ASYMPTOTIC DISTRIBUTION OF THE MAXIMAL LIKELIHOOD ESTIMATE IN HIGH DIMENSIONAL LOGISTIC REGRESSION

Xiaoyi Mai [1], *Zhenyu Liao* [1] *and Romain Couillet* [2,1]

[1]CentraleSupélec, University Paris-Saclay and [2]GIPSA-lab, University of Grenoble-Alpes

ABSTRACT

Logistic regression, as one of the most popular machine learning algorithms today, has been long believed to be unbiased in performing binary classification tasks. In this paper, we consider a "hard" classification problem of separating high dimensional Gaussian data, where the data dimension p and the sample size n are both large. Build on recent advances of random matrix theory and high dimensional statistics, we evaluate the asymptotic distribution of the logistic regression classifier and consequently, provide the associated test performance. This brings new insights into the internal mechanism of logistic regression classifier, including the possible bias in the separating hyperplane, as well as on practical issues such as hyper-parameter tuning.

***Index Terms*—** high-dimensional statistic, logistic regression, machine learning, random matrices

1. INTRODUCTION

Most of classical results in the statistical learning theory concern the regime where the sample size n is overwhelmingly larger than the feature dimension p. Following the extensive findings derived under this traditional setting of $n \gg p$, understanding statistical learning methods when n and p are commensurately large is of growing interest among researchers, and became the subject of a series of contributions [1, 2, 3, 4, 5, 6, 7, 8]. Such investigations are particularly relevant under the current big data paradigm, where it can be inaccurate to assume the validity of classical regime $n \gg p$.

Indeed, as already shown in the literature, some long-held common beliefs supported by classical results break down when n, p are comparably large. For instance, recent study [9] sheds new light on the high dimensional behavior of logistic regression, or more precisely on the maximal likelihood estimate $\boldsymbol{\beta}$ obtained by maximizing the posterior probability $\mathbb{P}(y|\mathbf{x}) = \sigma(y\boldsymbol{\beta}_*^\mathsf{T}\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^p$ being the feature vector, $y \in \{-1, 1\}$ the binary target variable, and $\sigma(t) = \frac{1}{1+e^{-t}}$ the *logistic sigmoid* function, over a set of training data $\{\mathbf{x}_i, y_i\}$ where $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The authors of [9] show that, for commensurately large n, p, the maximal likelihood estimate is biased, contradicting the expectation of classical theory that the maximal likelihood estimate is asymptotically unbiased. Also, its variability is greater than commonly predicted. Consequently, the commonly used procedure for testing the significance of the regression coefficients need to be adjusted for an improved accuracy.

Inspired by the work of [9], this article aims to provide the asymptotic distributions of $\boldsymbol{\beta}$ under a more practical data model with no constraint on the independence of features and a natural mixture structure of linking components to the classes (as presented in Section 2). Additionally, in order to incorporate into the analysis framework the situations where the maximal likelihood estimation is an ill-posed problem without unique solution, a Tikhonov regularization term of adjustable weight λ is included to the objective function. According to [10], such situations are bound to occur in high dimensions when the dimensionality ratio p/n is above a certain threshold c_{th}.

Besides corroborating the findings in [9] about the bias and inflated variability of $\boldsymbol{\beta}$ in an extended setting, the theoretical results in this paper also point out that even when the estimation problem is well-posed, it is beneficial to the generalization performance to use a regularized solution, in spite of a even more biased $\boldsymbol{\beta}$. More surprisingly, when the data covariance is identity matrix, the optimal generalization performance is actually achieved at $\lambda \to \infty$, suggesting the usage of a significantly emphasized regularization in practice.

Analyzing learning systems in the regime where n is comparable to p often involve advanced technical arguments, the learned parameters no longer converge to deterministic limits as in the classical setting of $n \gg p$, and is instead intricately related to the input training data, especially when the learning system admits no explicit solution like logistic regression. To capture the asymptotic statistical property of implicit learning methods, many related works [1, 11, 9] rely on the "double leave-one-out" approach, based on the procedure of eliminating one data sample/feature from the input dataset. The applicability of the leave-one-feature-out discussion in these works however depends on the independence and statistical equivalence of data features, which does not hold for the generalized model under study. A new strategy combining arguments from random matrix theory and the leave-one-observation-out procedure is therefore derived for this analysis.

