The Asymptotic Distribution of the MLE in High-dimensional Logistic Models: Arbitrary Covariance

Qian Zhao* Pragya Sur† Emmanuel J. Candès*‡

January 25, 2020

Abstract

We study the distribution of the maximum likelihood estimate (MLE) in high-dimensional logistic models, extending the recent results from [14] to the case where the Gaussian covariates may have an arbitrary covariance structure. We prove that in the limit of large problems holding the ratio between the number p of covariates and the sample size n constant, every finite list of MLE coordinates follows a multivariate normal distribution. Concretely, the jth coordinate $\hat{\beta}_j$ of the MLE is asymptotically normally distributed with mean $\alpha_\star \beta_j$ and standard deviation $\sigma_\star/\tau_j$; here, β_j is the value of the true regression coefficient, and τ_j the standard deviation of the jth predictor conditional on all the others. The numerical parameters $\alpha_\star > 1$ and $\sigma_\star$ only depend upon the problem dimensionality p/n and the overall signal strength, and can be accurately estimated. Our results imply that the MLE's magnitude is biased upwards and that the MLE's standard deviation is greater than that predicted by classical theory. We present a series of experiments on simulated and real data showing excellent agreement with the theory.

1 Introduction

Logistic regression is the most widely applied statistical model for fitting a binary response from a list of covariates. This model is used in a great number of disciplines ranging from social science to biomedical studies. For instance, logistic regression is routinely used to understand the association between the susceptibility of a disease and genetic and/or environmental risk factors.

A logistic model is usually constructed by the method of maximum likelihood (ML) and it is therefore critically important to understand the properties of ML estimators (MLE) in order to test hypotheses, make predictions and understand their validity. In this regard, assuming the logistic model holds, classical ML theory provides the asymptotic distribution of the MLE when the number of observations n tends to infinity while the *number p of variables remains constant.* In a nutshell, the MLE is asymptotically normal with mean equal to the true vector of regression coefficients and variance equal to $\mathcal{I}_{\boldsymbol{\beta}}^{-1}$, where $\mathcal{I}_{\boldsymbol{\beta}}$ is the Fisher information evaluated at true coefficients [7, Appendix A], [16, Chapter 5]. Another staple of classical ML theory is that the extensively used likelihood ratio test (LRT) asymptotically follows a chi-square distribution under the null, a result known as Wilk's theorem [17] [16, Chapter 16]. Again, this holds in the limit where p is fixed and $n \to \infty$ so that the dimensionality p/n is vanishingly small.

*Department of Statistics, Stanford University, Stanford, CA 94305
†Center for Research on Computation and Society, SEAS, Harvard University, Cambridge, MA 02138
‡Department of Mathematics, Stanford University, Stanford, CA 94305

1.1 High-dimensional maximum-likelihood theory

Against this background, a recent paper [14] showed that the classical theory does not even approximately hold in large sample sizes if p is not negligible compared to n. In more details, empirical and theoretical analyses in [14] establish the following conclusions:

1. The MLE is biased in that it overestimates the true effect magnitudes.
2. The variance of the MLE is larger than that implied by the inverse Fisher information.
3. The LRT is not distributed as a chi-square variable; it is stochastically larger than a chi-square.

Under a suitable model for the covariates, [14] developed formulas to calculate the asymptotic bias and variance of the MLE under a limit of large samples where the ratio p/n between the number of variables and the sample size has a positive limit κ. Operationally, these results provide an excellent approximation of the distribution of the MLE in large logistic models in which the number of variables obey $p \approx \kappa\, n$ (this is the same regime as that considered in random matrix theory when researchers study the eigenvalue distribution of sample covariance matrices in high dimensions). Furthermore, [14] also proved that the LRT is asymptotically distributed as a fixed multiple of a chi-square, with a multiplicative factor that can be determined.

1.2 This paper

The asymptotic distribution of the MLE in high-dimensional logistic regression briefly reviewed above holds for models in which the covariates are *independent* and Gaussian. This is the starting point of this paper: since features typically encountered in applications are not independent, it is important to describe the behavior of the MLE under models with arbitrary covariance structures. In this work, we shall limit ourselves to Gaussian covariates although we believe our results extend to a wide class of distributions with sufficiently light tails (we provide numerical evidence supporting this claim).

To give a glimpse of our results, imagine we have n independent pairs of observations $(\boldsymbol{x}_i, y_i)$, where the features $\boldsymbol{x}_i \in \mathbb{R}^p$ and the class label $y_i \in \{-1, 1\}$. We assume that the $\boldsymbol{x}_i$'s follow a multivariate normal distribution with mean zero and arbitrary covariance, and that the likelihood of the class label y_i is related to $\boldsymbol{x}_i$ through the logistic model

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = 1/(1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}). \tag{1}$$

Denote the MLE for estimating the parameters $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and consider centering and scaling $\hat{\boldsymbol{\beta}}$ via

$$T_j = \frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j)}{\sigma_\star / \tau_j}; \tag{2}$$

here, τ_j is the standard deviation of the jth feature variable (the jth component of $\boldsymbol{x}_i$) conditional on all the other variables (all the other components of $\boldsymbol{x}_i$), whereas $\alpha_\star > 1$ and $\sigma_\star$ are numerical parameters we shall determine in Section 3.2. Then after establishing a stochastic representation for the MLE which is valid for every finite n and p, this paper proves two distinct asymptotic results (both hold in the same regime where n and p diverge to infinity in a fixed ratio).

The first concerns **marginals of the MLE**. Under some conditions on the magnitude of the regression coefficient β_j, we show that[1]

$$T_j \xrightarrow{\text{d}} \mathcal{N}(0,1), \tag{3}$$

and demonstrate an analogous statement for the distribution of any finite collection of coordinates. The meaning is clear; if the statistician is given several data sets from the model above and computes a given regression coefficient for each via ML, then the histogram of these coefficients across all data sets will look approximately Gaussian.

This state of affairs extends [14, Theorem 3] significantly, which established the joint distribution of a finite collection of *null* coordinates, in the setting of independent covariates. Specifically,

1. Eqn. (3) is true when covariates have arbitrary covariance structure.
2. Eqn. (3) holds for both null and non-null coordinates.

The second asymptotic result concerns **the empirical distribution of the MLE in a single data set/realization**: we prove that the empirical distribution of the T_j's converges to a standard normal in the sense that,

$$\frac{\#\{j : T_j \le t\}}{p} \xrightarrow{\text{P}} \mathbb{P}(\mathcal{N}(0,1) \le t). \tag{4}$$

This means that if we were to plot the histogram of all the T_j's obtained from a single data set, we would just see a bell curve. Another consequence is that for sufficiently nice functions $f(\cdot)$, we have

$$\frac{1}{p}\sum_{j=1}^{p} f(T_j) \xrightarrow{\text{P}} \mathbb{E}\left[f(Z)\right], \tag{5}$$

where $Z \sim \mathcal{N}(0,1)$. For instance, taking f to be the absolute value—we use the caveat that f is not uniformly bounded—we would conclude that

$$\frac{1}{p}\sum_{j=1}^{p} \sqrt{n}\,\tau_j |\hat{\beta}_j - \alpha_\star \beta_j| \xrightarrow{\text{P}} \sigma_\star \sqrt{2/\pi}.$$

Taking f to be the indicator function of the interval $[-1.96, 1.96]$, we would see that

$$\frac{1}{p}\sum_{j=1}^{p} 1\left\{-1.96 \le \frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j)}{\sigma_\star/\tau_j} \le 1.96\right\} \xrightarrow{\text{P}} 0.95.$$

Hence, the miscoverage rate (averaged over all variables) of the confidence intervals

$$[\hat{\beta}_j^-, \hat{\beta}_j^+], \qquad \hat{\beta}_j^{\pm} = \frac{\hat{\beta}_j \pm 1.96\sigma_\star/\sqrt{n}\tau_j}{\alpha_\star},$$

in a single experiment would approximately be equal to 5%.

Finally, this paper extends the LRT asymptotics to the case of arbitrary covariance.

[1] Throughout, $\xrightarrow{\text{d}}$ (resp. $\xrightarrow{\text{P}}$) is a shorthand for convergence in distribution (resp. probability).

2 A stochastic representation of the MLE

We consider a setting with n independent observations $(\boldsymbol{x}_i, y_i)_{i=1}^n$ such that the covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ follow a multivariate normal distribution $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, and the response $y_i \in \{-1, 1\}$ follows the logistic model (1) with regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. We assume that $\boldsymbol{\Sigma}$ has full column rank so that the model is identifiable. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ optimizes the log-likelihood function

$$\ell(\boldsymbol{b}; \boldsymbol{X}, \boldsymbol{y}) = \sum_{i=1}^{n} -\log(1 + \exp(-y_i \boldsymbol{x}_i^\top \boldsymbol{b})) \tag{6}$$

over all $\boldsymbol{b} \in \mathbb{R}^p$. (Here and below, $\boldsymbol{X}$ is the $n \times p$ matrix of covariates and $\boldsymbol{y}$ the $n \times 1$ vector of responses.)

2.1 From dependent to independent covariates

We begin our discussion by arguing that the aforementioned setting of dependent covariates can be translated to that of independent covariates. This follows from the invariance of the Gaussian distribution with respect to linear transformations.

Proposition 2.1. *Fix any matrix $\boldsymbol{L}$ obeying $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, and consider the vectors*

$$\hat{\boldsymbol{\theta}} := \boldsymbol{L}^\top \hat{\boldsymbol{\beta}} \qquad \text{and} \qquad \boldsymbol{\theta} := \boldsymbol{L}^\top \boldsymbol{\beta}. \tag{7}$$

Then $\hat{\boldsymbol{\theta}}$ is the MLE in a logistic model with regression coefficient $\boldsymbol{\theta}$ and covariates drawn i.i.d. from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$.

Proof. Because the likelihood (6) depends on the $\boldsymbol{x}_i$'s and $\boldsymbol{b}$ only through their inner product,

$$\ell(\boldsymbol{b}; \boldsymbol{X}, \boldsymbol{y}) = \ell(\boldsymbol{L}^\top \boldsymbol{b}; \boldsymbol{X}\boldsymbol{L}^{-\top}, \boldsymbol{y}) \tag{8}$$

for every $\boldsymbol{b} \in \mathbb{R}^p$. If $\hat{\boldsymbol{\beta}}$ is the MLE of the original model, then $\hat{\boldsymbol{\theta}} = \boldsymbol{L}^\top \hat{\boldsymbol{\beta}}$ is the MLE of a logistic model whose covariates are i.i.d. draws from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, and true regression coefficients given by $\boldsymbol{\theta} = \boldsymbol{L}^\top \boldsymbol{\beta}$. □

Proposition 2.1 has a major consequence—for an arbitrary variable j, we may choose $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ to be a Cholesky factorization of the covariance matrix, such that $\boldsymbol{x}_i$ can be expressed as

$$\underbrace{\begin{bmatrix} \star \\ \star \\ \star \\ x_{i,j} \end{bmatrix}}_{\boldsymbol{x}_i} = \underbrace{\begin{bmatrix} \star & & & \\ \star & \star & & \\ \star & \star & \star & \\ \star & \star & \star & \tau_j \end{bmatrix}}_{\boldsymbol{L}} \underbrace{\begin{bmatrix} \star \\ \star \\ \star \\ z_{i,j} \end{bmatrix}}_{\boldsymbol{z}_i}, \tag{9}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and $\tau_j^2 = \mathrm{Var}(x_{i,j} | \boldsymbol{x}_{i,-j})$. This can be seen from the triangular form:

$$\mathrm{Var}(x_{i,j} | \boldsymbol{x}_{i,-j}) = \mathrm{Var}(x_{i,j} | \boldsymbol{z}_{i,-j}) = \mathrm{Var}(\tau_j z_{i,j} | \boldsymbol{z}_{i,-j}) = \tau_j^2.$$

Then the equations in (7) tell us that

$$\hat{\theta}_j = \tau_j \hat{\beta}_j, \qquad \theta_j = \tau_j \beta_j, \tag{10}$$

and, therefore, for any pair (α, σ),

$$\tau_j \frac{\hat{\beta}_j - \alpha\beta_j}{\sigma} = \frac{\hat{\theta}_j - \alpha\theta_j}{\sigma}. \tag{11}$$

In particular, if we can find α and σ so that the RHS is approximately $\mathcal{N}(0,1)$, then (11) says that the LHS is approximately $\mathcal{N}(0,1)$ as well. We will use the equivalence (7) whenever possible, as to leverage as many results from [14] as possible.

