbol{I}_k)$, independent of $\hat{\boldsymbol{\beta}}_{[k]}$. From (17) and the weak law of large numbers, we have $\|\hat{\boldsymbol{\beta}}_{[k]}\|/\|\boldsymbol{Z}\| \xrightarrow{\mathbb{P}} \sigma_\star$, leading to $\hat{\beta}_j \xrightarrow{\text{d}} \mathcal{N}(0,\sigma_\star^2)$.

4.3 The distribution of the LRT

Once the distribution of $\hat{\beta}_j$ for a null j is known, the distribution of the LRT is a stone throw away, at least conceptually; that is to say, if we are willing to ignore some technical difficulties and leverage existing work. Indeed, following a reduction similar to that in [52], one can establish that

$$2\Lambda_j = \frac{\kappa}{\lambda_{[-j]}}\hat{\beta}_j^2 + o_P(1), \tag{18}$$

where $\lambda_{[-j]} := \text{Tr}\left[\nabla^2(\ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]}))^{-1}\right]/n$ in which $\ell_{[-j]}$ is the negative log-likelihood with the j-th variable removed and $\hat{\boldsymbol{\beta}}_{[-j]}$ the corresponding MLE. Put $\lambda = \text{Tr}[\nabla^2(\ell(\hat{\boldsymbol{\beta}}))^{-1}]/n$. Then following an approach similar to that in [52, Appendix I], it can be established that $\lambda_{[-j]} = \lambda + o_P(1) \xrightarrow{\mathbb{P}} \lambda_\star$. This gives that $2\Lambda_j$ is a multiple of a χ_1^2 variable with multiplicative factor given by $\kappa\sigma_\star^2/\lambda_\star$.

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

5 Broader Implications and Future Directions

This paper shows that in high-dimensions, classical ML theory is unacceptable. Among other things, classical theory predicts that the MLE is approximately unbiased when in reality it seriously overestimates effect magnitudes. Since the purpose of logistic modeling is to estimate the risk of a specific disease given a patient's observed characteristics, say, the bias of the MLE is extremely problematic. As we have seen, an immediate consequence of the strong bias is that the MLE either dramatically overestimates, or underestimates, the chance of being sick. The issue becomes increasingly severe as either the dimensionality or the signal strength, or both, increase. This, along with the fact that p-values computed from classical approximations are misleading, 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 new theory which gives the asymptotic distribution of the MLE and the LRT in a model with independent covariates. As seen in Section 2.7, 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. Although our method works extremely well, it would be of interest to study others as well.

Finally, we conclude this paper with two promising directions for future work: (1) It would be of great interest to develop corresponding results in the case where the predictors are correlated. (2) It would be of interest to extend the results from this paper to other generalized linear models.

Acknowledgements

P. S. was partially supported by the Ric Weiland Graduate Fellowship in the School of Humanities and Sciences, Stanford University. E. 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. would like to thank Andrea Montanari and Subhabrata Sen for helpful discussions on AMP. We thank Małgorzata Bogdan, Nikolaos Ignatiadis, Asaf Weinstein, Lucas Janson and Chiara Sabatti for useful comments about an early version of the paper.

References

[1] JA Anderson and SC Richardson. Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1):71–78, 1979.

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

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

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

[5] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[6] Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2013.

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

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

[9] SB Bull, WW Hauck, and CMT Greenwood. Two-step jackknife bias reduction for logistic regression mles. *Communications in Statistics-Simulation and Computation*, 23(1):59–88, 1994.

[10] Florentina Bunea et al. Honest variable selection in linear and logistic regression models via $\ell 1$ and $\ell 1+\ell 2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[11] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

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

[13] John B Copas. Binary regression models for contaminated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 225–265, 1988.

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

[15] Gauss M Cordeiro and Francisco Cribari-Neto. *An introduction to Bartlett correction and bias reduction.* Springer, 2014.

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

[17] Gauss M Cordeiro and Peter McCullagh. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 629–643, 1991.

[18] Delphine S Courvoisier, Christophe Combescure, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V Perneger. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of clinical epidemiology*, 64(9):993–1000, 2011.

[19] Thomas M Cover. Geometrical and statistical properties of linear threshold devices. *Ph.D. thesis*, May 1964.

[20] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[21] Francisco Cribari-Neto and Gauss M Cordeiro. On bartlett and bartlett-type corrections francisco cribari-neto. *Econometric reviews*, 15(4):339–367, 1996.

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

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

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

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

[26] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

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

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

[29] Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

[30] Yingying Fan, Emre Demirkaya, and Jinchi Lv. Nonuniformity of p-values can occur early in diverging dimensions. *https://arxiv.org/abs/1705.03604*, May 2017.

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

[32] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

[33] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[34] Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821, 1973.

[35] Jana Janková and Sara van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

[36] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[37] Dennis E Jennings. Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81(394):471–476, 1986.

[38] Sham Kakade, Ohad Shamir, Karthik Sindharan, and Ambuj Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388, 2010.

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

[40] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses.* Springer Science & Business Media, 2006.

[41] Peter McCullagh and James A Nelder. Generalized linear models. *Monograph on Statistics and Applied Probability*, 1989.

[42] GJ McLachlan. A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics*, 22(4):621–627, 1980.

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

[44] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

[45] Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. i. consistency. *The Annals of Statistics*, pages 1298–1309, 1984.

[46] Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; ii. normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985.

[47] Stephen Portnoy et al. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16(1):356–366, 1988.

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

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

[50] Robert L Schaefer. Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, 2(1):71–78, 1983.

[51] Pragya Sur and Emmanuel J Candès. Additional supplementary materials for: A modern maximum-likelihood theory for high-dimensional logistic regression. *https://statweb.stanford.edu/~candes/papers/proofs_LogisticAMP.pdf*, 2018.

