High Dimensional Classification via Empirical Risk Minimization: Improvements and Optimality

Xiaoyi Mai [*1,2] **Zhenyu Liao** [*1,2]

Abstract

In this article, we investigate a family of classification algorithms defined by the principle of empirical risk minimization, in the high dimensional regime where the feature dimension p and data number n are both large and comparable. Based on recent advances in high dimensional statistics and random matrix theory, we provide under mixture data model a unified stochastic characterization of classifiers learned with different loss functions. Our results are instrumental to an in-depth understanding as well as practical improvements on this fundamental classification approach. As a main outcome, we demonstrate the existence of a universally optimal loss function which yields the best high dimensional performance at any given n/p ratio.

1. Introduction

Consider the following general classification problem: given a training set of n pre-labelled samples with feature vectors of dimension p, the objective is to predict the class label y (e.g., $y = \pm 1$) of a new observation $\mathbf{x}$ based on the knowledge of these training samples. The basic setup of a large number of classification algorithms is to obtain the class label y of a new instance by combining its feature vector $\mathbf{x}$ with a vector of weights $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $y = \text{sign}(\boldsymbol{\beta}^\mathsf{T}\mathbf{x})$. The weight vector $\boldsymbol{\beta}$ is usually learned from fitting the known class of training samples, for example, by minimizing the classification error (also known as the 0–1 loss) on the given training set. Despite being a natural choice, the minimization of the non-convex 0–1 loss is known to be NP-hard (Ben-David et al., 2003). To address this issue, the empirical risk minimization principle (Vapnik, 1992) suggests to obtain $\boldsymbol{\beta}$ by minimizing a certain *convex* surrogate of the 0–1 loss on the training set. Within this framework, the comparison between different designs of loss functions have been long discussed in the literature (Vapnik, 1992; Rosasco et al., 2004; Masnadi-Shirazi & Vasconcelos, 2009), mostly in the setting where the number of training data n largely exceeds their dimension p (i.e., p is considered small while n goes large to infinity). Besides the computational convenience, the usage of convex loss functions is also supported by their property of leading to the same Bayes optimal solution that minimizes the 0–1 loss in the limit of $n \gg p$ (Rosasco et al., 2004). In spite of this remark, the classification accuracy can significantly depend on the choice of loss function when n is not exceedingly larger than p. While it is crucial to know in practice which loss function to use for a given number of training samples, little is known in the regime of finite n/p.

The behavior of machine learning algorithms at finite n/p ratio is particularly important in the modern setting of big data, where the manipulation of hundreds of features or even more is constantly required, bringing about the inadequacy of considering p to be vanishingly small compared to n. Understandably, the statistical properties of algorithms at finite n/p are much less tractable than in the limit of $n \gg p$, due to the instability of learning systems. Nonetheless, it has recently come to attention that in high dimensions, such analyses can be rendered accessible by exploiting extra degrees of freedom induced by numerous features.

As such, the high dimensional investigation of machine learning methods for comparably large n, p is receiving an unprecedented research interest. The high dimensional performance of classification methods with explicit solutions were evaluated in several works (Liao & Couillet, 2017; Elkhalil et al., 2017; Dobriban et al., 2018), relying mostly on techniques from random matrix theory. To capture the statistical behavior of M-estimators with no closed-form, the authors of (El Karoui et al., 2013; Donoho & Montanari, 2016) adopted a "double leave-one-out" approach, hinging on the intuition that the outcomes of algorithms remain unchanged after excluding one sample from the training set or one feature from the feature vectors. Based on the same

[*]Equal contribution [1]Laboratoire des Signaux et Systèmes, CentraleSupélec, Université Paris-Saclay, France; [2]G-STATS Data Science Chair, GIPSA-lab, University Grenobles-Alpes, France. Correspondence to: Xiaoyi Mai <xiaoyi.mai@l2s.centralesupelec.fr>, Zhenyu Liao <zhenyu.liao@l2s.centralesupelec.fr>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

technique, the more recent line of works (Sur & Candès, 2018; Candès & Sur, 2018) investigated the logistic regression model for classification by imposing the existence of a linear classifier for non-structured data of i.i.d. Gaussian features (i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$), therefore cannot be applied to feature vectors with class-structured patterns.

In this work, we combine the advantage of the "leave-one-out" procedure for tackling implicit learning systems and the convenience of random matrix theory in handling class-structured patterns. As results, we derive, in the regime of finite n/p, a unified stochastic description of the (generally implicit) optimization solution obtained from minimizing the empirical risk of any convex and smooth loss, under a high dimensional mixture model of multivariate normal feature vectors. Our analysis is of both theoretical and practical values, as it allows not only to explain and predict empirical results, but also to propose practical improvements and discover the optimal solutions.

To begin with, the maximal likelihood principle (Lehmann & Romano, 2006; McCullagh & Nelder, 1989; Portnoy et al., 1984) states that the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ given by the negative log-likelihood loss function is a consistent estimator of the true parameter vector $\boldsymbol{\beta}_*$ underlying the conditional class probability $P(y|\mathbf{x})$, and often provides the best efficiency compared to other loss functions at $n \gg p$. However, it is empirically observed (in simulations that will be shown subsequently) that at finite n/p: 1) $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is a biased estimator of $\boldsymbol{\beta}_*$, up to a factor depending on n/p; 2) higher classification accuracy can be achieved with other losses, going against the natural use of maximal likelihood methods in high dimensions. These empirical evidences raise the questions on the possibility of bias-correcting as well as the optimal choice of loss function for finite n/p. From an ensemble learning perspective, it is also found that the classification accuracy can be improved by linearly combining solutions learned with different loss functions, as long as the weights assigned to the member solutions are properly chosen. It would thus be of interest to investigate on the condition and the limit of this improvement. Driven by these empirically motivated questions, our main findings are summarized as follows:

- Besides $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, all solutions $\hat{\boldsymbol{\beta}}$ within the present framework are aligned with $\boldsymbol{\beta}_*$ in expectation. The rescaling factor α that renders $\alpha\hat{\boldsymbol{\beta}}$ an unbiased estimator of $\boldsymbol{\beta}_*$ in high dimensions is given as an explicit function of $\hat{\boldsymbol{\beta}}$ and the training samples.
- The square loss, rather than the negative log-likelihood loss, is proved to yield the best classification accuracy for the high dimensional mixture model under study. This optimality holds universally for all n/p ratios and is irrespective of the model parameters.
- The performance gain from linearly combining different solutions can be achieved under certain condition. However, it is impossible to surpass the solution of square loss in terms of classification accuracy.

In the remainder of this article, we introduce the objects of interest in Section 2. Our main technical results are presented in Section 3, based on which we propose the aforementioned high dimensional improvements. In Section 4 we discuss the optimal choice of loss function and the limit of ensemble method. To complete our theoretical results, we provide in Section 5 an asymptotic deterministic description of the system performance. The article closes with concluding remarks and envisioned extensions in Section 6.

2. Preliminaries

Let us start by introducing some notations that will be employed throughout the article. Boldface lowercase (uppercase) characters stand for vectors (matrices). The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices. We follow the convention to use $o_P(1)$ for a sequence of random variables that convergences to zero in probability and $\xrightarrow{d}$ for the convergence in distribution. We say that an event occurs with high probability if it happens with a probability arbitrarily close to one for sufficiently large n, p.

As commonly supposed in popular statistical methods as linear discriminant analysis and logistic regression, each data instance $(\mathbf{x}, y)$, with feature vector $\mathbf{x} \in \mathbb{R}^p$ and class label $y = \pm 1$, is considered here to be drawn independently from a distribution $\mathcal{D}$ of the following mixture model:

$$y = -1 \Leftrightarrow \mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{C}),$$
$$y = +1 \Leftrightarrow \mathbf{x} \sim \mathcal{N}(+\boldsymbol{\mu}, \mathbf{C}),$$

with balanced class priors for some mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$. The training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is composed of n independent observations from the aforementioned model. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ be the feature matrix of training set, and $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^p$ the class label vector.

This model satisfies the hypotheses of both logistic regression and linear discriminant analysis, and has its conditional class probability given by:

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

with $s(t) = \frac{1}{1+e^{-t}}$ the *logistic sigmoid* function and

$$\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}. \tag{1}$$

As such, we shall refer to $\boldsymbol{\beta}_*$ as the vector of true parameters throughout this paper, which allows to recover the exact conditional class probability for a given $\mathbf{x}$.

