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

By

Emmanuel J. Candès
Pragya Sur

Technical Report No. 2018-03
April 2018

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

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

By

Emmanuel J. Candès
Pragya Sur
Stanford University

Technical Report No. 2018-03
April 2018

This research was supported in part by Office of Naval Research grant N00014-16-1-2712 and by National Science Foundation grant DMS 1712800.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

The Phase Transition for the Existence of the Maximum Likelihood Estimate in High-dimensional Logistic Regression

Emmanuel J. Candès*† Pragya Sur*

April 25, 2018

Abstract

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 $\kappa > h_{\text{MLE}}$, then the MLE does not exist with probability one. Conversely, if $\kappa < h_{\text{MLE}}$, the MLE asymptotically exists with probability one.

1 Introduction

Logistic regression [12, 13] is perhaps the most widely used and studied non-linear 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 $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, 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.

*Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
†Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.

1.1 Data geometry and the existence of the MLE

The delicacy of ML 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

$$y_i \boldsymbol{x}_i' \boldsymbol{b} > 0, \text{ for all } i, \tag{1}$$

then the MLE does not exist. To be clear, (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 $\boldsymbol{b} \neq \boldsymbol{0}$, there is at least one data point that is classified correcly ($y_i \boldsymbol{x}_i' \boldsymbol{b} > 0$) 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 $\boldsymbol{b} \neq \boldsymbol{0}$,

$$y_i \boldsymbol{x}_i' \boldsymbol{b} \geq 0, \text{ for all } i, \tag{2}$$

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 [15], 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 [14] 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, 16]. 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; i.e. $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = 1/2$. 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, [17] 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 [17] 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 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 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:

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

where the covariance $\boldsymbol{\Sigma}$ is non-singular 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

$$\mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta}) \to \gamma_0^2. \tag{3}$$

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,

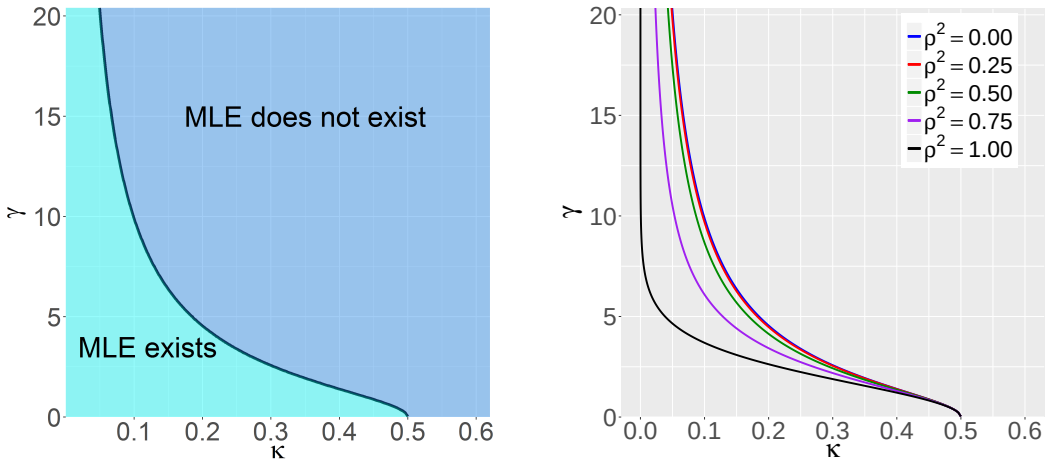
$$\sqrt{\mathbb{E}(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta})^2} \to \sqrt{\beta_0^2 + \gamma_0^2} =: \gamma. \tag{4}$$

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

Our main result is that there is an explicit function h_{MLE} given in (6) such that

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

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



(a) (b)

Figure 1: Theoretical predictions from (6). (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 $\rho^2 = 0.25$.

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.

