A LARGE SCALE ANALYSIS OF LOGISTIC REGRESSION: ASYMPTOTIC PERFORMANCE AND NEW INSIGHTS

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

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

ABSTRACT

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

Index Terms— High dimensional statistic, logistic regression, machine learning, random matrix theory.

1. INTRODUCTION

Most theoretical results and analyses in statistical learning are derived under the assumption that the sample size n is overwhelmingly larger than the feature dimension p. Under the current big data paradigm, where it is more accurate to assume $n \sim p$ (or even $p \gg n$), understanding the resulting impact of standard statistical learning methods when n and p are commensurately large is becoming a growing research concern in modern statistics [1–8].

Indeed, as already shown several times in the literature, some long-held beliefs supported by classical results (and intuitions) break down when n, p are comparably large, a problem often related to the *curse of dimensionality*. For instance, the recent study [9] sheds new light on the high dimensional behavior of logistic regression, or more precisely on the maximal likelihood estimator $\boldsymbol{\beta}$ obtained by maximizing the posterior probability $P(y|\mathbf{x}) = \sigma(y\boldsymbol{\beta}_*^{\mathsf{T}}\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^p$ 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 $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$, for $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The authors of [9] show that, for commensurately large n, p, the maximal likelihood estimator is biased, contradicting the expectation of classical theory that the maximal likelihood estimator is asymptotically unbiased for $n \gg p$. Also, its variability is greater than commonly predicted. Consequently, the commonly used procedure for significance test of the regression coefficients needs to be adjusted for 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 in the objective function. According to [10], such situations are bound to occur in high dimensional problems when the dimensionality ratio p/n is above a certain threshold c_{th}.

Besides corroborating the findings of [9] in an extended setting, the theoretical results in this paper notably 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 an even more biased $\boldsymbol{\beta}$. More surprisingly, when the data covariance is identity, the optimal generalization performance is actually achieved at $\lambda \to \infty$, suggesting the usage of a significant regularization in practice.

Analyzing learning systems in the regime where n is comparable to p often involves advanced technical arguments: the learned parameters no longer converge to deterministic limits as when $n \gg p$ but are instead intricately related to the training data, especially when the learning system admits no explicit solution as in the case of logistic regression. To capture the asymptotic statistical properties of implicit learning methods, many related works [1,9,11] rely on a "double leave-one-out" approach, based on the procedure of eliminating one data sample and one data feature from the input dataset. The applicability of the leave-one-*feature*-out approach in these works however depends on the independence and statistical equivalence of data features, which no longer holds 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.

Couillet's work is supported by the GSTATS UGA IDEX DataScience chair and the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

The remainder of the article is organized as follows. We establish the system model under study in Section 2 and present our main results in Section 3, while a brief review of the technical approach is deferred to Section 4.

2. PRELIMINARIES

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independent vectors from two distribution classes $\mathcal{C}_1, \mathcal{C}_2$ with balanced class priors. We assume the data vectors $\mathbf{x}_i \in \mathcal{C}_a$ for $a \in \{1, 2\}$ follow a Gaussian mixture model such that

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

for some mean $\boldsymbol{\mu}_p \in \mathbb{R}^p$ and covariance $\mathbf{C}_p \in \mathbb{R}^{p \times p}$ 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 rate, we shall (as in [12]) work under the following assumptions:

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

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

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

with $\boldsymbol{\beta}_* = 2\mathbf{C}_p^{-1}\boldsymbol{\mu}_p$ and $\sigma(t) = \frac{1}{1+e^{-t}}$ the logistic sigmoid function. Note that the existence of the true parameter vector $\boldsymbol{\beta}_*$ arises naturally from the above mixture model. This significantly differs from [13] where the existence of $\boldsymbol{\beta}_*$ is artificially imposed as the data vectors $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ are not statistically separable. The maximal *quadratically regularized* 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 $\rho(t) = \ln(1 + e^{-t})$, $\tilde{\mathbf{x}}_i = y_i \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_p, \mathbf{C}_p)$.

