Journal of Statistical Mechanics:
Theory and Experiment

PAPER

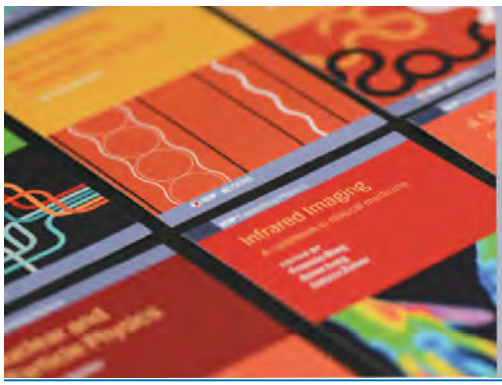
Large dimensional analysis of general margin based classification methods

To cite this article: Hanwen Huang and Qinglong Yang *J. Stat. Mech.* (2021) 113401

View the article online for updates and enhancements.

You may also like

- Data augmentation for self-paced motor imagery classification with C-LSTM
Daniel Freer and Guang-Zhong Yang
- Learning image context for segmentation of the prostate in CT-guided radiotherapy
Wei Li, Shu Liao, Qianjin Feng et al.
- ON MACHINE-LEARNED CLASSIFICATION OF VARIABLE STARS WITH SPARSE AND NOISY TIME-SERIES DATA
Joseph W. Richards, Dan L. Starr, Nathaniel R. Butler et al.



IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 198.137.20.27 on 11/10/2022 at 17:27

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

PAPER: Interdisciplinary statistical mechanics

Large dimensional analysis of general 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 8 June 2021
Accepted for publication 12 October 2021
Published 3 November 2021

Online at stacks.iop.org/JSTAT/2021/113401
https://doi.org/10.1088/1742-5468/ac2edd

Abstract. Margin-based classifiers have been popular in both machine learning and statistics for classification problems. Since a large number of classifiers are available, one natural question is which type of classifiers should be used given a particular classification task. We answer this question by investigating the asymptotic performance of a family of large-margin classifiers under the two component mixture models in situations where the data dimension p and the sample n are both large. This family covers a broad range of classifiers including support vector machine, distance weighted discrimination, penalized logistic regression, and large-margin unified machine as special cases. The asymptotic results are described by a set of nonlinear equations and we observe a close match of them with Monte Carlo simulation on finite data samples. Our analytical studies shed new light on how to select the best classifier among various classification methods as well as on how to choose the optimal tuning parameters for a given method.

Keywords: learning theory, machine learning, statistical inference

*Author to whom any correspondence should be addressed.

© 2021 IOP Publishing Ltd and SISSA Medialab srl
1742-5468/21/113401+30$33.00

J. Stat. Mech. (2021) 113401

Contents

1. Introduction 2
2. The margin-based classification method 4
3. Asymptotic performance 6
4. Phase transition 11
5. Estimation of data parameters 12
6. Numerical analysis 13
7. Real data 17
8. Conclusion 18
Acknowledgments 19
Appendix A 19
A1. Derivation of claim 4 27
References 28

1. Introduction

Classification is a very useful statistical tool that has been widely used in many disciplines and has achieved a lot of success. Its goal is to build a classification rule based on a training set that includes both covariates and class labels. Then for new objects whose covariates are available the classification rule can be used for class label prediction.

Since a large number of classifiers are available on the shelf, one natural question to ask is which type of classifiers should be used given a particular classification task. It is commonly agreed upon that there is no single method working best for all problems. The choice of classifiers really depends on the nature of the data set and the primary learning goal. Cross validation (CV) is a practically useful strategy for handling this task; its basic concept is to evaluate the prediction error by examining the data under control. Smaller values of the CV error are expected to be better in expressing the generative model of the data. However, the implementation of many classification methods involves tuning open parameters for achieving optimal performances, e.g. for regularized classification methods, one needs to deal with tuning parameters that control the trade-off between data fitting and principle of parsimony. Therefore, conducting CV incurs high computational costs, which makes it difficult in practice.

The purpose of this paper is to answer the above question by investigating the asymptotic performance of a family of large-margin classifiers in the limit of both sample

J. Stat. Mech. (2021) 113401

size n and dimension p going to infinity with fixed rate $\alpha = n/p$. We are motivated by the comparison between two commonly used classification methods: support vector machine (SVM) and distance weighted discrimination (DWD).

SVM is a state-of-the-art powerful classification method proposed by Vapnik (1995). Its has been demonstrated in Fernández-Delgado *et al* (2014) as one of the best performers in the pool of 179 commonly used classifiers. 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 (where n is much less than p) due to data-piling problem. They proposed DWD as a superior alternative to SVM. Both SVM and DWD are margin-based classification methods in the sense that they build the classifier through finding a decision boundary to separate the classes. DWD is different from SVM in that it seeks to maximize a notion of average distance instead of minimum distance between the classes. Thus, DWD allows all data points, rather than only the support vectors, to have a direct impact on the separating hyperplane. It gives high significance to those points that are close to the hyperplane, with little impact from points that are farther away. DWD is specifically designed for HDLSS situations. Many previous simulations and real data studies have shown 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, 2018). However, all previous studies are empirical and there is no theoretical justification about this phenomenon yet.

Recent rapid advances in statistical theory about the asymptotic performance of many classic machine learning algorithms in the limit of both large n and p have shed some light on this issue. There has been considerable effort to establish asymptotic results for different classification methods under the assumption that p and n grow at the same rate, that is, $n/p \to \alpha > 0$. The asymptotic results for SVM have been studied in Huang (2017) and Mai and Couillet (2018) under mixture models in which the data are assumed to be generated from a mixture distribution with two components, one for each class. The covariance matrix is assumed to follow a structure consisting of a pure background noise spiked with a few significant eigenvalues. The asymptotic results for DWD and logistic regression have been studied in Huang (2018) and Mai *et al* (2019) respectively. In Huang (2017, 2018), the spike eigenvector is assumed to be aligned with the signal direction while in Mai and Couillet (2018); Mai *et al* (2019) this assumption is relaxed. But all papers assume that the two classes have the same background noise.

In the present work, we derive the asymptotic results for a general family of large-margin classifiers in the limit of $p, n \to \infty$ at fixed $\alpha = n/p$ under the two component mixture models. The family covers a broad range of margin-based classifiers including SVM, DWD, penalized logistic regression (PLR), and large-margin unified machine (LUM). The results in Huang (2017, 2018), Mai and Couillet (2018), Mai *et al* (2019); Mai and Liao (2019), and Mai *et al* (2019) are all special cases of this general result. We also consider more general settings in the sense that the signals are not necessarily aligned with the spiked eigenvectors and the background noises of two classes are not necessarily the same. We derive the analytical results using the replica method developed

J. Stat. Mech. (2021) 113401

in statistical mechanics. All analytical results are confirmed by numerical experiments on finite-size systems and thus our formulas are verified to be correct.

Related work. The sharp asymptotics for hard margin SVM and unregularized logistic regression have been studied in Montanari *et al* (2019) and Candès and Sur (2020) respectively under the single Gaussian models in which the data are assumed to be generated from a single Gaussian distribution. Dobriban and Wager (2018) provide asymptotic analysis of the predictive risk of regularized discriminant analysis. Deng *et al* (2019) studied hard margin SVM and unregularized logistic regression under Gaussian mixture models in which the data are assumed to be generated from Gaussian mixture distribution with two components, one for each class. Gerace *et al* (2020) studied the classification error for PLR and SVM for single Gaussian model with two layer neural network covariance structure. The analogous results for Gaussian mixture models with standard Gaussian components was provided in Mignacco *et al* (2020). Wang and Thrampoulidis (2021) studied both max-margin SVM classifiers and min-norm interpolating classifiers under the popular generative Gaussian mixture model. The sharp asymptotics of generic convex generalized linear models was studied in Gerbelot *et al* (2020) for rotationally invariant Gaussian data and in Loureiro *et al* (2021a) for block-correlated Gaussian data. The multi-class classification for mixture of Gaussians was also provided in Loureiro *et al* (2021b) recently. Paralleling to classification, there has been considerable effort to establish the sharp asymptotics for regression. Examples include LASSO with i.i.d. setting (Bayati and Montanari 2011), LASSO with correlated data (Berthier *et al* 2020; Celentano *et al* 2020), ridgeless least squares (Hastie *et al* 2019), generalized linear model (Barbier *et al* 2019), and many others.