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

$$(Y, V) \sim F_{\beta_0,\gamma_0} \quad \text{if} \quad (Y, V) \overset{\mathrm{d}}{=} (Y, YX), \tag{5}$$

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

$$h_{\mathrm{MLE}}(\beta_0, \gamma_0) = \min_{t_0,t_1\in\mathbb{R}} \left\{\mathbb{E}(t_0 Y + t_1 V - Z)_+^2\right\}, \tag{6}$$

where $x_+ = \max(x, 0)$ and we write $x_+^2 = (x_+)^2$ for short. Then in the setting from Section 1.4,

$$\begin{aligned} \kappa > h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad &\Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 0, \\ \kappa < h_{\mathrm{MLE}}(\beta_0, \gamma_0) \quad &\Longrightarrow \quad \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 1. \end{aligned}$$

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

$$\begin{aligned} p/n > h_{\mathrm{MLE}}(\beta_0, \gamma_0) + \lambda_n n^{-1/2} &\implies \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 0, \\ p/n < h_{\mathrm{MLE}}(\beta_0, \gamma_0) - \lambda_n n^{-1/2} &\implies \lim_{n,p\to\infty} \mathbb{P}\{\text{MLE exists}\} = 1. \end{aligned}$$

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 non-negative 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 (6) 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 $\gamma_0 \to \infty$, $V \overset{\mathrm{d}}{\to} |Z'|$, $Z' \sim \mathcal{N}(0,1)$. Hence, plugging $t_0 = 0$ into (6) gives

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

  If $\beta_0 \to \infty$, $Y \overset{\mathrm{d}}{\to} 1$ and plugging $t_1 = 0$ into (6) 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 (6) in other special cases below.

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

$$h_{\mathrm{MLE}}(\beta_0, 0) = \min_{t \in \mathbb{R}} \left\{ \mathbb{E}(tY - Z)_+^2 \right\}. \tag{7}$$

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

$$h_{\mathrm{MLE}}(0, \gamma_0) = \min_{t \in \mathbb{R}} \left\{ \mathbb{E}(tV - Z)_+^2 \right\}. \tag{8}$$

Proof: Consider the first assertion. In this case, it follows from the definition (5) that $(Y, V) \stackrel{\mathrm{d}}{=} (Y, X)$ where Y and X are independent, $\mathbb{P}(Y = 1) = \sigma(\beta_0)$ and $X \sim \mathcal{N}(0, 1)$. Hence,

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

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

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

Hence, in this case, the miminum in (6) 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 1, only simpler. It is, therefore, omitted.

Theorem 2. *Assume $\beta_0 = 0$ and consider fitting a model without an intercept. If V has the marginal distribution from Theorem 1 and is independent from $Z \sim \mathcal{N}(0, 1)$, then the conclusions from Theorem 1 hold with the phase-transition curve given in (8). 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 setup, it can be shown that *quasi-complete separation* occurs with zero probability). To detect separability, we study whether the program [10]

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n y_i (b_0 + \boldsymbol{x}_i' \boldsymbol{b}) \\ \text{subject to} & y_i (b_0 + \boldsymbol{x}_i' \boldsymbol{b}) \geq 0,\ i = 1, \ldots, n \\ & -1 \leq b_0 \leq 1, -\mathbf{1} \leq \boldsymbol{b} \leq \mathbf{1} \end{array} \tag{9}$$

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 (9).

Below, each simulated data set follows a logistic model with $n = 4,000$, $p = \kappa\, n$, 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