To ensure a non-trivial misclassification rate in the high dimensional setting (i.e., the misclassification probability is neither 0 nor 1 for large p), we shall (as in (Couillet et al., 2018)) work under the following assumptions.

Assumption 1 (Growth rate)**.** *The sample ratio n/p is uniformly bounded in $(1, +\infty)$ for arbitrarily large p. Also, $\|\boldsymbol{\mu}\| = O(1)$, $\|\mathbf{C}\| = O(1)$ and $\|\mathbf{C}^{-1}\| = O(1)$ with respect to p.*

Following the empirical risk minimization principle, we consider the optimization problem as follows,

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

with $\rho : \mathbb{R} \mapsto \mathbb{R}$ some nonnegative loss function satisfying the following property,

Assumption 2 (Loss function)**.** *The function ρ is convex and at least twice differentiable.*

In particular, with the logistic loss $\rho(t) = \ln(1 + e^{-t})$ that gives the maximum likelihood estimate of $\boldsymbol{\beta}_*$, we obtain the logistic regression classifier. The least squares classifier is given by the square loss $\rho(t) = (t-1)^2$. Another popular choice is the exponential loss $\rho(t) = e^{-t}$, widely used in boosting algorithms (Freund et al., 1999; Rojas, 2009).

It is worth noting that, in the high dimensional setting of Assumption 1, the existence of the unique solution to (2) is not guaranteed for all n, p. A simple example is the case $n < p$ (as excluded from Assumption 1), for which one can show that (2) has non-unique solutions. Furthermore, it was shown in (Sur et al., 2017) that, in the case of logistic regression, $\|\hat{\boldsymbol{\beta}}\|$ is finite if and only if some dimensionality condition is met. The discussion on the existence condition is out of the scope of this work and we assume here that the learned classifier is "well-behaved" in the sense that the optimization problem (2) is well defined with a unique solution $\hat{\boldsymbol{\beta}}$ of finite norm.

3. Main Results and Improvements

Before introducing the main theoretical results, we define some random elements that will appear in the theorem. By cancelling the derivative of the convex loss function ρ, we obtain from (2) that $\mathbf{X}\mathbf{c} = \mathbf{0}$ with

$$\mathbf{c} = [c_1, \ldots, c_n]^\mathsf{T} \equiv [y_1\psi(y_1\mathbf{x}_1^\mathsf{T}\hat{\boldsymbol{\beta}}), \ldots, y_n\psi(y_n\mathbf{x}_n^\mathsf{T}\hat{\boldsymbol{\beta}})]^\mathsf{T}, \tag{3}$$

where we denote $\psi(t) \equiv -\frac{d\rho(t)}{dt}$ the negative derivative of the loss function ρ. Additionally, let

$$\mathbf{r} = [r_1, \ldots, r_n]^\mathsf{T} \equiv [\mathbf{x}_1^\mathsf{T}\hat{\boldsymbol{\beta}} - \kappa c_1, \ldots, \mathbf{x}_n^\mathsf{T}\hat{\boldsymbol{\beta}} - \kappa c_n]^\mathsf{T}, \tag{4}$$

with

$$\kappa = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i^\mathsf{T}\mathbf{Q}\mathbf{x}_i/n}{1 + \psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i^\mathsf{T}\mathbf{Q}\mathbf{x}_i/n}, \tag{5}$$

where $\mathbf{Q} = \left(-\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)^{-1}$. We denote by $\mathbf{r}_c$ a recentered version of $\mathbf{r}$ given as

$$\mathbf{r}_c = -\left(\mathbf{I}_n - \frac{1}{n}\mathbf{y}\mathbf{y}^\mathsf{T}\right)\mathbf{r}. \tag{6}$$

With the above notations, we are now in position to introduce the main technical result of this article, which concerns a stochastic description of the classifier $\hat{\boldsymbol{\beta}}$ defined in (2), in the following theorem.

Theorem 1. *Let Assumptions 1 and 2 hold. Then,*

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\| = o_P(1), \quad \tilde{\boldsymbol{\beta}} = \frac{1}{\alpha}\left(\boldsymbol{\beta}_* + \frac{2\sqrt{p}\|\mathbf{c}\|}{\mathbf{c}^\mathsf{T}\mathbf{y}}\mathbf{C}^{-\frac{1}{2}}\mathbf{u}\right)$$

for $\boldsymbol{\beta}_*$, $\mathbf{c}$, $\mathbf{r}_c$ defined respectively in (1), (3) and (6), $\mathbf{u} \in \mathbb{R}^p$ a random vector uniformly distributed on the unit sphere and

$$\alpha = \frac{2n\mathbf{c}^\mathsf{T}\mathbf{r}_c}{\mathbf{c}^\mathsf{T}\mathbf{y}\|\mathbf{r}_c\|^2}. \tag{7}$$

Leaving the proof to Supplementary Material, Theorem 1 gives a high dimensional equivalence $\tilde{\boldsymbol{\beta}}$ for the optimization solution $\hat{\boldsymbol{\beta}}$, so that the high dimensional performance of $\hat{\boldsymbol{\beta}}$ can be studied via $\tilde{\boldsymbol{\beta}}$. Indeed, consider the probability of misclassification

$$P(y\mathbf{x}^\mathsf{T}\boldsymbol{\beta} < 0|\boldsymbol{\beta}) \equiv M_{\mathrm{C}}(\boldsymbol{\beta}) \tag{8}$$

for some $(\mathbf{x}, y) \sim \mathcal{D}$ independent of $\boldsymbol{\beta}$, we deduce from Theorem 1 that

$$\mathcal{M}_{\mathrm{C}}(\hat{\boldsymbol{\beta}}) = Q\left(\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu} + \frac{p\|\mathbf{c}\|^2}{(\mathbf{c}^\mathsf{T}\mathbf{y})^2}}}\right) + o_P(1), \tag{9}$$

where $Q(t) \equiv \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$ denotes the Q-function of the standard Gaussian distribution. As is shown in Figure 1, the approximation of classification performance $\mathcal{M}_{\mathrm{C}}(\hat{\boldsymbol{\beta}})$ given by (9) is of high precision for moderately large n, p. Note also from Theorem 1 that, $\tilde{\boldsymbol{\beta}}$ is proportional to the true parameter $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$ in expectation, with an additive "noise" term $\frac{2\sqrt{p}\|\mathbf{c}\|}{\mathbf{c}^\mathsf{T}\mathbf{y}}\mathbf{C}^{-\frac{1}{2}}\mathbf{u}$ that is of random direction.[1] Clearly, one shall maximize the signal-to-noise ratio of $\tilde{\boldsymbol{\beta}}$ by minimizing $\frac{\|\mathbf{c}\|}{|\mathbf{c}^\mathsf{T}\mathbf{y}|}$. This conclusion can also be easily reached from (9).

Even though the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ obtained from $\rho(t) = \ln(1 + e^{-t})$ estimates exactly $\boldsymbol{\beta}_*$ in the limit

[1] Remark that α and $\sqrt{p}\|\mathbf{c}\|/\mathbf{c}^\mathsf{T}\mathbf{y}$ are both finite and away from zero with high probability.

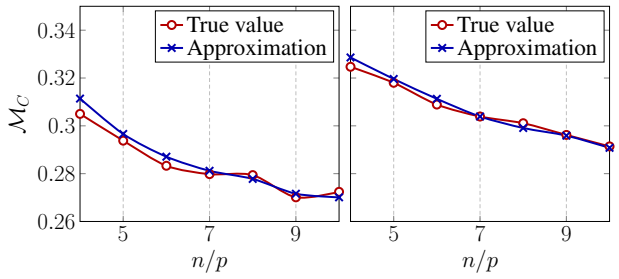

Figure 1. Comparison between the expected classification error $\mathcal{M}_{\mathrm{C}}$ and its approximation in (9) for $p = 256$, with $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$, $\rho(t) = \ln(1+e^{-t})$ (left) and $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{2p}$, $\mathbf{C}_{ij} = 0.1^{|i-j|}$, $\rho(t) = (t-1)^2/2$ (right).

of $n \gg p$, it is "biased" up to a constant factor in the high dimensional setting with finite n/p. Indeed, by estimating the expectation of $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ with the empirical mean $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ obtained over 500 independent realizations, we observe in Figure 2 that $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ is proportional, and clearly not equal, to the true parameter vector $\boldsymbol{\beta}_*$. According to Theorem 1, this bias can be eliminated by multiplying α given in (7). As corroborating evidence, the estimated expectation $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ of $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is given in Figure 2 to coincide with $\boldsymbol{\beta}_*$.