Note that most of the results in aforementioned literature are rigorous under Gaussian assumption. The rigorous analysis methods include convex random geometry (Candès and Sur 2020), random matrix theorem (Dobriban and Wager 2018), message-passing algorithms (Bayati and Montanari 2011; Berthier *et al* 2020; Loureiro *et al* 2021a, 2021b), convex Gaussian min-max theorem (Montanari *et al* 2019; Mignacco *et al* 2020; Deng *et al* 2019), and interpolation techniques (Barbier *et al* 2019). The present work focuses on mixture of two component under spiked covariance setting without Gaussian assumption. While it remains an open problem to derive a rigorous proof for our results, we shall use simulation on moderate system sizes to provide numerical support that the theoretical formula is indeed exact in the high-dimensional limit.

The rest of this paper is organized as follows: in section 2, we state the general framework for formulating the margin based classification methods. In section 3, the asymptotic results of the margin-based classifiers in the joint limit of large p and n for spiked population model are presented. Based on these asymptotic results, we study the separability phase transition in section 4. A method for estimating data parameters used in deriving the asymptotic results is provided in section 5. In section 6, we present numerical studies by comparing the theoretical results to Monte Carlo simulations on finite-size systems for several commonly used classification methods. An application of the proposed method to the breast cancer dataset is presented in section 7. The last section is devoted to the conclusion.

J. Stat. Mech. (2021) 113401

2. The margin-based classification method

In the binary classification problem, we are given a training dataset consisting of n observations $\{(\mathbf{x}_i, y_i)\,; i = 1, \ldots, n\}$ distributed according to some unknown joint probability distribution $P(\mathbf{x}, y)$. Here $\mathbf{x}_i \in \mathbb{R}^p$ 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. There are n_+ and n_- data in class $+$ and $-$ respectively.

The goal of linear classification is to calculate a function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + w_0$ such that $\text{sign}(f(\mathbf{x}))$ can be used as the classification rule. Here $\mathbf{w} \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ are parameters that need to be estimated. By definition of this classification rule, it is clear that correct classification occurs if and only if $yf(\mathbf{x}) > 0$. Therefore, the quantity $yf(\mathbf{x})$, commonly referred as the functional margin, plays a critical role in classification techniques. The focus of this paper is on large-margin classification methods, which can be fit in the regularization framework of loss + penalty. The loss function is used to keep the fidelity of the resulting model to the data while the penalty term in regularization helps to avoid overfitting of the resulting model. Using the functional margin, the regularization formulation of binary large-margin classifiers can be summarized as the following optimization problem

$$\min_{\mathbf{w}, w_0} \left\{ \sum_{i=1}^{n} V(y_i(\mathbf{x}_i^T\mathbf{w} + w_0)) + \sum_{j=1}^{p} J_\lambda(w_j) \right\}, \tag{1}$$

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

The general requirement for the loss function is convex decreasing with $V(u) \to 0$ as $u \to \infty$ and $V(u) \to \infty$ 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)), distance weighted discrimination (DWD; Marron *et al* (2007)), and large-margin unified machine (LUM; Liu *et al* (2011)). The loss functions of these classification methods are

$$\text{PLR}: V(u) = \log(1 + \exp(-u)),$$

$$\text{SVM}: V(u) = (1 - u)_+,$$

$$\text{DWD}: V(u) = \begin{cases} 1 - u & \text{if} \quad u \leqslant \dfrac{q}{q+1} \\ \dfrac{1}{u^q}\dfrac{q^q}{(q+1)^{q+1}} & \text{if} \quad u > \dfrac{q}{q+1} \end{cases},$$

$$\text{LUM}: V(u) = \begin{cases} 1 - u & \text{if} \quad u \leqslant \dfrac{c}{1+c} \\ \dfrac{1}{1+c}\left(\dfrac{a}{(1+c)u - c + a}\right)^a & \text{if} \quad u > \dfrac{c}{1+c} \end{cases},$$

where $q, a > 0$, and $c \geqslant 0$. It can be easily checked that SVM and DWD loss functions are special cases of the LUM loss function with appropriately chosen a and c

J. Stat. Mech. (2021) 113401

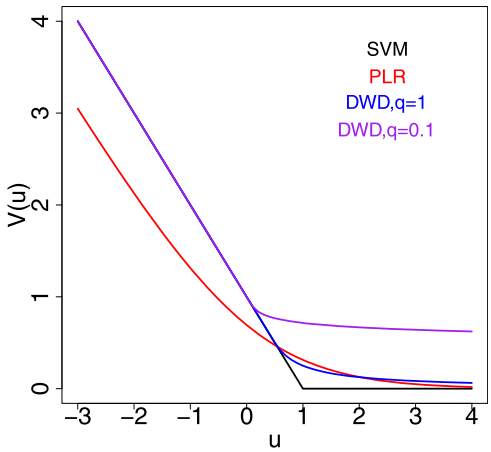


Figure 1. Plots of various loss functions.

(Liu *et al* 2011). For example, if we choose $a = c = q$, the LUM loss is the same as the DWD loss; if $a > 0$ and $c \to \infty$, the LUM loss is the same as the SVM loss. Besides the above methods, many other classification techniques can also be fit into the regularization framework, for example, 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 $J_\lambda(w) = \frac{\lambda}{2}w^2$ for L_2 regularization and $J_\lambda(w) = \lambda|w|$ for sparse L_1 regularization. In this paper, we focus on the standard L_2 regularization.

Figure 1 displays four loss functions: PLR, SVM, DWD with $q = 1$, and DWD with $q = 0.1$. 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 four loss functions, PLR has all order derivatives while DWD only has first order derivative. As $u \to -\infty$, $V(u) \to -u$ for all methods. As $u \to \infty$, $V(u)$ decays to 0 but with different speeds. The fastest one is SVM, followed by PLR, DWD with $q = 1$, and DWD with $q = 0.1$. We will see in section 6 that the decay speed of the loss function has a big influence on the classification performance in situations where λ is small. Also all classification methods have the same performance when λ is large enough due to the fact that $V(u)$ can be approximated by a linear function as $u \to 0$ for all loss functions.

3. Asymptotic performance

Now let us specify the joint probability distribution $P(\mathbf{x}, y)$. Conditional on $y = \pm 1$, assume that $\mathbf{x}$ follows a multivariate distribution $P(\mathbf{x}|y = \pm 1)$ with mean $\boldsymbol{\mu}_\pm$ and covariance matrix $\boldsymbol{\Sigma}_\pm$. Here $\boldsymbol{\mu}_\pm \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_\pm$ denote a $p \times p$ positive definite matrix. Without loss of generality, we take $\boldsymbol{\mu}_+ = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_- = -\boldsymbol{\mu}$.

We investigate the statistical behavior of the class separating hyperplane obtained from the optimization problem (1) in the limit of $n, p \to \infty$ with $n/p \to \alpha$. Let us begin by introducing some notations. Denote $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}/\mu$, where $\mu = \|\boldsymbol{\mu}\|$. For a given loss function

J. Stat. Mech. (2021) 113401

$V(u)$, define the proximal operator

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

where $b > 0$. It 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

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

For DWD loss with $q = 1$, we have

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

where $\tilde{u}$ is the solution of 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 the Newton–Raphson algorithm because the loss function has closed form second order derivatives.

Our main results are based upon the following Claim for the distributional limit of estimators $\hat{\mathbf{w}}$ obtained from (1).

Claim 1 The limiting distribution of $\hat{\mathbf{w}}$ is the same as the limiting distribution of $\hat{\boldsymbol{\mu}}$, which is defined as

$$\hat{\boldsymbol{\mu}} = (\xi^+ \boldsymbol{\Sigma}_+ + \xi^- \boldsymbol{\Sigma}_- + \lambda \mathbf{I}_p)^{-1} \left(\sqrt{\xi_0^+} \boldsymbol{\Sigma}_+^{1/2} \mathbf{z}_+ + \sqrt{\xi_0^-} \boldsymbol{\Sigma}_-^{1/2} \mathbf{z}_- + \sqrt{p} \tilde{R} \bar{\boldsymbol{\mu}} \right), \tag{3}$$

where $\mathbf{I}_p$ is p-dimensional identity matrix and $\mathbf{z}_\pm$ denote the vectors of length p whose elements are i.i.d. standard Gaussian random variables independent of $\boldsymbol{\mu}$, and

$$\xi^\pm = \frac{\alpha_\pm}{\sqrt{q_0^\pm} q^\pm} G_\pm, \quad \xi_0^\pm = \frac{\alpha_\pm}{(q^\pm)^2} H_\pm, \quad \text{and} \quad \tilde{R} = \frac{\alpha_+ \mu}{q^+} F_+ + \frac{\alpha_- \mu}{q^-} F_-.$$

Here $G_\pm$, $H_\pm$, and $F_\pm$ are functions of six quantities $q_0^\pm, q^\pm, R$, and w_0 defined as

$$\begin{aligned} F_\pm &= E_z \left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z \right), \\ G_\pm &= E_z \left\{ \left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z \right) z \right\}, \\ H_\pm &= E_z \left\{ \left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z \right)^2 \right\}, \end{aligned} \tag{4}$$

