Journal of Statistical Mechanics:
Theory and Experiment

SISSA

PAPER

Large scale analysis of generalization error in learning using margin based classification methods

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

Large scale analysis of generalization error in learning using margin based classification methods

Hanwen Huang[1] and **Qinglong Yang**[2,*]

[1] Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA 30602, United States of America
[2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, People's Republic of China
E-mail: huanghw@uga.edu and yangqinglong@zuel.edu.cn

Received 21 July 2020
Accepted for publication 28 September 2020
Published 30 October 2020

Online at stacks.iop.org/JSTAT/2020/103407
https://doi.org/10.1088/1742-5468/abbed5

J. Stat. Mech. (2020) 103407

Abstract. Large-margin classifiers are popular methods for classification. We derive the asymptotic expression for the generalization error of a family of large-margin classifiers in the limit of both sample size n and dimension p going to ∞ with fixed ratio $\alpha = n/p$. This family covers a broad range of commonly used classifiers including support vector machine, distance weighted discrimination, and penalized logistic regression. Our result can be used to establish the phase transition boundary for the separability of two classes. We assume that the data are generated from a single multivariate Gaussian distribution with arbitrary covariance structure. We explore two special choices for the covariance matrix: spiked population model and two layer neural networks with random first layer weights. The method we used for deriving the closed-form expression is from statistical physics known as the replica method. Our asymptotic results match simulations already when n, p are of the order of a few hundreds. For two layer neural networks, we reproduce the recently observed 'double descent' phenomenology for several classification models. We also discuss some statistical insights that can be drawn from these analysis.

Keywords: machine learning, statistical inference

[*] Author to whom any correspondence should be addressed.

1742-5468/20/103407+29$33.00

Contents

1. Introduction

Classification is a very useful supervised learning technique for information extraction from data. The goal of classification is to construct a classification rule based on a training set where both covariates and class labels are given. Once obtained, the classification rule can then be used for class prediction of new objects whose covariates are available. There are a large number of methods for classification in the literature. Examples include Fisher linear discrimination analysis, logistic regression, k-nearest neighbor, decision trees, neural networks, boosting, and many others. See Hastie *et al* (2001) for more comprehensive reviews of various classification methods. Among numerous classification techniques, margin-based classifiers have attracted tremendous attentions in recent years due to their competitive performance and ability in handling high dimensional data. The margin-based classifiers focus on the decision boundaries and bypass the requirement of estimating the class probability given input for discrimination.

The support vector machine (SVM) is one of the most well known large margin classifiers. Since its introduction, the SVM has gained much popularity in both machine learning and statistics. However, as pointed out by Marron *et al* (2007), SVM may suffer from a loss of generalization ability in the high-dimension-low-sample size (HDLSS) setting due to data-piling problem. They proposed distance weighted discrimination (DWD) as a superior alternative to SVM. Liu *et al* (2008) proposed a family of large-margin classifiers, namely, the large-margin unified machine (LUM) which embraces both SVM and DWD as special cases. Besides SVM, DWD, and LUM, there are a number of other large margin classifiers introduced in the literature. Examples include the penalized logistic regression (PLR) (Wahba 1999; Lin *et al* 2000), ψ-learning (Shen *et al* 2003), the robust SVM (Wu and Liu 2007), and so on.

Despite some known properties of these methods, a practitioner often needs to face one natural question: which method should one choose to solve the classification problem in hand? The choice can be difficult because typically the behaviors of different classifiers vary from setting to setting. Most of the previous studies in this area are empirical. For example, simulation and real data analysis indicate that DWD performs better than SVM especially in HDLSS cases, see e.g. Benito *et al* (2004); Qiao *et al* (2010); Qiao and Zhang (2015); Wang and Zou (2016); Wang and Zou (2017). Also simulation studies in Liu *et al* (2008) have shown that soft classifiers tend to give more accurate classification results when the true probability functions are relatively smooth. Despite such substantial effort, not too much theoretical studies have been conducted to quantitatively characterize the performance of different classification methods.

The objective of this paper is to follow up on a recent wave of research works aiming at providing sharp performance characterization of classical statistical learning methods including regression, classification, and principle component analysis. Particularly, we derive the asymptotic behavior of margin based classification methods in the limit of both large sample size n and large dimension p with fixed ratio $\alpha = p/n$. The main literature related to this work is represented by a series recent papers which derive asymptotic results for classification in the joint limit $p, n \to \infty$ with $n/p = \alpha$. Huang (2017); Mai and Couillet (2018) studied SVM under Gaussian mixture models in which the data are assumed to be generated from Gaussian mixture distribution with two components, one for each class. The covariance matrix is assumed to follow a spiked population model. Under the same setting, Mai *et al* (2019); Huang and Yang (2019) studied regularized logistic regression and general margin based classification methods respectively. Mignacco *et al* (2020) studied the classification error for PLR and SVM for Gaussian mixture models with standard Gaussian components. Montanari *et al* (2019) studied the hard margin SVM under the single Gaussian model in which the data are assumed to be generated from a single Gaussian distribution. Gerace *et al* (2020) studied the regularized logistic regression under the single Gaussian model with covariance structure generated from two layer neural network model with random first layer weights.

In this paper, we derive the asymptotic performance of the general margin based classification method under the single Gaussian model with arbitrary covariance structure. Our result is quite general in the sense that the family covers many of the aforementioned classifiers such as SVM, DWD, and PLR. Moreover, the covariance structure also includes spiked population model and two layer neural network model as special cases. We derive the analytical results using the heuristic replica method developed in statistical mechanics. Our result provides some insights on the behavior change among different classification methods. It also helps to shed some light on how to select the best model and optimal tuning parameter for a given classification task. As a corollary, we derive the phase transition boundary for the separability of two classes which embraces the previous results in Candès and Sur (2020) and Sifaou *et al* (2019) as special cases.

Moreover, for the two layer neural network covariance structure, our results exhibit the recently observed 'double descent' phenomenon which has been demonstrated empirically in Belkin *et al* (2019a). It is referred to as a peculiar behavior of the test error as a function of overparametrization ratio $\psi_1 = p/n$. Namely, the test error peaks at a critical value of ψ_1 where the training error vanishes, and descends again after that. This picture have been theoretically studied in Belkin et al (2019b); Belkin *et al* (2018); Hastie *et al* (2019) for simple least square estimators. It was also studied in Mei and Montanari (2019) for nonlinear regression and in Gerace *et al* (2020) for logistic regression. Here we can reproduce this phenomenon for general margin based classification methods.

Note that the replica method used in the present work is a non-rigorous calculation procedure that has proved successful in many difficult problems in machine learning. In particular, the replica method has been used to derive a number of fascinating results in the analysis of high-dimensional regression and classification. Rigorous analysis subsequently confirmed these heuristic calculations in several cases. For example, in aforementioned literature, the result in Candès and Sur (2020) was derived rigorously using convex random geometry and the results in Mei and Montanari (2019); Sifaou *et al* (2019); Mignacco *et al* (2020) were derived rigorously using Gaussian comparison methods based on Gordon's inequality. Other rigorous analysis methods include message-passing algorithms (Bayati and Montanari 2012; Gerbelot *et al* 2020) and interpolation techniques motivated from statistical physics (Barbier and Macris 2017). The rigorous work so far mainly focuses on 'i.i.d. randomness', corresponding to the case of standard Gaussian design. For arbitrary covariance structure models considered in the present work, while it remains an open problem to derive a rigorous proof for our results, we shall use simulations under finite size system to provide numerical support that the formula is indeed exact in the high-dimensional limit.

Most closely related to the current paper are results by Gerace *et al* (2020) that also use the heuristic replica method to derive the generalization error for PLR with non-i.i.d. covariance structure. However, the major difference is that Gerace *et al* (2020) only considers the two layer neural network covariance structure, while we focus on more general setting with arbitrary eigenvalue distribution and arbitrary signal decomposition in the basis of the eigenvectors.

The rest of this paper is organized as follows. In section 2, we first present the general result for the asymptotic generalization error of margin based classification methods and then apply it to two special covariance structures: spiked population model and two layer neural network model. The phase transition boundaries under different settings for the separability of two classes are also discussed. In section 3, we demonstrate the numerical analysis of prediction error and compare them with the simulation results based on finite size system. Some discussion is provided in section 4. The technical derivations are collected in the appendix.

2. Main analytical results

2.1. Overview of the margin-based classification method

In the binary classification problem, we are given a training dataset consisting of n observations $\{(\mathbf{x}_i, y_i)\,; i = 1, \cdots, n\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ represents the input vector and $y_i \in \{+1, -1\}$ denotes the corresponding output class label, n is the sample size, and p is the dimension. Assume that the data are drawn i.i.d from an unknown joint probability distribution $P(\mathbf{x}, y)$.

The goal of linear classification is to find a linear function $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$ with $\boldsymbol{\theta} \in \mathbb{R}$ and predict the class labels using $\mathrm{sign}(f(\mathbf{x}))$. Define the functional margin as $yf(\mathbf{x})$ which is larger than 0 if correct classification occurs. In this paper, we focus on large-margin classification methods which can be fit in the regularization framework of loss + penalty. The loss function is used to keep the goodness of fit to the data while the penalty term is to avoid overfitting. Using the functional margin, the regularization formulation of binary large-margin classifiers can be summarized as the following optimization problem

$$\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} V(y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + \sum_{j=1}^{p} J_\tau(\theta_j) \right\}, \tag{1}$$

where $V(\cdot) \geqslant 0$ is a loss function, $J_\tau(\cdot)$ is the regularization term, and $\tau > 0$ is the tuning parameter for penalty.

The general requirement for loss function is convex decreasing and $V(u) \to 0$ as $u \to \infty$. Many commonly used classification techniques can be fit into this regularization framework. The examples include penalized logistic regression (PLR; Lin *et al* (2000)), support vector machine (SVM; Vapnik (1995)), and distance weighted discrimination (DWD; Marron *et al* (2007)). The loss functions of these classification methods are

$$\begin{aligned} \mathrm{PLR}: &\quad V(u) = \log[1 + \exp(-u)], \\ \mathrm{SVM}: &\quad V(u) = (1 - u)_+, \\ \mathrm{DWD}: &\quad V(u) = \begin{cases} 1 - u & \text{if } u \leqslant \dfrac{1}{2} \\ \dfrac{1}{4u} & \text{if } u > \dfrac{1}{2} \end{cases}. \end{aligned}$$

Besides the above methods, many other classification techniques can also be fit into the regularization framework, for example, the large-margin unified machine (Liu *et al* 2011), the AdaBoost in boosting (Freund and Schapire 1997; Friedman *et al* 2000), the import vector machine (IVM; Zhu and Hastie (2005)), and ψ-learning (Shen *et al* 2003).

The commonly used penalty functions include $J_\tau(\theta) = \frac{\tau}{2}\theta^2$ for L_2 regularization and $J_\tau(\theta) = \tau|\theta|$ for sparse L_1 regularization. In this paper, we focus on the standard L_2 regularization.

Figure 1 displays three loss functions: PLR, SVM, and DWD. Note that all loss functions have continuous first order derivatives except the hinge loss of SVM which is not differentiable at $u = 1$. Among the three loss functions, PRL has all order derivatives while DWD only has first order derivative. As $u \to -\infty$, $V(u) \to -u$ for all methods. As

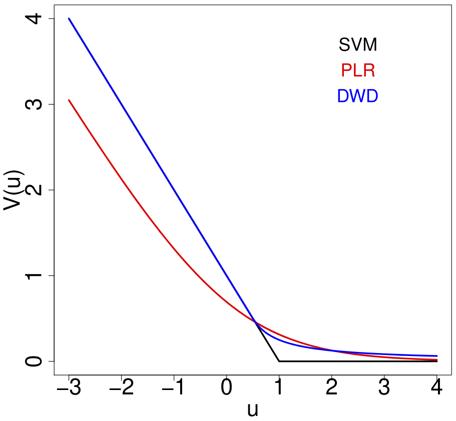

Figure 1. Plots of various loss functions.

$u \to \infty$, $V(u)$ decays to 0 but with different speeds. The fastest one is SVM, followed by PLR and DWD. We will see in section 3 that the decay speed of the loss function has big influence on the classification performance in situations where the tuning parameter τ is small.