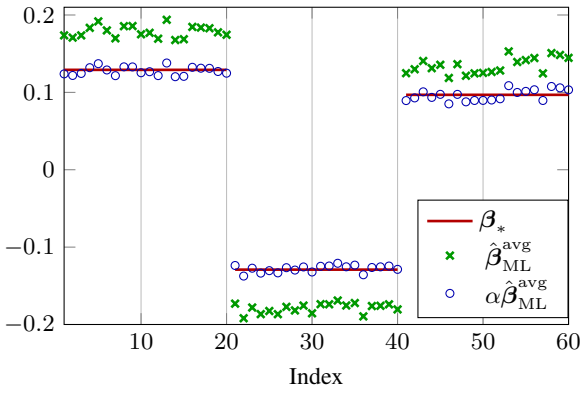

Figure 2. Comparison of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ (averaged over 500 realizations), the true parameter $\boldsymbol{\beta}_*$ and the rescaled classifier $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ defined in Theorem 1 with $\boldsymbol{\mu} = [\mathbf{1}_{p/3}, -\mathbf{1}_{p/3}, \frac{3}{4}\mathbf{1}_{p/3}]/\sqrt{p}$, $\mathbf{C} = 2\mathbf{I}_p$, for $p = 60$ and $n = 300$.

Although correcting the aforementioned bias with the rescaled solution $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ does not change the classification accuracy, it helps improve the conditional class probability estimation, which is required in many applications for risk management. We propose here an improvement strategy consisting in rescaling any solution $\hat{\boldsymbol{\beta}}$ (besides $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$) with its bias factor α for a more accurate class probability estimation, theoretically supported by the following corollary.

Corollary 1. *With the assumptions and notations of Theorem 1, we have*

$$\|\mathbb{E}[\alpha\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}_*\| = o_P(1).$$

Consider now the expected square loss of class probability estimation of a classifier $\boldsymbol{\beta}$ given by

$$\mathcal{M}_{\mathrm{E}}(\boldsymbol{\beta}) = \mathbb{E}[s(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}) - s(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_*)|\boldsymbol{\beta}]^2 \tag{10}$$

where $(\mathbf{x}, y) \sim \mathcal{D}$ is independent of $\boldsymbol{\beta}$, and $s(t) = \frac{1}{1+e^{-t}}$. We demonstrate in Figure 3 the utility of the proposed rescaling strategy with the significant performance gains measured by $\mathcal{M}_{\mathrm{E}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) - \mathcal{M}_{\mathrm{E}}(\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$, which are especially large at small n/p ratios. Moreover, note that both $\mathbf{c}$, $\mathbf{r}_c$ (and thus α) are fast computed once $\hat{\boldsymbol{\beta}}$ is obtained by solving (2). Therefore, the proposed rescaling scheme is computationally efficient in the sense that they induce little extra cost to the training of the original classifier $\hat{\boldsymbol{\beta}}$.

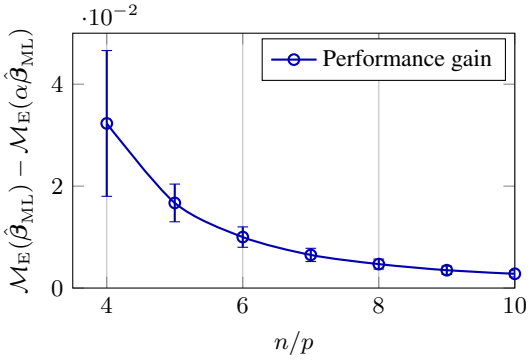

Figure 3. Performance gain $\mathcal{M}_{\mathrm{E}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) - \mathcal{M}_{\mathrm{E}}(\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

Corollary 1 indicates that all rescaled classifiers $\alpha\hat{\boldsymbol{\beta}}$ are equally efficient in expectation, it is then pertinent to ask whether it is possible to reduce the variance. One of the basic strategies in this aspect is to linearly combine several (rescaled) classifiers $\alpha_k\hat{\boldsymbol{\beta}}_k$ learned with different loss functions, to form an ensemble classifier (Fumera et al., 2008). In the following theorem (see Supplementary Material for its proof) we give a stochastic characterization of such ensemble classifier.

Theorem 2. *Let Assumptions 1 and 2 hold, and $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m$ stand respectively for classifiers learned with loss functions $\rho_1, \ldots, \rho_m$, m being some positive integer. For any set of m real-valued coefficients $\{w_1, \ldots, w_m\}$ such that $\sum_{k=1}^m w_k = 1$, define the ensemble classifier*

$$\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = \sum_{k=1}^{m} w_k\alpha_k\hat{\boldsymbol{\beta}}_k \tag{11}$$

with α_k the rescaling factor of $\hat{\boldsymbol{\beta}}_k$ given in (7). Then,

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{ES}} - \tilde{\boldsymbol{\beta}}_{\mathrm{ES}}\| = o_P(1),$$

with

$$\tilde{\boldsymbol{\beta}}_{\mathrm{ES}} = \boldsymbol{\beta}_* + 2\sqrt{p}\|\mathbf{c}_{\mathrm{ES}}\|\mathbf{C}^{-\frac{1}{2}}\mathbf{u}'$$

for $\mathbf{u}' \in \mathbb{R}^p$ a random vector uniformly distributed on the unit sphere and

$$\mathbf{c}_{\mathrm{ES}} = \sum_{k=1}^{m} \frac{w_k \mathbf{c}_k}{\mathbf{c}_k^{\mathsf{T}}\mathbf{y}} \tag{12}$$

for $\mathbf{c}_k$ defined in (3) with respect to the loss function ρ_k and the training set $(\mathbf{X}, \mathbf{y})$.

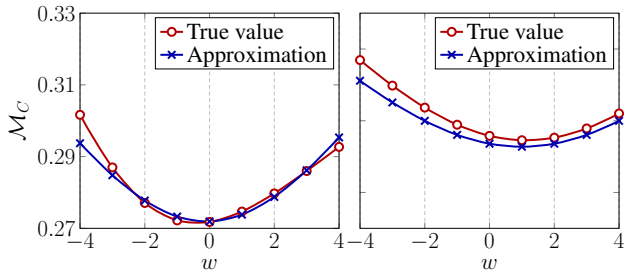

Figure 4. Comparison between the expected classification errors of the ensemble classifier $\mathcal{M}_{\mathrm{C}}(\hat{\boldsymbol{\beta}}_{\mathrm{ES}})$ and its approximation $\mathcal{M}_{\mathrm{C}}(\tilde{\boldsymbol{\beta}}_{\mathrm{ES}})$ as a function of w. For $\rho_1(t) = \ln(1+e^{-t})$, $\rho_2(t) = e^{-t}$, $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$ (left) and $\rho_1(t) = (t-1)^2/2$, $\rho_2(t) = \ln(1+e^{-t})$, $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{2p}$, $\mathbf{C}_{ij} = 0.1^{|i-j|}$ (right), $p = 256$, $n = 10p$.

In Figure 4 we consider the ensemble classifier $\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = w\alpha_1\hat{\boldsymbol{\beta}}_{\rho_1} + (1-w)\alpha_2\hat{\boldsymbol{\beta}}_{\rho_2}$ from different loss functions ρ_1, ρ_2, and compare its classification performance with its high dimensional equivalent $\tilde{\boldsymbol{\beta}}_{\mathrm{ES}}$ given in Theorem 2 as a function of the weight w. A close match is observed in both settings with different combinations of loss functions, suggesting that the optimal weights can be estimated with great precision from the vector $\mathbf{c}$ of its member classifiers. Indeed, it entails from Theorem 2 that the optimal weights w_k yielding the best performance can be obtained by minimizing $\|\mathbf{c}_{\mathrm{ES}}\|$. This remark is formally stated in the following corollary, where we also provide a necessary and sufficient condition under which $\hat{\boldsymbol{\beta}}_{\mathrm{ES}}$ is guaranteed to surpass all its (rescaled) member classifiers $\alpha_k\hat{\boldsymbol{\beta}}_k$ in terms of both classification accuracy and class probability estimation.