The article is organized as follows. The analysis frame-

work is presented in Section 2, along with a brief review of the technical approach leading up to the final results, which combines arguments from leave-one-out procedure, random matrix theory and convex optimization.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices), and scalars non-boldface respectively. The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices.

2. MODEL AND ASSUMPTIONS

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independent vectors from two balanced distribution classes $\mathcal{C}_1, \mathcal{C}_2$ (so that $\mathcal{C}_1$ and $\mathcal{C}_2$ both have cardinality $n/2$). We assume the data vectors $\mathbf{x}_i \in \mathcal{C}_a$ for $a \in \{1, 2\}$ follows a Gaussian mixture model such that

$$\mathbf{x}_i \sim \mathcal{N}((-1)^a \boldsymbol{\mu}, \mathbf{C})$$

for some mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ with associated labels $y_i = -1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $y_i = 1$ if $\mathbf{x}_i \in \mathcal{C}_2$. To achieve an asymptotically non-trivial misclassification rates, we shall (as in [12]) work under the following assumptions:

Assumption 1 (Growth rate)**.** *As $n \to \infty$, $p/n \to c > 0$. Besides, $\|\boldsymbol{\mu}\| = O(1)$ and $\|\mathbf{C}\| = O(1)$.*

Note that $\{\mathbf{x}_i, y_i\}$ follows a logistic regression model as

$$\begin{aligned} P(y_i|\mathbf{x}_i) &= \frac{P(y_i)P(\mathbf{x}_i|y_i)}{P(y_i)P(\mathbf{x}_i|y_i) + P(-y_i)P(\mathbf{x}_i|-y_i)} \\ &= \frac{1}{1+e^{2y_i\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\mathbf{x}_i}} = \sigma(y_i\boldsymbol{\beta}_*^\mathsf{T}\mathbf{x}_i) \end{aligned}$$

with $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$ and $\sigma(t) = \frac{1}{1+e^{-t}}$ the logistic sigmoid function. The maximal likelihood estimate $\boldsymbol{\beta}$ is thus the solution of

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{1}$$

where $\rho(t) = \ln(1+e^{-t})$, $\tilde{\mathbf{x}}_i = y_i\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$.

To investigate the asymptotic performance of the logistic regression classifier, it is of crucial importance to understand the statistical property of $\boldsymbol{\beta}$. The main technical difficulty of this analysis lies in the fact that $\boldsymbol{\beta}$, as the solution of an optimization problem, does not have an explicit form. Nonetheless, by vanishing the derivation of the loss function with respect to $\boldsymbol{\beta}$ we obtain the following implicit relation

$$\lambda\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n c_i\tilde{\mathbf{x}}_i, \quad c_i \equiv \psi(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i) \tag{2}$$

where we denote $\psi(t) \equiv -\frac{\partial\rho(t)}{\partial t} = \frac{1}{1+e^t}$.

Therefore, $\boldsymbol{\beta}$ can be seen as a linear combination of all $\tilde{\mathbf{x}}_i$'s, weighted by the coefficient c_i. The idea is to understand how $\tilde{\mathbf{x}}_i$ (and its statistical property) affects the corresponding coefficient c_i (or more precisely, $\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i$). However, as the solution of (1), $\boldsymbol{\beta}$ depends on all $\tilde{\mathbf{x}}_i$'s in an intricate manner. To handle this correlation, we first establish a "leave-one-out" version of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}_{-i}$, that is asymptotically close to $\boldsymbol{\beta}$ and independent of $\tilde{\mathbf{x}}_i$, by solving (1) for all $\tilde{\mathbf{x}}_j$, $j \neq i$. Then by further characterizing the relation between c_i and $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ (the latter being a Gaussian random variable since $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$), we obtain the asymptotic distribution of $\boldsymbol{\beta}$.

3. MAIN RESULTS

Before getting into our main theoretical results, we shall first set up the crucial system of equations that determines the asymptotic distribution of $\boldsymbol{\beta}$.

