The Annals of Statistics
2020, Vol. 48, No. 1, 27–42
https://doi.org/10.1214/18-AOS1789

THE PHASE TRANSITION FOR THE EXISTENCE OF THE MAXIMUM LIKELIHOOD ESTIMATE IN HIGH-DIMENSIONAL LOGISTIC REGRESSION

BY EMMANUEL J. CANDÈS[1,2] AND PRAGYA SUR[2]

[1] *Department of Mathematics, Stanford University, candes@stanford.edu*

[2] *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, pragya@seas.harvard.edu*

This paper rigorously establishes that the existence of the maximum likelihood estimate (MLE) in high-dimensional logistic regression models with Gaussian covariates undergoes a sharp "phase transition." We introduce an explicit boundary curve h_{MLE}, parameterized by two scalars measuring the overall magnitude of the unknown sequence of regression coefficients, with the following property: in the limit of large sample sizes n and number of features p proportioned in such a way that $p/n \to \kappa$, we show that if the problem is sufficiently high dimensional in the sense that $\kappa > h_{\mathrm{MLE}}$, then the MLE does not exist with probability one. Conversely, if $\kappa < h_{\mathrm{MLE}}$, the MLE asymptotically exists with probability one.

1. Introduction.

Logistic regression [12, 13] is perhaps the most widely used and studied nonlinear model in the multivariate statistical literature. For decades, statistical inference for this model has relied on likelihood theory, especially on the theory of maximum likelihood estimation and of likelihood ratios. Imagine we have n independent observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \dots, n$, where the response $y_i \in \{-1, 1\}$ is linked to the covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ via the logistic model

$$\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i'\boldsymbol{\beta}), \qquad \sigma(t) := \frac{e^t}{1+e^t};$$

here, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown vector of regression coefficients. In this model, the log-likelihood is given by

$$\ell(\boldsymbol{b}) = \sum_{i=1}^{n} -\log(1 + \exp(-y_i \boldsymbol{x}_i'\boldsymbol{b}))$$

and, by definition, the maximum likelihood estimate (MLE) is any maximizer of this functional.

1.1. *Data geometry and the existence of the MLE.*

The delicacy of maximum-likelihood theory is that the MLE does not exist in all situations, even when the number p of covariates is much smaller than the sample size n. This is a well-known phenomenon, which sparked several interesting series of investigation. One can even say that characterizing the existence and uniqueness of the MLE in logistic regression has been a classical problem in statistics. For instance, every statistician knows that if the n data points $(\boldsymbol{x}_i, y_i)$ are *completely separated* in the sense that that there is a linear decision boundary parameterized by $\boldsymbol{b} \in \mathbb{R}^p$ with the property

$$y_i \boldsymbol{x}_i'\boldsymbol{b} > 0 \qquad \text{for all } i, \tag{1.1}$$

Received October 2018.
MSC2010 subject classifications. 62J12, 62E20.
Key words and phrases. High-dimensional logistic regression, MLE phase transition.

then the MLE does not exist. To be clear, (1.1) means that the decision rule that assigns a class label equal to the sign of $\boldsymbol{x}_i'\boldsymbol{b}$ makes no mistake on the sample. Every statistician also knows that if the data points *overlap* in the sense that for every $\boldsymbol{b} \neq \mathbf{0}$, there is at least one data point that is classified correctly ($y_i\boldsymbol{x}_i'\boldsymbol{b} > 0$) and at least another that is classified incorrectly ($y_k\boldsymbol{x}_k'\boldsymbol{b} < 0$), then the MLE does exist. The remaining situation, where the data points are *quasi-completely separated*, is perhaps less well known to statisticians: this occurs when for any decision $\boldsymbol{b} \neq \mathbf{0}$,

$$y_i\boldsymbol{x}_i'\boldsymbol{b} \geq 0 \qquad \text{for all } i, \tag{1.2}$$

where equality above holds for some of the observations. A useful theorem of Albert and Anderson [1] states that the MLE does not exist in this case either. *Hence, the MLE exists if and only if the data points overlap.*

Historically, [1] follows earlier work of Silvapulle [17], who proposed necessary and sufficient conditions for the existence of the MLE based on a geometric characterization involving convex cones (see [1] for additional references). Subsequently, Santner and Duffy [15] expanded on the characterization from [1] whereas Kaufman [8] established theorems on the existence and uniqueness of the minimizer of a closed proper convex function. In order to detect separation, linear programming approaches have been proposed on multiple occasions; see, for instance, [1, 10, 18]. Detection of complete separation was studied in further detail in [9, 11]. Finally, [4] analyzes the notion of regression depth for measuring overlap in data sets.

1.2. *Limitations.* Although beautiful, the aforementioned geometric characterization does not concretely tell us when we can expect the MLE to exist and when we cannot. Instead, it trades one abstract notion, "there is an MLE," for another, "there is no separating hyperplane." To drive our point home, imagine that we have a large number of covariates $\boldsymbol{x}_i$, which are independent samples from some distribution F, as is almost always encountered in modern applications. Then by looking at the distribution F, the data analyst would like to be able to predict when she can expect to find the MLE and she cannot. The problem is that the abstract geometric separation condition does not inform her in any way; she would have no way to know a priori whether the MLE would go to infinity or not.

1.3. *Cover's result.* One notable exception against this background dates back to the seminal work of Cover [5, 6] concerning the separating capacities of decision surfaces. When applied to logistic regression, Cover's main result states the following: assume that the $\boldsymbol{x}_i$'s are drawn i.i.d. from a distribution F obeying some specific assumptions and that the *class labels are independent from* $\boldsymbol{x}_i$ and have equal marginal probabilities; that is, $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = 1/2$. Then Cover shows that as p and n grow large in such a way that $p/n \to \kappa$, the data points asymptotically overlap—with probability tending to one—if $\kappa < 1/2$ whereas they are separated—also with probability tending to one—if $\kappa > 1/2$. In the former case where the MLE exists, [20] refined Cover's result by calculating the limiting distribution of the MLE when the features $\boldsymbol{x}_i$ are Gaussian.

Hence, the results from [5, 6] and [20] describe a phase transition in the existence of the MLE as the dimensionality parameter $\kappa = p/n$ varies around the value 1/2. Therefore, a natural question is this:

Do phase transitions exist in the case where the class labels y_i actually $\underline{\text{depend}}$ on the features $\boldsymbol{x}_i$?

Since likelihood based inference procedures are used all the time, it is of significance to understand when the MLE actually exists. This paper is about this question.

1.4. *Motivation.* Our motivation behind the study of this problem is two-fold. First, knowing when the MLE exists has been a problem of fundamental importance to statistical science, as is evidenced by [5, 6] and the series of works mentioned in Section 1.1. Second, the results from this paper serve as a basis to derive a new theory of maximum-likelihood (ML) estimation in high-dimensional logistic regression [19]. In that work, the authors show that in the common modern setting, where the number of explanatory variables is not negligible compared to the sample size, classical ML theory breaks down. The MLE $\hat{\beta}$ is not close to being Gaussian with mean β (the true regression coefficient sequence) and covariance given by the inverse of the Fisher information matrix. In particular, the MLE is biased and systematically over-estimates effect sizes. Also, the variability of the MLE is greater than that estimated from the inverse Fisher information. Finally, the log-likelihood-ratio (LLR) statistic is far from a χ^2. In a nutshell, [19] proves that the distribution of a coordinate $\hat{\beta}_j$ of the MLE is in some sense equal to

$$\alpha_\star \beta_j + \sigma_\star Z,$$

where $Z \sim \mathcal{N}(0, 1)$ and independent of everything else. This holds provided that the covariates are independent Gaussian variables. For instance, a null coordinate $\hat{\beta}_j$ for which $\beta_j = 0$ follows $\hat{\beta}_j \sim \mathcal{N}(0, \sigma_\star^2)$. Above, $\alpha_\star$ and $\sigma_\star$ are parameters that can be calculated explicitly. In particular, $\alpha_\star > 1$ indicating a (possibly strong) bias of the MLE. Now this new asymptotic theory crucially builds on the results and methods from this paper and would be impossible to build without. To describe the distributional properties of the MLE, we first need to know when we can find it.