2.2. Asymptotic generalization error

For the training data, denote the design matrix as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^{\mathrm{T}}$ and the response vector as $\mathbf{y} = [y_1, \dots, y_n]$. Let the test error be defined by

$$\mathcal{E}(\mathbf{y}, \mathbf{X}) = P(y_{\text{new}} \mathbf{x}_{\text{new}}^{\mathrm{T}} \hat{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{X}) \leqslant 0),$$

where expectation is with respect to a fresh sample $(y_{\text{new}}, \mathbf{x}_{\text{new}})$ independent of the training data $(\mathbf{y}, \mathbf{X})$. We will sometimes refer to $\mathcal{E}(\mathbf{y}, \mathbf{X})$ as the prediction error. We will determine the precise asymptotics of the test error in the limit of $n, p \to \infty$ with $n/p \to \alpha \in (0, \infty)$.

We assume covariates $\mathbf{x}_i \sim N(0, \mathbf{\Sigma})$ to be independent draws from a p-dimensional centered Gaussian with covariance $\mathbf{\Sigma}$ and responses to be distributed according to

$$P(y_1 = +1 | \mathbf{x}_i) = 1 - P(y_1 = -1 | \mathbf{x}_i) = g(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}_\star) \tag{2}$$

for some vector $\boldsymbol{\theta}_\star \in \mathbb{R}^p$ and monotone nonlinear function $g(\cdot) \colon \mathbb{R} \to [0, 1]$. In what follows we will index sequence of instances by $n \in \mathbb{N}$, and it will be understood that $p = p_n$. In order for the limit to exist and be well defined, we need to make specific assumptions about the behavior of the covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_n$ and the true parameters vector $\boldsymbol{\theta}_\star = \boldsymbol{\theta}_{\star,n}$. Let $\mathbf{\Sigma}_n = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}}$ be the eigenvalue decomposition of $\mathbf{\Sigma}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ and $\mathbf{v}_j \in \mathbb{R}^p$ being orthonormal vectors for $1 \leqslant j \leqslant p$. Similar to Montanari *et al* (2019), our first assumption requires that $\mathbf{\Sigma}$ is well conditioned.

Assumption 1. Let $\lambda_{\min}(\mathbf{\Sigma}_n) = \lambda_p(\mathbf{\Sigma}_n)$ and $\lambda_{\max}(\mathbf{\Sigma}_n) = \lambda_1(\mathbf{\Sigma}_n)$, then $\lambda_1(\mathbf{\Sigma}_n) = O_p(1)$ and $\lambda_p(\mathbf{\Sigma}_n) = O_p(1)$.

Assumption 1 indicates that there exist constants $C_1, C_2 \in (0, \infty)$ such that,

$$C_1 \leqslant \lambda_{\min}\ (\boldsymbol{\Sigma}_n) \leqslant \lambda_{\max}\ (\boldsymbol{\Sigma}_n) \leqslant C_2.$$

Our second assumption concerns the eigenvalue distribution of $\boldsymbol{\Sigma}_n$ as well as the decomposition of $\boldsymbol{\theta}_{\star,n}$ in the basis of eigenvectors of $\boldsymbol{\Sigma}_n$.

Assumption 2. Let $\lim_{n\to\infty}\|\boldsymbol{\theta}_{\star,n}\|_2 = c$, $\rho_n = (\boldsymbol{\theta}_{\star,n}^{\mathrm{T}}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_{\star n})^{1/2}$, and $w_j = \sqrt{p\lambda_j}\boldsymbol{\theta}_{\star,n}^{\mathrm{T}}\mathbf{v}_j/\rho_n$. Then the empirical distribution of $\{(\lambda_j, w_j)\}_{1\leqslant j\leqslant p}$ converges to a probability distribution μ_{p} on $\mathbb{R}_{>0}\times\mathbb{R}$

$$\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j,w_j} \to \mu_{\mathrm{p}}.$$

In particular, $\int w^2\mu_{\mathrm{p}}(\mathrm{d}\lambda, \mathrm{d}w) = 1$, and $\rho_n \to \rho$, where $1/\rho^2 = \int (w^2/c\lambda)\mu_{\mathrm{p}}(\mathrm{d}\lambda, \mathrm{d}w)$.

Let us begin by introducing some functions. For a given loss function $V(u)$, we define the proximal operator function

$$\psi(a, b) = \mathrm{argmin}_u \left\{ V(u) + \frac{(u-a)^2}{2b} \right\}, \tag{3}$$

for $b > 0$ which can be considered as the solution of equation

$$\partial V(u) + \frac{u-a}{b} = 0,$$

where $\partial V(u)$ is one of the sub-gradients of $V(u)$. For convex $V(u)$, this equation has unique solution. Specifically, for SVM loss, we have closed form expression

$$\psi(a, b) = \begin{cases} a & \text{if } a \geqslant 1 \\ 1 & \text{if } 1-b \leqslant a < 1. \\ a+b & \text{if } a < 1-b \end{cases} \tag{4}$$

For DWD loss, we have

$$\psi(a, b) = \begin{cases} a+b & \text{if } a \leqslant 1/2 - b \\ \tilde{u} & \text{if } a > 1/2 - b \end{cases},$$

where $\tilde{u}$ is the solution of the cubic equation $4u^3 - 4au^2 - b = 0$. For other loss functions, we have to rely on certain numeric algorithms. Particularly for logistic loss, we can easily implement Newton-Raphson algorithm because the loss function has closed form second order derivatives.

Define functions $\phi_1(\cdot,\cdot,\cdot)$, $\phi_2(\cdot,\cdot,\cdot)$, and $\phi_3(\cdot,\cdot,\cdot)$ on $\mathbb{R}_{>0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$ as

$$\phi_1(c_1, c_2, q) = E\left\{[\psi(c_1YZ_1 + c_2YZ_2, q) - c_1YZ_1 - c_2YZ_2]YZ_1\right\},$$
$$\phi_2(c_1, c_2, q) = E\left\{[\psi(c_1YZ_1 + c_2YZ_2, q) - c_1YZ_1 - c_2YZ_2]YZ_2\right\},$$

$$\phi_3(c_1,c_2,q) = E\left\{[\psi(c_1YZ_1+c_2YZ_2,q)-c_1YZ_1-c_2YZ_2]^2\right\},$$

where

$$Z_2 \perp (Y,Z_1),\ Z_1\sim N(0,1),\ Z_2\sim N(0,1),$$
$$P(Y=+1|Z_1)=g(\rho Z_1),\ P(Y=-1|Z_1)=1-g(\rho Z_1).$$

We further define the asymptotic generalization error $\mathcal{E}^\star$ by

$$\mathcal{E}^\star(\mu_{\mathrm{p}},\alpha,\tau) = P\left(\frac{R^\star}{\sqrt{q_0^\star - R^{\star 2}}}YZ \leqslant 0\right), \tag{5}$$

where probability is over Z, Y with $Z\sim N(0,1)$ and $P(Y=+1|Z)=g(\rho Z)=1-P(Y=-1|Z)$ and $q_0^\star$ and $R^\star$ are the solution of the following equations:

$$\xi_0 = \frac{\alpha}{q^2}\phi_3\left(R,\sqrt{q_0-R^2},q\right), \tag{6}$$

$$\xi = -\frac{\alpha\phi_2\left(R,\sqrt{q_0-R^2},q\right)}{q\sqrt{q_0-R^2}}, \tag{7}$$

$$\hat{R} = \frac{\alpha}{q}\left[\phi_1\left(R,\sqrt{q_0-R^2},q\right) - \frac{R\phi_2\left(R,\sqrt{q_0-R^2},q\right)}{\sqrt{q_0-R^2}}\right], \tag{8}$$

$$q_0 = \xi_0 f_2(\xi,\tau) + \hat{R}^2 f_3(\xi,\tau), \tag{9}$$

$$R = \hat{R} f_1(\xi,\tau), \tag{10}$$

$$q = f_0(\xi,\tau), \tag{11}$$

where

$$f_0(\xi,\tau) = \int \frac{X}{\xi X+\tau}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W), \qquad f_1(\xi,\tau) = \int \frac{W^2X}{\xi X+\tau}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W),$$
$$f_2(\xi,\tau) = \int \frac{X^2}{(\xi X+\tau)^2}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W), \quad f_3(\xi,\tau) = \int \frac{W^2X^2}{(\xi X+\tau)^2}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W). \tag{12}$$

Our main mathematical results are based upon the following proposition for the asymptotic prediction error of the estimators $\hat{\boldsymbol{\theta}}$ obtained from (1).

Proposition 1. *Consider i.i.d. data* $(\mathbf{y},\mathbf{X}) = \{(y_i,\mathbf{x}_i)\}_{i\leqslant n}$ *where* $\mathbf{x}_i\sim N(0,\boldsymbol{\Sigma}_n)$ *and* $P(y_i=+1|\mathbf{x}_i) = g(\mathbf{x}_i^T\boldsymbol{\theta}_{\star,n})$. *Under assumptions 1 and 2, in the limit of* $n,p\to\infty$ *with* $n/p\to\alpha$ *for some positive constants* α. *Let* $\mathcal{E}_n(\mathbf{y},\mathbf{X}) = P(y_{\mathrm{new}}\mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\hat{\boldsymbol{\theta}}(\mathbf{y},\mathbf{X})\leqslant 0)$ *and* $\mathcal{E}^\star$ *be determined as per definition (5). Then we have, almost surely*

$$\lim_{n\to\infty}\mathcal{E}_n(\mathbf{y},\mathbf{X}) \to \mathcal{E}^\star(\mu_p,\alpha,\tau).$$

The derivation is given in the appendix A based on the replica method developed in statistical physics. Proposition 1 allows us to assess the performance of different classification methods and obtain the tuning parameter value of τ that yields the maximum precision for a given method.

2.3. Phase transition

In this section, we derive the phase transition for the non-regularized classification methods which solve the following optimization problem

$$\mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\sum_{i=1}^{n} V(y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta})\right\}. \tag{13}$$

A special case is that if one chooses logistic loss $V(\cdot)$, this is equivalent to the maximum likelihood estimator of logistic regression. It is well-known that the solution of (13) does not exist in all situations, even when the number of covariates p is much smaller than the sample size n. For instance, if the n data points $(\mathbf{x}_i, y_i)$ are completely linear separated in the sense that we can find a vector $\mathbf{b}\in\mathbb{R}^p$ with the property $y_i\mathbf{x}_i^{\mathrm{T}}\mathbf{b} > 0$, for all i, then the solution of (13) does not exist. If the data points overlap in the sense that for every $\mathbf{b}\neq 0$, there is at least one data point satisfying $y_i\mathbf{x}_i^{\mathrm{T}}\mathbf{b} > 0$ and at least another one satisfying $y_i\mathbf{x}_i^{\mathrm{T}}\mathbf{b} < 0$, the solution of (13) does exist. Therefore, the existence for the non-regularized classification methods undergoes a phase transition. Cover (1965) studied the phenomenon in special case where y_i is independent of $\mathbf{x}_i$. This result was recently generalized by Candès and Sur (2020) under the significantly more challenging setting in which $P(y_i = +1|\mathbf{x}_i) = 1/[1 + \exp(-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_\star)]$ and $\mathbf{x}_i$ is Gaussian. Here we derive a more general result. The following corollary allows one to characterize the minimum number of training samples per dimensions that are required in order for the non-regularized classification method (13) to have a solution.

Corollary 1. *Define $\alpha_{\min}(\rho)$ as*

$$1/\alpha_{\min}\ (\rho) = \min_{c\in\mathbb{R}} E\left\{(cYZ_1 + Z_2)_+^2\right\} \tag{14}$$

where $x_+ =\ \max(x, 0)$ and

$$Z_2\perp(Y, Z_1),\ Z_1\sim N(0,1),\ Z_2\sim N(0,1),$$
$$P(Y=+1|Z_1) = g(\rho Z_1),\ P(Y=-1|Z_1) = 1 - g(\rho Z_1).$$

In the setting from section 2.1, if the sample size is larger enough such that $\alpha > \alpha_{\min}$, then the solution of equation (13) asymptotically exists with probability one. Conversely, if $\alpha < \alpha_{\min}$, then the solution does not exist with probability one.

Corollary 1 is a generalization of the result of Candès and Sur (2020), which concerns the phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression, i.e. $g(x)$ defined in (2) is a logistic function. Corollary 1 applies to any cumulative distribution function $g(x)$.