2.2 A stochastic representation of the MLE

We work with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ in this section. The rotational invariance of the Gaussian distribution in this case yields a useful stochastic representation for the MLE $\hat{\boldsymbol{\theta}}$.

Lemma 2.1. *Let $\hat{\boldsymbol{\theta}}$ denote the MLE in a logistic model with regression vector $\boldsymbol{\theta}$ and covariates drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. Define the random variables*

$$\alpha(n) = \frac{\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} \rangle}{\|\boldsymbol{\theta}\|^2} \qquad \text{and} \qquad \sigma^2(n) = \|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|^2, \tag{12}$$

where $P_{\boldsymbol{\theta}^\perp}$ is the projection onto $\boldsymbol{\theta}^\perp$, which is the orthogonal complement of $\boldsymbol{\theta}$. Then

$$\frac{\hat{\boldsymbol{\theta}} - \alpha(n)\boldsymbol{\theta}}{\sigma(n)}$$

is uniformly distributed on the unit sphere lying in $\boldsymbol{\theta}^\perp$.

Proof. Notice that

$$\hat{\boldsymbol{\theta}} - \alpha(n)\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} - \langle \hat{\boldsymbol{\theta}}, \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \rangle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} = P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}},$$

the projection of $\hat{\boldsymbol{\theta}}$ onto the orthogonal complement of $\boldsymbol{\theta}$. We therefore need to show that for any orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ obeying $\boldsymbol{U}\boldsymbol{\theta} = \boldsymbol{\theta}$,

$$\frac{\boldsymbol{U} P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}}{\|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|} \overset{\mathrm{d}}{=} \frac{P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}}{\|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|}. \tag{13}$$

We know that

$$\hat{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}} + P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}} \implies \boldsymbol{U}\hat{\boldsymbol{\theta}} = \boldsymbol{U} P_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}} + \boldsymbol{U} P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}} + \boldsymbol{U} P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}, \tag{14}$$

where the last equality follows from the definition of $\boldsymbol{U}$. Now, $\boldsymbol{U}\hat{\boldsymbol{\theta}}$ is the MLE in a logistic model with covariates drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and regression vector $\boldsymbol{U}\boldsymbol{\theta} = \boldsymbol{\theta}$. Hence, $\boldsymbol{U}\hat{\boldsymbol{\theta}} \overset{\mathrm{d}}{=} \hat{\boldsymbol{\theta}}$ and (14) leads to

$$\boldsymbol{U} \frac{P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}}{\|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|} \overset{\mathrm{d}}{=} \frac{\hat{\boldsymbol{\theta}} - P_{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}}{\|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|} = \frac{P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}}{\|P_{\boldsymbol{\theta}^\perp} \hat{\boldsymbol{\theta}}\|}.$$

□

Note that, Lemma (2.1) provides an exact stochastic representation of $\hat{\boldsymbol{\theta}}$, which is valid for every choice of n and p.

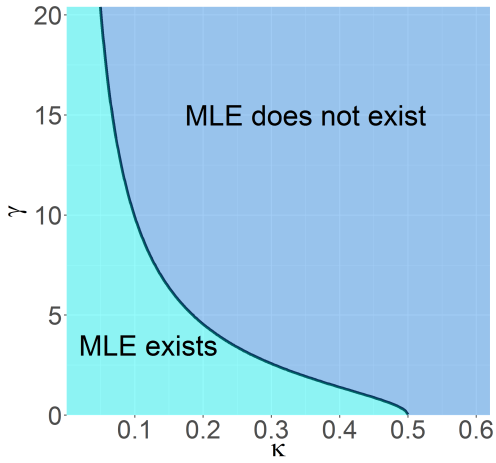


Figure 1: Boundary curve $\kappa \mapsto g_{\mathrm{MLE}}(\kappa)$ separating the regions where the MLE asymptotically exists and where it does not [2, Figure 1(a)].

3 The asymptotic distribution of the MLE in high dimensions

We now study the distribution of the MLE in the limit of a large number of variables and observations. We consider a sequence of logistic regression problems with n observations and $p(n)$ variables. In each problem instance, we have n independent observations $(\boldsymbol{x}_i, y_i)_{i=1}^n$ from a logistic model with covariates $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(n)), \boldsymbol{\Sigma}(n) \in \mathbb{R}^{p(n)\times p(n)}$, regression coefficients $\boldsymbol{\beta}(n)$, and response $y_i \in \{-1, 1\}$. As the sample size increases, we assume that the dimensionality $p(n)/n$ approaches a fixed limit in the sense that

$$p(n)/n \;\to\; \kappa \;> 0. \tag{15}$$

As in [14], we consider a scaling of the regression coefficients obeying

$$\mathrm{Var}(\boldsymbol{x}_i^\top \boldsymbol{\beta}(n)) = \boldsymbol{\beta}(n)^\top \boldsymbol{\Sigma}(n) \boldsymbol{\beta}(n) \;\to\; \gamma^2 < \infty. \tag{16}$$

This scaling keeps the "signal-to-noise-ratio" fixed. The larger γ, the easier it becomes to classify the observations. (If the parameter γ were allowed to diverge to infinity, we would have a noiseless problem in which we could correctly classify essentially all the observations.)

In the remainder of this paper, we will drop n from expressions such as $p(n)$, $\boldsymbol{\beta}(n)$, and $\boldsymbol{\Sigma}(n)$ to simplify the notation. We shall however remember that the number of variables p grows in proportion to the sample size n.

3.1 Existence of the MLE

An important issue in logistic regression is that the MLE does not always exist. In fact, the MLE exists if and only if the cases (the points $\boldsymbol{x}_i$ for which $y_i = 1$) and controls (those for which $y_i = -1$) cannot be linearly separated; linear separation here means that there is a hyperplane such that all the cases are on one side of the plane and all the controls on the other.

When both n and p are large, whether such a separating hyperplane exists depends only on the dimensionality κ and the overall signal strength γ. In the asymptotic setting described above, [2,

Theorem 1] demonstrated a phase transition phenomenon: there is a curve $\gamma = g_{\mathrm{MLE}}(\kappa)$ in the (κ, γ) plane that separates the region where the MLE exists from that where it does not, see Figure 1. Formally,

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

It is noteworthy that the phase-transition diagram only depends on whether $\gamma > g_{\mathrm{MLE}}(\kappa)$ and, therefore, does not depend on the details of the covariance $\boldsymbol{\Sigma}$ of the covariates. Since we are interested in the distribution of the MLE, we shall consider values of the dimensionality parameter κ and signal strength γ in the light blue region from Figure 1.

3.2 Finite-dimensional marginals of the MLE

We begin by establishing the asymptotic behavior of the random variables $\alpha(n)$ and $\sigma(n)$, introduced in (12). These limits will play a key role in the distribution of the MLE.

Lemma 3.1. *Consider a sequence of logistic models with covariates drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and regression vectors $\boldsymbol{\theta}$ satisfying $\lim_{n\to\infty} \|\boldsymbol{\theta}\|^2 \to \gamma^2$. Let $\hat{\boldsymbol{\theta}}$ be the MLE and define $\alpha(n)$ and $\sigma(n)$ as in (12). Then, if (κ, γ) lies in the region where the MLE exists asymptotically, we have that*

$$\alpha(n) \xrightarrow{\text{a.s.}} \alpha_\star \qquad \text{and} \qquad \sigma(n)^2 \xrightarrow{\text{a.s.}} \kappa\sigma_\star^2, \tag{17}$$

where $\alpha_\star$ and $\sigma_\star$ are numerical constants that only depend on κ and γ.

We defer the proof to Appendix A.1; here, we explain where the parameters $(\alpha_\star, \sigma_\star)$ come from. Along with an additional parameter $\lambda_\star$, the triple $(\alpha_\star, \sigma_\star, \lambda_\star)$ is the unique solution to the system of equations parameterized by (κ, γ) in three variables $(\alpha, \sigma, \lambda)$ given by

$$\begin{cases} \sigma^2 & = \dfrac{1}{\kappa^2}\mathbb{E}\left[2\rho'(Q_1)(\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2)))^2\right] \\ 0 & = \mathbb{E}\left[\rho'(Q_1)Q_1\lambda\rho'(\mathsf{prox}_{\lambda\rho}(Q_2))\right] \\ 1-\kappa & = \mathbb{E}\left[\dfrac{2\rho'(Q_1)}{1+\lambda\rho''(\mathsf{prox}_{\lambda\rho}(Q_2))}\right], \end{cases} \tag{18}$$

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

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

Above, the proximal operator is defined as

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

where $\rho(t) = \log(1+e^t)$ and Z is a standard Gaussian variable. This system of equations can be rigorously derived from the generalized approximate message passing algorithm [5], or by analyzing

the auxiliary optimization problem [10]. They can also be heuristically understood with an argument similar to that in [4]; we defer to [14] for a complete discussion. The important point here is that in the region where the MLE exists, the system (18) has a unique solution.

We are now in a position to describe the asymptotic behavior of the MLE. The proof is deferred to Appendix A.2.

Theorem 3.1. *Consider a logistic model with covariates $\boldsymbol{x}_i$ drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(n))$ and assume we are in the (κ, γ) region where the MLE exists asymptotically. Then for every coordinate whose regression coefficient satisfies $\sqrt{n}\tau_j\beta_j = O(1)$,*

$$\frac{\sqrt{n}(\hat{\beta}_j - \alpha_\star \beta_j)}{\sigma_\star/\tau_j} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{19}$$

Above $\tau_j^2 = \mathrm{Var}(x_{i,j} \,|\, \boldsymbol{x}_{i,-j})$ is the conditional variance of $x_{i,j}$ given all the other covariates. More generally, for any sequence of deterministic unit normed vectors $\boldsymbol{v}(n)$ with $\sqrt{n}\tau(\boldsymbol{v})\boldsymbol{v}(n)^\top \boldsymbol{\beta}(n) = O(1)$, we have that

$$\frac{\sqrt{n}\boldsymbol{v}^\top(\hat{\boldsymbol{\beta}} - \alpha_\star \boldsymbol{\beta})}{\sigma_\star/\tau(\boldsymbol{v})} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{20}$$

Here $\tau(\boldsymbol{v})$ is given by

$$\tau^2(\boldsymbol{v}) = \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{x}_i | P_{\boldsymbol{v}^\perp} \boldsymbol{x}_i) = \left(\boldsymbol{v}^\top \boldsymbol{\Theta}(n) \boldsymbol{v}\right)^{-1},$$

where $\boldsymbol{\Theta}(n)$ equals the precision matrix $\boldsymbol{\Sigma}(n)^{-1}$. A consequence is this: consider a finite set of coordinates $\mathcal{S} \subset \{1, \dots, p\}$ obeying $\sqrt{n}(\boldsymbol{\beta}_{\mathcal{S}}^\top \boldsymbol{\Theta}_{\mathcal{S}}^{-1} \boldsymbol{\beta}_{\mathcal{S}})^{\frac{1}{2}} = O(1)$. Then

$$\frac{\sqrt{n}\boldsymbol{\Theta}_{\mathcal{S}}^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \alpha_\star \boldsymbol{\beta}_{\mathcal{S}})}{\sigma_\star} \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{|\mathcal{S}|}). \tag{21}$$

Above $\boldsymbol{\beta}_{\mathcal{S}}$ is the slice of $\boldsymbol{\beta}$ with entries in $\mathcal{S}$ and, similarly, $\boldsymbol{\Theta}_{\mathcal{S}}$ is the slice of the precision matrix $\boldsymbol{\Theta}$ with rows and columns in $\mathcal{S}$.

Returning to the Introduction, we now see that the behavior of the MLE is different from that implied by the classical textbook result, which states that

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

We also see that Theorem 3.1 extends [14, Theorem 3] in multiple directions. Indeed, this prior work assumed standardized and independent covariates (i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$)—implying that $\tau_j^2 = 1$—and established $\sqrt{n}\hat{\beta}_j/\sigma_\star \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ only in the special case $\beta_j = 0$.