Corollary 2. *With the assumptions and notations in Theorem 2, the optimal ensemble classifier is given by*

$$\hat{\boldsymbol{\beta}}_{\mathrm{ES}}^{\mathrm{opt}} = \sum_{k=1}^{m} w_k^{\mathrm{opt}} \alpha_k \hat{\boldsymbol{\beta}}_k$$

with

$$\{w_1^{\mathrm{opt}}, \ldots, w_m^{\mathrm{opt}}\} = \mathrm{argmin}_{\{w_1,\ldots,w_m\}} \|\mathbf{c}_{\mathrm{ES}}\|, \tag{13}$$

then, with high probability,

$$\mathcal{M}(\hat{\boldsymbol{\beta}}_{\mathrm{ES}}^{\mathrm{opt}}) \geq \max_{k\in\{1,\ldots,m\}} \mathcal{M}(\alpha_k\hat{\boldsymbol{\beta}}_k) \tag{14}$$

for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$ as defined in (8) and (10), respectively. Furthermore, the inequality in (14) is strict if and only if, there exists $k \in \{1, \ldots, m\}$ such that,

$$\frac{|\mathbf{c}_k^{\mathsf{T}}\mathbf{c}_{k_*}|}{|\mathbf{y}^{\mathsf{T}}\mathbf{c}_k|} \neq \frac{\|\mathbf{c}_{k_*}\|^2}{|\mathbf{y}^{\mathsf{T}}\mathbf{c}_{k_*}|}, \quad k_* = \mathrm{argmin}_{k'\in\{1,\ldots,m\}} \frac{\|\mathbf{c}_{k'}\|}{|\mathbf{y}^{\mathsf{T}}\mathbf{c}_{k'}|}.$$

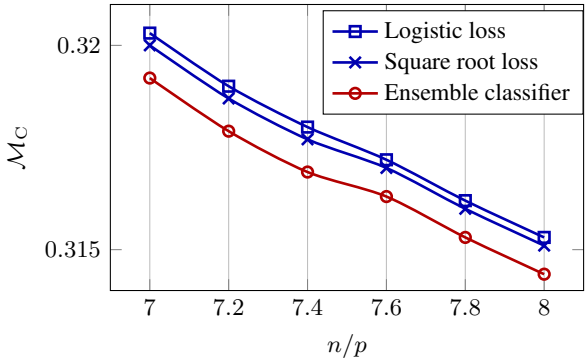

Figure 5. Comparison of classification error rate between the logistic loss $\rho(t) = \ln(1+e^{-t})$, the square root loss $\rho(t) = \sqrt{(t-1)^2+1}$ and the associated ensemble classifier given in Corollary 2 for $\boldsymbol{\mu} = [0.6, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 250$.

The simulations in Figure 5 confirm the benefits of this ensemble approach. Compared to its member classifiers, the ensemble classifier with the optimal weights w_k^{opt} given by (13) produces a similar effect as adding $p/5$ training samples in reducing classification error.

In this section we discussed two improvement strategies for high dimensional classification problem: 1) the rescaling method for obtaining an unbiased estimator of the true parameter vector $\boldsymbol{\beta}_*$, when the feature dimension p is comparable to the sample size n and 2) the ensemble scheme that helps improve the classification and estimation performance by linearly combining several classifiers obtained from different loss functions. Numerical evidences are also provided to support the advantages of these two methods. A natural question to ask then, is whether there exists a performance upper bound for these methods and when it can be attained. We answer this question in the next section.

4. Optimality

It has been shown in Corollary 1 that, regardless of the choice of loss function ρ, the true parameter vector $\boldsymbol{\beta}_*$ is attained by the rescaled classifier $\alpha\hat{\boldsymbol{\beta}}$, for α given by (7), in

the limit of $n \gg p$. Yet, it is still unclear as to the optimal choice of ρ at finite n/p, which is a far more interesting question to provide guidance in practice.

A default option, which is commonly believed to yield optimal learning results, would be to apply the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, obtained here with the logistic loss $\rho(t) = \ln(1 + e^{-t})$. However, as can be observed in Figure 6 where the classification performance of the maximal likelihood solution (in blue) is provided along with the results produced by the square loss $\rho(t) = (t-1)^2/2$ (in red), the maximum likelihood classifier is *consistently* surpassed by the least squares one, for n/p ranging from 4 to 10. In light of this empirical evidence which contradicts the maximal likelihood principle for not too large n/p, one may ask whether this observed superiority of square loss over logistic loss holds at all n/p ratios, or more generally, whether there exists a loss function providing the best high dimensional classification results for any given size of training samples.

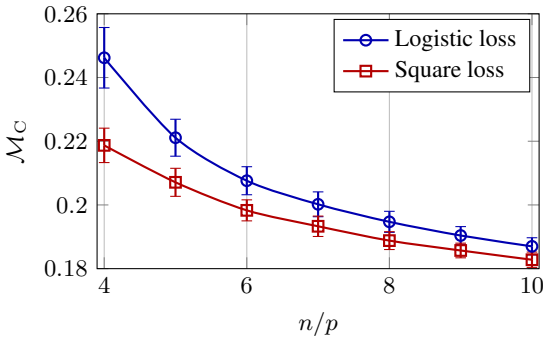

Figure 6. Comparison of the expected classification error rate between the logistic loss $\rho(t) = \ln(1 + e^{-t})$ and the square loss $\rho(t) = (t-1)^2/2$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

To answer these questions, note first that since $\alpha\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to $\alpha\tilde{\boldsymbol{\beta}}$ in high dimensions, it is straightforward to see (from the remarks following Theorem 1) that, with high probability,

$$\mathrm{argmin}_\rho \, \mathcal{M}(\alpha\hat{\boldsymbol{\beta}}) = \mathrm{argmin}_\rho \, \frac{\|\mathbf{c}\|}{|\mathbf{c}^\mathsf{T}\mathbf{y}|}$$

where $\mathcal{M}$ can be either the classification error function $\mathcal{M}_{\mathrm{C}}$ given by (8) or the estimation error function $\mathcal{M}_{\mathrm{E}}$ in (10).

To put it differently, the search for the optimal loss function ρ can be reduced to the minimization of $\frac{\|\mathbf{c}\|}{|\mathbf{c}^\mathsf{T}\mathbf{y}|}$ with respect to ρ. Now notice that we always have $\mathbf{X}\mathbf{c} = \mathbf{0}$ from (2) and $\mathbf{X} \in \mathbb{R}^{p\times n}$ is of rank p for $n > p$ with probability one. Then consider the singular value decomposition

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}$$

where $\mathbf{U} \in \mathbb{R}^{p\times p}$, $\mathbf{V} \in \mathbb{R}^{n\times n}$ are some unitary matrices such that $\boldsymbol{\Sigma} = \begin{bmatrix}\mathbf{S} & \mathbf{0}\end{bmatrix}$ with $\mathbf{S} \in \mathbb{R}^{p\times p}$ a diagonal matrix with positive diagonal entries. Write $\mathbf{V} = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}$ with $\mathbf{V}_1 \in \mathbb{R}^{n\times p}$ and $\mathbf{V}_2 \in \mathbb{R}^{n\times(n-p)}$. It follows from $\mathbf{X}\mathbf{c} = \mathbf{0}$ that $\mathbf{V}_1^\mathsf{T}\mathbf{c} = \mathbf{0}$. The vector $\mathbf{c} \in \mathbb{R}^n$ thus lies in the subspace spanned by the column vectors of $\mathbf{V}_2$, i.e., for vector $\mathbf{c}_\rho$ from any ρ, there exists a vector $\boldsymbol{\eta}_\rho \in \mathbb{R}^{n-p}$ such that

$$\mathbf{c}_\rho = \mathbf{V}_2\boldsymbol{\eta}_\rho. \tag{15}$$

Since $\frac{\|\mathbf{c}_\rho\|}{|\mathbf{c}_\rho^\mathsf{T}\mathbf{y}|} = \frac{\|\boldsymbol{\eta}_\rho\|}{|\boldsymbol{\eta}_\rho^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{y}|}$ and that $\frac{\|\boldsymbol{\eta}_\rho\|}{|\boldsymbol{\eta}_\rho^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{y}|}$ is minimized at $\boldsymbol{\eta}_* = a\mathbf{V}_2^\mathsf{T}\mathbf{y}$ for any non zero $a \in \mathbb{R}$, we infer that if there exists a loss function ρ_{opt} for which the vector $\mathbf{c}$ is of the form

$$\mathbf{c}_{\mathrm{opt}} = a\mathbf{V}_2\mathbf{V}_2^\mathsf{T}\mathbf{y}, \tag{16}$$

then the high dimensional performance (for both classification and class probability estimation) is optimized by the rescaled classifier $\alpha\hat{\boldsymbol{\beta}}$ obtained with $\rho = \rho_{\mathrm{opt}}$.