Note that our result is equivalent to establishing the maximum number of training samples per dimensions below which the hard-margin SVM can have solution as shown in Montanari *et al* (2019). The reason is that the hard-margin SVM can only be used if the two classes in the training data are linearly separable with a positive margin. If this was not the case, the optimization problem of the hard-margin SVM would be unfeasible. Such a situation is likely to occur as a larger number of training data is used.

For comparison, now we generalize the phase transition result for data drawn from a Gaussian mixture distribution studied in Sifaou *et al* (2019). Let us specify the joint probability distribution $P(\mathbf{x}, y)$ in that scenario. Conditional on $y = \pm 1$, $\mathbf{x}$ follows multivariate Gaussian distributions $P(\mathbf{x}|y = \pm 1)$ with mean $\pm\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$. Here $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ denotes the $p \times p$ positive definite matrices. From this model, we obtain the conditional distribution of y given $\mathbf{x}$ as

$$\begin{aligned} P(y = +1|\mathbf{x}) &= \frac{\exp\{-(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\}}{\exp\{-(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\} + \exp\{-(\mathbf{x}+\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}+\boldsymbol{\mu})/2\}} \\ &= \frac{1}{1+\exp(-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x})}, \end{aligned} \tag{15}$$

which, by comparing to (2), is equivalent to the logistic $g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_\star)$ with coefficient $\boldsymbol{\theta}_\star = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The following proposition characterizes the phase transition of this model in terms of the overall magnitude of the regression coefficient defined as $\rho^2 = \boldsymbol{\theta}_\star^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star = 4\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Proposition 2. *Define $\alpha_{\min}(\rho)$ as the solution of*

$$1 = \alpha\int_{-\infty}^{z_c}(z_c - z)^2 Dz + \left\{\alpha\rho\int_{-\infty}^{z_c}(z_c - z) Dz\right\}^2, \tag{16}$$

where $\Phi(z_c) = 1/\alpha$ and $Dz = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\mathrm{d}z$. In the above Gaussian mixture setting, if the sample size per dimensions is larger enough such that $\alpha > \alpha_{\min}$, then the solution of equation (13) asymptotically exists with probability one. Conversely, if $\alpha < \alpha_{\min}$, then the solution does not exist with probability one.

Note that the critical value $\alpha_{\min}$ depends on $\boldsymbol{\Sigma}$ only through the overall magnitude of the regression coefficient ρ. Proposition 2 generalizes the result of Sifaou *et al* (2019) for hard margin SVM which can be considered as a special case here if one chooses $\boldsymbol{\Sigma} = \mathbf{I}_p$, where $\mathbf{I}_p$ is p-dimensional identity matrix. Mignacco *et al* (2020) also studied the phase transition for the separability of the data drawn from Gaussian mixture distribution with $\boldsymbol{\Sigma} = \mathbf{I}_p$ but random $\boldsymbol{\mu} \sim N(0, \mathbf{I}_p)$.

2.4. Special examples

In this section we illustrate our main results presented in section 2 by considering a few special cases, namely special sequences of the true parameter vector $\boldsymbol{\theta}_{\star,n}$, and covariance matrix $\boldsymbol{\Sigma}_n$.

2.4.1. Spiked population model. We begin by considering data sets generated from the spiked covariance models which are particularly suitable for analyzing high dimensional statistical inference problems, because, for high dimensional data, typically only few components are scientifically important. The remaining structures can be considered as i.i.d. background noise. Therefore, we use a low-rank signal plus noise structure model (Ma 2013; Liu *et al* 2008), and assume that each observation vector $\mathbf{x}$ can be viewed as an independent sample from the generative models

$$\mathbf{x} = \sum_{k=1}^{K} \sqrt{\lambda_k}\mathbf{v}_k z_k + \boldsymbol{\epsilon}, \tag{17}$$

where $\lambda_k > 0$, $\mathbf{v}_k \in \mathbb{R}^p$ are orthonormal vectors, i.e. $\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_k = 1$ and $\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_{k'} = 0$ for $k \neq k'$. The random variables $z_1, \dots, z_K$ are i.i.d $N(0,\ 1)$. The elements of the p-vector $\boldsymbol{\epsilon} = \{\epsilon_1, \dots, \epsilon_p\}$ are i.i.d $N(0, 1)$ which are independent of z_k. In model (17), λ_k represents the strength of the k th signal component. The real signal is typically low-dimensional, i.e. $K \ll p$. Note that the eigenvalue λ_k is not necessarily decreasing in k and λ_1 is not necessarily the largest eigenvalue. From (17), the covariance matrix becomes

$$\boldsymbol{\Sigma} = \mathbf{I}_p + \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}}. \tag{18}$$

The kth eigenvalue of $\boldsymbol{\Sigma}$ is $1 + \lambda_k$ for $k = 1, \dots, K$ and 1 for $k = K + 1, \dots, p$.

Denote the projections of $\boldsymbol{\theta}_\star$ on eigenvectors as $R_k = \mathbf{v}_k^{\mathrm{T}}\boldsymbol{\theta}_\star/\|\boldsymbol{\theta}_\star\|$ for $k = 1, \dots, K$; $R_{K+1} = \sqrt{1 - \sum_{k=1}^{K} R_k^2}$; and $R_k = 0$ for $k = K + 2, \dots, p$. Substituting into (12), we have

$$f_0(\xi,\tau) = \frac{1}{\xi+\tau}, \qquad f_1(\xi,\tau) = \frac{1}{\sum_{k=1}^{K+1}(1+\lambda_k)R_k^2}\sum_{k=1}^{K+1}\frac{(1+\lambda_k)^2R_k^2}{(1+\lambda_k)\xi+\tau},$$

$$f_2(\xi,\tau) = \frac{1}{(\xi+\tau)^2}, \quad f_3(\xi,\tau) = \frac{1}{\sum_{k=1}^{K+1}(1+\lambda_k)R_k^2}\sum_{k=1}^{K+1}\frac{(1+\lambda_k)^3R_k^2}{[(1+\lambda_k)\xi+\tau]^2}.$$

2.4.2. A random features model. We next consider a special structure of $(\boldsymbol{\Sigma}, \boldsymbol{\theta}_\star)$ that captures the behavior of nonlinear random feature models, i.e. two-layers neural networks with random first layer weights. Random features methods were originally studied by Neal (1996), Balcan *et al* (2006), and Rahimi and Recht (2008). It was suggested in Goldt *et al* (2019); Aubin *et al* (2019); Mei and Montanari (2019); Gerace *et al* (2020) that the behavior of multilayer networks can be well approximated by certain random features model. Goldt *et al* (2020) proved that asymptotic behavior of the random feature models is the same as an appropriately chosen Gaussian feature model. Therefore, the two-layer neural network model can be fit within our general setting.

Assume that we perform classification on a training dataset consisting of n observations $\{(\mathbf{x}_i, y_i)\,; i = 1, \dots, n\}$ generated by the latent variable $\mathbf{z}_i \in N(0, \mathbf{I}_d)$ through the following mechanism. The features $\mathbf{x}_i$ are generated according to $x_{ij} = \sigma(\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_i)$ where $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear function and $\mathbf{w}_j$ are d-dimensional vectors drawn from $N(0, \mathbf{I}_d/\sqrt{d})$. The labels $y_i \in \{+1, -1\}$ are generated according to $P(y_i = +1|\mathbf{z}_i) = f_+(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_\star)$, where $\boldsymbol{\beta}_\star \sim N(0, \mathbf{I}_d/\sqrt{d})$. Denote $\mathbf{W} \in \mathbb{R}^{p\times d}$ the matrix with row $\mathbf{w}_j$, $1 \leqslant j \leqslant p$, we have $\mathbf{x}_i = \sigma(\mathbf{W}\mathbf{z}_i)$ which can be described as a two layers neural network with random first-layer weights $\mathbf{W}$.

Without loss of generality, we assume $E\{\sigma(Z)\}=0$ with $Z\sim N(0,1)$. According to Montanari *et al* (2019), the activation function can be decomposed as

$$\sigma(u)=\gamma_1 u+\gamma_\star\sigma_\perp(u),$$

where $\gamma_1=E\{Z\sigma(Z)\}$ and $\gamma_\star^2=E\{\sigma(Z)^2\}-E\{Z\sigma(Z)\}^2-E\{\sigma(Z)\}^2$. Then the above random feature model can be described as

$$\begin{aligned} x_{ij}&=\gamma_1\mathbf{w}_j^{\mathrm{T}}\mathbf{z}_i+\gamma_\star\xi_{ij},\quad \xi_{ij}\perp\mathbf{z}_i,\quad \xi_{ij}\sim N(0,1),\\ g_i&=\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_\star,\quad P(y_i=+1|g_i)=f_+(g_i). \end{aligned}$$

Note that under this model $\mathbf{x}_i$ and g_i are jointly Gaussian with $\mathbf{x}_i\sim N(0,\mathbf{\Sigma})$, and conditional on $\mathbf{x}_i$, g_i is normal with mean $\gamma_1\boldsymbol{\beta}_\star^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}_i$ and variance $\boldsymbol{\beta}_\star^{\mathrm{T}}\boldsymbol{\beta}_\star-\gamma_1^2\boldsymbol{\beta}_\star^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{W}\boldsymbol{\beta}_\star$, where $\mathbf{\Sigma}=\gamma_1^2\mathbf{W}\mathbf{W}^{\mathrm{T}}+\gamma_\star^2\mathbf{I}_p$. For sign activation function $y_i=$ sign(g_i), $\gamma_1=\sqrt{2/\pi}$ and $\gamma_\star=\sqrt{1-2/\pi}$, we have

$$f_+(g_i)=P(\mathrm{sign}(g_i)=+1)=E(g_i\geqslant 0)=\Phi(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_\star/\tilde{\tau}),\tag{19}$$

where $\boldsymbol{\theta}_\star=\gamma_1\mathbf{\Sigma}^{-1}\mathbf{W}\boldsymbol{\beta}_\star$, $\tilde{\tau}^2=\boldsymbol{\beta}_\star^{\mathrm{T}}\boldsymbol{\beta}_\star-\gamma_1^2\boldsymbol{\beta}_\star^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{W}\boldsymbol{\beta}_\star$, and $\Phi(\cdot)$ denotes the standard Gaussian distribution function. By Marchenko–Pastur's law, the empirical spectral distribution of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ converges to μ_s almost surely as $p,d\to\infty$ with $p/d\to\psi_1$, where

$$\begin{aligned} \mu_s(\mathrm{d}x)&=\begin{cases}(\psi_1-1)\delta_0+\nu_{1/\psi_1}(x)\mathrm{d}x & \text{if } \psi_1\geqslant 1\\ \nu_{\psi_1}(x)\mathrm{d}x & \text{if } \psi_1\in(0,1],\end{cases}\\ \nu_\lambda&=\frac{\sqrt{(\lambda_+-x)(x-\lambda_-)}}{2\pi\lambda x},\\ \lambda_\pm&=(1\pm\sqrt{\lambda})^2. \end{aligned}$$

Denote the decomposition of $\mathbf{W}$ as $\mathbf{W}=\sum_{i=1}^p\sqrt{s_i}\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}$, where the orthonormal vectors $\mathbf{v}\in\mathbb{R}^p$ and $\mathbf{u}\in\mathbb{R}^d$. Then we have $\mathbf{\Sigma}=\sum_{i=1}^p\lambda_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}$ with $\lambda_i=\gamma_1^2s_i+\gamma_\star^2$. According to the definition of $\rho^2=\boldsymbol{\theta}_\star^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\theta}_\star$ and $w_i=\sqrt{p\lambda_i}\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\theta}_\star/\rho$, we can derive

$$\begin{aligned} \rho^2&=\gamma_1^2\boldsymbol{\beta}_\star^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{W}\boldsymbol{\beta}_\star=\sum_{i=1}^p\frac{\gamma_1^2s_i(\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\beta}_\star)^2}{\gamma_1^2s_i+\gamma_\star^2}\to\psi_1E\frac{\gamma_1^2\tilde{X}}{\gamma_1^2\tilde{X}+\gamma_\star^2},\\ w_i&=\sqrt{p\lambda_i}\gamma_1\frac{\sqrt{s_i}(\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\beta}_\star)}{\rho\lambda_i}\to\frac{\gamma_1\sqrt{\psi_1\tilde{X}}Z}{\rho(\gamma_1^2\tilde{X}+\gamma_\star^2)^{1/2}},\\ \tilde{\tau}^2&\to 1-\psi_1E\frac{\gamma_1^2\tilde{X}}{\gamma_1^2\tilde{X}+\gamma_\star^2}=1-\rho^2, \end{aligned}$$

where $\tilde{X}\sim\mu_s$ independent of $Z\sim N(0,1)$. Then the joint distribution of λ,w converges to Law(X,W), where

$$X = \gamma_1^2 \tilde{X} + \gamma_\star^2, W = \frac{\gamma_1 \sqrt{\psi_1 \tilde{X}} Z}{\rho(\gamma_1^2 \tilde{X} + \gamma_\star^2)^{1/2}}.$$

3. Numerical analysis

In this section, we apply the general theoretical results derived in section 2 to three specific classification methods PLR, SVM, and DWD by numerically solving the non-linear equations (6)–(11) using the corresponding loss functions. The performance of a classification method is measured in terms of test error where the probability is over a fresh data point. Our theoretical results are verified using numerical simulations under finite size system. We aim at exploring and comparing different types of classifiers under various settings. In section 3.1, we present the phase transition boundary for the separability of two classes under several settings. Then we compare the test errors of three classification methods under the spiked population model in section 3.2 and the two layer neural network model in section 3.3.