1.5. *Phase transitions.* This work rigorously establishes the existence of a phase transition in the logistic model with Gaussian covariates, and computes the phase transition boundary explicitly.

Model. Since researchers routinely include an intercept in the fitted model, we consider such a scenario as well. Throughout the paper, we assume we have n samples $(\boldsymbol{x}_i, y_i)$ with Gaussian covariates unless otherwise mentioned:

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbb{P}(y_i = 1 | \boldsymbol{x}_i) = \sigma(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}) = 1 - \mathbb{P}(y_i = -1 | \boldsymbol{x}_i),$$

where the covariance $\mathbf{\Sigma}$ is nonsingular but otherwise arbitrary.

Peek at the result. To describe our results succinctly, assume the high-dimensional asymptotics from the previous section in which $p/n \to \kappa$ (assumed to be less than one throughout the paper). To get a meaningful result in diverging dimensions, we consider a sequence of problems with β_0 fixed and

$$\operatorname{Var}(\boldsymbol{x}_i' \boldsymbol{\beta}) \to \gamma_0^2. \tag{1.3}$$

This is set so that the log-odds ratio $\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}$ does not increase with n or p, so that the likelihood is not trivially equal to either 0 or 1. Instead,

$$\sqrt{\mathbb{E}(\beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta})^2} \to \sqrt{\beta_0^2 + \gamma_0^2} =: \gamma. \tag{1.4}$$

In other words, we put ourselves in a regime where accurate estimates of $\boldsymbol{\beta}$ translate into a precise evaluation of a nontrivial probability.

Our main result is that there is an explicit function h_{MLE} given in (2.2) such that

$$\kappa > h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad \Longrightarrow \quad \mathbb{P}\{\text{MLE exists}\} \to 0,$$
$$\kappa < h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad \Longrightarrow \quad \mathbb{P}\{\text{MLE exists}\} \to 1.$$

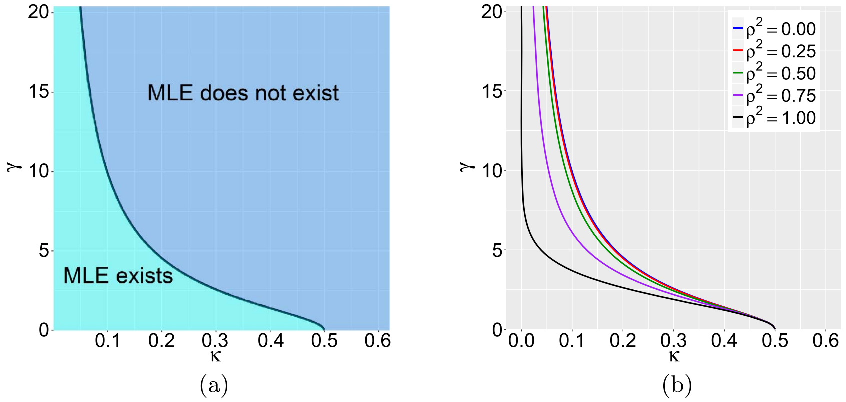

FIG. 1. *Theoretical predictions from* (2.2). (a) *Boundary curve* $\gamma \mapsto h_{\mathrm{MLE}}(0, \gamma)$ *separating the regions where the MLE asymptotically exists and where it does not* (*in this case* $\beta_0 = 0$). (b) *Boundary curves* $\gamma \mapsto h_{\mathrm{MLE}}(\rho\gamma, \sqrt{1-\rho^2}\gamma)$ *for various values of* ρ. *The curve with* $\rho = 0$ *shown in blue is that from* (a). *It is hardly visible because it is close to that with* $\rho^2 = 0.25$.

Hence, the existence of the MLE undergoes a sharp change: below the curves shown in Figure 1, the existence probability asymptotically approaches 1; above, it approaches 0. Also note that the phase-transition curve depends upon the unknown regression sequence $\boldsymbol{\beta} \in \mathbb{R}^p$ only through the intercept β_0 and $\gamma_0^2 = \lim_{n,p\to\infty} \mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta})$.

The formula for the phase transition h_{MLE} is new. As we will see, it is derived from ideas from convex geometry.

1.6. *Notation.* Throughout the paper, vectors and matrices are denoted by lower-case and upper-case bold symbols, respectively. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we write $\boldsymbol{u} \geq \boldsymbol{v}$ whenever the vector $\boldsymbol{u} - \boldsymbol{v}$ has nonnegative entries. Similarly, for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ whenever $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. Finally, $\|\boldsymbol{u}\|$ denotes the usual ℓ_2 norm of the vector $\boldsymbol{u}$ and $\|\boldsymbol{A}\|$ the operator norm of the matrix $\boldsymbol{A}$.

2. Main result.

2.1. *Model with intercept.* Throughout the paper, for each $\beta_0 \in \mathbb{R}$ and $\gamma_0 \geq 0$, we write

$$(Y, V) \sim F_{\beta_0,\gamma_0} \qquad \text{if } (Y, V) \overset{\mathrm{d}}{=} (Y, YX), \tag{2.1}$$

where $X \sim \mathcal{N}(0, 1)$, and $\mathbb{P}(Y = 1|X) = 1 - \mathbb{P}(Y = -1|X) = \sigma(\beta_0 + \gamma_0 X)$.

THEOREM 2.1. *Let* $(Y, V) \sim F_{\beta_0,\gamma_0}$ *and* $Z \sim \mathcal{N}(0, 1)$ *be independent random variables. Define*

$$h_{\mathrm{MLE}}(\beta_0, \gamma_0) = \min_{t_0,t_1\in\mathbb{R}} \{\mathbb{E}(t_0Y + t_1V - Z)_+^2\}, \tag{2.2}$$

where $x_+ = \max(x, 0)$ *and we write* $x_+^2 = (x_+)^2$ *for short. Then in the setting from Section* 1.5,

$$\kappa > h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad \Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\textit{MLE exists}\} = 0,$$

$$\kappa < h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad \Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\textit{MLE exists}\} = 1.$$

This result is proved in Section 3. As the reader will gather from checking our proof, our convergence result is actually more precise. We prove that the transition occurs in an interval of width $O(n^{-1/2})$: take any sequence $\lambda_n \to \infty$; then

$$p/n > h_{\mathrm{MLE}}(\beta_0, \gamma_0) + \lambda_n n^{-1/2} \quad \Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 0,$$

$$p/n < h_{\mathrm{MLE}}(\beta_0, \gamma_0) - \lambda_n n^{-1/2} \quad \Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 1.$$

It is not hard to see that h_{MLE} defined for values of $\beta_0 \in \mathbb{R}$ and $\gamma_0 \geq 0$ is symmetric in its first argument, $h_{\mathrm{MLE}}(\beta_0, \gamma_0) = h_{\mathrm{MLE}}(-\beta_0, \gamma_0)$. We thus only consider the case where $\beta_0 \geq 0$. Over the nonnegative orthant $\mathbb{R}^2_+$, $h_{\mathrm{MLE}}(\beta_0, \gamma_0)$ is a decreasing function of both β_0 and γ_0. Figure 1 shows a few phase-transition curves.

2.2. *Special cases.* It is interesting to check the predictions of formula (2.2) for extreme values of $\gamma := \sqrt{\beta_0^2 + \gamma_0^2}$, namely, $\gamma = 0$ (no signal) and $\gamma \to \infty$ (infinite signal).