Finite sample accuracy We study the finite sample accuracy of Theorem 3.1 through numerical examples. We consider an experiment with a fixed number of observations set to $n = 4,000$ and a number of variables set to $p = 800$ so that $\kappa = 0.2$. We set the signal strength to $\gamma^2 = 5$. (For this problem size, the asymptotic result for null variables has been observed to be very accurate when the covariates are independent [14].) We sample the covariates such that the correlation matrix comes from an AR(1) model with parameter $\rho = 0.5$. We then randomly sample half of the coefficients to be non-nulls, with equal and positive magnitudes, chosen to attain the desired

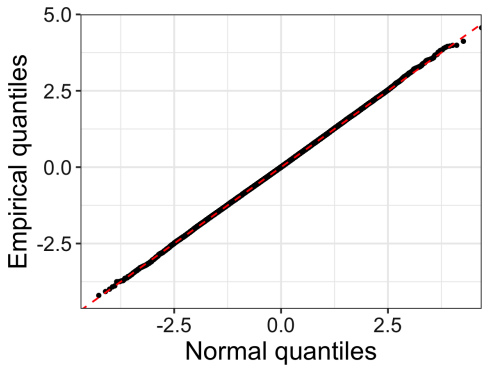


Figure 2: Comparison of quantiles. Quantiles of the empirical distribution of an appropriately centered and scaled MLE coordinate versus standard normal quantiles.

signal strength $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 5$. For a given non-null coordinate β_j, we calculate the centered and scaled MLE T_j (2), and repeat the experiment $B = 100,000$ times. Figure 2 shows a qqplot of the empirical distribution of T_j versus the standard normal distribution. Observe that the quantiles align perfectly, demonstrating the accuracy of (19).

We further examine the empirical accuracy of (19) through the lens of confidence intervals and finite sample coverage. Theorem 3.1 suggests that if $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$th quantile of a standard normal variable, β_j should lie within the interval

$$\left[\frac{1}{\alpha_\star} \left(\hat{\beta}_j - \frac{\sigma_\star}{\sqrt{n}\tau_j} z_{(1-\alpha/2)} \right), \frac{1}{\alpha_\star} \left(\hat{\beta}_j + \frac{\sigma_\star}{\sqrt{n}\tau_j} z_{(1-\alpha/2)} \right) \right] \tag{22}$$

about $(1-\alpha)B$ times. Table 1 shows the proportion of experiments in which β_j is covered by (22) for different choices of the confidence level $(1-\alpha)$, along with the respective standard errors. For every confidence level, the empirical coverage proportion lies extremely close to the corresponding target.

Nominal coverage $100(1-\alpha)$	99	98	95	90	80
Empirical coverage	98.97	97.96	94.99	89.88	79.88
Standard error	0.03	0.04	0.07	0.10	0.13

Table 1: Coverage proportion of a single variable. Each cell reports the proportion of times β_j falls within (22), calculated over $B = 100,000$ repetitions; the standard errors are provided as well.

3.3 The empirical distribution of all MLE coordinates (single experiment)

The preceding section tells us about the finite-dimensional marginals of the MLE, e.g. about the distribution of a given coordinate $\hat{\beta}_j$ when we repeat experiments. We now turn to a different type of asymptotics and characterize the limiting empirical distribution of *all* the coordinates calculated from a single experiment.

Theorem 3.2. *Let $c(n) = \lambda_{\max}(\mathbf{\Sigma}(n))/\lambda_{\min}(\mathbf{\Sigma}(n))$ be the condition number of $\mathbf{\Sigma}(n)$. Assume that $\limsup_{n\to\infty} c(n) < \infty$, and that (κ, γ) lies in the region where the MLE exists asymptotically. Then the empirical cumulative distribution function of the rescaled MLE (2) converges pointwise to that of a standard normal distribution, namely, for each $t \in \mathbb{R}$,*

$$\frac{1}{p}\sum_{i=1}^{p} \mathbb{I}\{T_j \le t\} \xrightarrow{\text{P}} \Phi(t). \tag{23}$$

The proof is in Appendix A.3. As explained in the Introduction, this says that empirical averages of functionals of the marginals have a limit (5). By Corollary A.2, this implies that the empirical distribution of $\boldsymbol{T}$ converges weakly to a standard Gaussian, in probability. In the remainder of this section, we study the empirical accuracy of Theorem 3.2 in finite samples through some simulated examples.

Finite sample accuracy We consider an experiment with dimensions n and p the same as that for Figure 2, and the regression vector sampled similarly. According to (23), about $(1-\alpha)$ of all the β_j's should lie within the corresponding intervals (22). Table 2 shows the proportion of β_j's covered by these intervals for a few commonly used confidence levels $(1-\alpha)$.[2] The proportions are as predicted for each of the confidence levels, and every covariance we simulated from.

The four columns in Table 2 correspond to different covariance matrices; they include a random correlation matrix (details are given below), a correlation matrix from an AR(1) model whose parameter is either set to $\rho = 0.8$ or $\rho = 0.5$, and a covariance matrix set to be the identity. The random correlation matrix is sampled as follows: we randomly pick an orthogonal matrix $\boldsymbol{U}$, and eigenvalues $\lambda_1, \ldots, \lambda_p$ i.i.d. from a chi-squared distribution with 10 degrees of freedom. We then form a positive definite matrix $\boldsymbol{B} = \boldsymbol{U}^\top \mathbf{\Lambda} \boldsymbol{U}$ from these eigenvalues, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$. $\mathbf{\Sigma}$ is the correlation matrix obtained from $\boldsymbol{B}$ by scaling the variables to have unit variance.

Nominal coverage $100(1-\alpha)$	Random	$\rho = 0.8$	$\rho = 0.5$	Identity
99	99.178 (0.002)	99.195 (0.002)	99.187 (0.002)	99.175 (0.002)
98	97.873 (0.003)	97.908 (0.003)	97.890 (0.003)	97.865 (0.003)
95	94.826 (0.005)	94.884 (0.005)	94.857 (0.005)	94.811 (0.005)
90	89.798 (0.007)	89.883 (0.007)	89.847 (0.007)	89.780 (0.007)
80	79.784 (0.009)	79.896 (0.009)	79.837 (0.009)	79.751 (0.009)

Table 2: Each cell reports the proportion of *all* the variables in each run falling within the corresponding intervals from (22), averaged over $B = 100,000$ repetitions; the standard deviation is given between parentheses.

4 The distribution of the LRT

Lastly, we study the distribution of the log-likelihood ratio (LLR) test statistic

$$\text{LLR}_j = \max_{\boldsymbol{b}} \ell(\boldsymbol{b}; \boldsymbol{X}, \boldsymbol{y}) - \max_{\boldsymbol{b}:\, b_j = 0} \ell(\boldsymbol{b}; \boldsymbol{X}, \boldsymbol{y}), \tag{24}$$

[2]Note that, in Table 1 we investigated a single coordinate across many replicates, whereas here we consider all the coordinates in each instance.

which is routinely used to test whether the jth variable is in the model or not; i. e. whether $\beta_j = 0$ or not.

Theorem 4.1. *Assume that we are in the (κ, γ) region where the MLE exists asymptotically. Then under the null (i.e. $\beta_j = 0$),*

$$2\,\mathrm{LLR}_j \xrightarrow{\mathrm{d}} \frac{\kappa\sigma_\star^2}{\lambda_\star}\chi_1^2.$$

Further, for every finite ℓ, twice the LLR for testing ℓ null hypotheses $\beta_{j_1} = \cdots = \beta_{j_\ell} = 0$ is asymptotically distributed as $(\kappa\sigma_\star^2/\lambda_\star)\ \chi_\ell^2$.

Invoking results from Section 2.1, we will show that this is a rather straightforward extension of [14, Theorem 4], which deals with independent covariates, see also [15, Theorem 1]. Choosing $\boldsymbol{L}$ to be the same as in (9) and setting $\boldsymbol{b}' = \boldsymbol{L}^\top \boldsymbol{b}$ tell us that $b'_j = 0$ if and only if $b_j = 0$. Hence, $\mathrm{LLR_j}$ reduces to

$$\mathrm{LLR}_j = \max_{\boldsymbol{b}} \ell(\boldsymbol{L}^\top \boldsymbol{b}; \boldsymbol{X}\boldsymbol{L}^{-\top}, \boldsymbol{y}) - \max_{\boldsymbol{b}: b_j = 0} \ell(\boldsymbol{L}^\top \boldsymbol{b}; \boldsymbol{X}\boldsymbol{L}^{-\top}, \boldsymbol{y}) \tag{25}$$

$$= \max_{\boldsymbol{b}'} \ell(\boldsymbol{b}'; \boldsymbol{X}\boldsymbol{L}^{-\top}, \boldsymbol{y}) - \max_{\boldsymbol{b}': b'_j = 0} \ell(\boldsymbol{b}'; \boldsymbol{X}\boldsymbol{L}^{-\top}, \boldsymbol{y}), \tag{26}$$

which is the log-likelihood ratio statistic in a model with covariates drawn i.i.d. from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and regression coefficient given by $\boldsymbol{\theta} = \boldsymbol{L}^\top \boldsymbol{\beta}$. This in turn satisfies $\theta_j = 0$ if and only if $\beta_j = 0$ so that we can think of the LLR above as testing $\theta_j = 0$. Consequently, the asymptotic distribution is the same as that given in [14, Theorem 4] with $\gamma^2 = \lim_{n\to\infty} \|\boldsymbol{\theta}\|^2 = \lim_{n\to\infty} \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$. The equality of the likelihood ratios implies that to study the finite sample accuracy of Theorem 4.1, we may just as well assume we have independent covariates; hence, we refer the readers to [14] for empirical results detailing the quality of the rescaled chi-square approximation in finite samples.

5 Accuracy with estimated parameters

In practice, the signal strength γ^2 and conditional variance τ_j^2 are typically not known a priori. In this section, we plug in estimates of these quantities and investigate their empirical performance. We focus on testing a null variable and constructing confidence intervals.

The parameters are the same as in Section 3.2. In brief, we set $n = 4,000$, $p = 800$ (so that $\kappa = 0.2$), and $\gamma^2 = 5$. The covariates follow an AR(1) model with $\rho = 0.5$ and $\Sigma_{jj} = 1$.

5.1 Estimating parameters

We here explain how to estimate the signal strength γ^2 and conditional variance τ_j^2 needed to describe the distribution of the LLR and MLE.

To estimate the signal strength, we use the *ProbeFrontier* method introduced in [14]. As we have seen in Section 3.1, for each γ, there is a corresponding problem dimension $\kappa(\gamma)$ on the phase transition curve, see Figure 1: once $\kappa > \kappa(\gamma)$, the MLE no longer exists asymptotically [2]. The *ProbeFrontier* method searches for the smallest κ such that the MLE ceases to exist by sub-sampling observations. Once we obtain $\hat{\gamma}$, we set $(\hat{\alpha}, \hat{\sigma}, \hat{\lambda})$ to be the solution to the system of equations with parameters $(\kappa, \hat{\gamma})$. Because the *ProbeFrontier* method only checks whether the points are separable,

the quality of the estimate $\hat{\gamma}$ does not depend upon whether the covariates are independent or not. We therefore expect good performance across the board.

As to the conditional variance, since the covariates are Gaussian, it can be estimated by a simple linear regression. Let $\boldsymbol{X}_{\bullet,-j}$ be the data matrix without the jth column, and consider the residual sum of squares RSS_j obtained by regressing the jth column $\boldsymbol{X}_{\bullet,j}$ onto $\boldsymbol{X}_{\bullet,-j}$. Then

$$\mathrm{RSS}_j \ \sim \tau_j^2 \ \chi^2_{n-p+1}.$$

Hence,

$$\hat{\tau}_j^2 = \frac{\mathrm{RSS}_j/n}{1-\kappa} \tag{27}$$

is nearly unbiased for τ_j^2.[3]

In our example, the covariates follow an AR(1) model and there is a natural estimate of ρ by maximum likelihood. This yields an estimated covariance matrix $\hat{\boldsymbol{\Sigma}}(\hat{\rho})$ parameterized by $\hat{\rho}$, which we then use to estimate the conditional variance $\hat{\tau}_j^2(\hat{\rho})$. Below, we use both the nonparametric estimates $\hat{\tau}_j$ and parametric estimates $\hat{\tau}_j(\hat{\rho})$.