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

[53] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[54] Sara A Van de Geer et al. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[55] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[56] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

[57] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

A Fisher information

We work with the model from Section 2 and introduce the Fisher information matrix defined as

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

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

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

From here on, a reasonably straightforward calculation gives

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

where

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

This implies that

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

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

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

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

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

B Properties of fixed points of the AMP algorithm

In this section, we elaborate on the connection between the fixed points of (12) and the MLE $\hat{\boldsymbol{\beta}}$. From (12), we immediately see that if $(\hat{\boldsymbol{\beta}}_\star, \boldsymbol{S}_\star)$ is a fixed point, the pair satisfies

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

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

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

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

C Refined analysis of the distribution of a null coordinate

The AMP analysis is useful to analyze the bulk behavior of the MLE; i.e. the expected behavior when averaging over all coordinates. It also helps in characterizing the distribution of a null coordinate when the limiting empirical cdf does not have a point mass at zero, as we have seen in Section 4.2. However, the study of the behavior of a single coordinate when there is an arbitrary number of nulls requires a more refined analysis. To this end, the proof uses a leave-one-out approach, as in [27, 28, 52]. The complete rigorous technical details are very involved and this is a reason why we only present approximate or non-rigorous heuristic calculations.

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

$$\boldsymbol{X}_i'\hat{\boldsymbol{\beta}} \approx \boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}; \tag{19}$$

here, $\boldsymbol{X}_{i,-j}$ is i-th row of the reduced matrix of predictors with the j-th column removed and $\hat{\boldsymbol{\beta}}_{[-j]}$ is the MLE for the reduced model. On the one hand, the approximation (19) suggests Taylor expanding $\rho'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})$ around the point $\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}$:

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

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

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

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

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

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

where $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})\boldsymbol{X}_{\bullet -j}(\nabla^2\ell_{-j}(\hat{\boldsymbol{\beta}}_{[-j]}))^{-1}\boldsymbol{X}_{\bullet -j}'\boldsymbol{D}^{1/2}(\hat{\boldsymbol{\beta}}_{[-j]})$ and $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_{[-j]})$ is an $n \times n$ diagonal matrix with $i-$th entry given by $\rho''(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})$. Above $\boldsymbol{X}_{\bullet j}$ is the j-th column of $\boldsymbol{X}$ and $\boldsymbol{X}_{\bullet -j}$ all the others. It was established in [52] that the denominator above is equal to $\kappa/\lambda_{[-j]} + o_P(1)$, where, we have see in Section 4.3 that

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

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

$$\hat{\beta}_j = \frac{\lambda_{[-j]}s_j}{\kappa}Z + o_P(1), \quad s_j^2 = \frac{1}{n}\sum_{i=1}^n\left(y_i - \rho'(\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]})\right)^2, \tag{20}$$

where Z is an independent standard normal. In Section 4.3, we saw that $\lambda_{[-j]} \xrightarrow{\mathbb{P}} \lambda_\star$. It remains to understand the behavior of s_j. Looking at s_j, the complicated dependence structure between $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{y}, \boldsymbol{X})$ makes this a potentially hard task. This is why we shall use a leave-one-out argument and seek to express the fitted values $\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-j]}$ in terms of $\boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-i],[-j]}$, where $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ is the MLE when both the j-th predictor and the i-th observation are dropped. The independence between $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ will simplify matters. Denote by $\nabla \ell_{[-i],[-j]}(\tilde{\boldsymbol{\beta}}_{[-i],[-j]}) = 0$ the reduced score equation and subtract it from the score equation for $\hat{\boldsymbol{\beta}}$ to obtain

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

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

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

Since $\boldsymbol{X}_{i,-j}$ and $\hat{\boldsymbol{\beta}}_{[-i],[-j]}$ are independent, by Hanson-Wright inequality [49, Theorem 1.1], the quadratic form above is approximately equal to $\mathrm{Tr}\left[\nabla^2 \ell_{[-i],[-j]}(\hat{\boldsymbol{\beta}}_{[-i],[-j]})^{-1}\right]$. Recall that $\lambda_{[-j]} = \mathrm{Tr}[\nabla^2 \ell_{[-j]}(\hat{\boldsymbol{\beta}}_{[-j]})^{-1}]$ and again, for a large number of i.i.d. observations, we expect these two quantities to be close. Hence, the fitted values can be approximated as

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

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

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

Further, by the triangle inequality we can show that

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

It can be shown that the residuals $\{y_i - \rho'(\mathsf{prox}_{\lambda_\star\rho}(\lambda_\star y_i + \boldsymbol{X}_{i,-j}'\hat{\boldsymbol{\beta}}_{[-i],[-j]}))\}_{i=1,\dots,n}$ are asymptotically independent among themselves, which implies that averaging over the squared residuals as in (20) should converge in probability to the corresponding expectation, leading to

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

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

$$\lim_{n\to\infty}\frac{\langle\hat{\boldsymbol{\beta}}_{[-i],[-j]} - \alpha_\star\boldsymbol{\beta}_{-j}, \boldsymbol{\beta}_{-j}\rangle}{n} = \kappa\lim_{t\to\infty}\lim_{n\to\infty}\frac{\langle\hat{\boldsymbol{\beta}}^t - \alpha_\star\boldsymbol{\beta}, \boldsymbol{\beta}\rangle}{p} = 0, \tag{21}$$

by an application of (16). Hence, writing $y_i = 1(U_i \leq \rho'(\boldsymbol{X}_{i,-j}'\boldsymbol{\beta}_{-j}))$, where the U_i's are i.i.d. $U(0,1)$ independent from anything else, we have

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

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

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

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