To investigate the asymptotic performance of the logistic regression classifier, it is of crucial importance to understand the statistical properties of $\boldsymbol{\beta}$. The main technical difficulty of this analysis lies in the fact that $\boldsymbol{\beta}$, as the solution of a non-trivial optimization problem, does not have an explicit form. Nonetheless, by cancelling the loss function derivative 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 $\psi(t) \equiv -\frac{\partial \rho(t)}{\partial t} = \frac{1}{1+e^t}$.

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 properties) affects the corresponding coefficient c_i (or more precisely, $\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i$). However, as a solution of (1), $\boldsymbol{\beta}$ depends on all $\tilde{\mathbf{x}}_i$'s in an intricate manner, we handle this correlation by establishing 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 $\tilde{\mathbf{x}}_j$, $j \neq i$. As it turns out in the technical development of Section 4, we can relate c_i to $\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_i$ (the latter being a Gaussian random variable since $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}_p, \mathbf{C}_p)$ and independent of $\boldsymbol{\beta}_{-i}$) through the proximal mapping (which is frequently used in convex optimization [14]), as stated in the lemma below.

Lemma 1. *Under Assumption 1, there exist two positive deterministic constants m, σ^2 such that, as $n, p \to \infty$,*

$$\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_i \xrightarrow{d} r \sim \mathcal{N}(m, \sigma^2)$$

where $\xrightarrow{d}$ denotes the convergence in distribution. Let function $g_\kappa : \mathbb{R} \mapsto \mathbb{R}$ be given by

$$g_\kappa(t) = \psi\left(\mathrm{prox}_\kappa(t)\right)$$

where the proximal mapping $\mathrm{prox}_\kappa(t)$ is defined by $\mathrm{prox}_\kappa(t) = \mathrm{argmin}_{z \in \mathbb{R}} \left(\kappa \rho(z) + \frac{1}{2}(z-t)^2\right)$. Then,

$$c_i \xrightarrow{d} c \sim g_\kappa(r)$$

where κ is a positive deterministic constant dependent of m, σ^2, determined by the following fixed-point equation

$$\kappa \equiv \frac{1}{n} \mathrm{tr} \left(\lambda \mathbf{I}_p - \mathbb{E}\left[\frac{\psi'(r + \kappa g_\kappa(r))}{1 - \kappa \psi'(r + \kappa g_\kappa(r))} \right] \mathbf{C}_p \right)^{-1} \mathbf{C}_p$$

with $\psi'(t) \equiv \frac{\partial \psi(t)}{\partial t} = -\frac{e^t}{(1+e^t)^2} < 0$ and where, for the above inverse is well defined for all $\lambda > 0$.

3. THEORETICAL ANALYSIS

3.1. Main results

Theorem 1 (Distribution of $\boldsymbol{\beta}$)**.** *Let Assumption 1 hold. Then, under the notations of Lemma 1,*

$$\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \to 0 \quad \textit{where} \quad (\lambda \mathbf{I}_p + \tau \mathbf{C}_p)\, \bar{\boldsymbol{\beta}} \sim \mathcal{N}(\eta \boldsymbol{\mu}_p, \gamma \mathbf{C}_p / n)$$

with $(\eta, \gamma, \tau) \in \mathbb{R}_+^3$ being the unique solution of the following system of equations

$$\eta = \mathbb{E}[g_\kappa(r)], \quad \gamma = \mathbb{E}[g_\kappa^2(r)], \quad \tau = \mathbb{E}[g_\kappa(r)(m-r)]/\sigma^2 \tag{3}$$

for some $r \sim \mathcal{N}(m, \sigma^2)$ with

$$\begin{aligned} m &\equiv \eta \boldsymbol{\mu}_p^\mathsf{T} \left(\lambda \mathbf{I}_p + \tau \mathbf{C}_p\right)^{-1} \boldsymbol{\mu}_p \\ \sigma^2 &\equiv \eta^2 \boldsymbol{\mu}_p^\mathsf{T} \left(\lambda \mathbf{I}_p + \tau \mathbf{C}_p\right)^{-1} \mathbf{C}_p \left(\lambda \mathbf{I}_p + \tau \mathbf{C}_p\right)^{-1} \boldsymbol{\mu}_p \\ &\quad + \gamma\, \mathrm{tr} \left[\left(\lambda \mathbf{I}_p + \tau \mathbf{C}_p\right)^{-1} \mathbf{C}_p \right]^2 / n. \end{aligned}$$