5.2 Empirical performance of a t-test

Imagine we want to use Theorem 3.1 to calibrate a test to decide whether $\beta_j = 0$ or not. After plugging in estimated parameters, a p-value for a two-sided test takes the form

$$\hat{p}_j = 2\bar{\Phi}(\sqrt{n}\hat{\tau}_j|\hat{\beta}_j|/\hat{\sigma}), \tag{28}$$

where $\bar{\Phi}(t) = \mathbb{P}(\mathcal{N}(0,1) > t)$. In Table 3, we report the proportion of p-values calculated from (28), below some common cutoffs. To control type-I errors, the proportion of p-values below 10% should be at most about 10% and similarly for any other level. The p-values computed from true parameters show a correct behavior, as expected. If we use estimated parameters, the p-values are also accurate and are as good as those obtained from true parameters. In comparison, p-values from classical theory are far from correct, as shown in Column 4.

	1 $(\hat{\tau}, \hat{\sigma})$	2 $(\hat{\tau}(\hat{\rho}), \hat{\sigma})$	3 $(\tau, \sigma_\star)$	4 Classical
$\mathbb{P}(\text{P-value} \le 10\%)$	10.09% (0.30%)	10.14% (0.30%)	10.22% (0.30%)	17.80% (0.38%)
$\mathbb{P}(\text{P-value} \le 5\%)$	5.20% (0.22%)	5.23% (0.22%)	5.24% (0.22%)	10.73% (0.31%)
$\mathbb{P}(\text{P-value} \le 1\%)$	1.16% (0.11%)	1.22% (0.11%)	1.33% (0.11%)	3.72% (0.19%)
$\mathbb{P}(\text{P-value} \le 0.5\%)$	0.68% (0.08%)	0.70% (0.08%)	0.74% (0.08%)	2.43% (0.15%)

Table 3: Empirical performance of a t-test from (28) Each cell reports the p-value probability and its standard error (in parentheses) estimated over $B = 10,000$ repetitions. The first two columns use *ProbeFrontier* to estimate the problem parameter $\hat{\sigma}$, and the two estimates of conditional variance from Section 5.1. The third column assumes knowledge of the signal-to-noise parameter γ. The last column uses the normal approximation from R.

[3] We also have $\mathrm{RSS}_j = 1/\Theta_{jj}$, $\boldsymbol{\Theta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

5.3 Coverage proportion

We proceed to check whether the confidence intervals constructed from the estimated parameters

$$\left[\frac{1}{\hat{\alpha}}\left(\hat{\beta}_j - \frac{\hat{\sigma}}{\sqrt{n}\hat{\tau}_j} z_{(1-\alpha/2)}\right), \frac{1}{\hat{\alpha}}\left(\hat{\beta}_j + \frac{\hat{\sigma}}{\sqrt{n}\hat{\tau}_j} z_{(1-\alpha/2)}\right)\right] \tag{29}$$

achieve the desired coverage property.

Nominal coverage $100(1-\alpha)$	1 $(\hat{\tau}, \hat{\sigma})$	2 $(\hat{\tau}(\hat{\rho}), \hat{\sigma})$	3 $(\tau, \sigma_\star)$
99.5	99.32 (0.08)	99.30 (0.08)	99.26 (0.09)
99	98.84 (0.11)	98.78 (0.11)	98.67 (0.11)
95	94.80 (0.22)	94.77 (0.22)	94.76 (0.22)
90	89.91 (0.30)	89.86 (0.30)	89.78 (0.30)

Table 4: Coverage proportion of a single variable. Each cell reports the proportion of times a variable β_j is covered by the corresponding confidence interval from (29), calculated over $B = 10,000$ repetitions; we chose the variable to be the same null coordinate as in Section 5.2. The standard errors are given between parentheses. The first two columns use estimated parameters, and the last one uses the true parameters.

Nominal coverage $100(1-\alpha)$	1 $(\hat{\tau}, \hat{\sigma})$	2 $(\hat{\tau}(\hat{\rho}), \hat{\sigma})$	3 $(\tau, \sigma_\star)$
98	97.96 (0.01)	97.95 (0.01)	97.85 (0.01)
95	95.01 (0.01)	95.00 (0.01)	94.85 (0.02)
90	89.92 (0.02)	89.91 (0.02)	89.72 (0.02)
80	79.99 (0.02)	79.99 (0.02)	79.77 (0.03)

Table 5: Proportion of variables inside the confidence intervals (29) Each cell reports the proportion of *all* the variables in each run falling within the corresponding confidence intervals from (29), averaged over $B = 10,000$ repetitions (standard errors in parentheses). The first two columns use estimated parameters, and the last one uses the true parameters.

We first test this in the context of Theorem 3.1, in particular (19). Table 4 reports the proportion of times a single coordinate lies in the corresponding confidence interval from (29). We observe that the coverage proportions are close to the respective targets, even with the estimated parameters.

Moving on, we study the accuracy of the estimated parameters in light of Theorem 3.2. This differs from our previous calculation: Table 4 focuses on whether a single coordinate is covered, but now we compute the proportion of *all* the $p = 800$ variables falling within the respective confidence intervals from (29), in each single experiment. We report the mean (Table 5) of these proportions, computed across 10,000 repetitions. Ideally, the proportion should be about the nominal coverage and this is what we observe.

5.4 Empirical performance of the LRT

Lastly, we examine p-values for the LRT when the signal strength γ^2 is unknown. The p-values take the form

$$\hat{p}_j = \mathbb{P}\left(\chi_1^2 \geq \frac{\hat{\lambda}}{\kappa\hat{\sigma}^2}\, 2\text{LLR}\right) \tag{30}$$

once we plug in estimated values for $\lambda_\star$ and $\sigma_\star$. Table 6 displays the proportion of p-values below some common cutoffs for the same null coordinate as in Table 3. Again, classical theory yields a gross inflation of the proportion of p-values in the lower tail. In contrast, p-values from either estimated or true parameters display the correct behavior.

	Estimated	True	Classical
$\mathbb{P}(\text{P-value} \leq 10\%)$	10.04% (0.30%)	10.06% (0.30%)	17.86% (0.38%)
$\mathbb{P}(\text{P-value} \leq 5\%)$	5.19% (0.22%)	5.25% (0.22%)	10.76% (0.31%)
$\mathbb{P}(\text{P-value} \leq 1\%)$	1.17 % (0.11%)	1.18% (0.11%)	3.75% (0.19%)
$\mathbb{P}(\text{P-value} \leq 0.5\%)$	0.68% (0.08%)	0.69% (0.08%)	2.49% (0.15%)

Table 6: Empirical performance of the LRT Each cell reports the p-value probability and its standard error (in parentheses) estimated over $B = 10,000$ repetitions. The first column uses *ProbeFrontier* estimated factor $\hat{\lambda}/\kappa\hat{\sigma}^2$ whereas the second uses $\lambda_\star/\kappa\sigma_\star^2$. The last column displays the results from classical theory.

6 A sub-Gaussian example

Our model assumes that the covariates arise from a multivariate normal distribution. As in [14, Section 4.g], however, we expect that our results apply to a broad class of covariate distributions, in particular, when they have sufficiently light tails. To test this, we consider a logistic regression problem with covariates drawn from a sub-Gaussian distribution that is inspired by genetic studies, and examine the accuracy of null p-values and confidence intervals proposed in this paper.

Since the signal strength γ^2 and conditional variances τ_j^2 are unknown in practice, we use throughout the *ProbeFrontier* method[4] and (27) to obtain accurate estimates.

6.1 Model setting

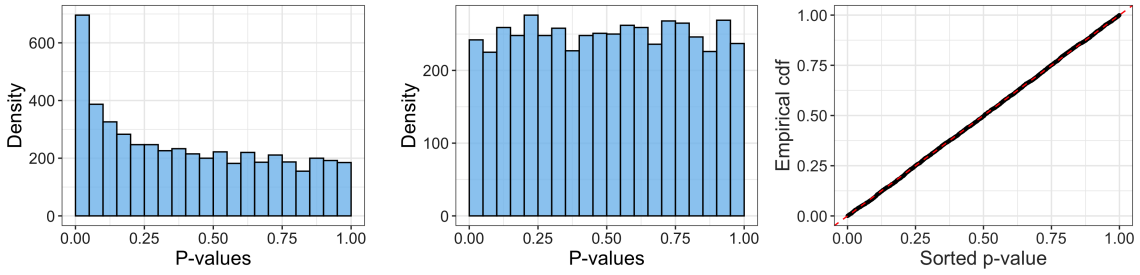
In genome-wide association studies (GWAS), one often wishes to determine how a binary response Y depends on single nucleotide polymorphisms (SNPs); here, each sample of the covariates measures the genotype of a collection of SNPs, and typically takes on values in $\{0, 1, 2\}^p$. Because neighboring SNPs are usually correlated, GWAS inspired datasets form an excellent platform for testing our theory. Hidden Markov Models (HMMs) are a broad class of distributions that have been widely used to characterize the behavior of SNPs [6, 8, 9, 11]. Here, we study the applicability of our theory when the covariates are sampled from a class of HMMs, and consider the specific model implemented in the fastPHASE software (see [9, Section 5] for details) that can be parametrized by three vectors $(\boldsymbol{r}, \boldsymbol{\eta}, \boldsymbol{\theta})$. We generate $n = 5000$ independent observations $(\boldsymbol{X}_i, y_i)_{1 \leq i \leq n}$ by first

[4] Here, we resample 10 times for each κ.

sampling $\boldsymbol{X}_i$ from an HMM with parameters $\boldsymbol{r} = \boldsymbol{r}_0, \boldsymbol{\eta} = \boldsymbol{\eta}_0, \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $p = 1454$, so that $\kappa = 0.29$, and then sampling $y_i \sim \text{Ber}(\sigma(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$. The `SNPknock` package [12] was used for sampling the covariates and the parameter values are available at https://github.com/zq00/logisticMLE. We then standardize the design matrix so that each column has zero mean and unit norm. The regression coefficients are obtained as follows: we randomly pick 100 coordinates to be i.i.d. draws from a mean zero normal distribution with standard deviation 10, and the remaining coordinates vanish. We repeat this experiment $B = 5000$ times.

6.2 Accuracy of null p-values

We focus on a single null coordinate and, across the B replicates, calculate p-values based on four test statistics—(a) the classical t-test, which yields the p-value formula $2\bar{\Phi}(\sqrt{n}|\hat{\beta}_j|/\hat{\sigma}_j)$; here $\hat{\sigma}_j$ is taken to be the estimate of the standard error from `R`, (b) the classical LRT, (c) the t-test suggested by Theorem 3.1; in this case, the formula is the same as in (a), except that $\hat{\sigma}_j = \hat{\sigma}/\hat{\tau}_j$, where $\hat{\sigma}$ is estimated from *ProbeFrontier* and $\hat{\tau}_j$ from (27), and finally, (d) the LRT based on Theorem 4.1; here again, the rescaling constant is specified via the estimates $\hat{\sigma}$, $\hat{\lambda}$ produced by *ProbeFrontier*. The histograms of the classical p-values are shown in Figures 3a and 4a—these are far from the uniform distribution, with severe inflation near the lower tail. The histograms of the two sets of p-values based on our theory are displayed in Figures 3b and 4b, whereas the corresponding empirical cdfs can be seen in Figures 3c and 4c. In both of these cases, we observe a remarkable proximity to the uniform distribution. Furthermore, Table 7 reports the proportion of null p-values below a collection of thresholds; both the t-test and the LRT suggested by our results provide accurate control of the type-I error. These empirical observations indicate that our theory likely applies to a much broader class of non-Gaussian distributions.