As a matter of fact, with the square loss function[2] $\rho(t) = (t-1)^2/2$, the optimization problem in (2) is of explicit solution

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\mathbf{y},$$

which is the least squares classifier. We obtain from (3) that

$$\mathbf{c}_{\mathrm{LS}} = \mathbf{y} - \mathbf{X}^\mathsf{T}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{V}_2\mathbf{V}_2^\mathsf{T}\mathbf{y}, \tag{17}$$

meeting the optimality condition given in (16). This remark, combined with the above arguments, leads to the following proposition on the optimal choice of loss function.

Proposition 1. *Let Assumptions 1 and 2 hold. Denote by $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ the solution of (2) with the square loss function $\rho(t) = (t-1)^2/2$, $\hat{\boldsymbol{\beta}}_{\rho'}$ the solution with some loss function ρ', and $\alpha_{\mathrm{LS}}, \alpha_{\rho'}$ respectively the rescaling factor of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}, \hat{\boldsymbol{\beta}}_{\rho'}$ given in (7). Then, for any given $\boldsymbol{\mu}, \mathbf{C}$ and n/p ratio, we have that*

$$\mathcal{M}(\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) \leq \mathcal{M}(\alpha_{\rho'}\hat{\boldsymbol{\beta}}_{\rho'})$$

with high probability, for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$, regardless of the choice of ρ'.

As we recall, the true parameter vector $\boldsymbol{\beta}_*$ is given by $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$, which can also be consistently estimated by

$$\hat{\boldsymbol{\beta}}_{\mathrm{LDA}} = 2\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}} \tag{18}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\mathbf{X}\mathbf{y}, \hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\mathsf{T}$ are respectively consistent estimators of the true mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$. This approach is commonly known as the linear discriminant analysis (Bishop, 2007). Actually, since

$$\hat{\boldsymbol{\beta}}_{\mathrm{LDA}} = \left[1 - \hat{\boldsymbol{\mu}}^\mathsf{T}(\mathbf{X}\mathbf{X}^\mathsf{T}/n)^{-1}\hat{\boldsymbol{\mu}}\right]^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{LS}},$$

[2] It can be shown that any square loss function of the type $\rho(t) = (t-a)^2/2$ for $a > 0$ yields the same classification performance. We consider $a = 1$ without loss of generality.

with $\hat{\boldsymbol{\mu}}^\mathsf{T}(\mathbf{X}\mathbf{X}^\mathsf{T}/n)^{-1}\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}^\mathsf{T}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}}{1+\hat{\boldsymbol{\mu}}^\mathsf{T}\mathbf{C}^{-1}\hat{\boldsymbol{\mu}}} < 1$ by Sherman-Morrison formula, we observe that $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ is in fact proportional to $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. As such, $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ leads to the same classification results as $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. However, when it comes to the prediction of class probability, the estimation error can be significantly reduced by using $\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ instead of $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$, thanks to the bias-correcting effect (as stated in Corollary 1) of the rescaling factor α_{LS} for finite n/p. This remark is confirmed in Figure 7, where the performance gain of $\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ over $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ in class probability estimation is reported.

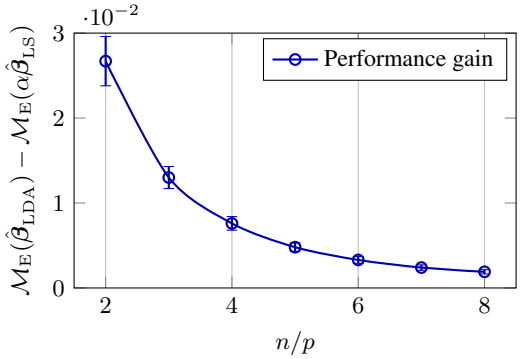

Figure 7. Performance gain $\mathcal{M}_{\mathrm{E}}(\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}) - \mathcal{M}_{\mathrm{E}}(\alpha\hat{\boldsymbol{\beta}}_{\mathrm{LS}})$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

After answering the question of optimality for individual classifiers, we move on to discuss the learning efficiency of the ensemble learning classifiers $\hat{\boldsymbol{\beta}}_{\mathrm{ES}}$ described in Theorem 2. As shown in Figure 5, the ensemble classifier yields superior results when compared to all of its member classifiers. However, since the learning process is always performed on the same training set, there exists certainly a limit for the performance gain achieved by this approach. To inquire into this limit, we develop the arguments below.

Similarly to the performance discussion on (rescaled) individual classifiers, it can be derived from Theorem 2 that

$$\operatorname{argmin}_{\hat{\boldsymbol{\beta}}_{\mathrm{ES}}} \mathcal{M}(\hat{\boldsymbol{\beta}}_{\mathrm{ES}}) = \operatorname{argmin}_{\hat{\boldsymbol{\beta}}_{\mathrm{ES}}} \|\mathbf{c}_{\mathrm{ES}}\|$$

with high probability, for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$. According to (12) and (15), we have

$$\mathbf{c}_{\mathrm{ES}} = \sum_{k=1}^{m} w_k \mathbf{c}_k/(\mathbf{c}_k^\mathsf{T}\mathbf{y}) = \sum_{k=1}^{m} w_k \boldsymbol{\eta}_k/(\mathbf{u}_k^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{y})$$

with $\sum_{k=1}^m w_k = 1$. By decomposing $\boldsymbol{\eta}_k$ as the sum of its projection and rejection on $\mathbf{V}_2^\mathsf{T}\mathbf{y}$, we have

$$\mathbf{c}_{\mathrm{ES}} = \frac{\mathbf{V}_2^\mathsf{T}\mathbf{y}}{\|\mathbf{V}_2^\mathsf{T}\mathbf{y}\|^2} + \sum_{k=1}^{m} w_k \left(\frac{\boldsymbol{\eta}_k}{\boldsymbol{\eta}_k^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{y}} - \frac{\mathbf{V}_2^\mathsf{T}\mathbf{y}}{\|\mathbf{V}_2^\mathsf{T}\mathbf{y}\|^2} \right)$$

where $\frac{\boldsymbol{\eta}_k}{\boldsymbol{\eta}_k^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{y}} - \frac{\mathbf{V}_2^\mathsf{T}\mathbf{y}}{\|\mathbf{V}_2^\mathsf{T}\mathbf{y}\|^2}$ is orthogonal to $\mathbf{V}_2^\mathsf{T}\mathbf{y}$. Therefore,

$$\|\mathbf{c}_{\mathrm{ES}}\| \geq \frac{1}{\|\mathbf{V}_2^\mathsf{T}\mathbf{y}\|}.$$

Moreover, since $\frac{1}{\|\mathbf{V}_2^\mathsf{T}\mathbf{y}\|} = \frac{\|\mathbf{c}_{\mathrm{LS}}\|}{|\mathbf{c}_{\mathrm{LS}}^\mathsf{T}\mathbf{y}|}$ with $\mathbf{c}_{\mathrm{LS}}$ given in (17), we deduce that the norm of $\mathbf{c}_{\mathrm{ES}}$ reaches its minimum at $\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = \alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, and thus conclude on the performance limit of ensemble learning classifier as follows.

Proposition 2. *Let Assumptions 1 and 2 hold, and $\hat{\boldsymbol{\beta}}_{\mathrm{ES}}$ be any ensemble learning classifier of the form (11). Denote by $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ the solution of (2) with the square loss function $\rho(t) = (t-1)^2/2$, and α_{LS} its rescaling factor as defined in (7). Then, for any given $\boldsymbol{\mu}, \mathbf{C}$ and n/p ratio, we have*

$$\mathcal{M}(\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) \leq \mathcal{M}(\hat{\boldsymbol{\beta}}_{\mathrm{ES}})$$

with high probability, for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$.