The proof sketch of Theorem 1 can be found in Section 4. Note that Theorem 1 allows also to determine the unknown parameters m, σ^2 in Lemma 1, thereby giving directly the generalization error of $\tilde{\mathbf{x}}_i$, obtained by leaving it out from the training process and treating it as unseen data.

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

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

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

Figure 1 reports the empirical distribution of $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ versus a Gaussian distribution $\mathcal{N}(m, \sigma^2)$ from Theorem 1 for one realization. We observe a close match of the asymptotic results on finite data samples *already for not too large* n, p.

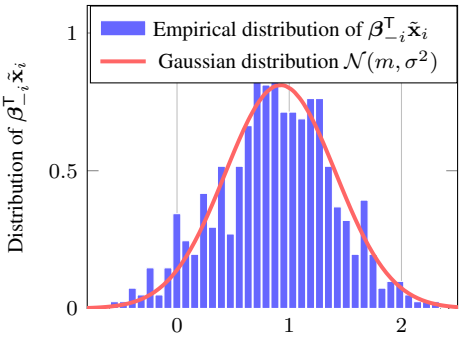

Fig. 1. Comparison between $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ and a Gaussian distribution $\mathcal{N}(m, \sigma^2)$ as defined in Theorem 1 with $\boldsymbol{\mu}_p = [2,\ \mathbf{0}_{p-1}]$, $\mathbf{C}_p = \mathbf{I}_p$, for $\lambda = 1$, $p = 256$ and $n = 512$.

3.2. Interpretation

The unregularized solution which, if well-defined, is retrieved by taking $\lambda = 0$ in the results of Theorem 1, gives rise to

$$\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \to 0 \quad \text{where} \quad \bar{\boldsymbol{\beta}} \sim \mathcal{N}\left(\frac{\eta}{2\tau}\boldsymbol{\beta}_*, \frac{\gamma}{n\tau^2}\mathbf{C}_p^{-1}\right)$$

where we recall $\boldsymbol{\beta}_* = 2\mathbf{C}_p^{-1}\boldsymbol{\mu}_p$ is the vector of the true parameters. A first remark is that the high dimensional maximum likelihood $\boldsymbol{\beta}$ is biased in the sense that its expectation converges to a rescaled version of the true $\boldsymbol{\beta}_*$, which is reminiscent of the conclusions in [9] for $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

Turning to the regularized solutions, it can be observed from Theorem 1 that when $\lambda \neq 0$, the asymptotic expectation of $\boldsymbol{\beta}$ is different from $\boldsymbol{\beta}_*$ in both scale and direction. From a classification viewpoint, using a rescaled $\boldsymbol{\beta}_*$ achieves the same oracle test performance as $\boldsymbol{\beta}_*$, meaning that only biases 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 γ/η^2, which is an *indicator of the variability of* $\boldsymbol{\beta}$, is minimal for extremely regularized solutions ($\lambda \to \infty$). Indeed, according to (3), η, γ equal respectively the first and second moment of the random variable $c = g_\kappa(r)$ (as defined in Lemma 1). It can then be shown that the random variable c converges to $1/2$ as $\lambda \to \infty$, where γ/η^2 reaches a minimal value of 1.