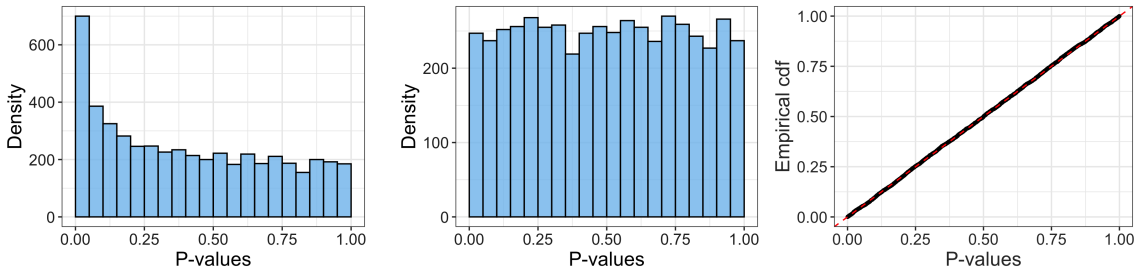


(a) **(b)** **(c)**

Figure 3: Distribution of null p-values from a two-sided t-test Histograms of p-values are calculated by $p_j = 2\bar{\Phi}(\sqrt{n}|\hat{\beta}_j|/\hat{\sigma}_j)$. (a) $\hat{\sigma}_j$ is taken to be the standard error from R. (b) $\hat{\sigma}_j = \hat{\sigma}/\hat{\tau}_j$, where $\hat{\sigma}$ is estimated by *ProbeFrontier* and $\hat{\tau}_j$ is from (27) (c) Empirical cdf of the p-values in (b).

6.3 Coverage proportion

We proceed to check the accuracy of the confidence intervals described by (29). We consider a single coordinate β_j (we chose $\beta_j \neq 0$) and report the proportion of times (29) covers β_j across the B repetitions (Table 8). At each level, the empirical coverage proportion agrees with the desired target level, validating the marginal distribution (19) in non-Gaussian settings. To investigate the



(a) **(b)** **(c)**

Figure 4: Distribution of null p-values calculated from the LRT (a) Histogram of p-values based on the chi-squared distribution (with 1 degree of freedom). (b) Histogram of p-values based on the re-scaled chi-squared distribution; the re-scaling factor is estimated by *ProbeFrontier*. (c) Empirical cdf of p-values from (b).

	t-test	LRT
$\mathbb{P}(\text{P-value} \leq 10\%)$	9.34% (0.41%)	9.68% (0.41%)
$\mathbb{P}(\text{P-value} \leq 5\%)$	4.84% (0.30%)	4.94% (0.30%)
$\mathbb{P}(\text{P-value} \leq 1\%)$	0.96% (0.14%)	0.94% (0.14%)
$\mathbb{P}(\text{P-value} \leq 0.1\%)$	0.08% (0.04%)	0.08% (0.04%)

Table 7: Empirical performance of testing a null Each cell reports the p-value probability and its standard error (in parentheses) estimated over $B = 5,000$ repetitions. The p-values are calculated from a two sided t-test (as in Figure 3c) and the LRT (as in Figure 4c).

efficacy of (19) further, we calculate the standardized versions of the MLE given by

$$\hat{T}_j = \frac{\sqrt{n}(\hat{\beta}_j - \hat{\alpha}\beta_j)}{\hat{\sigma}/\hat{\tau}_j} \tag{31}$$

for each run of the experiment; recall that the estimates $\hat{\alpha}, \hat{\sigma}, \hat{\tau}_j$ arise from the *ProbeFrontier* method and (27). Figure 5 displays a qqplot of the empirical quantiles of $\hat{T}_j$ versus the standard normal quantiles, and once again, we observe a remarkable agreement.

Nominal coverage $100(1-\alpha)$	99	98	95	90	80
Empirical coverage	99.04	97.98	94.98	89.9	80.88
Standard error	0.2	0.2	0.3	0.4	0.6

Table 8: Coverage proportion of a single variable. Each cell reports the proportion of times β_j falls within (29), calculated over $B = 5,000$ repetitions; the standard errors are provided as well. The unknown signal strength γ^2 is estimated by *ProbeFrontier*.

Finally, we turn to study the performance of Theorem 3.2. In each experiment, we compute the proportion of variables covered by the corresponding intervals in (29) and report the mean (Table 9) across the 5000 replicates. Once again, the average coverages remained close to the

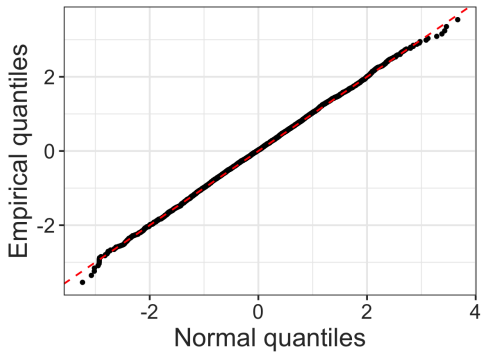


Figure 5: Comparison of quantiles. Quantiles of the empirical distribution of the MLE coordinate from Table 8, standardized as in (31), versus standard normal quantiles.

desired thresholds for all the levels considered, demonstrating the applicability of the bulk result (23) beyond the setting of Gaussian covariates.

Nominal coverage $100(1-\alpha)$	98	95	90	80
Empirical coverage	98.03	95.14	90.1	80.2
Standard error	0.01	0.01	0.02	0.02

Table 9: Proportion of variables inside the confidence intervals (29) Each cell reports the average coverage estimated over $B = 5,000$ repetitions. The standard errors are shown as well.

7 Is this all real?

We have seen that in logistic models with Gaussian covariates of moderately high dimensions, (a) the MLE overestimates the true effect magnitudes, (b) the classical Fisher information formula underestimates the true variability of the ML coefficients, and (c) classical ML based null p-values are far from uniform. We introduced a new maximum likelihood theory, which accurately amends all of these issues and demonstrated empirical accuracy on non-Gaussian light-tailed covariate distributions. We claim that the issues with ML theory apply to a broader class of covariate distributions; in fact, we expect to see similar *qualitative* phenomena in real datasets.

Consider the wine quality data [18], which contains 4898 white wine samples from northern Portugal. The dataset consists of 11 numerical variables from physico-chemical tests measuring various characteristics of the wine, such as density, pH and volatile acidity, while the response records a wine quality score that takes on values in $\{0, \ldots, 10\}$. We define a binary response by thresholding the scores, so that a wine receives a label $y = 0$, if the corresponding score is below 6, and a label $y = 1$, otherwise. We log-transform two of the explanatory variables as to make their distribution more symmetrical and concentrated. We also center the variables so that each has mean zero.

We explore the behavior of the classical logistic MLE for the variable "volatile acidity" (va) at

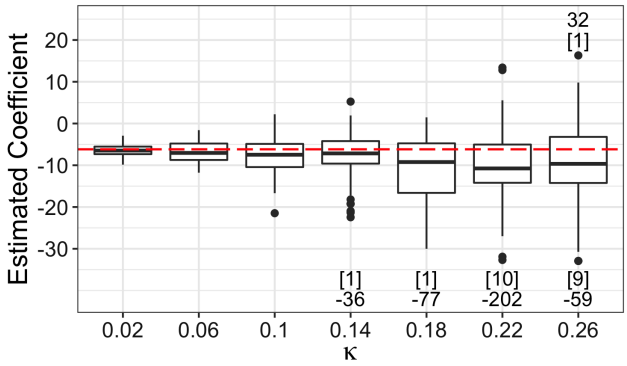


Figure 6: Estimated coefficient $\hat{\beta}_{\text{va}}$ of the variable "volatile acidity", obtained from $B = 100$ sub-samples of size p/κ. The red dashed line shows the MLE using all the observations. The numbers of outliers outside of range are those between squared brackets. The minimum/maximum value of these outliers is given by the accompanying integer.

a grid of values of the problem dimension $\kappa \in K$. For each $\kappa \in K$, we construct $B = 100$ sub-samples containing $n = p/\kappa$ observations and calculate the MLE $\hat{\beta}_{\text{va}}$ from each subsample. Figure 6 shows the boxplots of these estimated coefficients. Although the ground truth is unknown, the red dashed line plots the MLE $\hat{\beta}_{\text{va}} = -6.18$ calculated over all 4898 observations so that it is an accurate estimate of the corresponding parameter. Noticeably, the ML coefficients move further from the red line, as the dimensionality factor increases, exhibiting a strong bias. For instance, when κ is in $\{0.10, 0.18, 0.26\}$, the median MLE is repectively equal to $\{-7.48, -9.43, -10.58\}$; that is, $\{1.21, 1.53, 1.71\}$ times the value of the MLE (in magnitude) from the full data. These observations support our hypothesis that, irrespective of the covariate distribution, the MLE increasingly overestimates effect magnitudes in high dimensions.

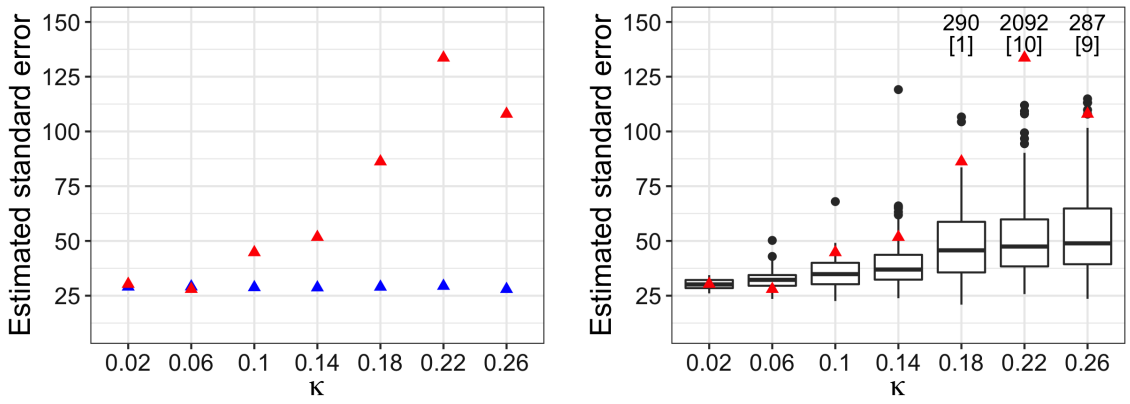
Next, Figure 7 compares the standard deviation (sd) in high dimensions with the corresponding prediction from the classical Fisher information formula.[5] Theorem 3.1 states that when n and p are both large and the covariates follow a multivariate Gaussian distribution, $\hat{\beta}_j$ approximately obeys

$$\hat{\beta}_j = \alpha_\star(\kappa)\beta_j + \sigma_\star(\kappa)Z/\sqrt{n}, \tag{32}$$

where $Z \sim \mathcal{N}(0,1)$. If we reduce n by a factor of 2, the standard deviation should increase by a factor of $\sigma_\star(2\kappa)/\sigma_\star(\kappa) \times \sqrt{2}$. Thus, in order to evidence the interesting contribution, namely, the factor of $\sigma_\star(2\kappa)/\sigma_\star(\kappa)$, we plot $\sqrt{n} \times \text{sd}(\hat{\beta}_{\text{va}})$, where n is the sample size used to calculate our estimate.

With the sample size adjustment, we see in Figure 7a that the variance of the ML coefficient is much higher than the corresponding classical value, and that the mismatch increases as the problem dimension increases. Thus, we see once more a "variance inflation" phenomenon similar to that observed for Gaussian covariates (see also, [3, 4]). To be complete, we here approximate/estimate the (inverse) Fisher information as follows: for each $\kappa \in K$, we form $\hat{\mathcal{I}}(\boldsymbol{\beta}) = \frac{1}{B}\sum_{j=1}^{B} \boldsymbol{X}_j' \boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{X}_j$, where $\boldsymbol{X}_j$ is the covariate matrix from the j-th subsample, and for $\boldsymbol{\beta}$, we plug in the MLE from the full data.

[5] The Fisher information here is given by $\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{X}]$, where $\boldsymbol{D}(\boldsymbol{\beta})$ is a diagonal matrix with the i-th diagonal entry given by $\rho''(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}/(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})^2$.