5. Asymptotic Deterministic Description

As discussed in Section 3, the high dimensional classification performance of $\hat{\boldsymbol{\beta}}$ can be computed with the associated random vector $\mathbf{c}$ via (9). The distribution of $\mathbf{c}$ thus provides a direct access to the classification performance of $\hat{\boldsymbol{\beta}}$ for any loss function. However, as $\mathbf{c}$ is a function of the predicted scores $\mathbf{x}_1^\mathsf{T}\hat{\boldsymbol{\beta}}, \ldots, \mathbf{x}_n^\mathsf{T}\hat{\boldsymbol{\beta}}$ on all training samples, its statistical behavior is difficult to capture since $\hat{\boldsymbol{\beta}}$ depends on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a (generally) implicit manner through the optimization problem in (2). Nonetheless, by considering the regime of large n, p, one can link (as detailed in Supplementary Material) the distribution of $\mathbf{c}$ (and that of the random vector $\mathbf{r}$ defined in (4)) to the predicted score of new data as specified in the following theorem.

Theorem 3. *Let Assumptions 1 and 2 hold, then there exist two positive constants m, σ such that $y\mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(m, \sigma^2)$ for some $(\mathbf{x}, y) \sim \mathcal{D}$ independent of $\hat{\boldsymbol{\beta}}$. For random vectors $\mathbf{c}, \mathbf{r}$ defined in (3) and (4), we have that for all $i \in \{1, \ldots, n\}$,*

$$(y_i r_i, y_i c_i) \xrightarrow{d} (r, g_{\bar{\kappa}}(r))$$

with $r \sim \mathcal{N}(m, \sigma^2)$, the function $g_{\bar{\kappa}} : \mathbb{R} \mapsto \mathbb{R}$ defined as

$$g_{\bar{\kappa}}(t) \equiv \psi\left(\mathrm{prox}_{\bar{\kappa}}(t)\right) \tag{19}$$

where we denote the proximal operator (with respect to ρ) $\mathrm{prox}_{\bar{\kappa}}(t) \equiv \operatorname{argmin}_{z\in\mathbb{R}} \left(\bar{\kappa}\rho(z) + \frac{1}{2}(z-t)^2\right)$ for $\bar{\kappa}$ the unique positive solution of the following fixed point equation

$$\bar{\kappa} = \frac{p/n}{(p/n-1)\mathbb{E}\left[\psi'(\mathrm{prox}_{\bar{\kappa}}(r))\right]}$$

for $\psi'(t)$ the derivative of $\psi(t)$. Moreover, m, σ can be determined by the following system of equations

$$m = \frac{\mathbb{E}[g_{\bar\kappa}(r)]\sigma^2}{m\mathbb{E}[g_{\bar\kappa}(r)] - \mathbb{E}[r g_{\bar\kappa}(r)]} \boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}, \tag{20}$$

$$\sigma = \frac{m}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}}} + \sqrt{\frac{p}{n}} \frac{\sigma^2\sqrt{\mathbb{E}[g_{\bar\kappa}(r)^2]}}{m\mathbb{E}[g_{\bar\kappa}(r)] - \mathbb{E}[r g_{\bar\kappa}(r)]}. \tag{21}$$

Since m, σ introduced in Theorem 3 are given as the solutions of the two deterministic equations (20) and (21), we can obtain the high dimensional classification performance directly from the parameters of data model and the n/p ratio without the actual training of classifier.

Corollary 3. *Under the conditions and notations of Theorem 3, the expected classification error rate is given by*

$$\mathcal{M}_\mathrm{C}(\hat{\boldsymbol{\beta}}) = Q\left(\frac{m}{\sigma}\right) + o_P(1)$$

where we recall that $Q(t) \equiv \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$. Similarly, the training (classification) error is given by

$$P\left(y_i \mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}} < 0\right) = P\left(\mathrm{prox}_{\bar\kappa}(r) < 0\right) + o_P(1).$$

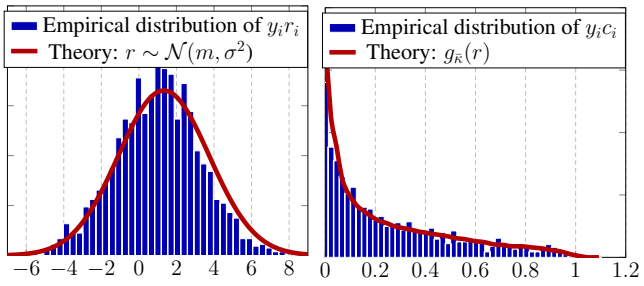

Figure 8. Comparison between the empirical distribution of $y_i r_i$ and $y_i c_i$ with their theoretical prediction given in Theorem 3. For logistic loss $\rho(t) = \ln(1 + e^{-t})$, $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$ and $p = 256$, $n = 6p$.

In Figure 8 we compare the empirical distribution of $y_i r_i$ and $y_i c_i$ with the theoretical predictions in Theorem 3. A close match is observed for $p = 256$ and $n = 6p$ which confirms our theoretical results. In Figure 9 we plot, as numerical validation to Corollary 3, the classification error rate $\mathcal{M}_\mathrm{C}$ and the associated values of $Q\left(\frac{m}{\sigma}\right)$ as a function of the n/p ratio, for logistic and square losses.

6. Conclusion

In this article, we investigated the problem of high dimensional classification within the general framework of empirical risk minimization. We showed that, for the high

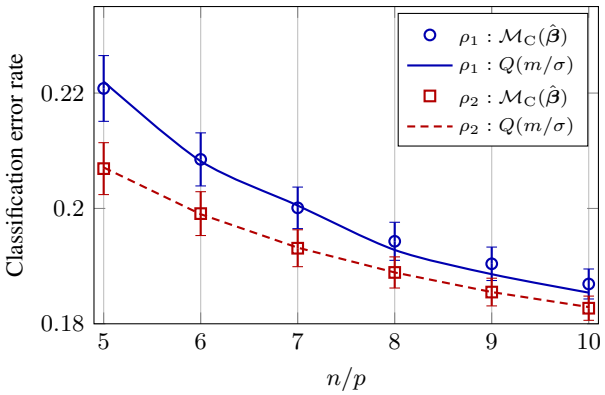

Figure 9. Comparison of classification error rate between the logistic loss $\rho_1(t) = \ln(1 + e^{-t})$, the square loss $\rho_2(t) = (t-1)^2/2$ and the theoretical results given in Corollary 3 with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{p}$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

dimensional mixture model under consideration, all classifiers $\hat{\boldsymbol{\beta}}$ given in (2) are aligned in expectation to the oracle direction, with different scaling factors that depends on the ratio n/p. Based on this result, we proposed the rescaling method to correct this high dimensional bias for an enhanced class probability estimation. We showed subsequently that the square loss solution, instead of the maximal likelihood solution given by the negative log-likelihood loss (i.e., the logistic loss), yields the best results in both classification and class probability estimation after being corrected by the proposed rescaling strategy. Our analysis served furthermore to statistically characterize linear combinations of classifiers learned with different loss functions, allowing to conclude on the possibility and limitation of this ensemble learning approach.

The proposed analysis framework is generalizable to more generic mixture models of non-Gaussian feature vectors, however at the cost of the readability and interpretability of the theoretical results. The extension to non-smooth and non-convex loss functions is on the other hand more technically challenging. The present study can also be further developed to encompass regularized solutions, as a means to explore the joint effect of loss functions and regularizations (e.g., ℓ_1 or ℓ_2 regularization). It is also of interest to track the evolution dynamics of the underlying optimization problem, for instance as a function of the number of descent steps when solved with gradient-based methods, which is closely related to the training of modern neural networks (Saxe et al., 2013).

Acknowledgments

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006) and by the IDEX GSTATS Chair at University Grenoble Alpes.

References

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Ben-David, S., Eiron, N., and Long, P. M. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2007.

Candès, E. J. and Sur, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

Couillet, R., Liao, Z., and Mai, X. Classification Asymptotics in the Random Matrix Regime. In *26th European Signal Processing Conference (EUSIPCO'2018)*. IEEE, 2018.

Dobriban, E., Wager, S., et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Donoho, D. and Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166 (3-4):935–969, 2016.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, pp. 201307842, 2013.

Elkhalil, K., Kammoun, A., Couillet, R., Al-Naffouri, T. Y., and Alouini, M.-S. A large dimensional analysis of regularized discriminant analysis classifiers. *arXiv preprint arXiv:1711.00382*, 2017.

Freund, Y., Schapire, R., and Abe, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

Fumera, G., Fabio, R., and Alessandra, S. A theoretical analysis of bagging as a linear combination of classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1293–1299, 2008.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Liao, Z. and Couillet, R. A large dimensional analysis of least squares support vector machines. *arXiv preprint arXiv:1701.02967*, 2017.