J. Stat. Mech. (2021) 113401

where z is a standard Gaussian random variable and the expectation $E_z = \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$. The $\hat{u}_\pm$ are also functions of $q_0^\pm, q^\pm, R$, and w_0 defined using (2) as

$$\hat{u}_\pm = \psi\left(R\mu \pm w_0 + \sqrt{q_0^\pm} z, q^\pm\right).$$

The values of $q_0^\pm, q^\pm, R, w_0$ can be obtained by solving the following six nonlinear equations:

$$q_0^\pm = \frac{1}{p} E_z(\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_\pm \hat{\boldsymbol{\mu}}), \tag{5}$$

$$R = \frac{1}{\sqrt{p}} E_z(\hat{\boldsymbol{\mu}}^T \bar{\boldsymbol{\mu}}), \tag{6}$$

$$\frac{\alpha_+}{q^+} F_+ = \frac{\alpha_-}{q^-} F_-, \tag{7}$$

$$\frac{q^+}{\sigma_+^2} = \frac{q^-}{\sigma_-^2}, \tag{8}$$

$$\frac{q^+ \lambda}{\sigma_+^2} = 1 + \alpha_+ G_+ + \alpha_- G_-. \tag{9}$$

From (3), the limiting distribution of $\hat{\mathbf{w}}$ is a multivariate normal with mean $(\xi^+ \boldsymbol{\Sigma}_+ + \xi^- \boldsymbol{\Sigma}_- + \lambda \mathbf{I}_p)^{-1} \left(\sqrt{p} \tilde{R} \bar{\boldsymbol{\mu}}\right)$ and covariance matrix $(\xi^+ \boldsymbol{\Sigma}_+ + \xi^- \boldsymbol{\Sigma}_- + \lambda \mathbf{I}_p)^{-1} (\xi_0^+ \boldsymbol{\Sigma}_+ + \xi_0^- \boldsymbol{\Sigma}_-) (\xi^+ \boldsymbol{\Sigma}_+ + \xi^- \boldsymbol{\Sigma}_- + \lambda \mathbf{I}_p)^{-1}$, where all the parameters can be determined by solving a set of six nonlinear equations (5)–(9). Note that two types of Gaussian random variables are introduced, i.e. primary variable $\hat{\boldsymbol{\mu}}$ and conjugate variable $\hat{u}$. The variances of these two random variables are controlled by $\xi_0^\pm$ and $q_0^\pm$ respectively. It is interesting to see that $\xi_0^\pm$ is determined by the expectation over a quadratic form of $\hat{u}$ i.e. (4), while $q_0^\pm$ is determined by the expectation over a quadratic form of $\hat{\boldsymbol{\mu}}$, i.e. (5).

The derivation of claim 1 is given in appendix A based on the replica method developed in statistical mechanics. The replica method is a non-rigorous but highly sophisticated calculation procedure that has been used to derive a number of fascinating results in probability theory and information theory, see e.g. Tanaka (2002); Dongning Guo and Verdu (2005); Wu and Verdu (2012).

Using the asymptotic statistical behavior of the classification estimators provided in claim 1, we are able to retrieve the asymptotic performance of the classification method (1). Denote $\hat{\mathbf{w}}, \hat{w}_0$ the solution of (1), the classification precision $P\{\pm(\mathbf{x}_\pm^T \hat{\mathbf{w}} + \hat{w}_0) \geqslant 0\}$ has an asymptotically deterministic behavior as given by the following claim.

Claim 2 For $\mathbf{x}_\pm$ generated from Class $\pm$, we have

$$P\{\pm(\mathbf{x}_\pm^T \hat{\mathbf{w}} + \hat{w}_0) \geqslant 0\} \xrightarrow{P} \Phi(\zeta_\pm), \tag{10}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of $N(0,1)$ and

$$\zeta_\pm = \frac{R\mu \pm w_0}{\sqrt{q_0^\pm}}.$$

J. Stat. Mech. (2021) 113401

The values of the quantities R, w_0, and $q_0^{\pm}$ can be obtained from solving equations (5)–(9) listed in claim 1.

Claim 2 allows us to assess the performance of different classification methods and obtain the value of λ that yields the maximum precision for a given method. If we consider $\mathbf{\Sigma}_+ = \mathbf{\Sigma}_-$ and PLR loss $V(u)$, claims 1 and 2 end up with the results of Mai *et al* (2019).

Now we consider datasets 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 the following:

Assumption 1 Each observation vector $\mathbf{x}_+$ (resp. $\mathbf{x}_-$) from class $+1$ (resp. class -1) can be viewed as an independent instantiation of the generative models

$$\mathbf{x}_+ = \boldsymbol{\mu} + \sum_{k=1}^{K} \sigma_+ \sqrt{\lambda_k^+} \mathbf{v}_k \varepsilon_k + \boldsymbol{\epsilon}^+ \quad \left(\text{resp.} \quad \mathbf{x}_- = -\boldsymbol{\mu} + \sum_{k=1}^{K} \sigma_- \sqrt{\lambda_k^-} \mathbf{v}_k \varepsilon_k + \boldsymbol{\epsilon}^- \right), \tag{11}$$

where $\lambda_k^{\pm} > 0$ and $\mathbf{v}_k \in \mathbb{R}^p$ are orthonormal vectors for $k = 1, \ldots, K$, i.e. $\mathbf{v}_k^T \mathbf{v}_k = 1$ and $\mathbf{v}_k^T \mathbf{v}_{k'} = 0$ for $k \neq k'$. The random variables $\varepsilon_1, \ldots, \varepsilon_K$ are i.i.d. with mean 0 and variance 1. The elements of the p-vector $\boldsymbol{\epsilon}^{\pm} = \{\epsilon_1^{\pm}, \ldots, \epsilon_p^{\pm}\}$ are i.i.d. random variables with $E(\epsilon_j^{\pm}) = 0$, $E\{(\epsilon_j^{\pm})^2\} = \sigma_{\pm}^2$, and $E\{(\epsilon_j^{\pm})^3\} < \infty$. The $\epsilon_j^{\pm}$s and ε_ks are independent from each other.

In model (11), $\lambda_k^{\pm}$ represents the strength of the kth signal component, and $\sigma_{\pm}^2$ represents the level of background noise. The real signal is typically low-dimensional, i.e. $K \ll p$. Here we use the most general assumption and allow different spiked covariances for different classes, e.g. we can have $\lambda_k^+ = 0$ and $\lambda_k^- \neq 0$. Note that the eigenvalue $\lambda_k^{\pm}$ is not necessarily decreasing in k and $\lambda_1^{\pm}$ is not necessarily the largest eigenvalue. From (11), the covariance matrix becomes

$$\mathbf{\Sigma}_{\pm} = \sigma_{\pm}^2 \left(\mathbf{I}_p + \sum_{k=1}^{K} \lambda_k^{\pm} \mathbf{v}_k \mathbf{v}_k^T \right). \tag{12}$$

The kth eigenvalue of $\mathbf{\Sigma}_{\pm}$ is $\sigma_{\pm}^2(1 + \lambda_k^{\pm})$ for $k = 1, \ldots, K$ and $\sigma_{\pm}^2$ for $k = K+1, \ldots, p$. Although the $\epsilon_j^{\pm}$s are i.i.d., we did not impose any parametric form for the distribution of $\epsilon_j^{\pm}$, which allows for very flexible covariance structures for $\mathbf{x}$, and thus the results are quite general. The requirement for the finite third order moment is to ensure Berry–Esseen central limit theorem applies. The assumption 1 is also called spiked population model and has been used in many situations, see Marcenko and Pastur (1967); Hastie *et al* (1995); Telatar (1999); Laloux *et al* (2000); Johnstone (2001); Sear and Cuesta (2003); Baik and Silverstein (2006) for examples.