For a standard Gaussian random variable $x \sim \mathcal{N}(0,1)$ define $c = g_\kappa(\sigma x + m)$ as the unique solution of

$$c = \psi(\sigma x + m + c\kappa) \tag{3}$$

with $(m, \sigma^2, \kappa) \in \mathbb{R}_+^3$ the solution of the following system of fixed-point equations

$$\begin{cases} m \equiv \eta\boldsymbol{\mu}^\mathsf{T}\left(\lambda\mathbf{I}_p + \tau\mathbf{C}\right)^{-1}\boldsymbol{\mu} \\ \sigma^2 \equiv \eta^2\boldsymbol{\mu}^\mathsf{T}\left(\lambda\mathbf{I}_p + \tau\mathbf{C}\right)^{-1}\mathbf{C}\left(\lambda\mathbf{I}_p + \tau\mathbf{C}\right)^{-1}\boldsymbol{\mu} \\ \quad + \frac{\gamma}{n}\operatorname{tr}\left(\left(\lambda\mathbf{I}_p + \tau\mathbf{C}\right)^{-1}\mathbf{C}\right)^2 \\ \tau \equiv -\frac{1}{\sigma}\mathbb{E}_x g_\kappa\left(\sigma x + m\right)x = -\mathbb{E}_x g_\kappa'(\sigma x + m) \\ \eta \equiv \mathbb{E}_x g_\kappa\left(\sigma x + m\right) \\ \gamma \equiv \mathbb{E}_x g_\kappa^2\left(\sigma x + m\right) \\ \kappa \equiv \frac{1}{n}\operatorname{tr}\left(\lambda\mathbf{I}_p - \mathbb{E}_x\left[\frac{\psi'(\sigma x+m+\kappa g_\kappa(\sigma x+m))}{1-\kappa\psi'(\sigma x+m+\kappa g_\kappa(\sigma x+m))}\right]\mathbf{C}\right)^{-1}\mathbf{C} \end{cases} \tag{4}$$

where $\psi'(t) \equiv \frac{\partial\psi(t)}{\partial t} = -\frac{e^t}{(1+e^t)^2}$.

We are now in position to present the main technical result of this article as follows.

Theorem 1 (Distribution of $\boldsymbol{\beta}$)**.** *Let Assumption 1 holds, then we have*

$$\left(\lambda\mathbf{I}_p + \tau\mathbf{C}\right)^{-1}\boldsymbol{\beta} \xrightarrow{d} \mathcal{N}(\eta\boldsymbol{\mu}, c\gamma\mathbf{C})$$

with $(\tau, \eta, \gamma) \in \mathbb{R}_+^3$ given by (4)*.*

Sketch of proof. As discussed at the end of Section 2, we shall connect $\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i$ to c_i by establishing a "leave-one-out" version of $\boldsymbol{\beta}$ that is independent of $\mathbf{x}_i$. To this end, we denote $\boldsymbol{\beta}_{-i}$ the solution of (1) with $\tilde{\mathbf{X}}_{-i} \equiv [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{i-1}, \tilde{\mathbf{x}}_{i+1}, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{p\times(n-1)}$, all $n-1$ training data except $\{\mathbf{x}_i, y_i\}$ such that $\boldsymbol{\beta}_{-i} = \frac{1}{n}\sum_{j\neq i}\psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\tilde{\mathbf{x}}_j$. The difference $\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}$ is therefore given by

$$\begin{aligned} \lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}) &= \frac{1}{n}\sum_{j\neq i}\left(c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\right)\tilde{\mathbf{x}}_j + \frac{1}{n}c_i\tilde{\mathbf{x}}_i \\ &= \frac{1}{n}\tilde{\mathbf{X}}_{-i}\boldsymbol{\Delta}\mathbf{c}_{-i} + \frac{1}{n}c_i\tilde{\mathbf{x}}_i \end{aligned} \tag{5}$$

with $\boldsymbol{\Delta}\mathbf{c}_{-i} \in \mathbb{R}^{n-1}$ the vector with its j-th entry equal to $c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)$ for all $j \neq i$.