Masnadi-Shirazi, H. and Vasconcelos, N. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pp. 1049–1056, 2009.

McCullagh, P. and Nelder, J. A. Generalized linear models, vol. 37 of monographs on statistics and applied probability, 1989.

Portnoy, S. et al. Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. i. consistency. *The Annals of Statistics*, 12(4):1298–1309, 1984.

Rojas, R. Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep*, 2009.

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

Sur, P., Chen, Y., and Candès, E. J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv preprint arXiv:1706.01191*, 2017.

Vapnik, V. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pp. 831–838, 1992.

Supplementary Material

High Dimensional Classification via Empirical Risk Minimization: Improvements and Optimality

A. Sketch of Proofs for Theorems

As discussed in Section 1, we shall connect $\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}$ to c_i by establishing a "leave-one-out" version of $\hat{\boldsymbol{\beta}}$ that is independent of $\mathbf{x}_i, y_i$. To this end, we denote $\hat{\boldsymbol{\beta}}_{-i}$ the solution of the original optimization problem in (2) for $\mathbf{X}_{-i}\mathbf{y}_{-i} \equiv [\mathbf{x}_1 y_1, \ldots, \mathbf{x}_{i-1}y_{i-1}, \mathbf{x}_{i+1}y_{i+1}, \ldots, \mathbf{x}_n y_n] \in \mathbb{R}^{p\times(n-1)}$, all training data except the pair $(\mathbf{x}_i, y_i)$, such that by cancelling the derivative we obtain

$$\frac{1}{n}\sum_{j\neq i} y_j \psi(y_j \mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i})\mathbf{x}_j = 0. \tag{22}$$

Recall the definition of $\mathbf{c}$ in (3) and the fact that $\frac{1}{n}\sum_{i=1}^n c_i\mathbf{x}_i = 0$, a simple subtraction from (22) yields

$$\frac{1}{n}\sum_{j\neq i}\left(c_j - y_j\psi(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i})\right)\mathbf{x}_j + \frac{1}{n}c_i\mathbf{x}_i = 0. \tag{23}$$

Since both $\|\hat{\boldsymbol{\beta}}\|$ and $\|\hat{\boldsymbol{\beta}}_{-i}\|$ are bounded and that the difference $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}\| = O(n^{-1/2})$, by performing a Taylor expansion of $\psi(t)$ around $t = y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i}$ we obtain

$$c_j - y_j\psi(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i}) = \psi'(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T}\mathbf{x}_j + O(n^{-1/2})$$

with $\psi'(t) \equiv \frac{d\psi(t)}{dt} < 0$. Plugging the above estimate back into (23) we deduce

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i} = \left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\frac{1}{n}c_i\mathbf{x}_i + O(n^{-1/2})$$

with $\mathbf{D}_{-i} \in \mathbb{R}^{n-1}$ a diagonal matrix with its (j,j)-entry equal to $\psi'(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{-i})$ and the notation $O(n^{-1/2})$ stands for an entry-wise difference of order $O(n^{-1/2})$ with high probability. As a consequence, the inner product $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T}\mathbf{x}_i$ gives

$$(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T}\mathbf{x}_i = \frac{c_i}{n}\mathbf{x}_i^\mathsf{T}\left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\mathbf{x}_i + o_P(1). \tag{24}$$

The right hand side of (24) is a quadratic form $\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{M}\mathbf{x}_i$ for some $\mathbf{M}$ of bounded operator norm (with high probability) and independent of $\mathbf{x}_i$, classical random matrix theory results yield the following approximation.

Lemma 1 (Asymptotic approximation of quadratic form)**.** *Let Assumption 1 and 2 holds. Then,*

$$\frac{1}{n}\mathbf{x}_i^\mathsf{T}\left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\mathbf{x}_i = \bar{\kappa} + o_P(1), \quad \bar{\kappa} = \frac{p/n}{(p/n-1)\mathbb{E}[\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})]}.$$

Proof. We start by computing the expectation of $\frac{1}{n}\mathbf{x}_i^\mathsf{T}\left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\mathbf{x}_i$ as

$$\frac{1}{n}\mathbb{E}\left[\mathbf{x}_i^\mathsf{T}\left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\mathbf{x}_i\right] = \frac{1}{n}\operatorname{tr}\left[\mathbb{E}\left(-\frac{1}{n}\mathbf{X}_{-i}\mathbf{D}_{-i}\mathbf{X}_{-i}^\mathsf{T}\right)^{-1}\mathbf{C}\right] + o_P(1) = \frac{1}{n}\operatorname{tr}(\mathbb{E}[\mathbf{Q}]\mathbf{C}) + o_P(1)$$

where we use, for the first equality the fact that $\mathbf{x}_i$ is independent of the inverse and $\|\boldsymbol{\mu}\|$ is of order $O(1)$ according to Assumption 1, and for the second equality the fact that a rank one perturbation does not change asymptotically the trace of the inverse (see for example Theorem A.43 in (Bai & Silverstein, 2010)). We recall the definition $\mathbf{Q} = \left(-\frac{1}{n}\sum_{i=1}^n \psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)^{-1}$.

Now we move on to compute the expectation $\mathbb{E}[\mathbf{Q}]$. In fact by denoting[3]

$$\bar{\mathbf{Q}} \equiv \left(\mathbb{E}\left[\frac{-\psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}})}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \tilde{\kappa}} \right] (\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathbf{C}) \right)^{-1}$$

with $\tilde{\kappa} = \frac{1}{n} \operatorname{tr}(\bar{\mathbf{Q}}\mathbf{C})$, we can show in the sequel that the operator norm $\|\bar{\mathbf{Q}} - \mathbb{E}[\mathbf{Q}]\| = o_P(1)$. To this end, with the resolvent identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ we have

$$\begin{aligned}
\mathbb{E}[\bar{\mathbf{Q}} - \mathbf{Q}] &= \bar{\mathbf{Q}} \mathbb{E}\left[\left(-\frac{1}{n} \sum_{j=1}^n \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} - \mathbb{E}\left[\frac{-\psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}})}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \tilde{\kappa}} \right] (\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathbf{C}) \right) \mathbf{Q} \right] \\
&= \bar{\mathbf{Q}} \mathbb{E}\left[-\frac{1}{n} \sum_{j=1}^n \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} \mathbf{Q} \right] - \bar{\mathbf{Q}} \mathbb{E}\left[\frac{-\psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}})}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \tilde{\kappa}} \right] (\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathbf{C}) \mathbb{E}[\mathbf{Q}] \\
&= \bar{\mathbf{Q}} \mathbb{E}\left[-\frac{1}{n} \sum_{j=1}^n \frac{\psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} \mathbf{Q}_{-j}}{1 - \frac{1}{n} \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j^\mathsf{T} \mathbf{Q}_{-j} \mathbf{x}_j} \right] - \bar{\mathbf{Q}} \mathbb{E}\left[\frac{-\psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}})}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \tilde{\kappa}} \right] (\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathbf{C}) \mathbb{E}[\mathbf{Q}]
\end{aligned}$$

where we denote $\mathbf{Q}_{-j} \equiv \left(-\frac{1}{n} \sum_{k \neq j} \psi'(y_k \mathbf{x}_k^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_k \mathbf{x}_k^\mathsf{T} \right)^{-1}$ and apply the Sherman-Morrison formula to obtain the last equation. Note here that in the denominator we obtain again the quadratic form $\frac{1}{n} \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j^\mathsf{T} \mathbf{Q}_{-j} \mathbf{x}_j$. Since we have $\|\mathbf{Q} - \mathbf{Q}_{-j}\| = o_P(1)$ and $\mathbb{E}[\mathbf{x}_j \mathbf{x}_j^\mathsf{T}] = \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \mathbf{C}$ for all j, it remains to show that the quadratic form concentrates around its expectation such that

$$\frac{1}{n} \mathbf{x}_j^\mathsf{T} \mathbf{Q}_{-j} \mathbf{x}_j = \frac{1}{n} \mathbb{E}[\mathbf{x}_j^\mathsf{T} \mathbf{Q}_{-j} \mathbf{x}_j] + o_P(1).$$