Denote the projections of eigenvectors on the signal direction $\bar{\boldsymbol{\mu}}$ as $R_k = \mathbf{v}_k^T \bar{\boldsymbol{\mu}}$ for $k = 1, \ldots, K$; $R_{K+1} = \sqrt{1 - \sum_{k=1}^{K} R_k^2}$; and $R_k = 0$ for $k = K+2, \ldots, p$. After integration

J. Stat. Mech. (2021) 113401

over $\mathbf{z}_\pm$ in (5) and (6), we have the explicit formulas for $q_0^\pm$ and R as

$$q_0^\pm = \alpha_+ H_+ + \alpha_- H_- + \left(\frac{\alpha_\pm \mu}{\sigma_\pm} F_\pm + \frac{\alpha_\mp \mu \sigma_\pm}{\sigma_\mp^2} F_\mp\right) \sum_{k=1}^{2K+1} \frac{(1+\lambda_k^\pm) R_k^2}{\left\{1 - \frac{\alpha_+ \lambda_k^+ G_+}{\sqrt{q_0^+}} - \frac{\alpha_- \lambda_k^- G_-}{\sqrt{q_0^-}}\right\}^2},$$

$$R = \left(\frac{\alpha_+ \mu}{\sigma_+} F_+ + \frac{\alpha_- \mu \sigma_+}{\sigma_-^2} F_-\right) \sum_{k=1}^{K+1} \frac{R_k^2}{\left\{1 - \frac{\alpha_+ \lambda_k^+ G_+}{\sqrt{q_0^+}} - \frac{\alpha_- \lambda_k^- G_-}{\sqrt{q_0^-}}\right\}}.$$

Note that if $\sigma_+ = \sigma_-$, we get $q^+ = q^-$ directly from (8). If we further assume $\mathbf{\Sigma}_+ = \mathbf{\Sigma}_-$ and $\alpha_+ = \alpha_-$, we get $w_0 = 0$, $q_0^+ = q_0^-$, $q^+ = q^-$ from (5) and (7). In this case, we only need to estimate three parameters and the results are much simpler.

Claim 3 Under assumption 1, assume that $\sigma_\pm$, $\lambda_k^\pm$, R_k, and K are fixed irrespective of p. Further assume that $\mathbf{\Sigma}_+ = \mathbf{\Sigma}_- = \mathbf{\Sigma}$, $\alpha_+ = \alpha_- = \alpha$, $\sigma_+ = \sigma_- = \sigma$, and $\lambda_k^+ = \lambda_k^- = \lambda_k$ for $k = 1, \ldots, K$. Denote $q^+ = q^- = q$ and $q_0^+ = q_0^- = q_0$. Then the limiting distribution of $\hat{\mathbf{w}}$ is the same as that of

$$(\xi \mathbf{\Sigma} + \lambda \mathbf{I}_p)^{-1} \left(\sqrt{\xi_0} \mathbf{\Sigma}^{1/2} \mathbf{z} + \sqrt{p} \tilde{R} \bar{\boldsymbol{\mu}}\right), \tag{13}$$

which leads to the asymptotic precision

$$P\{\pm(\mathbf{x}_\pm^T \hat{\mathbf{w}}) \geqslant 0\} \to \Phi\left(\frac{R\mu}{\sqrt{q_0}}\right). \tag{14}$$

Here $\mathbf{z}$ denotes the vectors of length p whose elements are i.i.d. standard Gaussian random variables, and

$$\xi = -\frac{\alpha}{\sqrt{q_0} q} f_2(q_0, q, R), \qquad \xi_0 = \frac{\alpha}{q^2} f_3(q_0, q, R), \qquad \tilde{R} = \frac{\alpha \mu}{q} f_1(q_0, q, R),$$

where the three functions f_1, f_2, f_3 are defined as

$$f_1(q_0, q, R) = E_z[(\hat{u} - R\mu - \sqrt{q_0} z)],$$

$$f_2(q_0, q, R) = E_z[(\hat{u} - R\mu - \sqrt{q_0} z) z],$$

$$f_3(q_0, q, R) = E_z[(\hat{u} - R\mu - \sqrt{q_0} z)^2].$$

Here the expectation is with respect to the standard Gaussian measure $z \sim N(0, 1)$ and

$$\hat{u} = \psi(R\mu + \sqrt{q_0} z, q).$$

J. Stat. Mech. (2021) 113401

The three parameters q_0, q, and R are determined by the following three nonlinear equations

$$\frac{\sigma R}{\sqrt{q_0}} = \alpha \frac{\mu}{\sigma} \frac{f_1(q_0, q, R)}{\sqrt{q_0}} \sum_{k=1}^{K+1} \frac{R_k^2}{1 - \alpha \lambda_k f_2(q_0, q, R)/\sqrt{q_0}}, \tag{15}$$

$$1 = \alpha \frac{f_3(q_0, q, R)}{q_0} + \left\{ \alpha \frac{\mu}{\sigma} \frac{f_1(q_0, q, R)}{\sqrt{q_0}} \right\} \sum_{k=1}^{2K+1} \frac{(1 + \lambda_k) R_k^2}{\{1 - \alpha \lambda_k f_2(q_0, q, R)/\sqrt{q_0}\}^2}, \tag{16}$$

$$\frac{q\lambda}{\sigma^2} = 1 + \alpha \frac{f_2(q_0, q, R)}{\sqrt{q_0}}. \tag{17}$$

Note that for SVM loss $V(u)$, claim 3 is reduced to the results of Huang (2017) and Mai and Couillet (2018) while for DWD loss $V(u)$, it is reduced to the results of Huang (2018). Therefore, the results presented here are more general and can be used to compare the performance of different classification methods for a given problem.

4. Phase transition

Based on the asymptotic results in section 3, in this section, we derive the phase transition for the non-regularized classification methods that solve the following optimization problem

$$\text{argmin}_{\mathbf{w} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} V(y_i \mathbf{x}_i^T \mathbf{w}) \right\}. \tag{18}$$

As shown in Candès and Sur (2020), the existence for the non-regularized classification methods undergoes a phase transition, i.e. the solution of (18) does not exist in situations when the two classes of n data points $(\mathbf{x}_i, y_i)$ are completely linear separated and it does exist if the data points overlap. This is equivalent to establishing the the maximum number of training samples per dimensions below which the hard-margin SVM can have solution as shown in Montanari *et al* (2019); Sifaou *et al* (2019); Deng *et al* (2019).

Consider the asymptotic regime where $n, p \to \infty$ such that

$$p/n \to \kappa \in (0, \tau], \tag{19}$$

where $\tau > 0$ and κ is called the overparametrization ratio. To quantify its effect on the test error, we study the problem of increasing dimensions as in (19) that further satisfy

$$\mu^2 = s^2 \kappa / \tau, \qquad \sigma^2 = s^2 (1 - \kappa/\tau). \tag{20}$$

The following claim characterizes the phase transition of the model (18) in terms of κ and τ.

Claim 4 Define $\kappa_{\min}(\tau)$ as the solution of

$$1 = \frac{1}{\kappa} \int_{-\infty}^{z_c} (z_c - z)^2 Dz + \frac{1}{\kappa(\tau - \kappa)} \sum_{k=1}^{K+1} \frac{R_k^2}{1 + \lambda_k} \left\{ \int_{-\infty}^{z_c} (z_c - z) Dz \right\}^2, \tag{21}$$

J. Stat. Mech. (2021) 113401

where $Dz = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\mathrm{d}z$ and $\Phi(z_c) = \kappa$, i.e. z_c is the κth quantile of standard normal distribution. If the overparametrization ratio is smaller enough such that $\kappa < \kappa_{\min}$, then the solution of equation (18) asymptotically exists with probability one. Conversely, if $\kappa > \kappa_{\min}$, then the solution does not exist with probability one.

Note that claim 4 generalizes the result of Deng *et al* (2019) for hard margin SVM, which can be considered as a special case here if one chooses $\mathbf{\Sigma} = \mathbf{I}_p$, where $\mathbf{I}_p$ is p-dimensional identity matrix.

5. Estimation of data parameters

So far we assumed that the design covariance $\mathbf{\Sigma}_\pm$ and other data parameters are known. In practice, we need to estimate K, $\mu, \sigma_\pm$, $\lambda_k^\pm$, and R_k for $k = 1, \ldots, K$ from the data. The problem of estimating covariance matrices in high-dimensional setting has attracted considerable attention in the past. Since the covariance estimation problem is not the focus of our paper, we will test the above approach using a simple covariance estimation method based the application of random matrix theory to spiked population model.

To estimate the background noise level $\sigma_\pm^2$, we use a robust variance estimate based on the full matrix of data values (Liu *et al* 2008); that is, for the full set of $n_\pm p$ entries of the original $n_\pm \times p$ data matrix $\mathbf{X}^\pm$, we calculate the robust estimate of scale, the median absolute deviation from the median (MAD), to estimate $\sigma_\pm$ as

$$\hat{\sigma}_\pm = \frac{\mathrm{MAD}_{\mathbf{X}^\pm}}{\mathrm{MAD}_{N(0,1)}}. \tag{22}$$

Here $\mathrm{MAD}_{\mathbf{X}^\pm} = \mathrm{median}(|x_{ij}^\pm - \mathrm{median}(\mathbf{X}^\pm)|)$ and $\mathrm{MAD}_{N(0,1)} = \mathrm{median}(|r_i - \mathrm{median}(\mathbb{R})|)$, where $\mathbb{R}$ is a $n_\pm p$-dimensional vector whose elements are i.i.d. samples from $N(0,1)$ distribution.