Note that under Assumption 1, $\|\tilde{\mathbf{x}}_i\|$ is of order $O(\sqrt{p})$ with high probability, so that if we assume all c_i's to be of order $O(1)$, we deduce from (2) that $\|\boldsymbol{\beta}\|$ (so as $\|\boldsymbol{\beta}_{-i}\|$) is of order $O(1)$ and from (5) that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_{-i}\|$ is of order $O(1/\sqrt{p})$. Moreover, use the fact that $\psi(t)$ is Lipschitz continuous we have $c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j) = O(1/\sqrt{p})$ that is smaller compared to c_j, which further allows for the following estimate,

$$\begin{aligned} c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j) &= \psi'(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)(\boldsymbol{\beta}-\boldsymbol{\beta}_{-i})^\mathsf{T}\tilde{\mathbf{x}}_j + O(1/p) \\ &= \psi'(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\frac{1}{n\lambda}\left(\tilde{\mathbf{x}}_j^\mathsf{T}\tilde{\mathbf{X}}_{-i}\boldsymbol{\Delta}\mathbf{c}_{-i} + c_i\tilde{\mathbf{x}}_j^\mathsf{T}\tilde{\mathbf{x}}_i\right) + O(1/p) \end{aligned}$$

by performing a Taylor expansion of $\psi(t)$ around $t = \boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j$, with $\psi'(t) \equiv \frac{\partial\psi(t)}{\partial t} = -\frac{e^t}{(1+e^t)^2} < 0$. Assembling the $n-1$ equations for $j \neq i$ and we get

$$\boldsymbol{\Delta}\mathbf{c}_{-i} = \left(\lambda\mathbf{I}_{n-1} - \frac{1}{n}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\tilde{\mathbf{X}}_{-i}\right)^{-1}\frac{1}{n}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i c_i + o_p(1)$$

with $\mathbf{D}_{-i} \in \mathbb{R}^{n-1}$ the diagonal matrix with its (j,j)-entry equal to $\psi'(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)$. Plugging in the above expression of $\boldsymbol{\Delta}\mathbf{c}_i$ into (5) we derive

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_{-i})^\mathsf{T}\tilde{\mathbf{x}}_i = c_i\frac{1}{n}\tilde{\mathbf{x}}_i^\mathsf{T}\left(\lambda\mathbf{I}_p - \frac{1}{n}\tilde{\mathbf{X}}_{-i}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\right)^{-1}\tilde{\mathbf{x}}_i + o_p(1). \tag{6}$$

Note that the RHS of (6) is the quadratic form of $\frac{1}{n}\tilde{\mathbf{x}}_i^\mathsf{T}\mathbf{M}\tilde{\mathbf{x}}_i$ for some $\mathbf{M}$ of bounded operator norm and independent of $\tilde{\mathbf{x}}_i$, classical RMT results yield the following approximation.

Lemma 1 (Asymptotic approximation of quadratic form)**.** *Let Assumption 1 holds, then with probability one,*

$$\frac{1}{n}\tilde{\mathbf{x}}_i^\mathsf{T}\left(\lambda\mathbf{I}_p - \frac{1}{n}\tilde{\mathbf{X}}_{-i}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\right)^{-1}\tilde{\mathbf{x}}_i - \kappa \to 0$$

where κ is the unique solution of $\kappa = \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{C})$ with $\bar{\mathbf{Q}} \equiv \left(\lambda I_p - \mathbb{E}\left[\frac{\psi'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}})}{1-\kappa\psi'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}})}\right]\mathbf{C}\right)^{-1}$.

From Lemma 1 and (2)-(5) we obtain the implicit relation $c_i = \psi(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i) = \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i + c_i\kappa)$, the unique solution of which, if exists, is defined via the following function g_κ as[1]

$$c_i = g_\kappa(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i)$$

and therefore the re-expression of $\boldsymbol{\beta}$ from (2) as

$$\lambda\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n g_\kappa\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i\right)\tilde{\mathbf{x}}_i. \tag{7}$$