- At $\gamma = 0$, Y and V are independent, and Y is a Rademacher variable whereas V is a standard Gaussian. The variable $t_0 Y + t_1 V - Z$ is, therefore, symmetric and

  $$h_{\mathrm{MLE}}(0, 0) = \min_{t_0, t_1} \frac{1}{2} \mathbb{E}(t_0 Y + t_1 V - Z)^2 = \min_{t_0, t_1} \frac{1}{2}(t_0^2 + t_1^2 + 1) = \frac{1}{2}.$$

  Hence, this recovers and extends Cover's result: in the limit where $\beta_0^2 + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \to 0$ (this includes the case where y_i is symmetric and independent of $\boldsymbol{x}_i$ as in [5, 6]), we obtain that the phase transition is at $\kappa = 1/2$.

- When $\gamma_0 \to \infty$, $V \xrightarrow{\mathrm{d}} |Z'|$, $Z' \sim \mathcal{N}(0, 1)$. Hence, plugging $t_0 = 0$ into (2.2) gives

  $$\lim_{t_1 \to -\infty} \mathbb{E}(t_1|Z'| - Z)_+^2 = 0.$$

  If $\beta_0 \to \infty$, $Y \xrightarrow{\mathrm{d}} 1$ and plugging $t_1 = 0$ into (2.2) gives

  $$\lim_{t_0 \to -\infty} \mathbb{E}(t_0 - Z)_+^2 = 0.$$

  Either way, this says that in the limit of infinite signal strength, we must have $p/n \to 0$ if we want to guarantee the existence of the MLE.

  We simplify (2.2) in other special cases below.

LEMMA 1. *In the setting of Theorem* 2.1, *consider the special case* $\gamma_0 = 0$, *where the response does not asymptotically depend on the covariates*: *we have*

$$h_{\mathrm{MLE}}(\beta_0, 0) = \min_{t \in \mathbb{R}} \{\mathbb{E}(tY - Z)_+^2\}. \tag{2.3}$$

In the case $\beta_0 = 0$ *where the marginal probabilities are balanced*, $\mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2$,

$$h_{\mathrm{MLE}}(0, \gamma_0) = \min_{t \in \mathbb{R}} \{\mathbb{E}(tV - Z)_+^2\}. \tag{2.4}$$

PROOF. Consider the first assertion. In this case, it follows from the definition (2.1) that $(Y, V) \overset{\mathrm{d}}{=} (Y, X)$ where Y and X are independent, $\mathbb{P}(Y = 1) = \sigma(\beta_0)$ and $X \sim \mathcal{N}(0, 1)$. Hence,

$$\begin{aligned} h_{\mathrm{MLE}}(\beta_0, 0) &= \min_{t_0, t_1} \mathbb{E}\Big(t_0 Y - \sqrt{1 + t_1^2} Z\Big)_+^2 = \min_{t_0, t_1} (1 + t_1^2) \mathbb{E}\Big(t_0/\sqrt{1 + t_1^2} Y - Z\Big)_+^2 \\ &= \min_{t_0', t_1} (1 + t_1^2) \mathbb{E}(t_0' Y - Z)_+^2 \end{aligned}$$

and the minimum is clearly achieved at $t_1 = 0$. For the second assertion, a simple calculation reveals that Y and V are independent and $\mathbb{P}(Y = 1) = 1/2$. By convexity of the mapping $Y \mapsto (t_0 Y + t_1 V - Z)_+^2$, we have that

$$\mathbb{E}\{(t_0 Y + t_1 V - Z)_+^2 \mid V, Z\} \geq \big(\mathbb{E}\{t_0 Y | V, Z\} + t_1 V - Z\big)_+^2 = (t_1 V - Z_+)^2.$$

Hence, in this case, the miminum in (2.2) is achieved at $t_0 = 0$. □

2.3. *Model without intercept.* An analogous result holds for a model without intercept. Its proof is the same as that of Theorem 2.1, only simpler. It is, therefore, omitted.

THEOREM 2.2. *Assume $\beta_0 = 0$ and consider fitting a model without an intercept. If V has the marginal distribution from Theorem* 2.1 *and is independent from $Z \sim \mathcal{N}(0, 1)$, then the conclusions from Theorem* 2.1 *hold with the phase-transition curve given in* (2.4). *Hence, the location of the phase transition is the same whether we fit an intercept or not.*

2.4. *Comparison with empirical results.* We compare our asymptotic theoretical predictions with the results of empirical observations in finite samples. For a given data set, we can numerically check whether the data is separated by using linear programming techniques; see Section 1.1. (In our set-up, it can be shown that *quasi-complete separation* occurs with zero probability). To detect separability, we study whether the program [10]

$$\begin{array}{lll} \text{maximize} & \displaystyle\sum_{i=1}^{n} y_i(b_0 + \boldsymbol{x}_i'\boldsymbol{b}) & \\ \text{subject to} & y_i(b_0 + \boldsymbol{x}_i'\boldsymbol{b}) \geq 0, & i = 1, \ldots, n, \\ & -1 \leq b_0 \leq 1, & -\mathbf{1} \leq \boldsymbol{b} \leq \mathbf{1} \end{array} \tag{2.5}$$

has a solution or not. For any triplet $(\kappa, \beta_0, \gamma_0)$, we can thus estimate the probability $\hat{\pi}(\kappa, \beta_0, \gamma_0)$ that complete separation does not occur (the MLE exists) by repeatedly simulating data with these parameters and solving (2.5).

Below, each simulated data set follows a logistic model with $n = 4000$, $p = \kappa n$, i.i.d. Gaussian covariates with identity covariance matrix (note that our results do not depend on the covariance $\boldsymbol{\Sigma}$) and $\boldsymbol{\beta}$ selected appropriately so that $\mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \gamma_0^2$. We consider a fixed rectangular grid of values for the pair (κ, γ) where the κ are equispaced between 0 and 0.6 and the γ's—recall that $\gamma = \sqrt{\beta_0^2 + \gamma_0^2}$—are equispaced between 0 and 10. For each triplet $(\kappa, \beta_0, \gamma_0)$, we estimate the chance that complete separation does not occur (the MLE exists) by averaging over 50 i.i.d. replicates.

Figure 2(a) shows empirical findings for a model without intercept; that is, $\beta_0 = 0$, and the other regression coefficients are here selected to have equal magnitude. Observe that the MLE existence probability undergoes a sharp phase transition, as predicted. The phase transition curve predicted from our theory (red) is in excellent agreement with the boundary between high and low probability regions. Figure 2(b) shows another phase transition in the setting where $\gamma_0 = 0$ so that $\beta_0 = \gamma$. The y-axis is here chosen to be the marginal distribution of the response, that is, $\mathbb{P}(y_i = 1) = e^{\gamma}/(1 + e^{\gamma})$. Once again, we observe the sharp phase transition, as promised, and an impeccable alignment of the theoretical and empirical phase transition curves. We also see that when the response distribution becomes increasingly asymmetric, the maximum dimensionality κ decreases, as expected. If y_i has a symmetric distribution, we empirically found that the MLE existed for all values of κ below 0.5 in all replications. For $\mathbb{P}(y_i = 1) = 0.9$, however, the MLE existed (resp., did not exist) if $\kappa < 0.24$ (resp., if $\kappa > 0.28$) in all replications. For information, the theoretical value of the phase transition boundary at $\mathbb{P}(y_i = 1) = 0.9$ is equal to $\kappa = 0.255$.