3.1. Phase transition

Figure 2 displays the phase transition boundaries in the plane of ρ and $1/\alpha$ for the separability of the two classes under different settings. Above the curve is the region where the probability of separating the two classes tends to one and below is the region where the probability of separating the two classes tends to zero. It can be seen that under the same α, the single Gaussian model needs larger ρ value in order to be separated than the two Gaussian mixture model. This indicates that the data generated from a two Gaussian mixture model are easier to be separated than from a single Gaussian model. For the single Gaussian model, the data generated based on a probit distribution is easier to be separated than the data generated based on a logit distribution.

3.2. Spiked population model

To examine the validity of our analysis and to determine the finite-size effect, we first present some Monte Carlo simulations to confirm that our theoretical estimation derived in section 2.2 is reliable. Figure 3 plots the test error as a function of tuning parameter τ. The comparison between our asymptotic estimations and simulations on finite dimensional datasets are also provided. We use the *R* packages *kernlab*, *glmnet*, and *DWDLargeR* for solving SVM, PLR, and DWD classification problem respectively. Here the dimension of the simulated data $p = 300$ and the data are generated according to (17) for spiked population model with i.i.d standard normal noise. We repeat the simulation 20 times for each parameter setting. The mean and standard errors over 20 replications are presented. From figure 3, we can see that our analytical curves show fairly good agreement with the simulation experiment. Thus our analytical formula (5) provides reliable estimates for average precision even under moderate system sizes.

Figure 4 compares the performance of three classification methods after optimally tuning the parameter τ. Define $\mu = \|\boldsymbol{\theta}_\star\|$. The left panel represents the dependence on

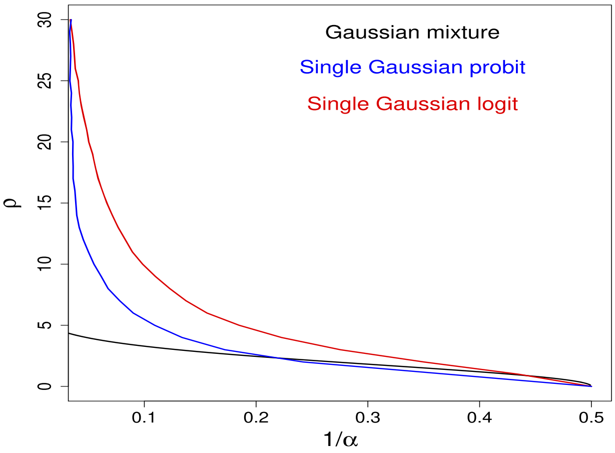

Figure 2. Theoretical prediction for the phase transition curves. The black curve represents the boundary for Gaussian mixture model. The blue and red curves represent the boundaries for single Gaussian model with the distribution functions being probit and logit respectively.

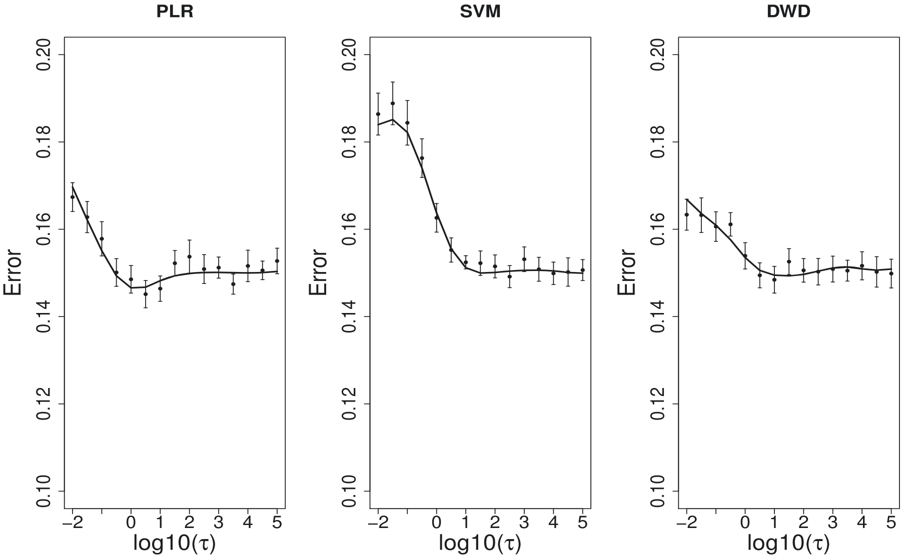

Figure 3. Dependence of generalization error on the tuning parameter τ for different methods under spiked population model. Here $\alpha = 2$ and the number of spikes $K = 2$. The two spiked eigenvalues $\lambda_1 = \lambda_2 = 4$. The two projections $R_1 = 1/\sqrt{2}$ and $R_2 = 0$. The simulations are based on 20 samples with dimension $p = 300$.

α with μ fixed while the right panel represents the dependence on μ with α fixed. In both cases except for small μ, PLR performs the best and SVM performs the worst while DWD is in between. For small μ with fixed α, SVM is slightly better than DWD as shown by the inset in the right plot. We have tried other settings for the spiked covariance structure and the conclusions are very similar.

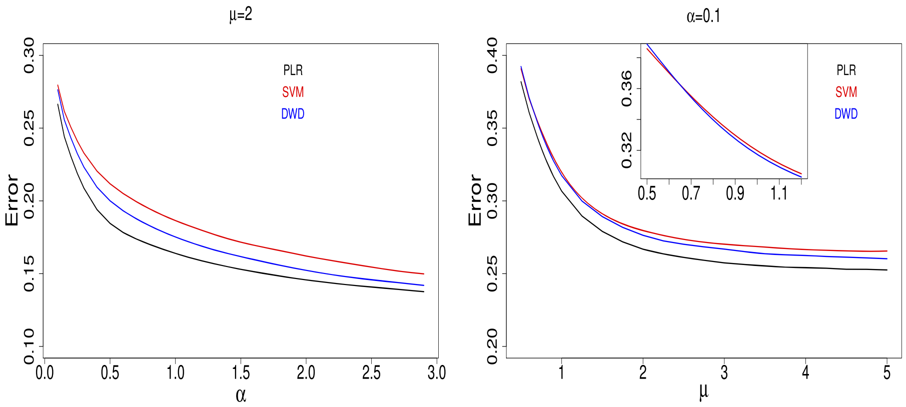

Figure 4. Performance comparison of three classifiers at optimal tuning τ under spiked population model. Here the number of spikes $K = 2$. The two spiked eigenvalues $\lambda_1 = \lambda_2 = 4$. The two projections $R_1 = 1/\sqrt{2}$ and $R_2 = 0$. The inset in the right plot shows the comparison of SVM and DWD for small μ.

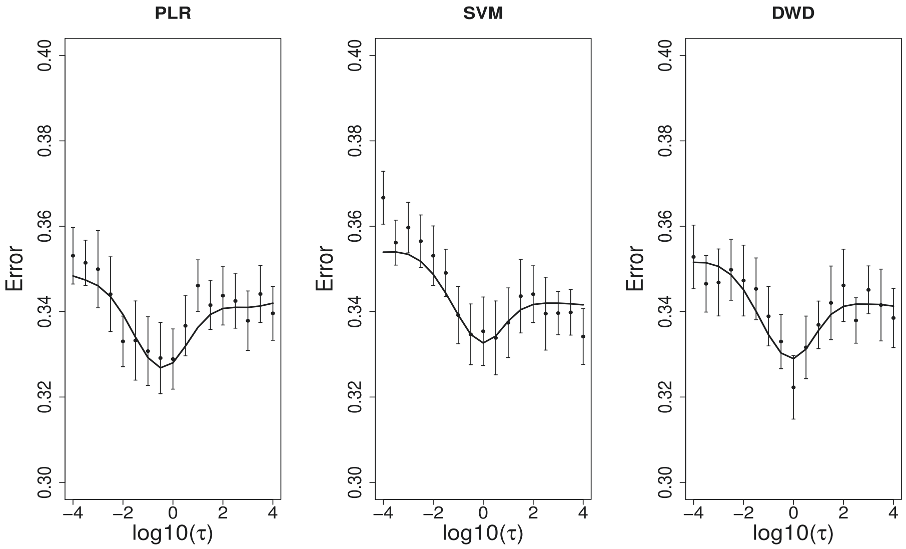

Figure 5. Dependence of generalization error on the tuning parameter τ for different methods under the two layer neural network model. Here $\psi_1 = p/d = 1$, $\psi_2 = n/d = 3$. The simulations are based on 20 samples with $d = 200$. Sign activation function is used thus $\gamma_1 = \sqrt{2/\pi}$ and $\gamma_\star = \sqrt{1 - 2/\pi}$.

The settings of figures 3 and 4 are quite general in such that the spike vectors $\mathbf{v}_k (k = 1, \ldots, K)$ are neither aligned with nor orthogonal to the signal vector $\boldsymbol{\theta}_\star$.

3.3. Two layer neural network model

Figure 5 shows the dependence of generalization error on the tuning parameter τ for the two layer neural network model. The comparisons with numerical simulations are also

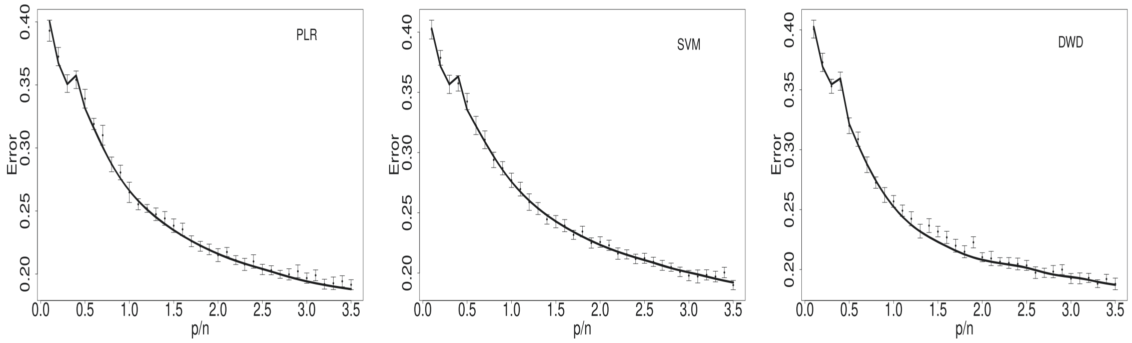

Figure 6. Generalization error plotted against the number of features per sample at small tuning parameter $\tau = 10^{-4}$. Here $\psi_2 = n/d = 3$. The simulations are based on 20 samples with $d = 200$. Sign activation function is used thus $\gamma_1 = \sqrt{2/\pi}$ and $\gamma_\star = \sqrt{1 - 2/\pi}$.