As discussed at the end of Section 2, since $\boldsymbol{\beta}_{-i}$ is independent of $\tilde{\mathbf{x}}_i$, $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ is a Gaussian random variable of mean $\boldsymbol{\mu}^\mathsf{T}\mathbb{E}[\boldsymbol{\beta}_{-i}]$ and variance $\mathbb{E}[\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{C}\boldsymbol{\beta}_{-i}]$ that are asymptotically close to

$$m \equiv \boldsymbol{\mu}^\mathsf{T}\mathbb{E}[\boldsymbol{\beta}], \quad \sigma^2 \equiv \mathbb{E}[\boldsymbol{\beta}^\mathsf{T}\mathbf{C}\boldsymbol{\beta}] = \operatorname{tr}\left(\mathbf{C}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}]\right) \tag{8}$$

so that the statistical property of $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ is naturally connected to those of $\boldsymbol{\beta}$.

However, note that in (7) we have $g_\kappa(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i)$ highly depends on $\tilde{\mathbf{x}}_i$ so that $\mathbb{E}[\boldsymbol{\beta}]$ is still not easily accessible. To address this issue, we further "separate" the dependence of $\tilde{\mathbf{x}}_i = \boldsymbol{\mu} + \mathbf{z}_i$ from $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ by writing

$$\mathbf{z}_i = \tilde{\mathbf{z}}_i + \frac{\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{C}\boldsymbol{\beta}_{-i}}\mathbf{C}\boldsymbol{\beta}_{-i} = \tilde{\mathbf{z}}_i + \frac{\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{z}_i}{\sigma^2}\mathbf{C}\boldsymbol{\beta}_{-i} + o_p(1)$$

so that $\mathbb{E}[(\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{z}_i)\tilde{\mathbf{z}}_i] = \mathbf{0}$ with $\mathbb{E}[\tilde{\mathbf{z}}_i] = \mathbf{0}$, $\mathbb{E}[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^\mathsf{T}] = \mathbf{C} - \frac{1}{\sigma^2}\mathbf{C}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}]\mathbf{C}$. As a consequence, (7) can be decomposed as

$$\lambda\boldsymbol{\beta} - \frac{1}{n}\sum_{i=1}^n g_\kappa\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i\right)\frac{\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{z}_i}{\sigma^2}\mathbf{C}\boldsymbol{\beta}_{-i} = \frac{1}{n}\sum_{i=1}^n g_\kappa\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i\right)(\boldsymbol{\mu}+\tilde{\mathbf{z}}_i)$$

which yields the following relation

$$(\lambda\mathbf{I}_p + \tau\mathbf{C})\,\boldsymbol{\beta} = \eta\boldsymbol{\mu} + \mathbf{u} + o_p(1)$$

with $(\tau, \eta, \gamma) \in \mathbb{R}_+^3$ given by (4), $\mathbf{u} \equiv \frac{1}{n}\sum_{i=1}^n c_i\tilde{\mathbf{z}}_i$ and

$$\mathbb{E}[\mathbf{u}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T}] = \frac{\gamma}{n}\left(\mathbf{C} - \frac{1}{\sigma^2}\mathbf{C}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}]\mathbf{C}\right)$$

and therefore

$$\boldsymbol{\beta} = \eta\,(\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\,\boldsymbol{\mu} + (\lambda\mathbf{I}_p + \tau\mathbf{C})^{-1}\,\mathbf{u} + o_p(1)$$

with concludes the proof of Theorem 1. □

A direct consequence of Theorem 1 is that, for a logistic regression classifier determined by $\boldsymbol{\beta}_{-i}$, the misclassification probability for an unseen $\tilde{\mathbf{x}}_i$ is given by the probability $P\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i < 0\right)$ that has an asymptotically deterministic behavior as given in the following corollary.

Corollary 1 (Test performance)**.** *Let Assumption 1 holds. Then the test performance of the classifier measured by the misclassification rate is given by*

$$P\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i < 0\right) - Q\left(\frac{m}{\sigma}\right) = o_p(1)$$

with the Q-function: $Q(t) \equiv \frac{1}{\sqrt{2\pi}}\int_t^\infty \exp(-u^2/2)du$.

Fig 3 reports the empirical distribution of $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ versus a Gaussian distribution $\mathcal{N}(m, \sigma^2)$ derived from (4) for one realization. We observe a close match of the asymptotic results on finite data samples with not too large n, p.