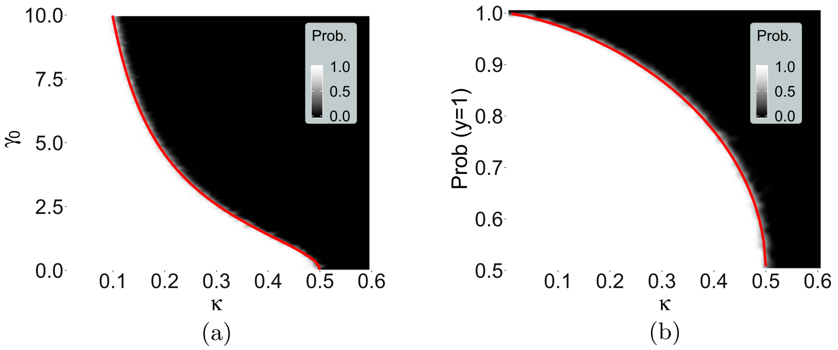

FIG. 2. *Empirical probability that the MLE exists* (*black is zero, and white is one*) *estimated from* 50 *independent trials for each "pixel."* (a) *Model without intercept in which* $\beta_0 = 0$ *and* $\gamma_0 = \gamma$, *with theoretical phase transition curve from* (2.4) *in red* (*this is the same curve as in Figure* 1(a)). (b) *Model with* $\gamma_0 = 0$, $\beta_0 = \gamma$ *and the theoretical phase transition curve from* (2.3) *in red. The* y*-axis is here chosen to be the marginal probability* $\mathbb{P}(y_i = 1) = e^{\gamma}/(1 + e^{\gamma})$.

Finally, Figure 3 shows the phase transition in the setting $\beta_0 = \rho\gamma$, $\gamma_0 = \sqrt{1 - \rho^2}\gamma$ and $\rho = \sqrt{0.75}$. The regression coefficients are chosen to have equal magnitudes. Once again, we observe perfect agreement between the theoretical and empirical phase transition curves.

3. Proof of main theorem. The proof of Theorem 2.1 comprises three main steps as outlined below:

1. Recall that the nonexistence of the MLE is characterized by the geometric conditions (1.1) and (1.2). In Section 3.1, we recast these geometric conditions as whether a random subspace intersects a random convex cone.
2. The motivation underlying the equivalence above is that there exist deep results in high-dimensional stochastic geometry that characterize precisely when a random subspace intersects a given convex cone. We leverage these ideas in Section 3.2 to obtain a mathematically tractable approximation for the probability that the MLE exists.
3. We simplify this approximation through a series of arguments in Section 3.2 to obtain the final result.

3.1. *Conic geometry.* This section introduces ideas from conic geometry and proves our main result. We shall use the characterization from Albert and Anderson [1] reviewed in Section 1.1; recall that the MLE does not exist if and only if there is $(b_0, \boldsymbol{b}) \neq \mathbf{0}$ such that $y_i(b_0 + \boldsymbol{x}_i'\boldsymbol{b}) \geq 0$ for all $i = 1, \ldots, n$. In passing, the same conclusion holds for the probit model and a host of related models.

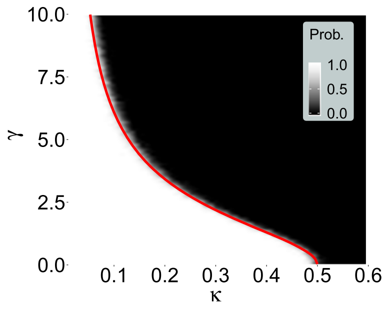

FIG. 3. *Same plot as Figure* 2 *but with* $\beta_0 = \rho\gamma$ *and* $\gamma_0 = \sqrt{1 - \rho^2}\gamma$ *and* $\rho = \sqrt{0.75}$. *The theoretical phase transition curve from* (2.2) *in red is the same as the magenta curve in Figure* 1(b).

3.1.1. *Gaussian covariates.* Write $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} z_i$, where $z_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Doing this, we immediately see that the MLE does not exist if and only if there is $(b_0, \boldsymbol{b}) \neq \mathbf{0}$ such that

$$y_i(b_0 + z_i' \boldsymbol{\Sigma}^{1/2} \boldsymbol{b}) \geq 0 \qquad \forall i.$$

This is equivalent to the existence of $(b_0, \boldsymbol{\theta}) \neq \mathbf{0}$ such that $y_i(b_0 + z_i' \boldsymbol{\theta}) \geq 0$ for all i. In words, multiplication by a nonsingular matrix preserves the existence of a separating hyperplane; that is to say, there is a hyperplane in the "z coordinate" system (where the variables have identity covariance) if and only if there is a separating hyperplane in the "x coordinate" system (where the variables have general nonsingular covariance). Therefore, it suffices to assume that the covariance is the identity matrix, which we do from now on.

We thus find ourselves in a setting where the p predictors are independent standard normal variables and the regression sequence is fixed so that $\mathrm{Var}(\boldsymbol{x}'\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2 = \gamma_0^2$ (the theorem assumes that this holds in the limit but this does not matter). By rotational invariance, we can assume without loss of generality that all the signal is in the first coordinate; that is,

$$\mathbb{P}(y_i = 1 | \boldsymbol{x}_i) = \sigma(\beta_0 + \gamma_0 x_{i1})$$

since this leaves invariant the joint distribution of $(\boldsymbol{x}_i, y_i)$.

At this point, it is useful to introduce some notation. Let $(X_1, \ldots, X_p)$ be independent standard normals. Then

$$(\boldsymbol{x}_i, y_i) \stackrel{\mathrm{d}}{=} (X_1, \ldots, X_p; Y),$$

where $\mathbb{P}(Y = 1 | X_1, \ldots, X_p) = \sigma(\beta_0 + \gamma_0 X_1)$. It thus follows that

$$(3.1) \qquad (y_i, y_i \boldsymbol{x}_i) \stackrel{\mathrm{d}}{=} (Y, V, X_2, \ldots, X_p), \qquad \begin{aligned} & Y, V \sim F_{\beta_0, \gamma_0}, \\ & (X_2, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{p-1}), \\ & (Y, V) \perp\!\!\!\perp (X_2, \ldots, X_p). \end{aligned}$$

This yields a useful characterization.

PROPOSITION 1. *Let the n-dimensional vectors $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ be n i.i.d. copies of $(Y, V, X_2, \ldots, X_p)$ distributed as in* (3.1). *Then if $p < n - 1$,*

$$(3.2) \qquad \mathbb{P}\{\textit{no MLE}\} = \mathbb{P}\{\mathrm{span}(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) \cap \mathbb{R}_+^n \neq \{\mathbf{0}\}\}.$$

Here and below, $\mathbb{R}_+^n$ is the nonnegative orthant.

PROOF. We have seen that there is no MLE if there exists $(b_0, b_1, \ldots, b_p) \neq \mathbf{0}$ such that

$$(3.3) \qquad b_0 \boldsymbol{Y} + b_1 \boldsymbol{V} + b_2 \boldsymbol{X}_2 + \cdots + b_p \boldsymbol{X}_p \geq \mathbf{0}.$$

By (3.1), this says that the chance there is no MLE is the chance of the event (3.3). Under our assumptions, the probability that the $(p - 1)$ dimensional subspace spanned by $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$ nontrivially intersects a fixed subspace of dimension 2 is zero. Since $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ are independent, this means that we have equality in (3.3) with probability zero. □

3.1.2. *Convex cones.* We are interested in rewriting (3.2) in a slightly different form. For a fixed subspace $\mathcal{W} \subset \mathbb{R}^n$, introduce the convex cone

$$\mathcal{C}(\mathcal{W}) = \{\boldsymbol{w} + \boldsymbol{u} : \boldsymbol{w} \in \mathcal{W}, \boldsymbol{u} \geq \boldsymbol{0}\}. \tag{3.4}$$

This is a polyhedral cone, which shall play a crucial role in our analysis. As we will see, the MLE does not exist if $\operatorname{span}(\boldsymbol{X}_2, \dots, \boldsymbol{X}_p)$ intersects the cone $\mathcal{C}(\operatorname{span}(\boldsymbol{Y}, \boldsymbol{V}))$ in a nontrivial way.