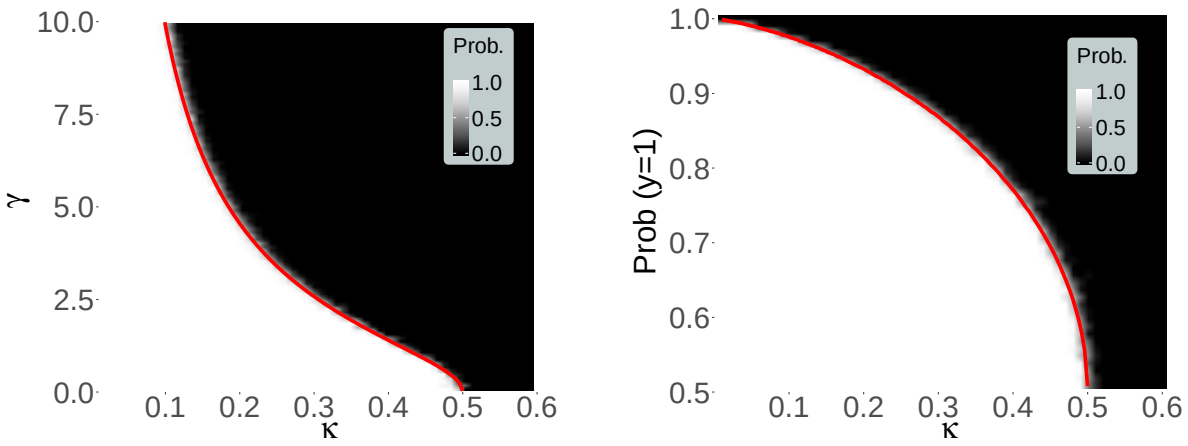


Figure 2: Empirical probability that the MLE exists (black is zero, and white is one) estimated from 50 independent trials for each 'pixel'. (a) Model without intercept in which $\beta_0 = 0$ and $\gamma_0 = \gamma$, with theoretical phase transition curve from (8) in red (this is the same curve as in Figure 1(a)). (b) Model with $\gamma_0 = 0$, $\beta_0 = \gamma$ and the theoretical phase transition curve from (7) in red. The y-axis is here chosen to be the marginal probability $\mathbb{P}(y_i = 1) = e^{\gamma}/(1 + e^{\gamma})$.

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, i.e. $\mathbb{P}(y_i = 1) = e^{\gamma}/(1 + e^{\gamma})$. 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 $\kappa = 0.255$.

3 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.

3.1 Gaussian covariates

Write $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. 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

$$y_i \left(b_0 + \boldsymbol{z}_i' \boldsymbol{\Sigma}^{1/2} \boldsymbol{b}\right) \geq 0, \quad \forall i.$$

This is equivalent to the existence of $(b_0, \boldsymbol{\theta}) \neq \mathbf{0}$ such that $y_i \left(b_0 + \boldsymbol{z}_i' \boldsymbol{\theta}\right) \geq 0$ for all i. In words, multiplication by a non-singular 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 non-singular 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

$$(y_i, y_i\, \boldsymbol{x}_i) \stackrel{\mathrm{d}}{=} (Y, V, X_2, \ldots, X_p), \qquad \begin{array}{l} Y, V \sim F_{\beta_0, \gamma_0}, \\ (X_2, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{p-1}), \\ (Y, V) \perp\!\!\!\perp (X_2, \ldots, X_p). \end{array} \tag{10}$$

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 (10). Then if $p < n - 1$,*

$$\mathbb{P}\{\text{no MLE}\} = \mathbb{P}\{\mathrm{span}(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) \cap \mathbb{R}^n_+ \neq \{\mathbf{0}\}\}. \tag{11}$$

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

$$b_0 \boldsymbol{Y} + b_1 \boldsymbol{V} + b_2 \boldsymbol{X}_2 + \ldots + b_p \boldsymbol{X}_p \geq \mathbf{0}. \tag{12}$$

By (10), this says that the chance there is no MLE is the chance of the event (12). Under our assumptions, the probability that the $(p-1)$ dimensional subspace spanned by $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$ non-trivially 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 (12) with probability zero. ■

3.2 Convex cones

We are interested in rewriting (11) in a slightly different form. For a fixed subspace $\mathcal{W} \subset \mathbb{R}^n$, introduce the convex cone

$$\mathcal{C}(\mathcal{W}) = \{\boldsymbol{w} + \boldsymbol{u} : \boldsymbol{w} \in \mathcal{W}, \boldsymbol{u} \geq \boldsymbol{0}\}. \tag{13}$$