To summarize, while the variability of $\boldsymbol{\beta}$ is at its lowest in the limit $\lambda \to \infty$, the undesired direction bias is only eliminated at the other extreme $\lambda \to 0$. 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 even though the classification error is usually minimized at finite λ, as in the case of the $\mathbf{C}_2$ covariance in Figure 2, it continually decreases as $\lambda \to \infty$ when the data covariance equals $\mathbf{I}_p$. A highly regularized solution is therefore favorable in this special case, as confirmed in Figure 2 for $\mathbf{C}_1 = 2\mathbf{I}_p$. 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 $\mathbf{C}_p = \mathbf{I}_p$; it then remains to minimize the variability of $\boldsymbol{\beta}$, which can be achieved for $\lambda \to \infty$. It is thus of interest for future investigation to construct an estimator of the optimal λ by capitalizing on the theoretical results in this paper and employing the arguments from random matrix theory (for example in [5, 15]).

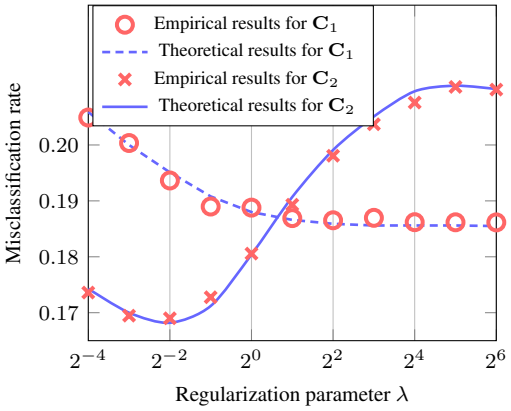

Fig. 2. Misclassification error as a function of λ, with $\boldsymbol{\mu}_p = [1,\ 1,\ \mathbf{0}_{p-2}]$, $\mathbf{C}_1 = 2\mathbf{I}_p$ and $\mathbf{C}_2 = \text{diag}[1,\ 5,\ \mathbf{1}_{p-2}]$, where $p = 128$, $n = 512$ and with number of test samples $n_{\text{test}} = 512$. Empirical results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

4. TECHNICAL APPROACH

As explained in the introduction, the technical approach derived for this work is essentially different from the analysis of [10], where the authors capitalized on the *double leave-one-out* procedure, developed earlier in [1]. Similarly to the

approach in [1], we consider first the *leave-one-observation-out* solution $\boldsymbol{\beta}_{-i}$ given by excluding the i-th training sample from the training set. Upon this consideration, we find (as will be shown subsequently), in analogy with the results of [1], that $\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i$ is linked to $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ through a proximal mapping with a unknown constant κ. The authors of [1] proceeded to determine this constant κ by introducing a *leave-one-feature-out* version of $\boldsymbol{\beta}$ obtained when one feature is eliminated from the training data vectors. This step is however not adaptable to more elaborate (and practical) situations where the features are correlated (i.e., $\mathbf{C}_p \neq \mathbf{I}_p$), as considered in this article. To tackle the correlations between features, we propose here to employ results from random matrix theory for the determination of κ (as will be described in Lemma 2).

Since $\boldsymbol{\beta}_{-i} = \frac{1}{n}\sum_{j\neq i} \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\tilde{\mathbf{x}}_j$ by definition, the difference $\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}$ is therefore given by

$$\begin{aligned}\lambda(\boldsymbol{\beta}-\boldsymbol{\beta}_{-i}) &= \frac{1}{n}\sum_{j\neq i}\left(c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\right)\tilde{\mathbf{x}}_j + \frac{1}{n}c_i\tilde{\mathbf{x}}_i \\ &= \frac{1}{n}\tilde{\mathbf{X}}_{-i}\boldsymbol{\Delta}\mathbf{c}_{-i} + \frac{1}{n}c_i\tilde{\mathbf{x}}_i \end{aligned} \tag{4}$$

with $\boldsymbol{\Delta}\mathbf{c}_{-i} \in \mathbb{R}^{n-1}$ the vector with 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. Establishing that the c_i's are of order $O(1)$, we deduce from (2) that $\|\boldsymbol{\beta}\|$ (and $\|\boldsymbol{\beta}_{-i}\|$) is of order $O(1)$ and from (4) that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}\|$ is of order $O(p^{-1/2})$. Moreover, using the fact that $\psi(t)$ is Lipschitz continuous, we have $c_j - \psi(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j) = O(p^{-1/2})$ that is smaller compared to c_j, which further allows for the following estimate