PROPOSITION 2. *Set* $\mathcal{L} = \operatorname{span}(\boldsymbol{X}_2, \dots, \boldsymbol{X}_p)$ *and* $\mathcal{W} = \operatorname{span}(\boldsymbol{Y}, \boldsymbol{V})$. *Let* {*No MLE Single*} *be the event that we can either completely or quasi-separate the data points by using the intercept and the first variable only*: *that is*, $\mathcal{W} \cap \mathbb{R}^n_+ \neq \{\boldsymbol{0}\}$. *We have*

$$\begin{aligned}\mathbb{P}\{\textit{no MLE}\} &= \mathbb{P}\bigl\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\} \textit{ and } \{\textit{No MLE Single}\}^c\bigr\} \\ &\quad + \mathbb{P}\{\textit{No MLE Single}\}.\end{aligned} \tag{3.5}$$

An immediate consequence is this:

$$0 \leq \mathbb{P}\{\textit{no MLE}\} - \mathbb{P}\bigl\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\bigr\} \leq \mathbb{P}\{\textit{No MLE Single}\}. \tag{3.6}$$

PROOF. If {No MLE Single} occurs, the data is separable and there is no MLE. Assume, therefore, that {No MLE Single} does not occur. We know from Proposition 1 that we do not have an MLE if and only if we can find a nonzero vector $(b_0, b_1, \dots b_p)$ such that

$$b_0\boldsymbol{Y} + b_1\boldsymbol{V} + b_2\boldsymbol{X}_2 + \cdots + b_p\boldsymbol{X}_p = \boldsymbol{u}, \qquad \boldsymbol{u} \geq \boldsymbol{0}, \boldsymbol{u} \neq \boldsymbol{0}.$$

By assumption, $b_0\boldsymbol{Y} + b_1\boldsymbol{V} = \boldsymbol{u}$ cannot hold. Therefore, $b_2\boldsymbol{X}_2 + \cdots + b_p\boldsymbol{X}_p$ is a nonzero element of $\mathcal{C}(\mathcal{W})$. This gives (3.5) from which (3.6) easily follows. □

We have thus reduced matters to checking whether $\mathcal{L}$ intersects $\mathcal{C}(\mathcal{W})$ in a nontrivial way. This is because we know that under our model assumptions, the chance that we can separate the data via a univariate model is exponentially decaying in n; that is, the chance that there is $(b_0, b_1) \neq 0$ such that $y_i(b_0 + b_1x_{i1}) \geq 0$ for all i is exponentially small. We state this formally below.

LEMMA 2. *In the setting of Theorem* 2.1, *the event* {*No MLE Single*} *occurs with exponentially small probability.*

PROOF. We only sketch the argument. We are in a univariate model with $\mathbb{P}(y_i = 1|x_i) = \sigma(\beta_0 + \gamma_0x_i)$ and x_i i.i.d. $\mathcal{N}(0, 1)$. Fix $t_0 \in \mathbb{R}$. Then it is easy to see that the chance that t_0 separates the x_i's is exponentially small in n. However, when the complement occurs, the data points overlap and no separation is possible. □

It follows from Lemma 2 and (3.6) that $\mathbb{P}(\text{no MLE}) \to 0$ if and only if $\mathbb{P}\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\} \to 0$.

3.2. *Proof of Theorem* 2.1. To prove our main result, we need to understand when a random subspace $\mathcal{L}$ with uniform orientation intersects $\mathcal{C}(\operatorname{span}(\boldsymbol{Y}, \boldsymbol{V}))$ in a nontrivial way. For a fixed subspace $\mathcal{W} \subset \mathbb{R}^n$, the approximate kinematic formula [2], Theorem I, from the literature on convex geometry tells us that for any $\varepsilon \in (0, 1)$,

$$\begin{aligned} p - 1 + \delta\bigl(\mathcal{C}(\mathcal{W})\bigr) > n + a_\varepsilon\sqrt{n} &\quad\Longrightarrow\quad \mathbb{P}\bigl\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\bigr\} \geq 1 - \varepsilon, \\ p - 1 + \delta\bigl(\mathcal{C}(\mathcal{W})\bigr) < n - a_\varepsilon\sqrt{n} &\quad\Longrightarrow\quad \mathbb{P}\bigl\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\bigr\} \leq \varepsilon. \end{aligned} \tag{3.7}$$

We can take $a_\varepsilon = \sqrt{8\log(4/\varepsilon)}$. Above, $\delta(\mathcal{C})$ is the *statistical dimension* of a convex cone $\mathcal{C}$ defined as

$$\delta(\mathcal{C}) := \mathbb{E}\|\Pi_{\mathcal{C}}(\boldsymbol{Z})\|^2 = n - \mathbb{E}\|\boldsymbol{Z} - \Pi_{\mathcal{C}}(\boldsymbol{Z})\|^2, \qquad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \tag{3.8}$$

where $\Pi_{\mathcal{C}}$ is the projection onto $\mathcal{C}$.

We develop a formula for the statistical dimension of the cone $\mathcal{C}(\mathcal{W})$ of interest to us.

LEMMA 3. *Fix $\mathcal{W} \subset \mathbb{R}^n$. Then with $\boldsymbol{Z}$ distributed as in* (3.8),

$$\delta(\mathcal{C}(\mathcal{W})) = n - \mathbb{E}\Big\{\min_{\boldsymbol{w}\in\mathcal{W}} \|(\boldsymbol{w} - \boldsymbol{Z})_+\|^2\Big\}. \tag{3.9}$$

PROOF. By definition, $\delta(\mathcal{C}(\mathcal{W})) = n - \mathbb{E}\,\mathrm{dist}^2(\boldsymbol{Z}, \mathcal{C}(\mathcal{W}))$, where for a fixed $\boldsymbol{z} \in \mathbb{R}^n$, $\mathrm{dist}^2(\boldsymbol{z}, \mathcal{C}(\mathcal{W}))$ is the optimal value of the quadratic program

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{z} - \boldsymbol{w} - \boldsymbol{u}\|^2 \\ \text{subject to} & \boldsymbol{w} \in \mathcal{W}, \\ & \boldsymbol{u} \geq \boldsymbol{0}. \end{array}$$

For any $\boldsymbol{w} \in \mathcal{W}$, the optimal value of $\boldsymbol{u}$ is given by $(\boldsymbol{z} - \boldsymbol{w})_+$. Hence, the optimal value of the program is

$$\min\nolimits_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{z} - \boldsymbol{w} - (\boldsymbol{z} - \boldsymbol{w})_+\|^2 = \min\nolimits_{\boldsymbol{w}\in\mathcal{W}} \|(\boldsymbol{w} - \boldsymbol{z})_+\|^2. \qquad \square$$

We claim that this lemma combined with the theorem below establish Theorem 2.1.

THEOREM 3.1. *Let $(\boldsymbol{Y}, \boldsymbol{V})$ be n i.i.d. samples from F_{β_0,γ_0}. The random variable*

$$Q_n := \min_{t_0,t_1\in\mathbb{R}} \frac{1}{n} \|(t_0\boldsymbol{Y} + t_1\boldsymbol{V} - \boldsymbol{Z})_+\|^2$$

obeys

$$Q_n \xrightarrow{\mathbb{P}} h_{\mathrm{MLE}}(\beta_0, \gamma_0) = \min_{t_0,t_1}\{\mathbb{E}(t_0Y + t_1V - Z)_+^2\}. \tag{3.10}$$

In fact, we establish the stronger statement $Q_n = h_{\mathrm{MLE}}(\beta_0, \gamma_0) + O_P(n^{-1/2})$.