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

Proposition 2. *Set* $\mathcal{L} = \text{span}(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ *and* $\mathcal{W} = \text{span}(\boldsymbol{Y}, \boldsymbol{V})$. *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: i.e.* $\mathcal{W} \cap \mathbb{R}^n_+ \neq \{\boldsymbol{0}\}$. *We have*

$$\mathbb{P}\{\text{no MLE}\} = \mathbb{P}\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\} \text{ and } \{\text{No MLE Single}\}^c\} + \mathbb{P}\{\text{No MLE Single}\}. \tag{14}$$

An immediate consequence is this:

$$0 \leq \mathbb{P}\{\text{no MLE}\} - \mathbb{P}\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\} \leq \mathbb{P}\{\text{No MLE Single}\}. \tag{15}$$

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 $(b_0, b_1, \ldots b_p)$ such that

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

By assumption, $b_0\boldsymbol{Y} + b_1\boldsymbol{V} = \boldsymbol{u}$ cannot hold. Therefore, $b_2\boldsymbol{X}_2 + \ldots + b_p\boldsymbol{X}_p$ is a non-zero element of $\mathcal{C}(\mathcal{W})$. This gives (14) from which (15) easily follows. ■

We have thus reduced matters to checking whether $\mathcal{L}$ intersects $\mathcal{C}(\mathcal{W})$ in a non-trivial 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 $y_i(b_0 + b_1 x_{i1}) \geq 0$ for all i is exponentially small. We state this formally below.

Lemma 2. *In the setting of Theorem 1, the event* {No MLE Single} *occurs with exponentially small probability.*

Proof: We only sketch the argument. We are in a univariate model with $\mathbb{P}(y_i = 1|x_i) = \sigma(\beta_0 + \gamma_0 x_i)$ 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. ■

3.3 Proof of Theorem 1

To prove our main result, we need to understand when a random subspace $\mathcal{L}$ with uniform orientation intersects $\mathcal{C}(\text{span}(\boldsymbol{Y}, \boldsymbol{V}))$ 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 $\epsilon \in (0, 1)$

$$\begin{aligned} p - 1 + \delta(\mathcal{C}(\mathcal{W})) > n + a_\epsilon\sqrt{n} \quad &\Longrightarrow \quad \mathbb{P}\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\} \geq 1 - \epsilon \\ p - 1 + \delta(\mathcal{C}(\mathcal{W})) < n - a_\epsilon\sqrt{n} \quad &\Longrightarrow \quad \mathbb{P}\{\mathcal{L} \cap \mathcal{C}(\mathcal{W}) \neq \{\boldsymbol{0}\}\} \leq \epsilon. \end{aligned} \tag{16}$$

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

$$\delta(\mathcal{C}) := \mathbb{E}\,\|\Pi_{\mathcal{C}}(\boldsymbol{Z})\|^2 = n - \mathbb{E}\,\|\boldsymbol{Z} - \Pi_{\mathcal{C}}(\boldsymbol{Z})\|^2, \quad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \tag{17}$$

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 (17),*

$$\delta(\mathcal{C}(\mathcal{W})) = n - \mathbb{E}\left\{\min_{\boldsymbol{w}\in\mathcal{W}} \|(\boldsymbol{w} - \boldsymbol{Z})_+\|^2\right\}. \tag{18}$$

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

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

■

We claim that this lemma combined with the theorem below establish Theorem 1.