(a) **(b)**

Figure 7: Comparison of standard errors adjusted for the sample size. Throughout, the red triangles represent an estimate of the standard deviation (sd) of the MLE for "volatile acidity", obtained by creating folds of sample size p/κ and computing the sd across these folds. To evidence, the bias, the estimated standard deviations are multiplied by the root sample size (see text). (a) Estimated sd given by the inverse Fisher information (adjusted for sample size) averaged over $B = 100$ subsamples for each value of κ. (b) Standard error (adjusted for sample size) from `R` calculated in each subsample. (The meaning of the numbers in between square brackets and their accompanying integer is as in Figure 6.)

Standard errors obtained from software packages are different from those shown in Figure 7a, since these typically use the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ from the data set at hand as a plug-in for $\boldsymbol{\beta}$, and in addition, do not take expectation over the randomness of the covariates. However, since these estimates are widely used in practice, it is of interest to contrast them with the true standard deviations. Figure 7b presents standard errors of $\hat{\beta}_{\mathrm{va}}$ (adjusted for sample size) as obtained from `R`. Observe that for large values of κ, these also severely underestimate the true variability.

8 Discussion

This paper establishes a maximum likelihood theory for high-dimensional logistic models with arbitrarily correlated Gaussian covariates. In particular, we establish a stochastic representation for the MLE that holds for finite sample sizes. This in turn yields a precise characterization of the finite-dimensional marginals of the MLE, as well as the average behavior of its coordinates. Our theory relies on the unknown signal strength parameter γ, which can be accurately estimated by the *ProbeFrontier* method. This provides a valid procedure for constructing p-values and confidence intervals for any finite collection of coordinates. Furthermore, we observe that our procedure produces reliable results for moderate sample sizes, even in the absence of Gaussianity—in particular, when the covariates are light-tailed.

We conclude with a few directions of future research—it would be of interest to understand (a) the class of covariate distributions for which our theory, or a simple modification thereof, continues to apply, (b) the class of generalized linear models for which analogous results hold, and finally, (c) the robustness of our proposed procedure to model misspecifications.

A Proofs

A.1 Proof of Lemma 3.1

We claim that Lemma 3.1 is a direct consequence of [14, Theorem 2], which states this: in the regime where the MLE exists asymptotically, the MLE $\hat{\boldsymbol{\eta}}$ in a logistic model with covariates drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and regression vector $\boldsymbol{\eta}$ obeys

$$\frac{1}{p}\sum_{j=1}^{p}\psi\left(\sqrt{n}(\hat{\eta}_j - \tilde{\alpha}\eta_j), \sqrt{n}\eta_j\right) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi\left(\tilde{\sigma}Z, \tilde{\eta}\right)\right] \tag{33}$$

for every pseudo-Lipschitz function ψ of order 2.[6] Above, $(\tilde{\alpha}, \tilde{\sigma})$ are defined as follows: together with another scalar $\tilde{\lambda}$, the triple $(\tilde{\alpha}, \tilde{\sigma}, \tilde{\lambda})$ forms the unique solution to (18), when we plug in $\gamma^2 = \lim_{n\to\infty}\|\boldsymbol{\eta}\|^2$.This result holds under the following conditions: (a) $1/p\sum_{j=1}^{p}\delta_{\sqrt{n}\eta_j} \xrightarrow{\text{d}} \Pi$ for some probability distribution Π (convergence of the empirical distribution), and (b) $\sum_{j=1}^{p}(\sqrt{n}\eta_j)^2/p \to \mathbb{E}\left[\Pi^2\right]$, the second moment of Π (convergence of the second moment). In (33), $\tilde{\eta} \sim \Pi$ is independent of $Z \sim \mathcal{N}(0,1)$.

To apply this result, set any orthogonal matrix $\boldsymbol{U}$ such that

$$\boldsymbol{U}\boldsymbol{\theta} = \frac{1}{\sqrt{p}}(\|\boldsymbol{\theta}\|, \ldots, \|\boldsymbol{\theta}\|).$$

Upon setting $\boldsymbol{\eta} = \boldsymbol{U}\boldsymbol{\theta}$, we see that $\boldsymbol{\eta}$ obeys the conditions above with $\Pi = \delta_{\gamma/\sqrt{\kappa}}$, and that $\lim_{n\to\infty}\|\boldsymbol{\eta}\|^2 = \lim_{n\to\infty}\|\boldsymbol{\theta}\|^2 = \gamma^2$, where γ is defined via (16). Applying (33) with $\psi(t,u) = tu$ gives

$$\frac{1}{\kappa}\langle \boldsymbol{U}\hat{\boldsymbol{\theta}} - \alpha_\star \boldsymbol{U}\boldsymbol{\theta}, \boldsymbol{U}\boldsymbol{\theta}\rangle = \frac{1}{\kappa}\langle \hat{\boldsymbol{\theta}} - \alpha_\star\boldsymbol{\theta}, \boldsymbol{\theta}\rangle \xrightarrow{\text{a.s.}} 0.$$

This implies

$$\alpha(n) = \frac{\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle}{\|\boldsymbol{\theta}\|^2} \xrightarrow{\text{a.s.}} \alpha_\star.$$

Similarly, choosing $\psi(t,u) = t^2$, we obtain

$$\frac{1}{\kappa}\|\boldsymbol{U}\hat{\boldsymbol{\theta}} - \alpha_\star\boldsymbol{U}\boldsymbol{\theta}\|^2 = \frac{1}{\kappa}\|\hat{\boldsymbol{\theta}} - \alpha_\star\boldsymbol{\theta}\|^2 \xrightarrow{\text{a.s.}} \sigma_\star^2.$$

This gives

$$\sigma^2(n) = \|P_{\boldsymbol{\theta}^\perp}\hat{\boldsymbol{\theta}}\|^2 = \|\hat{\boldsymbol{\theta}} - \alpha(n)\boldsymbol{\theta}\|^2 \xrightarrow{\text{a.s.}} \kappa\sigma_\star^2.$$

A.2 Proof of Theorem 3.1

To begin with, we recall from (11) that, or any arbitrary coordinate j,

$$\tau_j\frac{\hat{\beta}_j - \alpha_\star\beta_j}{\sigma_\star} = \frac{\hat{\theta}_j - \alpha_\star\theta_j}{\sigma_\star},$$

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

so that we only need to study the RHS. Above, $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ are defined as in (7), with the choice of $\boldsymbol{L}$ specified in (9). Now, we have

$$\sqrt{n}\frac{\hat{\theta}_j - \alpha_\star\theta_j}{\sigma_\star} = \sqrt{n}\frac{\hat{\theta}_j - \alpha(n)\theta_j}{\sigma(n)}\,\frac{\sigma(n)}{\sigma_\star} + \sqrt{n}\frac{(\alpha(n)-\alpha_\star)\theta_j}{\sigma_\star},$$

where $\alpha(n)$ and $\sigma(n)$ are defined as in (12). Lemma 2.1 states that for every problem dimension p,

$$\frac{\hat{\boldsymbol{\theta}} - \alpha(n)\boldsymbol{\theta}}{\sigma(n)} \overset{d}{=} \frac{P_{\boldsymbol{\theta}^\perp}\boldsymbol{Z}}{\|P_{\boldsymbol{\theta}^\perp}\boldsymbol{Z}\|}, \tag{34}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_p) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Upon expanding the projection as

$$P_{\boldsymbol{\theta}^\perp}\boldsymbol{Z} = \boldsymbol{Z} - \langle \boldsymbol{Z}, \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}\rangle \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \tag{35}$$

and using the fact that $\|P_{\boldsymbol{\theta}^\perp}\boldsymbol{Z}\|/\sqrt{p} \xrightarrow{\text{a.s.}} 1$, a quick calculation gives

$$\sqrt{n}\frac{\hat{\theta}_j - \alpha(n)\theta_j}{\sigma(n)} \overset{d}{=} \frac{1}{\sqrt{\kappa}}\sigma_j\, Z + o_P(1), \qquad \sigma_j^2 = 1 - \frac{\theta_j^2}{\|\boldsymbol{\theta}\|^2},$$

where $Z \sim \mathcal{N}(0,1)$. Since $\alpha(n) \xrightarrow{\text{a.s.}} \alpha_\star$ and $\sigma(n) \xrightarrow{\text{a.s.}} \sigma_\star$, (19) holds if, additionally, $\sqrt{n}\theta_j = O(1)$, which is equivalent to $\sqrt{n}\tau_j\beta_j = O(1)$, from the relation (10).

The general result (20) follows immediately by rotational invariance. To be precise, consider an orthogonal matrix $\boldsymbol{U}$ with the first row given by $\boldsymbol{U}_{1\bullet} = \boldsymbol{v}^\top$. Recall that $\hat{\boldsymbol{\zeta}} := \boldsymbol{U}\hat{\boldsymbol{\beta}}$ is the MLE in a logistic model with covariates $\boldsymbol{U}\boldsymbol{x_i}$ and regression vector $\boldsymbol{\zeta} = \boldsymbol{U}\boldsymbol{\beta}$. Applying (19) with $j = 1$ we obtain that, under the condition $\sqrt{n}\tau(\boldsymbol{v})\boldsymbol{v}^\top\boldsymbol{\beta} = O(1)$,

$$\frac{\sqrt{n}\boldsymbol{v}^\top(\hat{\boldsymbol{\beta}} - \alpha_\star\boldsymbol{\beta})}{\sigma_\star/\tau(\boldsymbol{v})} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where $\tau^2(\boldsymbol{v}) = \mathrm{Var}(\boldsymbol{v}^\top\boldsymbol{x}_i \,|\, P_{\boldsymbol{v}^\perp}\boldsymbol{x}_i) = (\boldsymbol{v}^\top\boldsymbol{\Theta}\boldsymbol{v})^{-1}$. This follows since the inverse of the covariance matrix of $\boldsymbol{U}\boldsymbol{x_i}$ is given by $\boldsymbol{U}\boldsymbol{\Theta}\boldsymbol{U}^\top$, which completes the proof.

A.3 Proof of Theorem 3.2

We seek to establish the asymptotic behavior of the empirical cdf of $\boldsymbol{T}$, where

$$T_j = \frac{\tau_j}{\sigma_\star}\sqrt{n}(\hat{\beta}_j - \alpha_\star\beta_j).$$

Instead of analyzing the MLE $\hat{\boldsymbol{\beta}}$ directly, we use once more the rotational invariance of the Gaussian, to relate the distribution of $\boldsymbol{T}$ with a *correlated* Gaussian, and study the asymptotic behavior of this Gaussian vector. We state below the correspondence and note that this is in the same spirit as Lemma 2.1.

Proposition A.1. *Set* $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_\Sigma = \boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{y}$ *and* $\|\boldsymbol{x}\|_\Sigma^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle_\Sigma$, *and let*

$$\boldsymbol{\xi} = \boldsymbol{Z} - \langle \boldsymbol{Z}, \frac{\boldsymbol{\beta}}{\gamma(n)}\rangle_\Sigma \frac{\boldsymbol{\beta}}{\gamma(n)}, \tag{36}$$

where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta})$ *is independent of everything else, and* $\gamma(n)^2 = \|\boldsymbol{\beta}\|_\Sigma^2 = \mathrm{Var}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. *Define*

$$\tilde{\alpha}(n) := \frac{\langle \hat{\boldsymbol{\beta}}, \boldsymbol{\beta} \rangle_\Sigma}{\|\boldsymbol{\beta}\|_\Sigma^2}, \qquad \tilde{\sigma}^2(n) := \|\hat{\boldsymbol{\beta}}\|_\Sigma^2 - \frac{\langle \hat{\boldsymbol{\beta}}, \boldsymbol{\beta} \rangle_\Sigma^2}{\|\boldsymbol{\beta}\|_\Sigma^2}. \tag{37}$$

Then for every n *and* p, *the MLE* $\hat{\boldsymbol{\beta}}$ *obeys*

$$\frac{\hat{\boldsymbol{\beta}} - \tilde{\alpha}(n)\boldsymbol{\beta}}{\tilde{\sigma}(n)} \overset{d}{=} \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_\Sigma}.$$

Proof. We start from (34) and (35), which hold for every logistic MLE with independent Gaussian covariates. In particular, (34) holds for $\boldsymbol{\theta} = \boldsymbol{L}^\top \boldsymbol{\beta}$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{L}^\top \hat{\boldsymbol{\beta}}$, where $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ is any Cholesky factorization of the covariance matrix. The claim now follows from multiplying both sides of (34) by $\boldsymbol{L}^{-\top}$. We omit the details. □

Proposition A.1 directly implies the following equivalence.