[1] This is equivalent to solve $\frac{1}{1+e^x} = ax + b$ for x, the solution of which exists and is unique for a away from zero.

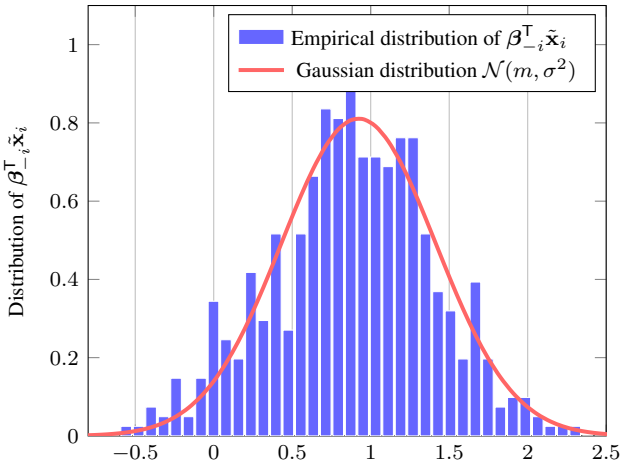

Fig. 1. Comparison between $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ and Gaussian distribution $\mathcal{N}(m,\sigma^2)$ as defined in (4) with $\boldsymbol{\mu}=[2,\ \mathbf{0}_{p-1}]$, $\mathbf{C}=\mathbf{I}_p$ for $\lambda=1$, $p=256$ and $n=512$.

4. DISCUSSIONS

4.1. Solution without regularization

The unregularized solution, which, if exists, is retrieved by taking $\lambda=0$ in the results of Theorem 1, giving rise to

$$\boldsymbol{\beta} \xrightarrow{d} \mathcal{N}\big((\eta/\tau)\mathbf{C}^{-1}\boldsymbol{\mu}, (\gamma/n\tau^2)\mathbf{C}^{-1}\big),$$

where η,τ,γ, as recall, are positive deterministic values given by the system of equations in (4). Denote $\boldsymbol{\beta}^{\infty}$ the maximal likelihood estimate obtained in the limit of large samples in which p is fixed and $n\to\infty$. It is well known [13] that

$$\boldsymbol{\beta}^{\infty} \xrightarrow{d} \mathcal{N}\big(\boldsymbol{\beta}_*, (1/n)\mathcal{I}(\boldsymbol{\beta}_*)^{-1}\big),$$

where $\mathcal{I}(\boldsymbol{\beta}_*)=(1/4)\mathbf{C}$ is Fisher information matrix evaluated at the true $\boldsymbol{\beta}_*=2\mathbf{C}^{-1}\boldsymbol{\mu}$. A first remark is that the high dimensional maximal likelihood $\boldsymbol{\beta}$ is biased in the manner that its expectation converges to a rescaled version of the true $\boldsymbol{\beta}_*$, as opposed to the unbiased large samples estimate $\boldsymbol{\beta}^{\infty}$. Secondly, as

$$\frac{\mathrm{Cov}(\boldsymbol{\beta})}{\mathbb{E}(\boldsymbol{\beta})\mathbb{E}(\boldsymbol{\beta})^{\mathsf{T}}} = \frac{\gamma}{\eta^2}\frac{\mathrm{Cov}(\boldsymbol{\beta}^{\infty})}{\mathbb{E}(\boldsymbol{\beta}^{\infty})\mathbb{E}(\boldsymbol{\beta}^{\infty})^{\mathsf{T}}}$$

where $\gamma/\eta^2>1$ for η,γ equal respectively the first and second moment of g_κ according to the forth and fifth equations of (4), we observe an inflated variability of the maximal likelihood estimate in high dimensions. Both remarks are reminiscent of the conclusions of [9] in the setting of $\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$.