Denote $\bar{\boldsymbol{\mu}}_c = \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-$, where $\bar{\mathbf{x}}_+ = \frac{1}{n_+}\sum_{i=1}^{n_+} \mathbf{x}_{+,i}$ and $\bar{\mathbf{x}}_- = \frac{1}{n_-}\sum_{i=1}^{n_-} \mathbf{x}_{-,i}$ represent the sample means for class $+1$ and class -1, respectively. Then, according to Huang (2017), we estimate μ as

$$\hat{\mu} = \frac{1}{2}\sqrt{\|\bar{\boldsymbol{\mu}}_c\|^2 - \frac{\hat{\sigma}_+^2}{\alpha_+} - \frac{\hat{\sigma}_-^2}{\alpha_-}}.$$

Denote $\tilde{\mathbf{\Sigma}}_\pm$ the sample covariance matrix for class ± 1. Store all eigenvalues of $\tilde{\mathbf{\Sigma}}_\pm$ greater than $(1 + \sqrt{1/\alpha_\pm})^2 - 1$ as $[\tilde{\lambda}_1^\pm, \ldots, \tilde{\lambda}_{\hat{K}_\pm}^\pm]$ and their corresponding eigenvectors as $[\tilde{\mathbf{v}}_1^\pm, \ldots, \tilde{\mathbf{v}}_{\hat{K}_\pm}^\pm]$. Let $\hat{K} = \hat{K}_+ + \hat{K}_-$. By concatenating the spiked eigenvalues and eigenvectors from the two classes together, we obtain $\hat{K}$ spiked eigenvalues and their corresponding $\hat{K}$ eigenvectors. Then we relabel them and assign label $k \in [1, \ldots, \hat{K}_+]$ to class $+1$ and label $k \in [\hat{K}_+ + 1, \ldots, \hat{K}]$ to class -1. To estimate $\lambda_k^\pm$ and R_k for $k = 1, \ldots, \hat{K}$, we use the results from Baik and Silverstein (2006). Define the function

J. Stat. Mech. (2021) 113401

$P(u,v) = \sqrt{(1-1/uv^2)(1+1/uv)}$. For $k \in [1,\ldots,\hat{K}_+]$, we have

$$\hat{\lambda}_k^+ = \frac{1}{2}\left(\tilde{\lambda}_k^+ - \frac{1}{\alpha_+} + \sqrt{\left(\tilde{\lambda}_k^+ - \frac{1}{\alpha_+}\right)^2 - \frac{4}{\alpha_+}}\right)$$

$$\tilde{R}_k = \frac{\bar{\boldsymbol{\mu}}_c^T \tilde{\mathbf{v}}_k^+}{\|\bar{\boldsymbol{\mu}}_c\| P(\|\tilde{\mathbf{v}}_k^+\|, \alpha_+)} \tag{23}$$

and $\hat{\lambda}_k^- = 0$. For $k \in [\hat{K}_+ + 1,\ldots,\hat{K}]$, we have

$$\hat{\lambda}_k^- = \frac{1}{2}\left(\tilde{\lambda}_{k-\hat{K}_+}^- - \frac{1}{\alpha_-} + \sqrt{\left(\tilde{\lambda}_{k-\hat{K}_+}^- - \frac{1}{\alpha_-}\right)^2 - \frac{4}{\alpha_-}}\right)$$

$$\tilde{R}_k = \frac{\bar{\boldsymbol{\mu}}_c^T \tilde{\mathbf{v}}_{k-\hat{K}_+}^-}{\|\bar{\boldsymbol{\mu}}_c\| P(\|\tilde{\mathbf{v}}_{k-\hat{K}_+}^-\|, \alpha_+)} \tag{24}$$

and $\hat{\lambda}_k^+ = 0$.

If we consider homogeneous situation where the two classes have the same covariance matrix, the results are much simpler. In this case, we need to combine two matrices $\mathbf{X}_+$ and $\mathbf{X}_-$ together to get a common set of spiked eigenvalues and eigenvectors. Then similar to (23) and (24), we use the results from Baik and Silverstein (2006) to estimate the common eigenvalues λ_k and projecting coefficients R_k.

6. Numerical analysis

In this section, we apply the general theoretical results derived in section 3 to several specific classification methods by numerically solving the nonlinear equations for the corresponding loss functions. We aim to exploring and comparing different types of classifiers under various settings. Here we focus on homogeneous situations with $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_-$ and $\alpha_+ = \alpha_-$ because in these situations the Bayes optimal classifiers are also linear and the classification performance can be exactly retrieved by the average precision derived in claim 2.

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 3 is reliable. The performance of a classification method is assessed through the average precision computed based on (14). Figure 2 shows the comparison between our asymptotic estimations and simulations on finite dimensional datasets. We use the *R* packages *kernlab*, *glmnet*, *DWD*, and *DWDLargeR* for solving SVM, PLR, DWD ($q = 1$), and DWD ($q = 2$) classification problem respectively. We did not present simulation results for LUM and generalized DWD with non-integer q because we cannot find reliable software package for solving this problem. The software package *DWDLargeR* that is based on the algorithm developed in Lam *et al* (2018) does not

J. Stat. Mech. (2021) 113401

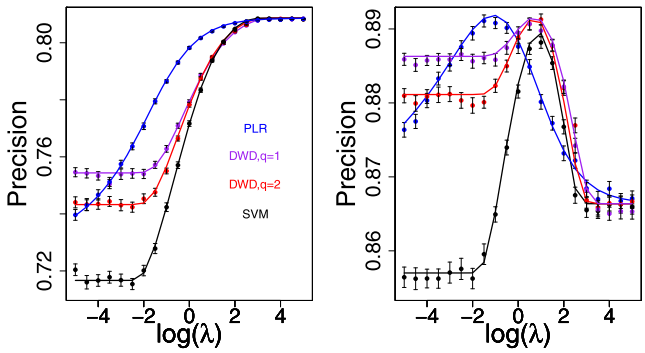


Figure 2. Theoretical and empirical precision as a function of λ for SVM, $\text{DWD}_{q=1}$, $\text{DWD}_{q=2}$, and PLR with $\alpha = 1,\ \mu = 2$. The empirical precision is taken by averaging over 100 simulated datasets with $p = 500$. (Left panel) The spike vector is aligned with μ with $K = 2$, $\lambda_1 = \lambda_2 = 4$, $R_1 = 1$, $R_2 = 0$. (Right panel) The spike vector is not aligned with μ with $K = 2$, $\lambda_1 = \lambda_2 = 4$, $R_1 = 1/\sqrt{2}$, $R_2 = 0$.

provide the option for non-integer q. Here the dimension of the simulated data is $p = 500$ and the data are generated according to (11) in assumption 1 with i.i.d. normal noise. We repeat the simulation 100 times for each parameter setting. The mean and standard errors over 100 replications are presented.

From figure 2, we can see that our analytical curves show fairly good agreement with the simulation experiment. Thus our analytical formula (14) provides reliable estimates for average precision even under moderate system sizes.

It is interesting to see that for small regularization parameter λ, the four patterns are quite different and SVM yields much smaller precision comparing to other three methods. On the other hand, if the tuning parameter λ is large enough, the precision of all four methods approaches the same value. This is easy to understand because for large λ, the solution of (1) is determined by the behavior of the loss function $V(u)$ at values of $u \to 0$, which turns out to be $1 - u$ for SVM, DWD, LUM, and $\log(2) - u/2$ for PLR. Therefore, as $\lambda \to \infty$, the asymptotic results of (1) are approximately equal to the solution of

$$\text{argmin}_{\mathbf{w}} \left\{ \sum_{i=1}^{n} (c_1 - c_2 y_i \mathbf{x}_i^T \mathbf{w}) + \sum_{j=1}^{p} \frac{\lambda w_j^2}{2} \right\}, \tag{25}$$

which is proportional to the weighted sample mean difference between two classes, i.e.

$$\hat{\mathbf{w}} \sim \alpha_+ \bar{\mathbf{x}}_+ - \alpha_- \bar{\mathbf{x}}_-. \tag{26}$$

Here c_1 and c_2 are two constants, and $\bar{\mathbf{x}}_+$ and $\bar{\mathbf{x}}_-$ are the sample means for class $+1$ and class -1 respectively. On the other hand, for small λ, the solution of (1) is also determined by the tail behavior of the loss function $V(u)$ at large u values. Since the decay rates of different loss functions are quite different, this ends up with different behaviors at small λ values as shown in figure 2.