Below, we let $\mathcal{F}$ be the σ-algebra generated by $\boldsymbol{Y}$ and $\boldsymbol{V}$. Set $\varepsilon_n = n^{-\alpha}$ for some positive α, $a_n = \sqrt{8\alpha\log(4n)}$, and define the events

$$A_n = \{p/n > \mathbb{E}\{Q_n|\mathcal{F}\} + a_n n^{-1/2}\}, \qquad E_n = \{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\}.$$

We first show that if $\kappa > h_{\mathrm{MLE}}(\beta_0, \gamma_0)$, then $\mathbb{P}\{\text{no MLE}\} \to 1$ or, equivalently, $\mathbb{P}\{E_n\} \to 1$. Our geometric arguments (3.7) tell us that if A_n occurs, then $\mathbb{P}\{E_n \mid \mathcal{F}\} \geq 1 - \varepsilon_n$. This means that

$$\mathbb{1}\{A_n\} \leq \mathbb{1}\{\mathbb{P}\{E_n \mid \mathcal{F}\} \geq 1 - \varepsilon_n\} \leq \mathbb{P}\{E_n \mid \mathcal{F}\} + \varepsilon_n.$$

Taking expectation gives

$$\mathbb{P}\{E_n\} \geq \mathbb{P}\{A_n\} - \varepsilon_n.$$

Next, we claim that

$$\mathbb{E}\{Q_n|\mathcal{F}\} \xrightarrow{\mathbb{P}} h_{\mathrm{MLE}}(\beta_0, \gamma_0). \tag{3.11}$$

This concludes the proof since (3.11) implies that $\mathbb{P}\{A_n\} \to 1$ and, therefore, $\mathbb{P}\{E_n\} \to 1$. The argument showing that if $\kappa < h_{\mathrm{MLE}}(\beta_0, \gamma_0)$, then $\mathbb{P}\{\text{no MLE}\} \to 0$ is entirely similar and omitted.

It remains to justify (3.11). Put $h = h_{\mathrm{MLE}}(\beta_0, \gamma_0)$ for short (this is a nonrandom quantity), and note that $Q_n - h$ is uniformly integrable (this is because Q_n is the minimum of an average of n i.i.d. subexponential variables). Hence, if Q_n converges in probability, it also converges in mean in the sense that $\mathbb{E}|Q_n - h| \to 0$. Since

$$\big|\mathbb{E}\{Q_n|\mathcal{F}\} - h\big| \le \mathbb{E}\big\{|Q_n - h| \mid \mathcal{F}\big\},$$

we see that taking expectation on both sides yields that $\mathbb{E}\{Q_n|\mathcal{F}\}$ converges to h in mean and, therefore, in probability (since convergence in means implies convergence in probability).

3.3. *About the approximate kinematic formula.* The approximate kinematic formula (3.7) [2], Theorem I, arises from a set of deep ideas in conic integral geometry. A classical problem in this field is to study when a randomly rotated convex cone shares a ray with a fixed convex cone. Although explicit formulas were established in [16], they were not immediately mathematically tractable. Several works subsequently developed variants that are more useful and [2] ultimately derived an approximate kinematic formula by utilizing the statistical dimension of cones. This approach is connected to Gordon's escape through the mesh lemma [7]; we refer to [2] for a detailed exposition.

4. Sub-Gaussian designs. While this paper was under review, the referees asked whether our main conclusions apply more broadly, and we indeed expect Theorem 2.1 to hold under a class of covariate distributions with sub-Gaussian tails. We sketch the proof of a simple extension but do not examine more complicated situations in this paper. We also provide some numerical simulations offering support for our belief.

4.1. *Empirical support.* To begin with, consider features $\boldsymbol{x}_i$ with entries drawn i.i.d. from the Rademacher distribution (each entry is equally likely to take on the values 1 and -1). Keeping everything else as in Section 2.4, we repeat the experiments of Figure 2 and the results are shown in Figure 4. We observe a perfect agreement of the theoretical and empirical phase transition curves in both cases, which corroborates our belief.

We next study the phase transition behavior under a more general covariate distribution, which is loosely inspired by genome-wide association studies. In such studies, the features are single nucleotide polymorphisms (SNPs), which count the number of occurrences of a reference allele at various locations along the genome (the reference allele of course depends on the location). When the jth SNP is in Hardy–Weinberg equilibrium, the chances of

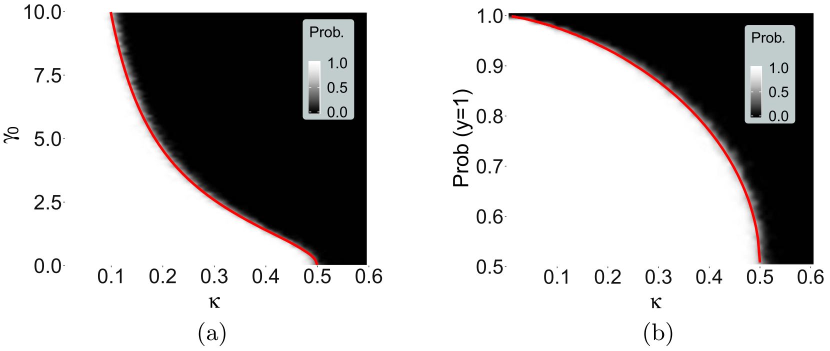

FIG. 4. *Empirical probability that the MLE exists* (*black is zero, and white is one*) *estimated from* 50 *independent trials for each "pixel" under the Rademacher model. Panels* (a) *and* (b) *mirror Figures* 2(a) *and* 2(b).

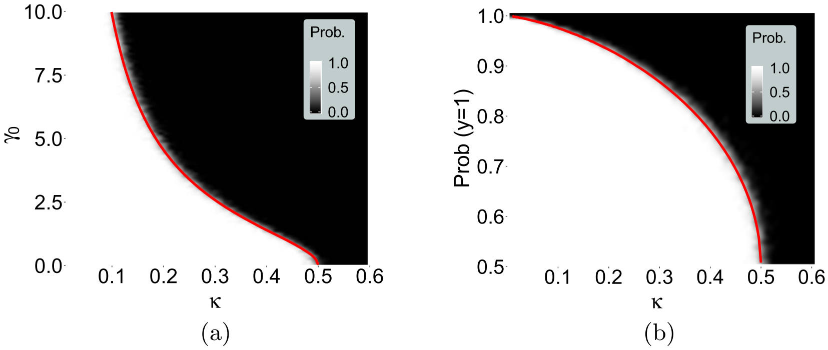

FIG. 5. *Empirical probability that the MLE exists* (*black is zero, and white is one*) *estimated from* 50 *independent trials for each "pixel" under a GWAS inspired model. Panels* (a) *and* (b) *mirror Figures* 2(a) *and* 2(b).

observing 0, 1 and 2 are respectively p_j^2, $2p_j(1-p_j)$ and $(1-p_j)^2$, where $p_j \in (0,1)$. Consider independent features generated from such marginal distributions with parameters p_j varying in $[0.25, 0.75]$, and center each feature to have zero mean. This results in independent variables X_j, each taking on three possible values and obeying $\mathbb{E}X_j = 0$, $\mathrm{Var}(X_j) = 2p_j(1-p_j)$. From here on, repeat the experiments from Section 2.4 with regression coefficients $\beta_j = \sqrt{\gamma_0^2/p\,\mathrm{Var}(X_j)}$ so that $\mathrm{Var}(\boldsymbol{x_i}'\boldsymbol{\beta}) = \gamma_0^2$ (recall $\gamma = \sqrt{\beta_0^2 + \gamma_0^2}$). The results are shown in Figure 5. Once again, we observe a perfect agreement between the theoretical and empirical phase transition curves.

4.2. *Theoretical support.* Our methods establish that our results continue to hold in the special case of independent symmetric sub-Gaussian variables—we assume the explanatory variables are nondegenerate, that is, are not identically zero—and $\gamma_0 = 0$; that is to say, $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = \sigma(\beta_0)$ (this is the setting of the first part of Lemma 1). This is possible because the central piece of our argument is the approximate kinematic formula (3.7), which has been generalized to a broad class of independent variables.