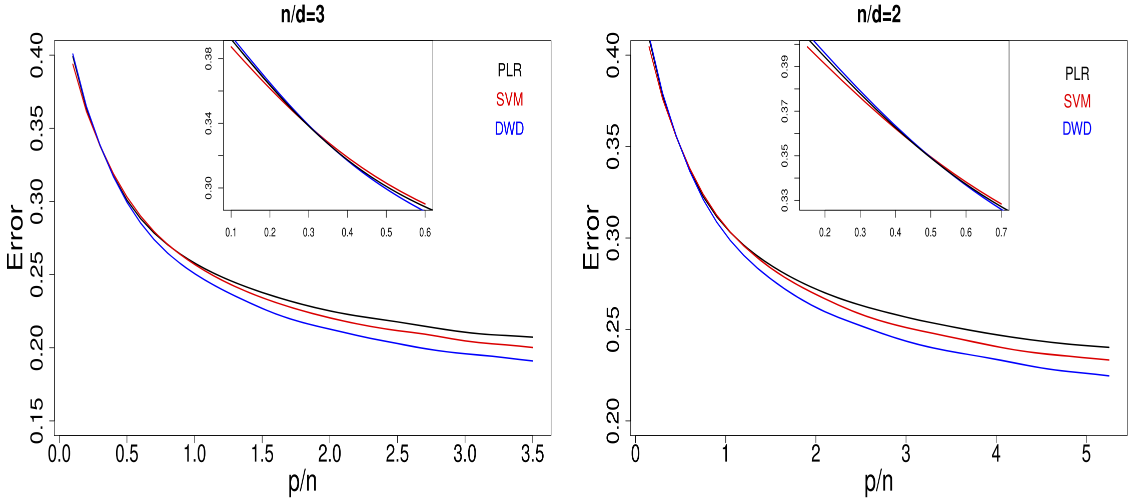

Figure 7. Performance comparison of three classifiers at optimal tuning τ under the two layer neural network model. Sign activation function is used thus $\gamma_1 = \sqrt{2/\pi}$, $\gamma_\star = \sqrt{1 - 2/\pi}$. The insets in the plots show the comparison of the three methods at low p/n.

included. The results show a fairly good agreement between theoretical prediction and Monte Carlo simulations which indicates the correctness of our analytical derivation.

In figure 6, we plot the value of the generalization error as a function of p/n with fixed $\psi_2 = n/d$ at small values of the regularization parameter $\tau = 10^{-4}$. We show the so-called double descent behavior for all three classification methods with a peak at the threshold value where the data become linearly separable. This finding agrees with the recently observed ‘double descent’ phenomenology for hard margin SVM in Montanari *et al* (2019) and logistic regression in Goldt *et al* (2019).

Figure 7 compares the performance of three classification methods after optimally tuning the parameter τ for two layer neural network model. For two fixed ratios between the number of samples and dimension d, the generalization errors of the three methods are very close at small value of overparametrization ratio p/n as shown by the insets in the plots. For large p/n, DWD performs the best and PLR performs the worst while SVM is in between. This is different from the performance under the spiked population model as shown in figure 4. We have tried other values for the ratio of n/d and the conclusions are very similar.

4. Conclusion

Large margin classifiers are commonly used in practice. In this paper, we examine the limiting behavior of a general family of large-margin classifiers as $p, n \to \infty$ with fixed $\alpha = n/p$. This family is very general and it includes many popular classification methods as special cases. We illustrate our main results by considering two special covariance structures: spiked population model and two layer neural network model with random first layer weights. We explore the phase transition behavior for the separability of the two classes and our general conclusion covers several existing results as special cases. Although our theoretical results are asymptotic in the problem dimensions, numerical simulations have shown that they are accurate already on problems with a few hundreds of variables. Our main observations from the derived analytic formulas are

- Under the same condition, data generated from Gaussian mixture distribution are easier to be separated than from single Gaussian distribution.
- Except in the situation where the signal is very small, e.g. for small μ under spiked structure or small p/n under random feature structure, DWD usually yields a better performance than SVM. The performance of PLR depends on the choice of activation. For probit probability distribution and sign activation function, PLR performs the best under spiked population covariance structure and the worst under the two layer neural network covariance structure for large value of p/n.
- For two layer neural network covariance structure, we reproduce the double descent phenomenon for all three methods. We show that the test error peaks at a critical value of ψ_1 when the two classes become separable.

It is interesting to note that our findings provide theoretical confirmations to the empirical results observed in Marron *et al* (2007) that DWD yields superior performance to SVM in HDLSS situations. This statement has been confirmed in Huang and Yang (2019) for the Gaussian mixture model. Here it is also confirmed to be true for the single Gaussian model. Although our observations may not hold for all covariance structure, it can help us to understand the classification behaviors of different methods better.

We have tried other settings and found that our results are not sensitive to the choice of K, λ_k, and R_k for spiked population model. One of our future research topics is to study the dependence of the generalization errors on the choice of the probability distribution $g(\cdot)$, the activation function $\sigma(\cdot)$, and other types of covariance structure in order to provide some practical guidelines in real application. So far we assumed that the covariance structures and their associated parameters are all known, but in practice, we need to estimate them from the data.

Acknowledgments

The authors thank the editor and three referees for many helpful comments and suggestions which led to a much improved presentation. This research is supported in part by Division of Mathematical Sciences (National Science Foundation) Grant DMS-1916411 (Huang), the National Science Foundation of China under Grant 11671311 and 11971324 (Yang), the National Social Science Foundation of China under Grant 20BTJ039 (Yang), and the Chinese Ministry of education of Humanities and Social Science Project under Grant 19YJC910008 (Yang).

Appendix A

A.1. Derivation of proposition 1

This appendix outlines the replica calculation leading to propositions 1. We limit ourselves to the main steps. For a general introduction to the method and its motivation, we refer to Mezard *et al* (1987); Mézard and Montanari (2009); Krzakala *et al* (2012).

Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. We consider regularized classification of the form

$$\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}{\sqrt{p}} \right) + \sum_{j=1}^{p} J_\tau(\theta_j) \right\}. \tag{A1}$$

After suitable scaling, the terms inside the bracket $\{\cdot\}$ are exactly equal to the objective function of model (1) in the main text.

The replica calculation aims at estimating the following moment generating function (partition function)

$$Z_\beta(\mathbf{X}, \mathbf{y}) = \int \exp\left\{ -\beta \left[\sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}{\sqrt{p}} \right) + \sum_{j=1}^{p} J_\tau(\theta_j) \right] \right\} \mathrm{d}\boldsymbol{\theta} \tag{A2}$$

where $\beta > 0$ is a 'temperature' parameter. In the zero temperature limit, i.e. $\beta \to \infty$, $Z_\beta(\mathbf{X}, \mathbf{y})$ is dominated by the values of $\boldsymbol{\theta}$ which are the solution of (A1).

Within the replica method, it is assumed that the limits $p \to \infty$, $\beta \to \infty$ exist almost surely for the quantity $(p\beta)^{-1} \log Z_\beta(\mathbf{X}, \mathbf{y})$, and that the order of the limits can be exchanged. We therefore define the free energy

$$\mathcal{F} = -\lim_{\beta\to\infty} \lim_{p\to\infty} \frac{1}{p\beta} \log\ Z_\beta(\mathbf{X}, \mathbf{y}) = -\lim_{p\to\infty} \lim_{\beta\to\infty} \frac{1}{p\beta} \log\ Z_\beta(\mathbf{X}, \mathbf{y}).$$

It is also assumed that $p^{-1} \log Z_\beta(\mathbf{X}, \mathbf{y})$ concentrates tightly around its expectation so that the free energy can in fact be evaluated by computing

$$\mathcal{F} = -\lim_{\beta\to\infty} \lim_{p\to\infty} \frac{1}{p\beta} \langle \log\ Z_\beta(\mathbf{X}, \mathbf{y}) \rangle_{\mathbf{X},\mathbf{y}}, \tag{A3}$$

where the angle bracket stands for the expectation with respect to the distribution of training data $\mathbf{X}$ and $\mathbf{y}$. Notice that, by (A3) and using Laplace method in the integral

(A2), we have

$$\mathcal{F} = \lim_{p\to\infty} \frac{1}{p} \min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}{\sqrt{p}}\right) + \sum_{j=1}^{p} J_\tau(\theta_j) \right\}.$$

In order to evaluate the integration of a log function, we make use of the replica method based on the identity

$$\log\ Z = \lim_{k\to 0} \frac{\partial Z^k}{\partial k} = \lim_{k\to 0} \frac{\partial}{\partial k}\ \log\ Z^k, \tag{A4}$$

and rewrite (A3) as

$$\mathcal{F} = -\lim_{\beta\to\infty} \lim_{p\to\infty} \frac{1}{p\beta} \lim_{k\to 0} \frac{\partial}{\partial k}\ \log \Xi_k(\beta), \tag{A5}$$

where

$$\Xi_k(\beta) = \langle \{Z_\beta(\mathbf{X},\mathbf{y})\}^k \rangle_{\mathbf{X},\mathbf{y}} = \int \{Z_\beta(\mathbf{X},\mathbf{y})\}^k \prod_{i=1}^{n} P(\mathbf{x}_i, y_i) \mathrm{d}\mathbf{x}_i\ \mathrm{d}y_i. \tag{A6}$$

Equation (A5) can be derived by using the fact that $\lim_{k\to 0} \Xi_k(\beta) = 1$ and exchanging the order of the averaging and the differentiation with respect to k. In the replica method, we will first evaluate $\Xi_k(\beta)$ for integer k and then apply to real k and take the limit of $k \to 0$.

For integer k, in order to represent $\{Z_\beta(\mathbf{X},\mathbf{y})\}^k$ in the integrand of (A6), we use the identity

$$\left(\int f(x)\mu(\mathrm{d}x)\right)^k = \int f(x_1)\ldots f(x_k)\mu(\mathrm{d}x_1)\ldots\mu(\mathrm{d}x_k),$$

and obtain

$$\{Z_\beta(\mathbf{X},\mathbf{y})\}^k = \prod_{a=1}^{k} \left[\int \exp\left\{-\beta\left[\sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right) + \sum_{j=1}^{p} J_\tau(\theta_j^a)\right]\right\} \mathrm{d}\boldsymbol{\theta}^a\right] \tag{A7}$$

where we have introduced replicated parameters

$$\boldsymbol{\theta}^a \equiv [\theta_1^a, \ldots, \theta_p^a]^{\mathrm{T}}, \quad \text{for } a = 1, \ldots, k.$$

Exchanging the order of the two limits $p \to \infty$ and $k \to 0$ in (A5), we have

$$\mathcal{F} = -\lim_{\beta\to\infty} \frac{1}{\beta} \lim_{k\to 0} \frac{\partial}{\partial k} \left(\lim_{p\to\infty} \frac{1}{p}\ \log\ \Xi_k(\beta)\right). \tag{A8}$$

Define the measure $\nu(\mathrm{d}\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \mathbb{R}^p$ as follows

$$\nu(\mathrm{d}\boldsymbol{\theta}) = \exp\left\{-\beta \sum_{j=1}^{p} J_\tau(\theta_j)\right\} \mathrm{d}\boldsymbol{\theta}.$$

Similarly, define the measure $\nu(\mathrm{d}\mathbf{x})$ as $\nu(\mathrm{d}\mathbf{x}) = P(\mathbf{x})\mathrm{d}\mathbf{x}$. In order to carry out the calculation of $\Xi_k(\beta)$, we let $\nu^k(\mathrm{d}\boldsymbol{\theta}) \equiv \nu(\mathrm{d}\boldsymbol{\theta}^1) \times \cdots \times \nu(\mathrm{d}\boldsymbol{\theta}^k)$ be a measure over $(\mathbb{R}^p)^k$, with $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k \in \mathbb{R}^p$. Analogously $\nu^n(\mathrm{d}\mathbf{x}) \equiv \nu(\mathrm{d}\mathbf{x}_1) \times \cdots \times \nu(\mathrm{d}\mathbf{x}_n)$ with $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and $\nu^n(\mathrm{d}y) = \nu(\mathrm{d}y_1) \ldots \nu(\mathrm{d}y_n)$. With these notations, we have

$$\Xi_k(\beta) = \int \exp\left\{-\beta \sum_{i=1}^{n} \sum_{a=1}^{k} V\left(\frac{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right)\right\} \nu^k(\mathrm{d}\boldsymbol{\theta}) \nu^n(\mathrm{d}\mathbf{x}) \nu^n(\mathrm{d}y)$$
$$= \int \{I(\boldsymbol{\theta})\}^n \nu^k(\mathrm{d}\boldsymbol{\theta}), \tag{A9}$$

where

$$I(\boldsymbol{\theta}) = \iint \exp\left\{-\beta \sum_{a=1}^{k} V\left(\frac{y \mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right)\right\} \nu(\mathrm{d}\mathbf{x}) \nu(\mathrm{d}y)$$
$$= \int \left[\exp\left\{-\beta \sum_{a=1}^{k} V\left(\frac{\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right)\right\} f_+\left(\frac{\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}_\star}{\sqrt{p}}\right)\right.$$
$$\left. + \exp\left\{-\beta \sum_{a=1}^{k} V\left(\frac{-\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right)\right\} f_-\left(\frac{\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}_\star}{\sqrt{p}}\right)\right] \nu(\mathrm{d}\mathbf{x}), \tag{A10}$$

where $f_+(\frac{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_\star}{\sqrt{p}}) = g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_\star/\sqrt{p})$ and $f_-(\frac{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_\star}{\sqrt{p}}) = 1 - g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_\star/\sqrt{p})$ as shown in (2). Notice that above we used the fact that the integral over $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in (\mathbb{R}^p)^n$ factors into n integrals over $(\mathbb{R})^p$ with measure $\nu(\mathrm{d}\mathbf{x})$. We next use the identity