Corollary A.1. *Let* T_j^{approx} *denote the finite sample version of* T_j *given by*

$$T_j^{\mathrm{approx}} = \tau_j \sqrt{n} \frac{\hat{\beta}_j - \tilde{\alpha}(n)\beta_j}{\tilde{\sigma}(n)/\sqrt{\kappa}},$$

where $\tilde{\alpha}(n)$ *and* $\tilde{\sigma}(n)$ *are defined as in* (37). *Take* $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta})$ *and define* $Z_j^{\mathrm{scaled}} = \tau_j Z_j$ *so that*

$$\boldsymbol{Z}^{\mathrm{scaled}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}), \quad R_{ij} = \tau_i \tau_j \Theta_{ij}. \tag{38}$$

(Recall that $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ *so that* $\tau_i^2 = 1/\Theta_{ii}$.*) Furthermore, let*

$$\tilde{Z}_j^{\mathrm{scaled}} = \left(Z_j^{\mathrm{scaled}} - \frac{\langle \boldsymbol{Z}, \boldsymbol{\beta}/\gamma(n) \rangle_\Sigma}{\gamma(n)} \tau_j \beta_j \right) \frac{\sqrt{p}}{\|\boldsymbol{\xi}\|_\Sigma},$$

where $\boldsymbol{\xi}$ *is defined as in* (36). *Then for every* n,

$$\boldsymbol{T}^{\mathrm{approx}} \overset{d}{=} \tilde{\boldsymbol{Z}}^{\mathrm{scaled}}. \tag{39}$$

Notice that for every n, $\langle \boldsymbol{Z}, \boldsymbol{\beta}/\gamma(n) \rangle_\Sigma$ follows a standard Gaussian distribution, and, therefore, this sequence is stochastically bounded.[7] Also,

$$\frac{\|\boldsymbol{\xi}\|_\Sigma^2}{p} = \frac{\|\boldsymbol{Z}\|_\Sigma^2}{p} - \frac{1}{p} \langle \boldsymbol{Z}, \frac{\boldsymbol{\beta}}{\gamma(n)} \rangle_\Sigma^2 \overset{\text{a.s.}}{\longrightarrow} 1.$$

Finally, $\gamma(n) \to \gamma$ by assumption.

Now the proof consists of the following steps:

1. Using $\tilde{\alpha}(n) \overset{a.s.}{\longrightarrow} \alpha_\star$ and $\tilde{\sigma}(n)/\sqrt{\kappa} \overset{a.s.}{\longrightarrow} \sigma_\star$ from Lemma 3.1, we show that the empirical cdf of $\boldsymbol{T}$ is close to that of $\boldsymbol{T}^{\mathrm{approx}}$. Furthermore, we establish that $\boldsymbol{Z}^{\mathrm{scaled}}$ and $\tilde{\boldsymbol{Z}}^{\mathrm{scaled}}$ are close in a similar sense, under some regularity conditions on $\boldsymbol{\Sigma}$. These results are formalized in Lemma A.1. Together these imply that the empirical cdf of $\boldsymbol{T}$ must approximately equal that of the correlated Gaussian vector $\boldsymbol{Z}^{\mathrm{scaled}}$.

[7] We say that a sequence of random variables $\{X_n\}_{n \geq 1}$ is stochastically bounded or $O_p(1)$ if for every $\epsilon > 0$, there exist finite $M, N > 0$ such that $\mathbb{P}(|X_n| > M) < \epsilon, \ \forall n > N$.

2. Next, we show that the empirical cdf of $\boldsymbol{Z}^{\text{scaled}}$ converges to the standard normal cdf in Lemma A.2.

3. Putting these together yields the desired result.

We now proceed to detail these steps.

A.4 Approximating the rescaled MLE by a Gaussian vector

Lemma A.1. *Define* $c(n) := \lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma})$ *and suppose that*

$$\limsup_{n\to\infty} c(n) < \infty. \tag{40}$$

Then for any Lipschitz continuous function $\psi : \mathbb{R} \to \mathbb{R}$,

$$\frac{1}{p}\sum_{j=1}^{p}[\psi(Z_j^{\text{scaled}}) - \psi(\tilde{Z}_j^{\text{scaled}})] \xrightarrow{\mathbb{P}} 0.$$

The same holds for the pair $(\boldsymbol{T}, \boldsymbol{T}^{\text{approx}})$.

Proof. We first prove the result for $\boldsymbol{Z}^{\text{scaled}}$ and $\tilde{\boldsymbol{Z}}^{\text{scaled}}$. If L is the Lipschitz constant for ψ, it holds that

$$\begin{aligned}\left|\frac{1}{p}\sum_{j=1}^{p}\psi(Z_j^{\text{scaled}}) - \frac{1}{p}\sum_{j=1}^{p}\psi(\tilde{Z}_j^{\text{scaled}})\right| &\le \frac{L}{p}\sum_{i=1}^{p}\left|Z_j^{\text{scaled}} - \tilde{Z}_j^{\text{scaled}}\right| \\ &\le L\left|\frac{\sqrt{p}}{\|\boldsymbol{\xi}\|_\Sigma} - 1\right|\frac{1}{p}\sum_{j=1}^{p}\left|Z_j^{\text{scaled}}\right| + \frac{L\sqrt{p}}{\|\boldsymbol{\xi}\|_\Sigma}\frac{\langle \boldsymbol{Z}, \boldsymbol{\beta}/\gamma(n)\rangle_\Sigma}{\gamma(n)}\frac{1}{p}\sum_{j=1}^{p}|\tau_j\beta_j|.\end{aligned}$$

To establish that the RHS converges to zero in probability, it suffices to show that

$$\frac{1}{p}\sum_{j=1}^{p}\left|Z_j^{\text{scaled}}\right| = O_{\mathbb{P}}(1) \quad \text{and} \quad \frac{1}{p}\sum_{j=1}^{p}|\tau_j\beta_j| = o(1). \tag{41}$$

Note that

$$\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p}\left|Z_j^{\text{scaled}}\right|\right] = \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[\left|Z_j^{\text{scaled}}\right|\right] \le 1,$$

since $\mathbb{E}\left[|Z_j^{\text{scaled}}|\right]^2 \le \mathbb{E}\left[(Z_j^{\text{scaled}})^2\right] = 1$. For the second term in (41), it suffices to analyze $\|\boldsymbol{\tau}\|$ and $\|\boldsymbol{\beta}\|$. We know that

$$\|\boldsymbol{\tau}\|^2 = \sum_{j=1}^{p}\tau_j^2 \le p\tau_{\max}^2, \quad \text{where} \quad \tau_{\max}^2 = 1/\min\{\boldsymbol{\Theta}_{jj}\} \le 1/\lambda_{\min}(\boldsymbol{\Theta}) = \lambda_{\max}(\boldsymbol{\Sigma}), \tag{42}$$

and that

$$\lambda_{\min}(\boldsymbol{\Sigma})\,\|\boldsymbol{\beta}\|^2 \le \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} = \gamma(n)^2. \tag{43}$$

Thus, we obtain

$$\frac{1}{p}\sum_{i=1}^{p}|\tau_j\beta_j| \le \frac{1}{p}\|\boldsymbol{\tau}\|\,\|\boldsymbol{\beta}\| \le \frac{\gamma(n)}{\sqrt{p}}\left(\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)^{1/2} = o(1), \tag{44}$$

where the last equality follows from (40).

We now turn to establishing the result for $\boldsymbol{T}$ and $\boldsymbol{T}^{\text{approx}}$. Note that

$$\left|\frac{1}{p}\sum_{j=1}^{p}\psi(T_j) - \frac{1}{p}\sum_{j=1}^{p}\psi(T_j^{\text{approx}})\right| \le \frac{L}{p}\sum_{j=1}^{p}\left|T_j - T_j^{\text{approx}}\right|$$

$$\le L\sqrt{\frac{n\kappa}{p}}\left[\left|\frac{1}{\tilde{\sigma}(n)} - \frac{1}{\sqrt{\kappa}}\frac{1}{\sigma_\star}\right|\frac{1}{\sqrt{p}}\sum_{j=1}^{p}\left|\tau_j\hat{\beta}_j\right| + \left|\frac{\tilde{\alpha}(n)}{\tilde{\sigma}(n)} - \frac{1}{\sqrt{\kappa}}\frac{\alpha_\star}{\sigma_\star}\right|\frac{1}{\sqrt{p}}\sum_{j=1}^{p}|\tau_j\beta_j|\right].$$

The second term is $o_P(1)$ because

$$\left|\frac{\tilde{\alpha}(n)}{\tilde{\sigma}(n)} - \frac{1}{\sqrt{\kappa}}\frac{\alpha_\star}{\sigma_\star}\right| \xrightarrow{a.s.} 0 \quad \text{and} \quad \frac{1}{\sqrt{p}}\sum_{j=1}^{p}|\tau_j\beta_j| = O(1)$$

the same way as (44).

For the first term, we recall that in the (κ,γ) region where the MLE exists, $\hat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = O(1)$ with exponentially high probability [13, Theorem 4]. Once again by Cauchy Schwarz and an application of (43) with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$, we obtain that

$$\frac{1}{\sqrt{p}}\sum_{j=1}^{p}\left|\tau_j\hat{\beta}_j\right| = O_{\mathbb{P}}(1). \tag{45}$$

Since $\tilde{\sigma}(n)^2 \to \kappa\sigma_\star^2$ and $\tilde{\alpha}(n) \to \alpha_\star$ almost surely, this completes the proof.

□

Together with (39), Lemma A.1 implies that

$$\frac{1}{p}\sum_{j=1}^{p}[\psi(T_j) - \psi(Z_j^{\text{scaled}})] \xrightarrow{\mathbb{P}} 0. \tag{46}$$

In what follows, we establish that the empirical cdf of $\boldsymbol{Z}^{\text{scaled}}$ converges to that of a standard normal distribution, as long as $c(n)$ grows at a rate negligible compared to $\sqrt{p(n)}$. Then, we will transfer this property to $\boldsymbol{T}$ by approximating indicator functions via Lipschitz continuous functions.

A.5 Weakly correlated normal variables

Recall from (38) that $\boldsymbol{Z}^{\text{scaled}}$ follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{R}$. If $\boldsymbol{Z}^{\text{scaled}}$ is weakly correlated, then [1, Theorem 1] tells us that its empirical cdf will converge uniformly in $\mathcal{L}_2$ to a standard normal cdf. The notion of weak correlation is this:

Definition A.1. *([1, Definition 1]) Let $\{\xi_i\}_{i=1}^{\infty}$ be a sequence of standard normal variables with joint normal distribution. Denote the correlation matrix of $(\xi_1, \ldots, \xi_p)$ by $\boldsymbol{R}_p$. If $\|\boldsymbol{R}_p\|_1^{(p)} := \frac{1}{p^2}\sum_{ij}|r_{ij}| \to 0$, then $\{\xi_i\}_{i=1}^{\infty}$ is called* weakly correlated. *Else it is said to be* strongly correlated. *It can be shown that the condition $\|\boldsymbol{R}_p\|_1^{(p)} \to 0$ is equivalent to $\|\boldsymbol{R}_p\|_2^{(p)} := \frac{1}{p}\left(\sum_{i,j} r_{ij}^2\right)^{\frac{1}{2}} \to 0$.*

Lemma A.2. *Consider a sequence of random vectors $\boldsymbol{Z}^{\text{scaled}}(n) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}(n))$, such that the (i,j)th entry of $\boldsymbol{R}(n)$ is given by*

$$r_{ij} = \tau_i \tau_j \Theta_{ij} \quad \text{where} \quad \boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}. \tag{47}$$

Assume $\limsup_{n\to\infty} c(n)/\sqrt{p(n)} = 0$, where $c(n) = \lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma})$, then for every $t \in \mathbb{R}$, the empirical cdf of $\boldsymbol{Z}^{\text{scaled}}$ converges in probability to the standard normal cdf; that is,

$$\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}\left\{Z_j^{\text{scaled}} \le t\right\} \xrightarrow{\text{P}} \Phi(t). \tag{48}$$

Proof. We rely on [1, Theorem 1], which states that if $\boldsymbol{R}$ is a weak correlation matrix, then for every $s > 0$ and $t \in \mathbb{R}$, the empirical cdf $\hat{F}_p(t) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}\left\{Z_j^{\text{scaled}} \le t\right\}$ satisfies

$$\mathbb{P}\left(\left|\hat{F}_p(t) - \Phi(t)\right| \ge s\right) \le \frac{1}{s^2}\mathbb{E}\left[(\hat{F}_p(t) - \Phi(t))^2\right] \overset{p\to\infty}{\longrightarrow} 0.$$

We defer the readers to Appendix B.1 for a precise statement of the theorem.