The difference between the settings of left panel and right panel of figure 2 is that in the left panel, the spike vectors $\mathbf{v}_k (k = 1, \ldots, K)$ are either aligned with or orthogonal

J. Stat. Mech. (2021) 113401

to $\boldsymbol{\mu}$ but in the right panel, the spike vectors are neither aligned nor orthogonal to $\boldsymbol{\mu}$. This discrepancy causes different patterns of the precision curves. In the left panel, the Bayes optimal solution is proportional to $\boldsymbol{\mu}$, which can be estimated using the difference of sample means between two classes. In this situation, as λ increases, all solutions approach to the optimal one and thus we obtain increasing function for the precision. More specifically, it was pointed out in Huang (2017), that the asymptotic value we can achieve for the precision is $\Phi\left(\frac{\rho_c}{\sqrt{1+\lambda_1\rho_c^2}}\frac{\mu}{\sigma}\right)$, where $\rho_c = \sqrt{\frac{\alpha\left(\frac{\mu}{\sigma}\right)^2}{1+\alpha\left(\frac{\mu}{\sigma}\right)^2}}$, and λ_1 represents the spiked eigenvalue in the $\boldsymbol{\mu}$ direction. In the right panel where $\mathbf{v}_k$ is different in direction from $\boldsymbol{\mu}$, the Bayes optimal solution is proportional to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, thus the asymptotic solution as $\lambda \to \infty$ is no long the optimal one. In this situation, we need to tune λ so as to find the maximum precision for different methods. Note that this is consistent with the phenomenon observed in Dobriban and Wager (2018); Mignacco *et al* (2020), which show that in the case $\Sigma_\pm = I_p$ with balanced clusters, $\lambda = \infty$ gives the Bayes estimator, while in the unbalanced case the optimal regularization is finite. Because in balanced case, $\alpha_+ = \alpha_-$ and (26) is the Bayes estimator of $\boldsymbol{\mu}$, while in unbalanced case, $\alpha_+ \neq \alpha_-$ and (26) is not the Bayes estimator of $\boldsymbol{\mu}$.

In figure 3, we study the phase transition for the separability of two classes. The (left panel) of figure 3 displays the phase transition boundary for the separability of the two classes in the plane of κ and τ, which are defined in (19) and (20). Above the curve is the region where the probability of separating the two classes tends to 1 and below is the region where the probability of separating the two classes tends to 0. The prediction errors as a function of the overparametrization ratio κ with fixed $\tau = 1.5$ for the four classification methods under small regularization $\lambda = 10^{-5}$ are shown in the right panel of figure 3. The double descent behavior are found for all methods with peaks near the separability threshold $\kappa_{\min}(1.5)$. This phenomenon indicates that the prediction error descends again after the threshold. A similar study has been given in Deng *et al* (2019) for hard margin SVM and unregularized logistic regression under i.i.d. covariance structure setting. The curves in the right panel of figure 3 also show that, as dimension increasing, DWD and PLR perform better than SVM under the non-regularized setting, i.e. $\lambda \to 0$. This is consistent with the previous empirical observations in Marron *et al* (2007); Benito *et al* (2004). The reason is that the small λ behaviors of classification is determined by the decay speed of the corresponding loss function $V(u)$. The SVM hinge loss vanishes for the entire region of $u \geqslant 1$ but all the other loss functions decay to zero gradually as $u \to \infty$.

To further compare the performances of different methods, figure 4 plots the precision as a function of the parameters μ and α for optimal regularization, i.e. λ is tuned to obtain the maximum precision. We consider five different methods, which are SVM, PLR, DWD ($q = 1$), DWD ($q = 2$), and LUM ($a = 1$, $c = 2$). In figure 4, the left panel plots the precision as a function of α with fixed μ and the right panel plots the precision as a function of μ with fixed α. As it turns out, SVM performs worse than all the other four methods, but the discrepancy at optimal λ is smaller than at small λ as shown in figure 2. The performances of the other four methods are quite similar once λ is optimally tuned.

J. Stat. Mech. (2021) 113401

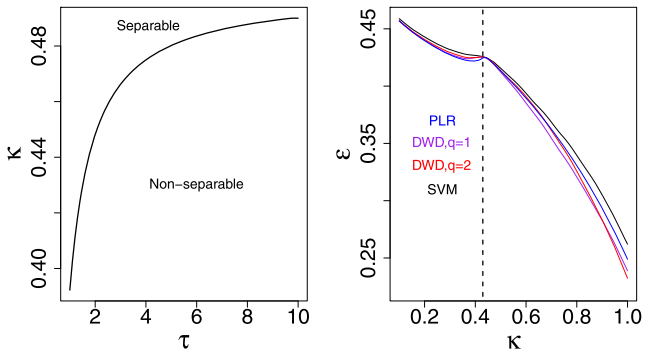


Figure 3. (Left) Theoretical prediction for the phase transition curves. (Right) Theoretical classification error against the number of features per sample $\kappa = p/n$ for four classification methods with fixed $\tau = 1.5$ under a small regularization setting $\lambda = 10^{-5}$. Here $K = 3$, $\lambda_1 = \lambda_2 = \lambda_3 = 4$, $R_1 = R_2 = 1/2, R_3 = 0$. The vertical dashed line represents the threshold $\kappa_{\min}(1.5)$ of linear separability of the dataset.

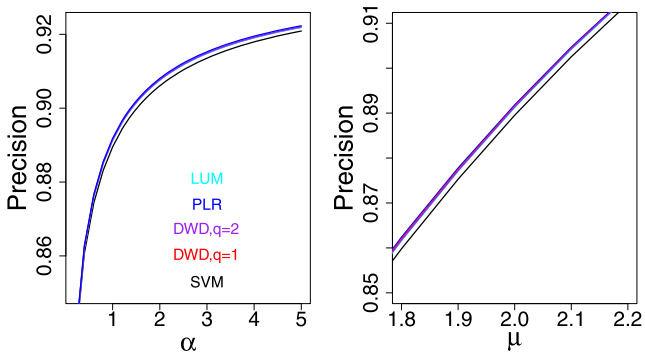


Figure 4. Theoretical precision as a function of α and μ at optimal λ for five different methods. Here $K = 3$, $\lambda_1 = \lambda_2 = \lambda_3 = 4$, $R_1 = R_2 = 1/2, R_3 = 0$. (Left) Theoretical precision as a function of α for fixed $\mu = 2$. (Right) Theoretical precision as a function of μ for fixed $\alpha = 1$.

We have performed numeric analysis under many other settings and the conclusions are quite similar. Overall, our analytical calculations agree well with the numerical simulations for moderate system sizes, and claim 2 provides reliable estimates for average precision. Our main observations from numeric analyses are

- All methods achieve the same performances for large enough λ.
- For situations where the spiked vectors are the same in direction with $\boldsymbol{\mu}$, the optimal solutions of all methods are the same, which is also equivalent to the limiting results as $\lambda \to \infty$.
- For situations where the spiked vectors are different in direction from $\boldsymbol{\mu}$, DWD and PLR are better than SVM especially when the regularization parameter λ is small.

J. Stat. Mech. (2021) 113401

This finding provides theoretical confirmations to the empirical results that have been observed in many previous simulation and real data studies.

- The previous empirical observation that DWD is better than SVM only holds at small λ. After carefully tuning for λ, the performance of all methods is quite similar. DWD, PLR, and LUM are slightly better than SVM at large α.
- The so-called double descent behaviors exist for all non-regularized margin-based classification methods with a peak at the separability threshold.

Note that the analytical demonstrations about the superior performance of DWD over SVM at small λ are consistent with many previous empirical findings. However, this does not mean that DWD is better than SVM at small λ in all situations because our numerical results are derived based on the spiked population assumption that may not always hold in practice.

7. Real data

We apply our methods to a breast cancer dataset from The Cancer Genome Atlas Research Network (TCGA 2010), which include two subtypes: LumA and LumB. As in Liu *et al* (2008); Huang *et al* (2015), we filter the genes using the ratio of the sample standard deviation and sample mean of each gene. After gene filtering, the dataset contained 235 patients with 169 genes. Among the 235 samples, there are 154 LumA samples and 81 LumB samples.

We consider LumA as class $+1$ and LumB as class -1. Assume the data are generated based on model (11), using the method discussed in section 5, we obtain the following parameter estimations: $\mu = 4.81$, $\sigma = 1.66$, $\alpha = 1.39$, $p = 169$, $n = 235$, $n_+ = 154$, $n_- = 81$, $K = 16$, $[\lambda_1, \ldots, \lambda_{16}] = [25.67, 12.00, 10.10, 9.37, 6.41, 4.92, 4.38, 4.19, 3.63, 3.09, 2.45, 1.96, 1.87, 1.69, 1.57, 0.98]$, and $[R_1, \ldots, R_{16}] = [-0.67, -0.11, 0.57, 0.04, 0.18, -0.06, -0.03, -0.03, -0.16, -0.32, 0.01, 0.03, -0.06, 0.06, 0.03, -0.05]$.