Theorem 3. *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

$$Q_n \xrightarrow{\mathbb{P}} h_{\mathrm{MLE}}(\beta_0, \gamma_0) = \min_{t_0,t_1} \left\{\mathbb{E}\,(t_0 Y + t_1 V - Z)_+^2\right\}. \tag{19}$$

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 $\epsilon_n = n^{-\alpha}$ for some positive α, $a_n = \sqrt{8\alpha\log(4n)}$, and define the events

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

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 (16) tell us that if A_n occurs, then $\mathbb{P}\{E_n \,|\, \mathcal{F}\} \geq 1 - \epsilon_n$. This means that

$$1\{A_n\} \leq 1\{\mathbb{P}\{E_n \,|\, \mathcal{F}\} \geq 1 - \epsilon_n\} \leq \mathbb{P}\{E_n \,|\, \mathcal{F}\} + \epsilon_n.$$

Taking expectation gives

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

Next we claim that

$$\mathbb{E}\{Q_n|\mathcal{F}\} \xrightarrow{\mathbb{P}} h_{\mathrm{MLE}}(\beta_0, \gamma_0). \tag{20}$$

This concludes the proof since (20) 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 (20). Put $h = h_{\mathrm{MLE}}(\beta_0, \gamma_0)$ for short (this is a non-random 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. sub-exponential 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

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

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).

4 Proof of Theorem 3

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

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

Both F and f are 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 asks us to prove that

$$F(\boldsymbol{\lambda}_\star) = f(\boldsymbol{\lambda}_0) + O_P(n^{-1/2}) \tag{21}$$

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

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

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

$$\nabla^2 f(\boldsymbol{\lambda}) = \begin{bmatrix} \mathbb{E}\{Y^2\Phi(\lambda_0 Y + \lambda_1 V)\} & \mathbb{E}\{YV\Phi(\lambda_0 Y + \lambda_1 V)\} \\ \mathbb{E}\{YV\Phi(\lambda_0 Y + \lambda_1 V)\} & \mathbb{E}\{V^2\Phi(\lambda_0 Y + \lambda_1 V)\} \end{bmatrix}, \tag{22}$$

where (Y, V) is distributed as in Theorem 1 and Φ is the cdf of a standard normal. We claim that for fixed (β_0, γ_0), it holds that

$$\alpha_0 \boldsymbol{I}_2 \preceq \nabla^2 f(\boldsymbol{\lambda}) \preceq \alpha_1 \boldsymbol{I}_2, \tag{23}$$

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 sub-exponential variables which are i.i.d. copies of $(\lambda_0 Y + \lambda_1 V - Z)_+^2$; classical bounds [18, Corollary 5.17] give

$$\mathbb{P}\{|F(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda})| \geq t\} \leq 2\exp\left(-c_0 n \min\left(\frac{t^2}{c_1^2(1+\|\boldsymbol{\lambda}\|^2)^2}, \frac{t}{c_1(1+\|\boldsymbol{\lambda}\|^2)}\right)\right), \tag{24}$$

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

$$\mathbb{P}\{\|\nabla F(\boldsymbol{\lambda}) - \nabla f(\boldsymbol{\lambda})\| \geq t\} \leq 2\exp\left(-c_2 n \min\left(\frac{t^2}{c_3^2(1+\|\boldsymbol{\lambda}\|^2)^2}, \frac{t}{c_3(1+\|\boldsymbol{\lambda}\|^2)}\right)\right), \tag{25}$$

where c_2, c_3 are numerical constants. In the sequel, we shall make a repeated use of the inequalities (24)–(25).

With these preliminaries in place, we can turn to the proof of (21). On the one hand, the convexity of F gives

$$F(\boldsymbol{\lambda}_\star) \geq F(\boldsymbol{\lambda}_0) + \langle \nabla F(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0 \rangle. \tag{26}$$

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

$$F(\boldsymbol{\lambda}_\star) \leq F(\boldsymbol{\lambda}_0) + \langle \nabla F(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0 \rangle + (1 + \|\boldsymbol{V}\|^2/n)\|\boldsymbol{\lambda}_\star - \boldsymbol{\lambda}_0\|^2. \tag{27}$$

Now observe that (24) gives that

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

Also, since $\nabla f(\boldsymbol{\lambda}_0) = \mathbf{0}$, (25) 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 (26) and (27) that (21) 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 $\boldsymbol{\lambda} \in \mathbb{R}^2$, (23) gives