To see why this special extension holds, we begin by rehearsing our result for Gaussian variables. Recall that the key is to understand whether a p-dimensional subspace $\mathcal{L} = \mathrm{span}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_p)$ spanned by the feature vectors intersects $\mathcal{C}(\mathrm{span}(\boldsymbol{Y}))$ in a nontrivial way (here, we may take $\mathcal{W}$ to be the span of $\boldsymbol{Y}$ because there is no signal). We have seen that the probability of a nontrivial intersection is essentially either 0 or 1 depending on the value of the statistical dimension of $\mathcal{C}(\mathrm{span}(\boldsymbol{Y}))$ (3.7). The crucial point is that a version of the same kinematic formula (3.7) continues to hold if the n-dimensional vectors $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_p)$ are i.i.d. copies of $(X_1, \dots, X_p)$ where our explanatory variables are independent, symmetric, nondegenerate and obey certain moment assumptions [14], Theorem I. Since the statistical dimension does not depend upon the distribution of the explanatory variables and that

$$1 - \frac{\delta(\mathcal{C}(\mathrm{span}(\boldsymbol{Y})))}{n} \xrightarrow{\mathbb{P}} \min_{t \in \mathbb{R}} \mathbb{E}(tY - Z_+)^2 = h_{\mathrm{MLE}}(\beta_0, 0),$$

we readily see that the same phase transition holds.

In sum, the arguments above prove that the phase transition curves from Figures 4(b) and 5(b) are asymptotically correct. Of course, extending our results to broader settings is likely to be considerably more involved.

5. Proof of Theorem 3.1. We begin by introducing some notation to streamline our exposition as much as possible. Define the mapping $J : \boldsymbol{x} \mapsto \|\boldsymbol{x}_+\|^2/2$ and let $\boldsymbol{A}$ be the $n \times 2$

matrix with $\boldsymbol{y}$ and $\boldsymbol{V}$ as columns. Next, define the random function F and its expectation f as

$$F(\boldsymbol{\lambda}) = n^{-1} J(\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{Z}), \qquad f(\boldsymbol{\lambda}) = \mathbb{E} F(\boldsymbol{\lambda}).$$

F is convex and it is not hard to see that f is strictly convex (we will see later that it is, in fact, strongly convex). Let $\boldsymbol{\lambda}_\star$ be any minimizer of F ($\boldsymbol{\lambda}_\star$ is a random variable) and $\boldsymbol{\lambda}_0$ be the unique minimizer of f ($\boldsymbol{\lambda}_0$ is not random and finite). With this notation, Theorem 3.1 asks us to prove that

$$F(\boldsymbol{\lambda}_\star) = f(\boldsymbol{\lambda}_0) + O_P(n^{-1/2}) \tag{5.1}$$

and in the rest of this section, we present the simplest argument we could think of.

We begin by recording some simple properties of F and f. It follows from $\nabla J(\boldsymbol{x}) = \boldsymbol{x}_+$ that ∇J is Lipschitz and obeys

$$\|\nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{x}_0)\| \le \|\boldsymbol{x} - \boldsymbol{x}_0\|.$$

Consequently, F is also Lipschitz with constant at most $n^{-1}\|\boldsymbol{A}\|^2 \le n^{-1}(\|\boldsymbol{y}\|^2 + \|\boldsymbol{V}\|^2) = 1 + n^{-1}\|\boldsymbol{V}\|^2$. It is also a straightforward calculation to see that f is twice differentiable with Hessian given by

$$\nabla^2 f(\boldsymbol{\lambda}) = n^{-1}\mathbb{E}\{\boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}\}, \qquad \boldsymbol{D} = \operatorname{diag}(\mathbb{1}\{\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{Z} \ge \boldsymbol{0}\}).$$

It follows that with $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)$, the Hessian is given by

$$\nabla^2 f(\boldsymbol{\lambda}) = \begin{bmatrix} \mathbb{E}\{Y^2\Phi(\lambda_0 Y + \lambda_1 V)\} & \mathbb{E}\{YV\Phi(\lambda_0 Y + \lambda_1 V)\} \\ \mathbb{E}\{YV\Phi(\lambda_0 Y + \lambda_1 V)\} & \mathbb{E}\{V^2\Phi(\lambda_0 Y + \lambda_1 V)\} \end{bmatrix}, \tag{5.2}$$

where (Y, V) is distributed as in Theorem 2.1 and Φ is the cdf of a standard normal. We claim that for fixed (β_0, γ_0), it holds that

$$\alpha_0 \boldsymbol{I}_2 \preceq \nabla^2 f(\boldsymbol{\lambda}) \preceq \alpha_1 \boldsymbol{I}_2, \tag{5.3}$$

uniformly over $\boldsymbol{\lambda}$, where α_0, α_1 are fixed positive numerical constant (that may depend on (β_0, γ_0)).

Next, we claim that for a *fixed* $\boldsymbol{\lambda}$, $F(\lambda)$ does not deviate much from its expectation $f(\boldsymbol{\lambda})$. This is because $F(\boldsymbol{\lambda})$ is an average of subexponential variables which are i.i.d. copies of $(\lambda_0 Y + \lambda_1 V - Z)_+^2$; classical bounds [21], Corollary 5.17, give

$$\begin{aligned} &\mathbb{P}\{|F(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda})| \ge t\} \\ &\quad \le 2\exp\left(-c_0 n \min\left(\frac{t^2}{c_1^2(1 + \|\boldsymbol{\lambda}\|^2)^2}, \frac{t}{c_1(1 + \|\boldsymbol{\lambda}\|^2)}\right)\right), \end{aligned} \tag{5.4}$$

where c_0, c_1 are numerical constants. Also, $\nabla F(\lambda)$ does not deviate much from its expectation $\nabla f(\boldsymbol{\lambda})$ either because this is also an average of sub-exponential variables. Hence, we also have

$$\begin{aligned} &\mathbb{P}\{\|\nabla F(\boldsymbol{\lambda}) - \nabla f(\boldsymbol{\lambda})\| \ge t\} \\ &\quad \le 2\exp\left(-c_2 n \min\left(\frac{t^2}{c_3^2(1 + \|\boldsymbol{\lambda}\|^2)^2}, \frac{t}{c_3(1 + \|\boldsymbol{\lambda}\|^2)}\right)\right), \end{aligned} \tag{5.5}$$

where c_2, c_3 are numerical constants. In the sequel, we shall make a repeated use of the inequalities (5.4)–(5.5).

With these preliminaries in place, we can turn to the proof of (5.1). On the one hand, the convexity of F gives

$$F(\boldsymbol{\lambda}_\star) \ge F(\boldsymbol{\lambda}_0) + \langle \nabla F(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0 \rangle. \tag{5.6}$$

On the other hand, since ∇F is Lipschitz, we have the upper bound

$$F(\boldsymbol{\lambda}_\star) \le F(\boldsymbol{\lambda}_0) + \langle \nabla F(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\rangle + (1 + \|\boldsymbol{V}\|^2/n)\|\boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\|^2. \tag{5.7}$$

Now observe that (5.4) gives that

$$F(\boldsymbol{\lambda}_0) = f(\boldsymbol{\lambda}_0) + O_P(n^{-1/2}).$$

Also, since $\nabla f(\boldsymbol{\lambda}_0) = \mathbf{0}$, (5.5) gives

$$\|\nabla F(\boldsymbol{\lambda}_0)\| = O_P(n^{-1/2}).$$

Finally, since $\|\boldsymbol{V}\|^2/n \xrightarrow{\mathbb{P}} \mathbb{E}V^2$, we see from (5.6) and (5.7) that (5.1) holds if $\|\boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\| = O_P(n^{-1/4})$.