Figure 5 plots the analytical curves of the average precision as functions of λ for three classification methods SVM, PLR, and DWD ($q = 1$). For comparison, the CV precision is computed by randomly splitting the data into two parts, 95% for training and 5% for test. The mean and standard deviation over 100 random splitting are presented. It can be seen that, at large λ, there are some discrepancies between the theoretical estimation and CV experiment. Note that, from (26), the solution of the margin based classification method (1) can be approximated by the sample mean estimation if λ is big enough. However, it is well known that the sample mean estimation method performs much worse than DWD and SVM if there are unbalanced sub-classes within each class as shown in Liu *et al* (2009). Therefore, our results indicate that the data might include more complicated sub-cluster structure than the mixture of two simple components. On the other hand, at small λ, our theoretical estimation are quite close to the CV analysis. Particularly, at optimal λ, the theoretical estimations are 80.9% (SVM), 81.0% (DWD), 81.1% (PLR), which are very close to the corresponding results based on CV analysis, which are 81.5% (SVM), 81.4% (DWD), 80.9% (PLR). Moreover, the maximum theoretical estimation for the three methods occurs at quite similar λ values as the corresponding CV experiment. Overall, our theoretical results on the asymptotic

J. Stat. Mech. (2021) 113401

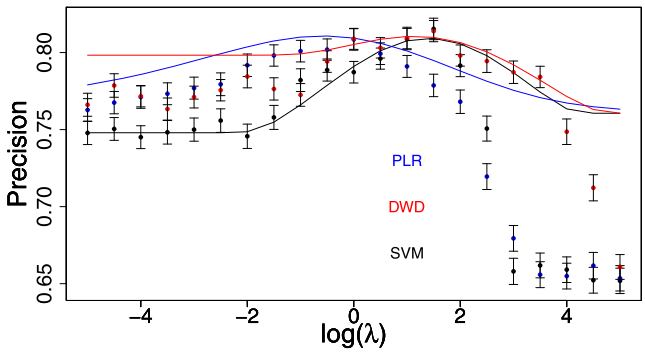


Figure 5. The classification precision as a function of λ based on theoretical estimation and CV analysis. The solid curves represent the theoretical results based on the parameters estimated from the breast cancer data. The error bars represent the CV experiment over 100 splittings.

precision can still provide reasonable guidelines on how to choose classification methods and tuning parameters for a given problem in practice.

8. Conclusion

Large-margin classifiers play an important role in classification problems. In this study, we examine the asymptotic behavior of a family of large-margin classifiers in the limit of $p, n \to \infty$ with fixed $\alpha = n/p$. This family includes many existing classifiers such as the SVM, DWD, PLR, and LUM as well as many new ones that can be built from the general convex loss function. Our focus is on the limiting distribution and classification precision of the estimators. On the basis of analytical evaluation, a method of selecting the best model and optimal tuning parameter is naturally developed for analyzing high dimensional data, which significantly reduces the computational cost. Although our theoretical results are asymptotic in the problem dimensions, numerical simulations have shown that they are already accurate on problems with a few hundreds of variables.

Our analytical analyses provide deeper theoretical evidence to support the empirical conclusion that hard margin DWD yields better classification performance than hard margin SVM in high dimensions. Certainly, our observations may not be valid for all classification problems because we have applied the mixture of two components assumption with spiked covariance structure in numerical studies, which cannot be true in all situations. Nevertheless, our analyses provide a convenient platform for deep investigation of the nature of margin-based classification methods and can also improve their practical use in various aspects as shown by the real data analysis in section 7. Note that our numerical analysis focus on homogeneous cases where $\Sigma_+ = \Sigma_-$. For non-homogeneous cases, the Bayes optimal solution is nonlinear and one possible solution is to use kernel based method. One of our future research topics is to derive the asymptotic behavior for the kernel based large margin classification methods. In situations where the spiked model cannot be applied or each class includes further sub-cluster structure, we plan to

J. Stat. Mech. (2021) 113401

study the generalized spiked population model (Bai and Yao 2012) or the classification methods that can incorporate the sub-cluster analysis.

Acknowledgments

The authors thank the editor and three referees for many helpful comments and suggestions that 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 Social Science Foundation of China under Grant 20BTJ039 (Yang), the National Science Foundation of China under Grant 11971324 (Yang), the Chinese Ministry of Education of Humanities and Social Science Project under Grant 19YJC910008 and the Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law under Grant 2722021AJ009 (Yang).

Appendix A

This appendix outlines the replica calculation leading to claim 1. Claims 2 and 3 are just direct applications of claim 1. We limit ourselves to the main steps. For a general introduction to the replica method and its motivation, we refer to Mezard *et al* (1987); Mézard and Montanari (2009).

Denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$. Among the n samples, let the first n_+ ones belong to class $+1$, i.e. $y_i = 1$ for $i \in \{1, \ldots, n_+\}$ and the last n_- ones belong to class -1, i.e. $y_i = -1$ for $i \in \{n_+ + 1, \ldots, n\}$. We consider regularized classification of the form

$$(\hat{\mathbf{w}}, \hat{w}_0) = \mathrm{argmin}_{\mathbf{w}, w_0} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{p}} + y_i w_0 \right) + \sum_{j=1}^{p} J_\lambda(w_j) \right\}. \tag{A1}$$

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)

$$\begin{aligned} Z_\beta(\mathbf{X}, \mathbf{y}) &= \int \exp\left\{ -\beta \left[\sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{p}} + y_i w_0 \right) + \sum_{j=1}^{p} J_\lambda(w_j) \right] \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}w_0 \\ &= \int \exp\left\{ -\beta \left[\sum_{i=1}^{n_+} V\left(\frac{\mathbf{x}_i^T \mathbf{w}}{\sqrt{p}} + w_0 \right) + \sum_{i=n_+ + 1}^{n} V\left(-\frac{\mathbf{x}_i^T \mathbf{w}}{\sqrt{p}} - w_0 \right) \right.\right. \\ &\qquad \left.\left. + \sum_{j=1}^{p} J_\lambda(w_j) \right] \right\} \mathrm{d}\mathbf{w}\, \mathrm{d}w_0, \end{aligned} \tag{A2}$$

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 $\mathbf{w}$ and w_0, which are the solution of (A1).

J. Stat. Mech. (2021) 113401

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

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

Notice that, by (A3) and using Laplace method in the integral (A2), we have

$$\mathcal{F} = \lim_{p\to\infty} \frac{1}{p} \min_{\mathbf{w}, w_0} \left\{ \sum_{i=1}^{n} V\left(\frac{y_i \mathbf{x}_i^T \mathbf{w}}{\sqrt{p}} + y_i w_0 \right) + \sum_{j=1}^{p} J_\lambda(w_j) \right\}.$$

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

$$\mathcal{F} = -\lim_{\beta\to\infty} \lim_{p\to\infty} \frac{1}{p\beta} \langle \log\, Z_\beta(\mathbf{X}, \mathbf{y}) \rangle_{\mathbf{X},\mathbf{y}}, \tag{A4}$$

where the angle bracket stands for the expectation with respect to the distribution of training data $\mathbf{X}$ and $\mathbf{y}$.

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

$$\log\, Z = \lim_{k\to 0} \frac{\partial Z^k}{\partial k} = \lim_{k\to 0} \frac{\partial}{\partial k} \log\, Z^k, \tag{A5}$$

and rewrite (A4) as

$$\mathcal{F} = -\lim_{\beta\to\infty} \lim_{p\to\infty} \frac{1}{p\beta} \lim_{k\to 0} \frac{\partial}{\partial k} \Xi_k(\beta), \tag{A6}$$

where

$$\Xi_k(\beta) = \langle \{Z_\beta(\mathbf{X}, \mathbf{y})\}^k \rangle_{\mathbf{X},\mathbf{y}} = \int \{Z_\beta(\mathbf{X}, \mathbf{y})\}^k \prod_{i=1}^{n} P(\mathbf{x}_i, y_i) \mathrm{d}\mathbf{x}_i\, \mathrm{d}y_i. \tag{A7}$$

Equation (A6) 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 (A7), we use the identity

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

where $\nu(\mathrm{d}x)$ denotes the measure over $x \in \mathbb{R}$. We obtain

J. Stat. Mech. (2021) 113401

$$\{Z_\beta(\mathbf{X},\mathbf{y})\}^k = \prod_{a=1}^{k}\left[\int \exp\left\{-\beta\left[\sum_{i=1}^{n} V\left(\frac{y_i\mathbf{x}_i^T\mathbf{w}^a}{\sqrt{p}} + y_i w_0^a\right) + \sum_{j=1}^{p} J_\lambda(w_j^a)\right]\right\}\mathrm{d}\mathbf{w}^a\,\mathrm{d}w_0^a\right] \tag{A8}$$

where we have introduced replicated parameters

$$\mathbf{w}^a \equiv [w_1^a,\ldots,w_p^a]^T \quad \text{and} \quad w_0^a, \quad \text{for } a=1,\ldots,k.$$