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

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

$$f(\boldsymbol{\lambda}) \geq f(\boldsymbol{\lambda}_0) + 3y, \qquad y = \frac{\alpha_0 x^2}{6\sqrt{n}}. \tag{28}$$

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

$$F(\boldsymbol{\lambda}_0) < z \quad \text{and} \quad \inf_{\lambda \in C(x)} F(\boldsymbol{\lambda}) > z. \tag{29}$$

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

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,

$$F(\boldsymbol{\lambda}) \geq F(\boldsymbol{\lambda}_i) + \langle \nabla F(\boldsymbol{\lambda}_i), \boldsymbol{\lambda} - \boldsymbol{\lambda}_i \rangle \geq F(\boldsymbol{\lambda}_i) - \|\nabla F(\boldsymbol{\lambda}_i)\| \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\|. \tag{30}$$

On the one hand, $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\| \leq \pi x n^{-1/4}/d$. On the other, by (25) we know that if we define B as

$$B := \left\{ \max_i \|\nabla F(\boldsymbol{\lambda}_i) - \nabla f(\boldsymbol{\lambda}_i)\|_2 \geq x n^{-1/2} \right\}$$

then

$$\mathbb{P}\{B^c\} \leq 2d \exp\left(-c_2 \min\left(\frac{x^2}{c_3^2 (1 + \max_i \|\boldsymbol{\lambda}_i\|^2)^2}, \frac{\sqrt{n} x}{c_3 (1 + \max_i \|\boldsymbol{\lambda}_i\|^2)} \right) \right). \tag{31}$$

Also, since $\|\nabla^2 f\|$ is bounded (23) and $\nabla f(\boldsymbol{\lambda}_0) = 0$,

$$\|\nabla f(\boldsymbol{\lambda}_i)\|_2 \leq \alpha_1 \|\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_0\| = \alpha_1 \, x n^{-1/4}.$$

For n sufficiently large, this gives that on B,

$$\|\nabla F(\boldsymbol{\lambda}_i)\| \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_i\| \leq C \, y/d$$

for some numerical constant C. Choose $d \geq C$. Then it follows from (30) that on B,

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

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 (28). Hence, the complement of the event E in (29) 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 (31) and that of (24) to the last two terms in the right-hand side concludes the proof. ■

5 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 3.3, 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.

Acknowledgements

P. S. was partially supported by the Ric Weiland Graduate Fellowship in the School of Humanities and Sciences, Stanford University. E. C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, by the National Science Foundation via DMS 1712800, by the Math + X Award from the Simons Foundation and by a generous gift from TwoSigma. E. C. would like to thank Stephen Bates and Nikolaos Ignatiadis for useful comments about an early version of the paper.

References

[1] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[3] Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

[4] Andreas Christmann and Peter J Rousseeuw. Measuring overlap in binary regression. *Computational Statistics & Data Analysis*, 37(1):65–75, 2001.

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

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

[7] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In Joram Lindenstrauss and Vitali D. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.

[8] Heinz Kaufmann. On existence and uniqueness of a vector minimizing a convex function. *Zeitschrift für Operations Research*, 32(6):357–373, 1988.

[9] John E Kolassa. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scandinavian Journal of Statistics*, 24(4):523–530, 1997.

[10] Kjell Konis. *Linear programming algorithms for detecting separated data in binary logistic regression models*. PhD thesis, University of Oxford, 2007.

[11] Emmanuel Lesaffre and Adelin Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society. Series B Methodological*, 51(1):109–116, 1989.

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

[13] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[14] Thomas J Santner and Diane E Duffy. A note on A. Albert and JA Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3):755–758, 1986.

[15] Mervyn J Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 310–313, 1981.

[16] Mervyn J Silvapulle and J Burridge. Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 100–106, 1986.

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

[18] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, pages 210 – 268, 2012.