which can be achieved either with concentration of measure arguments (Ledoux, 2001) or by bounding its variance with Nash-Poincare inequality and apply Chebyshev's inequality. Ultimately, note that $\tilde{\kappa}$ is a rank one perturbation of $\bar{\kappa}$ so that

$$\bar{\kappa} = \tilde{\kappa} + o_P(1)$$

and hence the conclusion of Lemma 1. □

In particular, recall the definition of κ in (5). We have with Sherman-Morrison formula and Lemma 1 that

$$\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q} \mathbf{x}_i = \frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i + \mathbf{x}_i^\mathsf{T} \frac{\frac{1}{n} \mathbf{Q}_{-i} \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i}}{1 - \frac{1}{n} \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{-i} \mathbf{x}_i} \mathbf{x}_i = \bar{\kappa} + \frac{\psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \bar{\kappa}^2}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \bar{\kappa}} + o_P(1) = \frac{\bar{\kappa}}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \bar{\kappa}} + o_P(1)$$

where we denote $\mathbf{Q}_{-i} = \left(-\sum_{j \neq i}^n \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} \right)^{-1}$. As such, we deduce from (5) that

$$\kappa = \frac{1}{n} \sum_{i=1}^n \frac{\frac{\bar{\kappa}}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \bar{\kappa}}}{1 + \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \frac{\bar{\kappa}}{1 - \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \bar{\kappa}}} + o_P(1) = \bar{\kappa} + o_P(1).$$

We move on to consider the random vectors $\mathbf{c}$ and $\mathbf{r}$, defined respectively in (3) and (4). With Lemma 1 we obtain from (24) the following approximation

$$(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^\mathsf{T} \mathbf{x}_i = \bar{\kappa} c_i + o_P(1)$$

and therefore by definition $r_i = \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i} + o_P(1)$. Also, for $c_i = y_i \psi(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}})$ we get the implicit relation $c_i = y_i \psi(y_i r_i + y_i c_i \kappa) + o_P(1)$, the solution of which can be given via the function $g_{\bar{\kappa}}$ defined in (19) as

$$y_i c_i = g_{\bar{\kappa}}(y_i r_i) + o_P(1) \tag{25}$$

[3] Recall that $\psi'(y_i \mathbf{x}_i \hat{\boldsymbol{\beta}})$ follows the same distribution for all i.

for $y_i = \pm 1$, $i = 1, \ldots, n$. By making the substitution (25) we get from $\frac{1}{n}\sum_{i=1}^n c_i \mathbf{x}_i = \mathbf{0}$ that

$$\frac{1}{n}\sum_{i=1}^n g_{\bar{\kappa}}(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}) y_i \mathbf{x}_i = o_P(1), \tag{26}$$

from which we wish to extract the statistical information of $\hat{\boldsymbol{\beta}}_{-i}$ that is asymptotically closed to that of the original solution $\hat{\boldsymbol{\beta}}$. To this end, we further "separate" explicitly the dependence of $y_i \mathbf{x}_i = \boldsymbol{\mu} + \mathbf{C}^{\frac{1}{2}} \mathbf{z}_i$ (so that $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I_p})$) from the random variable $y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}$ by writing

$$\mathbf{z}_i = \mathbf{z}_i^\perp + \frac{\mathbf{z}_i^\mathsf{T} \mathbf{C}^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_{-i}}{\hat{\boldsymbol{\beta}}_{-i}^\mathsf{T} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i}} \mathbf{C}^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_{-i}. \tag{27}$$

As such, conditioned on $\hat{\boldsymbol{\beta}}_{-i}$ that is independent of $\mathbf{z}_i \in \mathbb{R}^p$, $\mathbf{z}_i^\perp$ lies in the $(p-1)$-dimensional subspace that is orthogonal to $\mathbf{C}^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_{-i}$. As a consequence of the orthogonal invariance of the standard multivariate Gaussian distribution, we know that $\mathbf{z}_i^\perp$ is also Gaussian and independent of $y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}$.

Therefore, (26) can be developed as

$$\frac{1}{n}\sum_{i=1}^n g_{\bar{\kappa}}\left(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}\right) \left(\boldsymbol{\mu} + \mathbf{C}^{\frac{1}{2}} \mathbf{z}_i^\perp + \frac{\mathbf{z}_i^\mathsf{T} \mathbf{C}^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_{-i}}{\hat{\boldsymbol{\beta}}_{-i}^\mathsf{T} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i}} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i}\right) = o_P(1),$$

or equivalently

$$\frac{1}{n}\sum_{i=1}^n \frac{y_i c_i (\mathbb{E}[y_i r_i | \hat{\boldsymbol{\beta}}_{-i}] - y_i r_i)}{\mathrm{Var}[y_i r_i | \hat{\boldsymbol{\beta}}_{-i}]} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i} = \frac{1}{n}\sum_{i=1}^n y_i c_i \boldsymbol{\mu} + \frac{1}{n}\sum_{i=1}^n y_i c_i \mathbf{C}^{\frac{1}{2}} \mathbf{z}_i^\perp + o_P(1) \tag{28}$$

since $\mathbb{E}[y_i r_i | \hat{\boldsymbol{\beta}}_{-i}] = \boldsymbol{\mu}^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i} + o_P(1)$ and $\mathrm{Var}[y_i r_i | \hat{\boldsymbol{\beta}}_{-i}] = \hat{\boldsymbol{\beta}}_{-i}^\mathsf{T} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i} + o_P(1)$. Moreover, it can be deduced from (27) that, conditioned on $\boldsymbol{\beta}_{-i}$,

$$\mathbf{z}_i^\perp \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p - \frac{\mathbf{C}^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_{-i} \hat{\boldsymbol{\beta}}_{-i}^\mathsf{T} \mathbf{C}^{\frac{1}{2}}}{\hat{\boldsymbol{\beta}}_{-i}^\mathsf{T} \mathbf{C} \hat{\boldsymbol{\beta}}_{-i}}\right)$$

which, together with the fact that $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}\| = o_P(1)$, concludes the proof of Theorem 1.

To derive Theorem 2, it suffices to combine (28) for different loss functions $\rho_1, \ldots, \rho_m$ with the associated vectors $\mathbf{c}_k$ and $\mathbf{r}_k$ for $k = 1, \ldots, m$.

We now move on to prove Theorem 3. From (28), we deduce that $y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}$ converges in distribution to a Gaussian random variable $r \sim \mathcal{N}(m, \sigma^2)$, with the parameters m, σ^2 to be determined, for all i. Since $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}\| = o_P(1)$, we have

$$m - \boldsymbol{\mu}^\mathsf{T} \mathbb{E}[\hat{\boldsymbol{\beta}}] = o_P(1), \quad \sigma^2 - \mathbb{E}[\hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{C} \hat{\boldsymbol{\beta}}] = o_P(1) \tag{29}$$

so that m, σ^2 are naturally connected to the statistics of $\hat{\boldsymbol{\beta}}$. Denote the shortcut $\mathbf{z} \equiv \frac{1}{n}\sum_{i=1}^n g_{\bar{\kappa}}(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{-i}) \mathbf{C}^{\frac{1}{2}} \mathbf{z}_i^\perp$ so that

$$\mathbb{E}[\mathbf{z}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{z}\mathbf{z}^\mathsf{T}] = \frac{\mathbb{E}[g_{\bar{\kappa}}^2(r)]}{n}\left(\mathbf{C} - \frac{1}{\sigma^2}\mathbf{C}\mathbb{E}[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\mathsf{T}]\mathbf{C}\right) + o_P(1)$$

with respect to the operator norm. As such, we have, with the law of large numbers that

$$\frac{1}{\sigma^2}\mathbb{E}[(m-r) g_{\bar{\kappa}}(r)] \mathbf{C} \hat{\boldsymbol{\beta}} = \mathbb{E}[g_{\bar{\kappa}}(r)] \boldsymbol{\mu} + \mathbf{z} + o_P(1)$$

and hence

$$\hat{\boldsymbol{\beta}} = \frac{\sigma^2 \mathbb{E}[g_{\bar{\kappa}}(r)]}{\mathbb{E}[(m-r) g_{\bar{\kappa}}(r)]} \mathbf{C}^{-1} \boldsymbol{\mu} + \frac{\sigma^2}{\mathbb{E}[(m-r) g_{\bar{\kappa}}(r)]} \mathbf{C}^{-1} \mathbf{z} + o_P(1).$$

Plugging the above result back into (29) we conclude the proof of Theorem 3.