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(q) \mathrm{e}^{\mathrm{i}(q-x)\hat{q}} \, \mathrm{d}q \, \mathrm{d}\hat{q}. \tag{A11}$$

We apply this identity to (A10) and introduce integration variables $\mathrm{d}u^a, \mathrm{d}\hat{u}^a$ for $1 \leqslant a \leqslant k$. Letting $\nu^k(\mathrm{d}u) = \mathrm{d}u^1 \ldots \mathrm{d}u^k$ and $\nu^k(\mathrm{d}\hat{u}) = \mathrm{d}\hat{u}^1 \ldots \mathrm{d}\hat{u}^k$

$$I(\boldsymbol{\theta}) = \int \left[\exp\left\{-\beta \sum_{a=1}^{k} V(u^a)\right\} f_+(u^\star) + \exp\left\{-\beta \sum_{a=1}^{k} V(-u^a)\right\} f_-(u^\star)\right]$$
$$\times \exp\left\{\mathrm{i}\sqrt{p} \sum_{a=1}^{k} \left(u^a - \frac{\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}^a}{\sqrt{p}}\right) \hat{u}^a + \mathrm{i}\sqrt{p}\left(u^\star - \frac{\mathbf{x}^{\mathrm{T}} \boldsymbol{\theta}_\star}{\sqrt{p}}\right) \hat{u}^\star\right\}$$
$$\times \nu(\mathrm{d}\mathbf{x}) \nu^k(\mathrm{d}u) \nu^k(\mathrm{d}\hat{u}) \nu(\mathrm{d}u^\star) \nu(\mathrm{d}\hat{u}^\star)$$
$$= \int \left[\exp\left\{-\beta \sum_{a=1}^{k} V(u^a)\right\} f_+(u^\star) + \exp\left\{-\beta \sum_{a=1}^{k} V(-u^a)\right\} f_-(u^\star)\right]$$
$$\times \exp\left\{\mathrm{i}\sqrt{p} \sum_{a=1}^{k} u^a \hat{u}^a + \mathrm{i}\sqrt{p} u^\star \hat{u}^\star - \frac{1}{2} \sum_{ab} (\boldsymbol{\theta}^a)^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\theta}^b \hat{u}^a \hat{u}^b\right.$$

$$\left. -\frac{1}{2}(\boldsymbol{\theta}_\star)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\hat{u}^\star\hat{u}^\star - \sum_a(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\hat{u}^a\hat{u}^\star\right\}\nu^k(\mathrm{d}u)\nu^k(\mathrm{d}\hat{u})\mathrm{d}u^\star\ \mathrm{d}\hat{u}^\star. \tag{A12}$$

In deriving (A12), we have used the fact that the low-dimensional marginals of $\mathbf{x}$ can be approximated by Gaussian distribution based on multivariate central limit theorem.

Next we apply (A11) to (A9), and introduce integration variables $Q_{ab}, \hat{Q}_{ab}$ and $R^a, \hat{R}^a$ associated with $(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}^b/p$ and $(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star/p$ respectively for $1 \leqslant a, b \leqslant k$. Denote $\mathbf{Q} \equiv (Q_{ab})_{1\leqslant a,b\leqslant k}$, $\hat{\mathbf{Q}} \equiv (\hat{Q}_{ab})_{1\leqslant a,b\leqslant k}$, $\mathbf{R} \equiv (R^a)_{1\leqslant a\leqslant k}$, and $\hat{\mathbf{R}} \equiv (\hat{R}^a)_{1\leqslant a\leqslant k}$. Note that, constant factors can be applied to the integration variables, and we choose convenient factors for later calculations. Letting $\mathrm{d}\mathbf{Q} \equiv \prod_{a,b}\mathrm{d}Q_{ab}$, $\mathrm{d}\hat{\mathbf{Q}} \equiv \prod_{a,b}\mathrm{d}\hat{Q}_{ab}$, $\mathrm{d}\mathbf{R} \equiv \prod_a \mathrm{d}R^a$, and $\mathrm{d}\hat{\mathbf{R}} \equiv \prod_a \mathrm{d}\hat{R}^a$, we obtain

$$\begin{aligned}\Xi_k(\beta) = \int \{\hat{\xi}(\mathbf{Q},\mathbf{R})\}^n \exp\Bigg\{&\mathrm{i}\sum_{ab}pQ_{ab}\hat{Q}_{ab} + \mathrm{i}\sum_a pR_a\hat{R}_a - \mathrm{i}\sum_{ab}(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}^b\hat{Q}_{ab}\\ &- \mathrm{i}\sum_a(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\hat{R}_a\Bigg\}\mathrm{d}\mathbf{Q}\ \mathrm{d}\hat{\mathbf{Q}}\ \mathrm{d}\mathbf{R}\ \mathrm{d}\hat{\mathbf{R}}\nu^k(d\boldsymbol{\theta}),\end{aligned} \tag{A13}$$

where

$$\begin{aligned}\hat{\xi}(\mathbf{Q},\mathbf{R}) = \int &\left[\exp\left\{-\beta\sum_{a=1}^k V(u^a)\right\}f_+(u^\star) + \exp\left\{-\beta\sum_{a=1}^k V(-u^a)\right\}f_-(u^\star)\right]\\ &\exp\Bigg\{\mathrm{i}\sqrt{p}\sum_{a=1}^k u^a\hat{u}^a + \mathrm{i}\sqrt{p}u^\star\hat{u}^\star - \frac{1}{2}\sum_{ab}pQ_{ab}\hat{u}^a\hat{u}^b\\ &- \frac{1}{2}p\rho^2\hat{u}^\star\hat{u}^\star - \sum_a pR^a\hat{u}^a\hat{u}^\star\Bigg\}\nu^k(\mathrm{d}u)\nu^k(\mathrm{d}\hat{u})\mathrm{d}u^\star\ \mathrm{d}\hat{u}^\star.\end{aligned} \tag{A14}$$

Now we can rewrite (A13) as

$$\Xi_k(\beta) = \int \exp\left\{-p\mathcal{S}_k(\mathbf{Q},\hat{\mathbf{Q}},\mathbf{R},\hat{\mathbf{R}})\right\}\mathrm{d}\mathbf{Q}\ \mathrm{d}\hat{\mathbf{Q}}\ \mathrm{d}\mathbf{R}\ \mathrm{d}\hat{\mathbf{R}}, \tag{A15}$$

where

$$\begin{aligned}\mathcal{S}_k(\mathbf{Q},\hat{\mathbf{Q}},\mathbf{R},\hat{\mathbf{R}}) &= -\mathrm{i}\beta\left(\sum_{ab}Q_{ab}\hat{Q}_{ab} + \sum_a R^a\hat{R}^a\right) - \frac{1}{p}\ \log\ \xi(\hat{\mathbf{Q}},\hat{\mathbf{R}}) - \alpha\ \log\ \hat{\xi}(\mathbf{Q},\mathbf{R}),\\ \xi(\hat{\mathbf{Q}},\hat{\mathbf{R}}) &= \int \exp\left\{-\mathrm{i}\sum_{ab}\hat{Q}_{ab}(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}^b - \mathrm{i}\sum_a(\boldsymbol{\theta}^a)^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\hat{R}_a\right\}\nu^k(d\boldsymbol{\theta}).\end{aligned} \tag{A16}$$

Now we apply steepest descent method to the remaining integrations. According to Varadhan's proposition (Tanaka 2002), only the saddle points of the exponent of the integrand contribute to the integration in the limit of $p \to \infty$. We next use the saddle point method in (A15) to obtain

$$-\lim_{p\to\infty}\frac{1}{p}\Xi_k(\beta)=\mathcal{S}_k(\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star),$$

where $\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star$ are the saddle point location. Looking for saddle-points over all the entire space is in general difficult to perform. We assume replica symmetry for saddle-points such that they are invariant under exchange of any two replica indices a and b, where $a\neq b$. Under this symmetry assumption, the space is greatly reduced and the exponent of the integrand can be explicitly evaluated. The replica symmetry is also motivated by the fact that $\mathcal{S}_k(\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star)$ is indeed left unchanged by such change of variables. This is equivalent to postulating that $R^a=R$, $\hat{R}^a=\mathrm{i}\hat{R}$,

$$(Q_{ab})^\star=\begin{cases}q_1 & \text{if a = b}\\ q_0 & \text{otherwise}\end{cases},\quad\text{and}\quad(\hat{Q}_{ab})^\star=\begin{cases}\mathrm{i}\dfrac{\beta\xi_1}{2} & \text{if a = b}\\ \mathrm{i}\dfrac{\beta\xi_0}{2} & \text{otherwise}\end{cases},\tag{A17}$$

where the factor $\mathrm{i}\beta/2$ is for future convenience. The next step consists in substituting the above expressions for $\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star$ in $\mathcal{S}_k(\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star)$ and then taking the limit $k\to 0$. We will consider separately each term of $\mathcal{S}_k(\mathbf{Q}^\star,\hat{\mathbf{Q}}^\star,\mathbf{R}^\star,\hat{\mathbf{R}}^\star)$. Let us begin with the first term

$$-\mathrm{i}\beta\left(\sum_{ab}Q_{ab}\hat{Q}_{ab}+\sum_a R^a\hat{R}^a\right)=\frac{k\beta^2}{2}(\xi_1q_1-\xi_0q_0)+k\beta R\hat{R}.\tag{A18}$$

Let us consider $\log\,\xi(\hat{\mathbf{Q}},\hat{\mathbf{R}})$. For p-vectors $\mathbf{u},\mathbf{v}\in\mathbb{R}^p$ and $p\times p$ matrix $\mathbf{\Sigma}$, introducing the notation $\|\mathbf{v}\|^2_{\mathbf{\Sigma}}\equiv\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v}$ and $\langle\mathbf{u},\mathbf{v}\rangle\equiv\sum_{j=1}^p u_jv_j/p$, we have

$$\begin{aligned}\xi(\hat{\mathbf{Q}},\hat{\mathbf{R}})&=\int\exp\left\{\frac{\beta^2}{2}(\xi_1-\xi_0)\sum_{a=1}^k\|\boldsymbol{\theta}^a\|^2_{\mathbf{\Sigma}}+\frac{\beta^2\xi_0}{2}\sum_{a,b=1}^k(\boldsymbol{\theta}^a)^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\theta}^b\right.\\&\quad\left.+\beta\sum_{a=1}^k\hat{R}(\boldsymbol{\theta}^a)^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\theta}_\star\right\}\nu^k(\mathrm{d}\boldsymbol{\theta})\\&=E\int\exp\left\{\frac{\beta^2}{2}(\xi_1-\xi_0)\sum_{a=1}^k\|\boldsymbol{\theta}^a\|^2_{\mathbf{\Sigma}}+\beta\sqrt{\xi_0}\sum_{a=1}^k(\boldsymbol{\theta}^a)^{\mathrm{T}}\mathbf{\Sigma}^{1/2}\mathbf{z}\right.\\&\quad\left.+\beta\sum_{a=1}^k\hat{R}(\boldsymbol{\theta}^a)^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{\theta}_\star\right\}\nu^k(\mathrm{d}\boldsymbol{\theta}),\end{aligned}\tag{A19}$$

where expectation is with respect to $\mathbf{z}\sim N(0,\mathbf{I}_p)$. Notice that, given $\mathbf{z}\in\mathbb{R}^p$, the integrals over $\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^k$ factorize, whence

$$\xi(\hat{\mathbf{Q}},\hat{\mathbf{R}}) = E\left\{\left[\int \exp\left\{\frac{\beta^2}{2}(\xi_1-\xi_0)\|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}} + \beta\sqrt{\xi_0}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\mathbf{z} \right.\right.\right.$$
$$\left.\left.\left. + \beta\hat{R}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\right\}\nu(\mathrm{d}\boldsymbol{\theta})\right]^k\right\}.$$