4.2. Solution with regularization

It can be observed from Theorem 1 that when $\lambda\neq 0$, the asymptotic expectation of $\boldsymbol{\beta}$ are different from $\boldsymbol{\beta}_*$ in both scale and direction. From a viewpoint of classification, using a rescaled $\boldsymbol{\beta}_*$ achieves the same oracle test performance as $\boldsymbol{\beta}_*$, meaning that only bias in direction are detrimental to the classification performance. However, it is found that the performance is actually improved with regularization, despite the presence of a more severe bias. This is because the quantity γ/η^2, which is indicator of the variability of $\boldsymbol{\beta}$, goes to its minimum for extremely regularized solutions with $\lambda\to\infty$. There is thus a trade-off between learning the correct direction and reducing the randomness for $\boldsymbol{\beta}$ through the tuning of λ. It should be pointed out that while the classification error is usually minimized at a finite λ, as in the case of $\mathbf{C}_2$ covariance in Fig 4.2, it will continually decrease as $\lambda\to\infty$ if the data covariance is identity matrix. A highly regularized solution is therefore favorable in this special case, as observed in the case of $\mathbf{C}_1$ covariance in Fig 4.2. The underlying reason behind this conterintuitive phenomenon is easily understood with our results: the asymptotic expectation of $\boldsymbol{\beta}$ is always aligned to $\boldsymbol{\beta}_*$ for any λ when $\mathbf{C}=\mathbf{I}_p$, it then remains to minimize the variability of $\boldsymbol{\beta}$, which can be achieved for $\lambda\to\infty$.

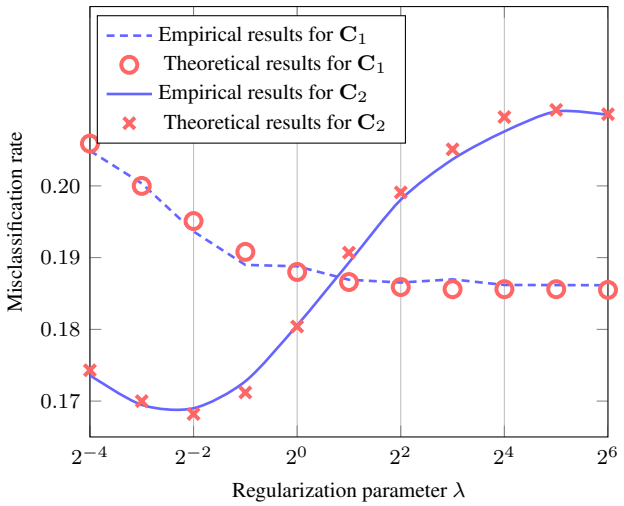

Fig. 2. Misclassification error as a function of λ, with $\boldsymbol{\mu}=[1,\ 1,\ \mathbf{0}_{p-2}]$, $\mathbf{C}_1=2\mathbf{I}_p$ and $\mathbf{C}_2=\mathrm{diag}[1,\ 5,\ \mathbf{1}_{p-2}]$ for $p=128$, $n=512$ and number of test sample $n_{\text{test}}=512$. Empirical results are obtained by averaging over 500 runs.

5. CONCLUDING REMARKS

See if still any space for conclusion.

6. REFERENCES

[1] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, p. 201307842, 2013.

[2] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu, "Optimal m-estimation in high-dimensional regression," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14563–14568, 2013.

[3] Noureddine El Karoui, "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators," *Probability Theory and Related Fields*, vol. 170, no. 1-2, pp. 95–175, 2018.

[4] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[5] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[6] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *arXiv preprint arXiv:1701.02967*, 2017.

[7] Hanwen Huang, "Asymptotic behavior of support vector machine for spiked population model," *Journal of Machine Learning Research*, vol. 18, no. 45, pp. 1–21, 2017.

[8] Cosme Louart, Zhenyu Liao, Romain Couillet, et al., "A random matrix approach to neural networks," *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1190–1248, 2018.

[9] Pragya Sur and Emmanuel J Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *arXiv preprint arXiv:1803.06964*, 2018.

[10] Emmanuel J Candès and Pragya Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *arXiv preprint arXiv:1804.09753*, 2018.

[11] David Donoho and Andrea Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing," *Probability Theory and Related Fields*, vol. 166, no. 3-4, pp. 935–969, 2016.

[12] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai, "Classification Asymptotics in the Random Matrix Regime," in *26th European Signal Processing Conference (EUSIPCO'2018)*. IEEE, 2018.

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