Now, it remains to check whether the covariance matrix from (47) is weakly correlated. We have

$$\begin{aligned}
\|\boldsymbol{R}_p\|_2^{(p)} = \frac{1}{p}\Big(\sum_{i,j} r_{ij}^2\Big)^{\frac{1}{2}} &= \frac{1}{p}\Big(\sum_{i,j} \tau_i^2 \tau_j^2 \Theta_{ij}^2\Big)^{1/2} \\
&\le \frac{1}{p}\tau_{\max}^2\Big(\sum_{i,j}\Theta_{ij}^2\Big)^{1/2} = \frac{1}{p}\tau_{\max}^2\Big(\sum_{i=1}^{p}\lambda_i^2(\boldsymbol{\Theta})\Big)^{1/2} \\
&\le \frac{1}{p}\tau_{\max}^2\sqrt{p}\,\lambda_{\max}(\boldsymbol{\Theta}) \\
&\le \frac{1}{\sqrt{p}}\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})} \\
&= \frac{c(n)}{\sqrt{p(n)}} \longrightarrow 0,
\end{aligned}$$

where the last inequality follows from (42) and the fact that $\lambda_{\max}(\boldsymbol{\Theta}) = 1/\lambda_{\min}(\boldsymbol{\Sigma})$. □

Lemma A.2 also guarantees that the empirical average of $\psi(Z_j^{\text{scaled}})$ for compactly supported continuous functions ψ converges to the corresponding expectation of a standard Gaussian.

Corollary A.2. *Let $\{W_{n,j}\}$ be a triangular array of random variables, with the n-th row given by $\boldsymbol{W}(n) = (W_{n,1}, \ldots, W_{n,p(n)})$. If the empirical cdf of $\boldsymbol{W}(n)$ converges in probability to a continuous cumulative distribution function F, that is,*

$$\frac{1}{p(n)}\sum_{j=1}^{p(n)} \mathbb{I}\{W_{n,j} \le t\} \xrightarrow{\text{P}} F(t), \tag{49}$$

then for any continuous function $\psi(t)$ with compact support,

$$\frac{1}{p(n)}\sum_{j=1}^{p(n)}\psi(W_{n,j}) - \mathbb{E}\left[\psi(Z)\right] \xrightarrow{p} 0, \tag{50}$$

where $Z \sim F$. In particular, this holds for functions of the form

$$\varphi_{t,L}(x) = \begin{cases} 1 & x \le t, \\ -Lx + (1 + Lt), & x \in [t, t + 1/L), \\ 0, & x > t + \frac{1}{L}. \end{cases} \tag{51}$$

The proof follows from basic analysis and is deferred to Appendix B.

A.6 Putting things together

We now proceed to complete the proof of Theorem 3.2. Fix $t \in \mathbb{R}$ and $\epsilon > 0$, we seek to show that there exists $\delta > 0, p_0 \in \mathbb{N}$ such that $\forall p \ge p_0$,

$$\mathbb{P}\left(\left|\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\left\{T_j \le t\right\} - \Phi(t)\right| > \varepsilon\right) < \delta.$$

To this end, we consider the functions $\varphi_{t,4/\varepsilon}(s)$ defined via (51). Since for every s, $\varphi_{t,4/\varepsilon}(s) \ge \mathbb{I}\left\{s \le t\right\}$, we have that

$$\mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\left\{T_j \le t\right\} - \Phi(t) > \varepsilon\right) \le \mathbb{P}\left(\frac{1}{p}\sum_{i=1}^{p}\varphi_{t,4/\varepsilon}(T_j) - \Phi(t) > \varepsilon\right).$$

On the other hand, $\varphi_{t,4/\varepsilon}(s)$ differs from $\mathbb{I}\left\{s \le t\right\}$ only in the interval $[t, t + \varepsilon/4)$, so that upon integrating with respect to the Gaussian cdf, we obtain

$$\mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z) - \mathbb{I}\left\{Z \le t\right\}\right] < \varepsilon/4,$$

where $Z \sim \mathcal{N}(0,1)$. Adding and subtracting $\mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z)\right]$, we see that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\varphi_{t,4/\varepsilon}(T_j) - \Phi(t) > \varepsilon\right) &\le \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\varphi_{t,4/\varepsilon}(T_j) - \mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z)\right] > \varepsilon/2\right) \\
&\le \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}[\varphi_{t,4/\varepsilon}(T_j) - \varphi_{t,4/\varepsilon}(T_j^{\text{approx}})] > \varepsilon/4\right) + \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\varphi_{t,4/\varepsilon}(T_j^{\text{approx}}) - \mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z)\right] > \varepsilon/4\right) \\
&= \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}[\varphi_{t,4/\varepsilon}(T_j) - \varphi_{t,4/\varepsilon}(T_j^{\text{approx}})] > \varepsilon/4\right) + \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\varphi_{t,4/\varepsilon}(\tilde{Z}_j^{\text{scaled}}) - \mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z)\right] > \varepsilon/4\right) \\
&\le \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}[\varphi_{t,4/\varepsilon}(T_j) - \varphi_{t,4/\varepsilon}(T_j^{\text{approx}})] > \varepsilon/4\right) + \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}[\varphi_{t,4/\varepsilon}(\tilde{Z}_j^{\text{scaled}}) - \varphi_{t,4/\varepsilon}(Z_j^{\text{scaled}})] > \varepsilon/8\right) \\
&\quad + \mathbb{P}\left(\frac{1}{p}\sum_{j=1}^{p}\varphi_{t,4/\varepsilon}(Z_j^{\text{scaled}}) - \mathbb{E}\left[\varphi_{t,4/\varepsilon}(Z)\right] > \varepsilon/8\right),
\end{aligned}$$

where the third equality follows from (39). From Lemma A.1 and Corollary A.2, one can obtain p_0 and δ such that for all $p \geq p_0$, the RHS above remains less than δ. This concludes our proof for one direction. The other direction can be established by arguments similar to the above, on considering the functions $\varphi_{t-4/\varepsilon,4/\varepsilon}$.

B Auxiliary results and proofs

Theorem B.1. *([1, Theorem 1]) Let $Z_1, \ldots, Z_p, \ldots$ be $\mathcal{N}(0,1)$ random variables with empirical cdfs given by*

$$\hat{F}_p(z) = \frac{1}{p} \sum_{i=1}^{p} \mathbb{I}\{Z_i \leq z\}.$$

Let $\boldsymbol{R}_p$ denote the covariance matrix of $(Z_1, \ldots, Z_p)$.

1. *(Sufficiency) If $\{Z_i\}_{i=1}^{\infty}$ is weakly correlated, then $\hat{F}_p(z)$ converges to $\Phi(z)$ in $\mathcal{L}_2$ uniformly:*

$$\sup_z \mathbb{E}\left\{\hat{F}_p(z) - \Phi(z)\right\}^2 \leq \frac{1}{4p} + C \left\|\boldsymbol{R}_p\right\|_1^{(p)} \to 0,$$

   where C is a universal constant.

2. *(Necessity) If $\{Z_i\}_{i=1}^{\infty}$ is not weakly correlated, that is, $\|\boldsymbol{R}_p\|_2^{(p)} \not\to 0$, then $\hat{F}_p(z)$ does not converge to $\Phi(z)$ in $\mathcal{L}_2$ for any $z \neq 0$, that is:*

$$\mathbb{E}\left\{\hat{F}_p(z) - \Phi(z)\right\}^2 \overset{p\to\infty}{\not\longrightarrow} 0, \quad \forall z \neq 0.$$

Proof of Corollary A.2 Assume w.l.o.g. that ψ is supported on $[0,1]$. Since ψ is continuous, for any $\varepsilon > 0$, we can partition the unit interval into $0 = t_0 < t_1 < \cdots < t_N = 1$ so that for any $x \in [t_{i-1}, t_i]$,

$$\left|\psi(x) - \psi\left(\frac{t_{i-1} + t_i}{2}\right)\right| \leq \varepsilon/8.$$

Then the step function

$$f(x) = \sum_{i=1}^{N-1} \psi\left(\frac{t_{i-1} + t_i}{2}\right) \mathbb{I}\{x \in [t_{i-1}, t_i)\}$$

satisfies

$$|\psi(x) - f(x)| \leq \varepsilon/8, \quad \forall x \in [0,1];$$

hence, for any random variable Z, $\mathbb{E}|\psi(Z) - f(Z)| \leq \varepsilon/8$. The LHS of (50) can be rearranged as

$$\frac{1}{p}\sum_{j=1}^{p} \psi(W_j) - \mathbb{E}[\psi(Z)] = \frac{1}{p}\sum_{j=1}^{p} (\psi(W_j) - f(W_j)) + \left(\frac{1}{p}\sum_{i=1}^{p} f(W_j) - \mathbb{E}[f(Z)]\right) + (\mathbb{E}[f(Z)] - \mathbb{E}[\psi(Z)]).$$

By the triangle inequality,

$$\mathbb{P}\left(\left|\frac{1}{p}\sum_{j=1}^{p} \psi(W_j) - \mathbb{E}[\psi(Z)]\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{1}{p}\sum_{j=1}^{p} f(W_j) - \mathbb{E}[f(Z)]\right| > 3\varepsilon/4\right).$$

Since $\frac{1}{p}\sum_{j=1}^{p} f(W_j) \overset{\mathbb{P}}{\to} f(Z)$ from (49), this concludes the proof.

Acknowledgements

E. C. was supported by the National Science Foundation via DMS 1712800 and via the Stanford Data Science Collaboratory OAC 1934578, and by a generous gift from TwoSigma. P.S. was supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences. Q. Z. would like to thank Stephen Bates for helpful comments about an early version of this paper.

References

[1] David Azriel and Armin Schwartzman. The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association*, 110(511):1217–1228, 2015.

[2] Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *to appear, The Annals of Statistics*, 2018+.

[3] David Donoho and Andrea Montanari. High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, 2016.

[4] 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.

[5] 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.

[6] Gad Kimmel and Ron Shamir. A block-free hidden markov model for genotypes and its application to disease association. *Journal of Computational Biology*, 12(10):1243–1260, 2005.

[7] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[8] Pasi Rastas, Mikko Koivisto, Heikki Mannila, and Esko Ukkonen. A hidden markov technique for haplotype reconstruction. In Rita Casadio and Gene Myers, editors, *Algorithms in Bioinformatics*, pages 140–151, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[9] Chiara Sabatti, Emmanuel J. Candès, and Matteo Sesia. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2018.

[10] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992, 2019.

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

[12] Matteo Sesia. Using snpknock with genetic data. *https://msesia.github.io/snpknock/articles/genotypes.html*, 2019. [Online; accessed 30-October-2019].

[13] Pragya Sur and Emmanuel J. Candès. Supporting information to: A modern maximum-likelihood theory for high-dimensional logistic regression. *https://www.pnas.org/content/pnas/suppl/2019/06/29/1810420116.DCSupplemental/pnas.1810420116.sapp.pdf*, 2018.

[14] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[15] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558, 2019.

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

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

[18] Àngela Nebot, Francisco Mugica, and Antoni Escobet. Modeling wine preferences from physicochemical properties using fuzzy techniques. In *Proceedings of the 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications - Volume 1: SIMULTECH,*, pages 501–507. INSTICC, SciTePress, 2015.