Finally, after integration over $\nu^k(\mathrm{d}\hat{u})$, (A14) becomes

$$\hat{\xi}(\mathbf{Q},\mathbf{R}) = \int\left[\exp\left\{-\beta\sum_{a=1}^{k}V(u^a)\right\}f_+(u^\star) + \exp\left\{-\beta\sum_{a=1}^{k}V(-u^a)\right\}f_-(u^\star)\right]$$
$$\exp\left\{\mathrm{i}\sqrt{p}u^\star\hat{u}^\star - \frac{1}{2}p\rho^2\hat{u}^\star\hat{u}^\star - \frac{1}{2}\sum_{ab}(u^a+\mathrm{i}\sqrt{p}R^a\hat{u}^\star)(\mathbf{Q}^{-1})_{ab}\right.$$
$$\left.\left(u^b+\mathrm{i}\sqrt{p}R^b\hat{u}^\star\right) - \frac{1}{2}\ \log\ \det\ \mathbf{Q}\right\}\nu^k(\mathrm{d}u)\mathrm{d}u^\star\ \mathrm{d}\hat{u}^\star. \qquad \text{(A20)}$$

We can next take the limit $\beta\to\infty$. The analysis of the saddle point parameters q_0, q_1, ξ_0, ξ_1 shows that q_0, q_1 have the same limit with $q_1 - q_0 = (q/\beta) + o(\beta^{-1})$ and ξ_0, ξ_1 have the same limit with $\xi_1 - \xi_0 = (-\xi/\beta) + o(\beta^{-1})$. Substituting the above expression in (A18) and (A19), in the limit of $k\to 0$, we then obtain

$$-\mathrm{i}\beta\left(\sum_{ab}Q_{ab}\hat{Q}_{ab} + \sum_{a}R^a\hat{R}^a\right) = \frac{k\beta}{2}(\xi_0 q - \xi q_0) + k\beta R\hat{R}, \qquad \text{(A21)}$$

and

$$\xi(\hat{\mathbf{Q}},\hat{\mathbf{R}}) = E\left\{\left[\int \exp\left\{-\frac{\beta\xi}{2}\|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}} + \beta\sqrt{\xi_0}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}\right.\right.\right.$$
$$\left.\left.\left. + \beta\hat{R}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\right\}\nu(\mathrm{d}\boldsymbol{\theta})\right]^k\right\}. \qquad \text{(A22)}$$

Similarly, using (A17), we obtain

$$\sum_{ab}(u^a+\mathrm{i}\sqrt{p}R^a\hat{u}^\star)(\mathbf{Q}^{-1})_{ab}(u^b+\mathrm{i}\sqrt{p}R^b\hat{u}^\star)$$
$$= \frac{\beta\sum_a(u^a+\mathrm{i}\sqrt{p}R^a\hat{u}^\star)^2}{q} - \frac{\beta^2 q_0\{\sum_a(u^a+\mathrm{i}\sqrt{p}R^a\hat{u}^\star)\}^2}{(q)^2},$$
$$\log\det\ \mathbf{Q} = \log\left[(q_1-q_0)^k\left(1+\frac{kq_0}{q_1-q_0}\right)\right] = \frac{k\beta q_0}{q},$$

where we retain only the leading order terms. Therefore, (A14) becomes

$$\hat{\xi}(\mathbf{Q},\mathbf{R}) = \int\left[\exp\left\{-\beta\sum_{a=1}^{k}V(u^a)\right\}f_+(u^\star) + \exp\left\{-\beta\sum_{a=1}^{k}V(-u^a)\right\}f_-(u^\star)\right]$$

$$
\begin{aligned}
&\exp\left\{\mathrm{i}\sqrt{p}u^\star\hat{u}^\star-\frac{1}{2}p\rho^2\hat{u}^\star\hat{u}^\star-\frac{\beta\sum_a(u^a)^2}{2q}-\frac{\mathrm{i}\sqrt{p}\beta\hat{u}^\star\sum_a u^aR^a}{q}\right.\\
&\left.+\frac{\beta^2q_0(\sum_a u^a)^2}{2q^2}-\frac{k\beta q_0}{2q}\right\}\nu^k(\mathrm{d}u)\\
&=E_{u^\star}\int\left[\exp\left\{-\beta\sum_{a=1}^{k}V(u^a)\right\}f_+(u^\star)+\exp\left\{-\beta\sum_{a=1}^{k}V(-u^a)\right\}f_-(u^\star)\right]\\
&\exp\left\{-\frac{\beta\sum_a(u^a)^2}{2q}+\frac{\beta^2(q_0-R^2/\rho^2)(\sum_a u^a)^2}{2q^2}+\frac{\beta Ru^\star\sum_a u^a}{q\rho^2}-\frac{k\beta q_0}{2q}\right\}\nu^k(\mathrm{d}u)\\
&=\exp\left(-\frac{k\beta q_0}{2q}\right)E_zE_{u^\star}\\
&\left[\left\{\int\exp\left\{-\beta V(u)-\frac{\beta u^2}{2q}+\frac{\beta\sqrt{q_0-R^2/\rho^2}zu}{q}+\frac{\beta Ru^\star u}{q\rho}\right\}\mathrm{d}u\right\}^kf_+(\rho u^\star)\right.\\
&\left.+\left\{\int\exp\left\{-\beta V(-u)-\frac{\beta u^2}{2q}+\frac{\beta\sqrt{q_0-R^2/\rho^2}zu}{q}+\frac{\beta Ru^\star u}{q\rho}\right\}\mathrm{d}u\right\}^kf_-(\rho u^\star)\right]\\
&=\exp\left(-\frac{k\beta q_0}{2q}\right)E_zE_{u^\star}E_{y^\star}\left(\int\exp\left\{-\beta V(u)-\frac{\beta(u-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z)^2}{2q}\right.\right.\\
&\left.\left.+\frac{\beta(\sqrt{q_0-R^2/\rho^2}y^\star z+y^\star u^\star R/\rho)^2}{2q}\right\}\mathrm{d}u\right)^k,
\end{aligned}
$$

where the expectation $z\perp u, z\sim N(0,1), u^\star\sim N(0,1)$, and $P(y^\star=\pm|u^\star)=f_\pm(\rho u^\star)$. Substituting this expression in (A16), we obtain

$$
\log\ \hat{\xi}(\mathbf{Q},\mathbf{R})=-k\beta E\left\{\min_u\left[V(u)+\frac{(u-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z)^2}{2q}\right]\right\},\tag{A23}
$$

where the expectation is with respect to z, $u^\star$, and $y^\star$. Putting (A21), (A22), and (A23) together into (A15) and then into (A5), we obtain

$$
\begin{aligned}
\mathcal{F}=&\frac{1}{2}(\xi_0q-\xi q_0)+R\hat{R}\\
&+\alpha E\left\{\min_u\left[V(u)+\frac{\left(u-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z\right)^2}{2q}\right]\right\}\\
&+\frac{1}{p}\mathrm{E}\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\xi}{2}\|\boldsymbol{\theta}\|_{\boldsymbol{\Sigma}}^2-\left\langle\sqrt{\xi_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z}+\hat{R}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star,\mathbf{w}\right\rangle+\sum_{j=1}^{p}J_\tau(\theta_j)\right\},
\end{aligned}\tag{A24}
$$

where the expectations are with respect to z, $u^\star$, and $y^\star$. Here ξ, ξ_0, q, q_0, R, $\hat{R}$ are order parameters which can be determined from the saddle point equations of $\mathcal{F}$. Define functions ϕ_1, ϕ_2, and ϕ_3 as

$$\phi_1 = E\left\{\left(\hat{u} - y^\star u^\star R/\rho - \sqrt{q_0 - R^2/\rho^2} y^\star z\right) y^\star u^\star\right\},$$
$$\phi_2 = E\left\{\left(\hat{u} - y^\star u^\star R/\rho - \sqrt{q_0 - R^2/\rho^2} y^\star z\right) y^\star z\right\},$$
$$\phi_3 = E\left\{\left(\hat{u} - y^\star u^\star R/\rho - \sqrt{q_0 - R^2/\rho^2} y^\star z\right)^2\right\},$$

where

$$\hat{u} = \mathrm{argmin}_{u\in\mathbb{R}}\left\{V(u) + \frac{\left(u - y^\star u^\star R/\rho - \sqrt{q_0 - R^2/\rho^2} y^\star z\right)^2}{2q}\right\}.$$

The result in (A24) is for general penalty function $J_\tau(w)$. For quadratic penalty $J_\tau(w) = \tau w^2$, we get the closed form limiting distribution of $\mathbf{w}$ as

$$\hat{\boldsymbol{\theta}} = (\xi\boldsymbol{\Sigma} + \tau\mathbf{I}_p)^{-1}\left(\sqrt{\xi_0}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \hat{R}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star\right). \tag{A25}$$

All the order parameters can be determined by the following saddle-point equations:

$$\xi_0 = \frac{\alpha}{q^2}\phi_3, \tag{A26}$$

$$\xi = -\frac{\alpha\phi_2}{q\sqrt{q_0 - R^2/\rho^2}}, \tag{A27}$$

$$\hat{R} = \frac{\alpha}{q}\left(\frac{\phi_1}{\rho} - \frac{R\phi_2}{\rho^2\sqrt{q_0 - R^2/\rho^2}}\right), \tag{A28}$$

$$q_0 = \frac{1}{p}E\|\hat{\boldsymbol{\theta}}\|^2_{\boldsymbol{\Sigma}}, \tag{A29}$$

$$q = \frac{1}{p\sqrt{\xi_0}}E\left\langle\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \hat{\boldsymbol{\theta}}\right\rangle \tag{A30}$$

$$R = \frac{1}{p}E\langle\Sigma\boldsymbol{\theta}_\star, \hat{\boldsymbol{\theta}}\rangle. \tag{A31}$$

Note that two types of Gaussian random variables are introduced, one is in primary $\hat{\boldsymbol{\theta}}$ and another one is in conjugate $\hat{u}$. The variances of these two random variables are controlled by ξ_0 and q_0 respectively. It is interesting to see that ξ_0 is determined by the expectation over a quadratic form of $\hat{u}$ while ξ_0 is determined by the expectation over a quadratic form of $\hat{\boldsymbol{\theta}}$.

The above formulas are for general positive definite covariance matrix $\boldsymbol{\Sigma}$. Then after applying the random features model and integrating over $\mathbf{z}$, we obtain the explicit nonlinear equations (A29)–(A31) for determining six parameters q_0, q, and R as

$$q_0 = \frac{1}{p}\xi_0\ \mathrm{Tr}\left(\boldsymbol{\Sigma}^{1/2}(\xi\boldsymbol{\Sigma} + \tau\mathbf{I}_p)^{-1}\boldsymbol{\Sigma}(\xi\boldsymbol{\Sigma} + \tau\mathbf{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\right) \tag{A32}$$

$$+\frac{1}{p}\hat{R}^2(\boldsymbol{\theta}_\star)^{\mathrm{T}}\boldsymbol{\Sigma}(\xi\boldsymbol{\Sigma}+\tau\mathbf{I}_p)^{-1}\boldsymbol{\Sigma}(\xi\boldsymbol{\Sigma}+\tau\mathbf{I}_p)^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}_\star \tag{A33}$$

$$= \xi_0 f_2(\xi,\tau)+\hat{R}^2\rho^2 f_3(\xi,\tau),$$

$$R=\hat{R}\rho^2 f_1(\xi,\tau),$$

$$q=f_0(\xi,\tau), \tag{A34}$$

where

$$f_0(\xi,\tau)=\int\frac{X}{\xi X+\tau}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W),\quad f_1(\xi,\tau)=\int\frac{W^2X}{\xi X+\tau}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W),$$

$$f_2(\xi,\tau)=\int\frac{X^2}{(\xi X+\tau)^2}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W),\quad f_1(\xi,\tau)=\int\frac{W^2X^2}{(\xi X+\tau)^2}\mu_{\mathrm{p}}(\mathrm{d}X,\mathrm{d}W).$$

After variable substitution $R/\rho\to R$ and $\rho\hat{R}\to\hat{R}$, we derive the equations (6)–(11) in the main text.