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

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[1]

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

with $\mathbf{D}_{-i} \in \mathbb{R}^{n-1}$ the diagonal matrix with (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 (4) we reach

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_{-i})^\mathsf{T}\tilde{\mathbf{x}}_i = \frac{c_i}{n}\tilde{\mathbf{x}}_i^\mathsf{T}\left(\lambda\mathbf{I}_p - \frac{1}{n}\tilde{\mathbf{X}}_{-i}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\right)^{-1}\tilde{\mathbf{x}}_i + o_p(1). \tag{5}$$

The RHS of (5) being a quadratic $\frac{1}{n}\tilde{\mathbf{x}}_i^\mathsf{T}\mathbf{M}\tilde{\mathbf{x}}_i$ for some $\mathbf{M}$ of bounded operator norm and independent of $\tilde{\mathbf{x}}_i$, classical RMT results yield the following approximation.

[1]Note that $\psi'(t) < 0$, $-\mathbf{D}_{-i}$ is positive definite so that both $\lambda\mathbf{I}_{n-1} - \frac{1}{n}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}\tilde{\mathbf{X}}_{-i}$ and $\lambda\mathbf{I}_p - \frac{1}{n}\tilde{\mathbf{X}}_{-i}\mathbf{D}_{-i}\tilde{\mathbf{X}}_{-i}^\mathsf{T}$ are invertible for $\lambda > 0$.

Lemma 2 (Asymptotic approximation of quadratic form)**.** *Let Assumption 1 hold. 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 $\kappa = \frac{1}{n}\operatorname{tr}(\bar{\mathbf{Q}}\mathbf{C}_p)$ with

$$\bar{\mathbf{Q}} \equiv \left(\lambda\mathbf{I}_p - \mathbb{E}\left[\frac{\psi'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}})}{1-\kappa\psi'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}})}\right]\mathbf{C}_p\right)^{-1}.$$

From Lemma 2 and (2)-(4) 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 solution of which is given via g_κ as defined in Lemma 1 as

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

which gives a new expression of $\boldsymbol{\beta}$ from (2) as

$$\lambda\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n g_\kappa\left(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i\right)\tilde{\mathbf{x}}_i. \tag{6}$$

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 $\boldsymbol{\mu}_p^\mathsf{T}\mathbb{E}[\boldsymbol{\beta}_{-i}]$ and variance $\mathbb{E}[\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{C}_p\boldsymbol{\beta}_{-i}]$ that are asymptotically close to

$$m \equiv \boldsymbol{\mu}_p^\mathsf{T}\mathbb{E}[\boldsymbol{\beta}], \quad \sigma^2 \equiv \mathbb{E}[\boldsymbol{\beta}^\mathsf{T}\mathbf{C}_p\boldsymbol{\beta}] = \operatorname{tr}\left(\mathbf{C}_p\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}]\right) \tag{7}$$

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

However, note that in (6), the term $g_\kappa(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i)$ 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 $\tilde{\mathbf{x}}_i = \boldsymbol{\mu}_p + \mathbf{z}_i$ from $\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_i$ by writing

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

so that $\mathbb{E}[(\boldsymbol{\beta}_{-i}^\mathsf{T}\mathbf{z}_i)\tilde{\mathbf{z}}_i] = \mathbf{0}$ with $\mathbb{E}[\tilde{\mathbf{z}}_i] = \mathbf{0}$, $\mathbb{E}[\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^\mathsf{T}] = \mathbf{C}_p - \frac{1}{\sigma^2}\mathbf{C}_p\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}]\mathbf{C}_p$. As a consequence, (6) can be rewritten as

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

which yields the following relation

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

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

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

and therefore

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

which concludes the proof of Theorem 1.

5. 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," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[6] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1065–1074, 2019.

[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] 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.

[14] Neal Parikh and Stephen Boyd, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[15] Hafiz Tiomoko Ali and Romain Couillet, "Improved spectral community detection in large heterogeneous networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8344–8392, 2017.