LEMMA 4. *We have* $\|\boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\| = O_P(n^{-1/4})$.

PROOF. The proof is inspired by an argument in [3]. For any $\boldsymbol{\lambda} \in \mathbb{R}^2$, (5.3) gives

$$f(\boldsymbol{\lambda}) \ge f(\boldsymbol{\lambda}_0) + \frac{\alpha_0}{2}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|^2.$$

Fix $x \ge 1$. For any $\boldsymbol{\lambda}$ on the circle $C(x) := \{\lambda \in \mathbb{R}^2 : \|\lambda - \lambda_0\| = xn^{-1/4}\}$ centered at $\boldsymbol{\lambda}_0$ and of radius $xn^{-1/4}$, we have

$$f(\boldsymbol{\lambda}) \ge f(\boldsymbol{\lambda}_0) + 3y, \qquad y = \frac{\alpha_0 x^2}{6\sqrt{n}}. \tag{5.8}$$

Fix $z = f(\boldsymbol{\lambda}_0) + y$ and consider the event E defined as

$$F(\boldsymbol{\lambda}_0) < z \quad \text{and} \quad \inf_{\lambda \in C(x)} F(\boldsymbol{\lambda}) > z. \tag{5.9}$$

By convexity of F, when E occurs, $\boldsymbol{\lambda}_\star$ must lie inside the circle and, therefore, $\|\boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\| \le xn^{-1/4}$.

It remains to show that E occurs with high probability. Fix d equispaced points $\{\boldsymbol{\lambda}_i\}_{i=1}^d$ on $C(x)$. Next, take any point $\boldsymbol{\lambda}$ on the circle and let $\boldsymbol{\lambda}_i$ be its closest point. By convexity,

$$F(\boldsymbol{\lambda}) \ge F(\boldsymbol{\lambda}_i) + \langle \nabla F(\boldsymbol{\lambda}_i), \boldsymbol{\lambda} - \boldsymbol{\lambda}_i\rangle \ge F(\boldsymbol{\lambda}_i) - \|\nabla F(\boldsymbol{\lambda}_i)\|\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\|. \tag{5.10}$$

On the one hand, $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\| \le \pi xn^{-1/4}/d$. On the other, by (5.5) we know that if we define B as

$$B := \Bigl\{\max_i \|\nabla F(\boldsymbol{\lambda}_i) - \nabla f(\boldsymbol{\lambda}_i)\|_2 \ge xn^{-1/2}\Bigr\}$$

then

$$\mathbb{P}\{B^c\} \le 2d \exp\biggl(-c_2 \min\biggl(\frac{x^2}{c_3^2(1 + \max_i \|\boldsymbol{\lambda}_i\|^2)^2}, \frac{\sqrt{n}x}{c_3(1 + \max_i \|\boldsymbol{\lambda}_i\|^2)}\biggr)\biggr). \tag{5.11}$$

Also, since $\|\nabla^2 f\|$ is bounded (5.3) and $\nabla f(\boldsymbol{\lambda}_0) = 0$,

$$\|\nabla f(\boldsymbol{\lambda}_i)\|_2 \le \alpha_1 \|\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_0\| = \alpha_1 xn^{-1/4}.$$

For n sufficiently large, this gives that on B,

$$\|\nabla F(\boldsymbol{\lambda}_i)\|\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\| \le Cy/d$$

for some numerical constant C. Choose $d \ge C$. Then it follows from (5.10) that on B,

$$\inf_{\lambda \in C(x)} F(\boldsymbol{\lambda}) \ge \min_i F(\boldsymbol{\lambda}_i) - y.$$

It remains to control the right-hand side above. To this end, observe that

$$F(\boldsymbol{\lambda}_i) > f(\boldsymbol{\lambda}_i) - y \quad \Longrightarrow \quad F(\boldsymbol{\lambda}_i) - y > f(\boldsymbol{\lambda}_0) + y = z$$

since $f(\boldsymbol{\lambda}_i) \geq f(\boldsymbol{\lambda}_0) + 3y$ by (5.8). Hence, the complement of the event E in (5.9) has probability at most

$$\mathbb{P}\{E^c\} \leq \mathbb{P}\{B^c\} + \mathbb{P}\{F(\boldsymbol{\lambda}_0) \geq f(\boldsymbol{\lambda}_0) + y\} + \sum_{i=1}^{d} \mathbb{P}\{F(\boldsymbol{\lambda}_i) \leq f(\boldsymbol{\lambda}_i) - y\}.$$

The application of (5.11) and that of (5.4) to the last two terms in the right-hand side concludes the proof. □

6. Conclusion.

In this paper, we established the existence of a phase transition for the existence of the MLE in a high-dimensional logistic model with Gaussian covariates. We derived a simple expression for the phase-transition boundary when the model is fitted with or without an intercept. Our methods use elements of convex geometry, especially the kinematic formula reviewed in Section 3.2, which is a modern version of Gordon's escape through a mesh theorem [7]. It is likely that the phenomena and formulas derived in this paper hold for more general covariate distributions, and we leave this to future research.

Acknowledgments. The first author would like to thank Stephen Bates and Nikolaos Ignatiadis for useful comments about an earlier version of the paper.

The first author was supported in part by Office of Naval Research Grant N00014-16-1-2712, NSF Grant DMS-1712800, by the Math + X Award from the Simons Foundation and by a generous gift from TwoSigma.

The second author was supported in part by the Ric Weiland Graduate Fellowship in the School of Humanities and Sciences, Stanford University.

REFERENCES

[1] ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

[2] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[3] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982 https://doi.org/10.1214/14-AOS1254

[4] CHRISTMANN, A. and ROUSSEEUW, P. J. (2001). Measuring overlap in binary regression. *Comput. Statist. Data Anal.* **37** 65–75. MR1862480 https://doi.org/10.1016/S0167-9473(00)00063-3

[5] COVER, T. M. (1964). Geometrical and statistical properties of linear threshold devices. Ph.D. thesis.

[6] COVER, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **3** 326–334.

[7] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric Aspects of Functional Analysis* (1986/87). *Lecture Notes in Math.* **1317** 84–106. Springer, Berlin. MR0950977 https://doi.org/10.1007/BFb0081737

[8] KAUFMANN, H. (1988). On existence and uniqueness of a vector minimizing a convex function. *Z. Oper.-Res.* **32** 357–373. MR0976377 https://doi.org/10.1007/BF01920035

[9] KOLASSA, J. E. (1997). Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scand. J. Stat.* **24** 523–530. MR1615343 https://doi.org/10.1111/1467-9469.00078

[10] KONIS, K. (2007). Linear programming algorithms for detecting separated data in binary logistic regression models. Ph.D. thesis, Univ. Oxford.

[11] LESAFFRE, E. and ALBERT, A. (1989). Partial separation in logistic discrimination. *J. Roy. Statist. Soc. Ser. B* **51** 109–116. MR0984997

[12] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition [of MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

[13] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A* **135** 370–384.

[14] Oymak, S. and Tropp, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Inf. Inference* **7** 337–446. MR3858331 https://doi.org/10.1093/imaiai/iax011

[15] Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73** 755–758. MR0897873 https://doi.org/10.1093/biomet/73.3.755

[16] Schneider, R. and Weil, W. (2008). *Stochastic and Integral Geometry*. Springer, Berlin. MR2455326 https://doi.org/10.1007/978-3-540-78859-1

[17] Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* **43** 310–313. MR0637943

[18] Silvapulle, M. J. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48** 100–106. MR0848055

[19] Sur, P. and Candès, E. J. (2018). A modern maximum-likelihood theory for high-dimensional logistic regression. ArXiv Preprint ArXiv:1803.06964.

[20] Sur, P., Chen, Y. and Candès, E. J. (2018). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Related Fields*. To Appear.

[21] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170