A.2. Derivation of corollary 1

Under $\tau=0$, from (A32)–(A34), we have

$$q_0=\frac{\xi_0+\hat{R}^2\rho^2}{\xi^2},\quad q=\frac{1}{\xi},\quad R=\frac{\hat{R}\rho^2}{\xi^2}.$$

Substitute into (A26)–(A28), we have

$$q_0-\frac{R^2}{\rho^2}=\alpha\phi_3, \tag{A35}$$

$$1=-\frac{\alpha\phi_2}{\sqrt{q_0-R^2/\rho^2}}, \tag{A36}$$

$$\frac{R}{\rho}=\alpha\left(\phi_1-\frac{R\phi_2}{\rho\sqrt{q_0-R^2/\rho^2}}\right). \tag{A37}$$

Substituting (A36) into (A37), we have $\phi_1=0$. From (A35), we have

$$q_0-\frac{R^2}{\rho^2}=\alpha E\left\{\left(\hat{u}-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z\right)\left(\hat{u}-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z\right)\right\},$$

where $u^\star\perp z, u^\star\sim N(0,1), z\sim N(0,1)$, and $P(y=+1|u^\star)=f_+(\rho u^\star)$. Substituting (A36) and (A37), we obtain

$$E\left\{\left(\hat{u}-y^\star u^\star R/\rho-\sqrt{q_0-R^2/\rho^2}y^\star z\right)\hat{u}\right\}=0.$$

Denote $r=R/\rho/\sqrt{q_0}$. For SVM, we get

$$
\begin{aligned}
0 = E\Big\{\Big(1-\sqrt{q_0}(ry^\star u^\star+\sqrt{1-r^2}y^\star z)\Big)\, I(1-q \leqslant \sqrt{q_0}(ry^\star u^\star+\sqrt{1-r^2}y^\star z) \leqslant 1)\Big\}\\
+ E\Big\{q\left(q+\sqrt{q_0}(ry^\star u^\star+\sqrt{1-r^2}y^\star z)\right) I(\sqrt{q_0}(ry^\star u^\star+\sqrt{1-r^2}y^\star z) \leqslant 1-q)\Big\}.
\end{aligned}
$$

We are interested in the separability, i.e. the behaviour of $q_0 \to \infty$. The above equation implies that $q/\sqrt{q_0} \to \infty$. Therefore from (A35) and (A37), we obtain

$$
1/\alpha = E\left\{\left(\frac{r}{\sqrt{1-r^2}}y^\star u^\star + y^\star z\right)_+^2\right\} \tag{A38}
$$

$$
0 = E\left\{\left(\frac{r}{\sqrt{1-r^2}}y^\star u^\star + y^\star z\right)_+ y^\star u^\star\right\}, \tag{A39}
$$

which is equivalent to find

$$
1/\alpha = \min_{c\in\mathbb{R}} E\left\{(cy^\star u^\star + z)_+^2\right\}.
$$

A.3. Derivation of proposition 2

From equations (14)–(16) in proposition 3 of Huang and Yang (2019), we obtain

$$
\begin{aligned}
q_0 - \frac{R^2}{\gamma^2} &= \alpha E\{(\hat{u}-a)^2\},\\
\frac{R}{\gamma^2} &= \alpha\mu E(\hat{u}-a),\\
1 &= -\frac{\alpha}{\sqrt{q_0}}E\{(\hat{u}-a)z\},
\end{aligned}
$$

where $a = R\mu + \sqrt{q_0}z$. For SVM, define $\gamma^2 = \hat{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$, $z_c = (1-R\mu)/\sqrt{q_0}$, $x = q/\sqrt{q_0}$, and $r = R/\sqrt{q_0}$, we have

$$
1 - \frac{r^2}{\gamma^2} = \alpha\left\{\int_{z_c-x}^{z_c}(z_c-z)^2 Dz + x^2\int_{-\infty}^{z_c-x} Dz\right\} \tag{A40}
$$

$$
\frac{r}{\gamma^2} = \alpha\mu\left\{\int_{z_c-x}^{z_c}(z_c-z) Dz + x\int_{-\infty}^{z_c-x} Dz\right\} \tag{A41}
$$

$$
1 = \alpha\int_{z_c-x}^{z_c} Dz. \tag{A42}
$$

From (A40) and (A41), we have

$$
\begin{aligned}
1 = \alpha\left\{\int_{z_c-x}^{z_c}(z_c-z)^2 Dz + x^2\int_{-\infty}^{z_c-x} Dz\right\}\\
+\left\{\alpha\gamma\mu\left(\int_{z_c-x}^{z_c}(z_c-z) Dz + x\int_{-\infty}^{z_c-x} Dz\right)\right\}^2.
\end{aligned}
$$

For fixed α, μ has upper bound in order to have solution. Because of (A42), the biggest value for μ we can achieve is when $x \to \infty$. Therefore the phase transition for Gaussian mixture model is determined by

$$1 = \alpha \int_{-\infty}^{z_c} (z_c - x)^2 Dz + \left\{ \alpha\gamma\mu \int_{-\infty}^{z_c} (z_c - z) Dz \right\}^2,$$

where $\Phi(z_c) = 1/\alpha$.

References

Aubin B, Maillard A, Barbier J, Krzakala F, Macris N and Zdeborová L 2019 The committee machine: computational to statistical gaps in learning a two-layers neural network *J. Stat. Mech.* 124023

Balcan M-F, Blum A and Vempala S 2006 Kernels as features: on kernels, margins, and low-dimensional mappings *Mach. Learn.* **65** 79–94

Barbier J and Macris N 2017 The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference (arXiv:1705.02780)

Bayati M and Montanari A 2012 The LASSO risk for Gaussian matrices *IEEE Trans. Inf. Theory* **58** 1997–2017

Belkin M, Hsu D and Mitra P P 2018 Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate *Proc. 32nd Int. Conf. on Neural Information Processing Systems, NIPS'18* (Red Hook, NY: Curran Associates Inc.) pp 2306–17

Belkin M, Hsu D, Ma S and Mandal S 2019a Reconciling modern machine-learning practice and the classical bias-variance trade-off *Proc. Natl Acad. Sci. USA* **116** 15849–54

Belkin M, Hsu D and Xu J 2019b Two models of double descent for weak features (arXiv:1903.07571)

Benito M, Parker J, Du Q, Wu J, Xiang D, Perou C M and Marron J S 2004 Adjustment of systematic microarray data biases *Bioinform.* **20** 105–14

Candès E J and Sur P 2020 The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression *Ann. Stat.* **48** 27–42

Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition *IEEE Trans. Electron. Comput.* **EC-14** 326–34

Freund Y and Schapire R E 1997 A decision-theoretic generalization of on-line learning and an application to boosting *J. Comput. Syst. Sci.* **55** 119–39

Friedman J, Hastie T and Tibshirani R 2000 Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors) *Ann. Stat.* **28** 337–407

Gerace F, Loureiro B, Krzakala F, Mézard M and Zdeborová L 2020 Generalisation error in learning with random features and the hidden manifold model (arXiv:2002.09339)

Gerbelot C, Abbara A and Krzakala F 2020 Asymptotic errors for convex penalized linear regression beyond Gaussian matrices (arXiv:2002.04372)

Goldt S, Mézard M, Krzakala F and Zdeborová L 2019 Modelling the influence of data structure on learning in neural networks: the hidden manifold model (arXiv:1909.11500)

Goldt S, Reeves G, Mézard M, Krzakala F and Zdeborová L 2020 The Gaussian equivalence of generative models for learning with two-layer neural networks (arXiv:2006.14709)

Hastie T, Montanari A, Rosset S and Tibshirani R J 2019 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)

Hastie T, Tibshirani R and Friedman J 2001 *The Elements of Statistical Learning* (*Springer Series in Statistics*) (Berlin: Springer)

Huang H 2017 Asymptotic behavior of support vector machine for spiked population model *J. Mach. Learn. Res.* **18** 1–21

Huang H and Yang Q 2019 Large dimensional analysis of general margin based classification methods (arXiv:1901.08057)

Krzakala F, Mézard M, Sausset F, Sun Y F and Zdeborová L 2012 Statistical-physics-based reconstruction in compressed sensing *Phys. Rev.* X **2** 021005

Lin X, Wahba G, Xiang D, Gao F, Klein R and Klein B 2000 Smoothing spline anova models for large data sets with Bernoulli observations and the randomized gacv *Ann. Stat.* **28** 1570–600

Liu Y, Hayes D N, Nobel A and Marron J S 2008 Statistical significance of clustering for high-dimension, low-sample size data *J. Am. Stat. Assoc.* **103** 1281–93

Liu Y, Zhang H H and Wu Y 2011 Hard or soft classification? Large-margin unified machines *J. Am. Stat. Assoc.* **106** 166–77

Ma Z 2013 Sparse principal component analysis and iterative thresholding *Ann. Stat.* **41** 772–801

Mai X and Couillet R 2018 Statistical analysis and improvement of large dimensional svm private communication

Mai X, Liao Z and Couillet R 2019 A large scale analysis of logistic regression: asymptotic performance and new insights *ICASSP 2019-2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* pp 3357–61

Marron J S, Todd M J and Ahn J 2007 Distance-weighted discrimination *J. Am. Stat. Assoc.* **102** 1267–71

Mei S and Montanari A 2019 The generalization error of random features regression: precise asymptotics and double descent curve (arXiv:1908.05355)

Mézard M and Montanari A 2009 *Information, Physics, and Computation* (*Oxford Graduate Texts*) (Oxford: Oxford University Press)

Mezard M, Parisi G and Virasoro M 1987 Spin glass theory and beyond: an introduction to the replica method and its applications *World Scientific Lecture Notes in Physics* (New York: World Scientific)

Mignacco F, Krzakala F, Lu Y M and Zdeborová L 2020 The role of regularization in classification of high-dimensional noisy Gaussian mixture (arXiv:2002.11544)

Montanari A, Ruan F, Sohn Y and Yan J 2019 The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime (arXiv:1911.01544)

Neal R M 1996 *Bayesian Learning for Neural Networks* (Berlin: Springer)

Qiao X, Zhang H H, Liu Y, Todd M J and Marron J S 2010 Weighted distance weighted discrimination and its asymptotic properties *J. Am. Stat. Assoc.* **105** 401–14

Qiao X and Zhang L 2015 Flexible high-dimensional classification machines and their asymptotic properties *J. Mach. Learn. Res.* **16** 1547–72

Rahimi A and Recht B 2008 Random features for large-scale kernel machines *Advances in Neural Information Processing Systems* ed J C Platt, D Koller, Y Singer and S T Roweis vol 20 (Red Hook, NY: Curran Associates, Inc) pp 1177–84

Shen X, Tseng G C, Zhang X and Wong W H 2003 On ψ-learning *J. Am. Stat. Assoc.* **98** 724–34

Sifaou H, Kammoun A and Alouini M 2019 Phase transition in the hard-margin support vector machines *2019 IEEE 8th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* pp 415–9

Tanaka T 2002 A statistical-mechanics approach to large-system analysis of cdma multiuser detectors *IEEE Trans. Inf. Theory* **48** 2888–910

Vapnik V N 1995 *The Nature of Statistical Learning Theory* (Berlin: Springer)

Wahba G 1999 *Support Vector Machines* (*Reproducing Kernel Hilbert Spaces, and Randomized GACV*) (Cambridge, MA: MIT Press) pp 69–88

Wang B and Zou H 2016 Sparse distance weighted discrimination *J. Comput. Graph. Stat.* **25** 826–38

Wang B and Zou H 2017 Another look at distance-weighted discrimination *J. Roy. Stat. Soc.* B **80** 177–98

Wu Y and Liu Y 2007 Robust truncated hinge loss support vector machines *J. Am. Stat. Assoc.* **102** 974–83

Zhu J and Hastie T 2005 Kernel logistic regression and the import vector machine *J. Comput. Graph. Stat.* **14** 185–205