Exchanging the order of the two limits $p\to\infty$ and $k\to 0$ in (A6), we have

$$\mathcal{F} = -\lim_{\beta\to\infty}\frac{1}{\beta}\lim_{k\to 0}\frac{\partial}{\partial k}\left(\lim_{p\to\infty}\frac{1}{p}\Xi_k(\beta)\right). \tag{A9}$$

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

$$\nu(\mathrm{d}\mathbf{w}) = \int \exp\left\{-\beta\sum_{j=1}^{p} J_\lambda(w_j)\right\}\mathrm{d}\mathbf{w}.$$

Similarly, define the measure $\nu_+(\mathrm{d}\mathbf{x})$ and $\nu_-(\mathrm{d}\mathbf{x})$ over $\mathbf{x}\in\mathbb{R}^p$ as

$$\nu_+(\mathrm{d}\mathbf{x}) = P(\mathbf{x}|y=+1)\mathrm{d}\mathbf{x} \quad \text{and} \quad \nu_-(\mathrm{d}\mathbf{x}) = P(\mathbf{x}|y=-1)\mathrm{d}\mathbf{x}.$$

In order to carry out the calculation of $\Xi_k(\beta)$, we let $\nu^k(\mathrm{d}\mathbf{w}) \equiv \nu(d\mathbf{w}^1)\times\cdots\times\nu(\mathrm{d}\mathbf{w}^k)$ be a measure over $(\mathbb{R}^p)^k$, with $\mathbf{w}^1,\ldots,\mathbf{w}^k\in\mathbb{R}^p$. Analogously $\nu^n(\mathrm{d}\mathbf{x}) \equiv \nu(\mathrm{d}\mathbf{x}_1)\times\cdots\times\nu(\mathrm{d}\mathbf{x}_n)$ with $\mathbf{x}_1,\ldots,\mathbf{x}_n\in\mathbb{R}^p$, $\nu^n(\mathrm{d}y) \equiv \nu(\mathrm{d}y_1)\times\cdots\times\nu(\mathrm{d}y_n)$ with $y_1,\ldots,y_n\in\{-1,1\}$, and $\nu^k(\mathrm{d}w_0) \equiv \nu(\mathrm{d}w_0^1)\times\cdots\times\nu(\mathrm{d}w_0^k)$ with $w_0^1,\ldots,w_0^k\in\mathbb{R}$. With these notations, we have

$$\begin{aligned}
\Xi_k(\beta) &= \int \exp\left\{-\beta\sum_{i=1}^{n}\sum_{a=1}^{k} V\left(\frac{y_i\mathbf{x}_i^T\mathbf{w}^a}{\sqrt{p}} + y_i w_0^a\right)\right\}\nu^k(\mathrm{d}\mathbf{w})\nu^k(\mathrm{d}w_0)\nu^n(\mathrm{d}\mathbf{y})\nu^n(\mathrm{d}\mathbf{x}) \\
&= \int\left[\int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}} + w_0^a\right)\right\}\nu_+(\mathrm{d}\mathbf{x})\right]^{n_+} \\
&\quad\times\left[\int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{-\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}} - w_0^a\right)\right\}\nu_-(\mathrm{d}\mathbf{x})\right]^{n_-}\nu^k(\mathrm{d}\mathbf{w})\nu^k(\mathrm{d}w_0) \\
&= \int \exp\{p(\alpha_+\log I_+ + \alpha_-\log I_-)\}\nu^k(d\mathbf{w})\nu^k(\mathrm{d}w_0),
\end{aligned} \tag{A10}$$

where $\alpha_\pm = n_\pm/p$ and

J. Stat. Mech. (2021) 113401

$$I_\pm = \int \exp\left\{-\beta\sum_{a=1}^{k} V\left(\frac{\pm\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}} \pm w_0^a\right)\right\}\nu_\pm(\mathrm{d}\mathbf{x}). \tag{A11}$$

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 $\nu_+(\mathrm{d}\mathbf{x})$ and n_- integrals over $(\mathbb{R})^p$ with measure $\nu_-(\mathrm{d}\mathbf{x})$. We next use the identity

$$\mathrm{e}^{f(x)} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}(q-x)\tilde{Q}+f(q)}\mathrm{d}q\,\mathrm{d}\tilde{Q}. \tag{A12}$$

We apply this identity to (A11) and introduce integration variables $\mathrm{d}u^a, \mathrm{d}\tilde{u}^a$ for $1\leqslant a\leqslant k$. Letting $\nu^k(\mathrm{d}u)=\mathrm{d}u^1\ldots\mathrm{d}u^k$ and $\nu^k(d\tilde{u})=d\tilde{u}^1\ldots\mathrm{d}\tilde{u}^k$

$$\begin{aligned}
I_\pm &= \int \exp\left\{-\beta\sum_{a=1}^{k} V\left(u^a\pm w_0^a\right) + \mathrm{i}\sqrt{p}\sum_{a=1}^{k}\left(u^a \mp \frac{\mathbf{x}^T\mathbf{w}^a}{\sqrt{p}}\right)\tilde{u}^a\right\}\nu_\pm(\mathrm{d}\mathbf{x})\nu^k(\mathrm{d}u)\nu^k(\mathrm{d}\tilde{u})\\
&= \int \exp\left\{-\beta\sum_{a=1}^{k} V\left(u^a\pm w_0^a\right) + \mathrm{i}\sqrt{p}\sum_{a=1}^{k} u^a\tilde{u}^a - \frac{1}{2}\sum_{ab}(\mathbf{w}^a)^T\boldsymbol{\Sigma}_\pm\mathbf{w}^b\tilde{u}^a\tilde{u}^b\right.\\
&\quad\left. - \mathrm{i}\sum_{a=1}^{k}(\mathbf{w}^a)^T\boldsymbol{\mu}\tilde{u}^a\right\}\nu^k(\mathrm{d}u)\nu^k(\mathrm{d}\tilde{u}).
\end{aligned} \tag{A13}$$

Note that, conditional on $y=\pm1$, $\mathbf{x}$ follows multivariate distributions with mean $\pm\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}_\pm$. In deriving (A13), 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 (A13) to (A10), and introduce integration variables $Q_{ab}^\pm, \tilde{Q}_{ab}^\pm$ and $R^a, \tilde{R}^a$ associated with $(\mathbf{w}^a)^T\boldsymbol{\Sigma}_\pm\mathbf{w}^b/p$ and $(\mathbf{w}^a)^T\bar{\boldsymbol{\mu}}/\sqrt{p}$ respectively for $1\leqslant a,b\leqslant k$. Denote $\mathbf{w}_0=(w_0^a)_{1\leqslant a\leqslant k}$, $\mathbf{Q}_\pm\equiv(Q_{ab}^\pm)_{1\leqslant a,b\leqslant k}$, $\tilde{\mathbf{Q}}_\pm\equiv(\tilde{Q}_{ab}^\pm)_{1\leqslant a,b\leqslant k}$, $\mathbf{R}\equiv(R^a)_{1\leqslant a\leqslant k}$, and $\tilde{\mathbf{R}}\equiv(\tilde{\mathbf{R}}^a)_{1\leqslant a\leqslant k}$. Note that, constant factors can be applied to the integration variables, and we choose convenient factors for later calculations. Letting $\mathrm{d}\mathbf{Q}^\pm\equiv\prod_{a,b}\mathrm{d}Q_{ab}^\pm$ $\mathrm{d}\tilde{\mathbf{Q}}_\pm\equiv\prod_{a,b}\mathrm{d}\tilde{Q}_{ab}^\pm$, $\mathrm{d}\mathbf{R}\equiv\prod_a\mathrm{d}R^a$, and $\mathrm{d}\tilde{\mathbf{R}}\equiv\prod_a\mathrm{d}\tilde{R}^a$, we obtain

$$\Xi_k(\beta) = \int \exp\left\{-p\mathcal{S}_k(\mathbf{Q}_\pm,\tilde{\mathbf{Q}}_\pm,\mathbf{R},\tilde{\mathbf{R}},\mathbf{w}_0)\right\}\mathrm{d}\mathbf{Q}_+\,\mathrm{d}\mathbf{Q}_-\,\mathrm{d}\tilde{\mathbf{Q}}_+\,\mathrm{d}\tilde{\mathbf{Q}}_-\mathrm{d}\mathbf{R}\,\mathrm{d}\tilde{\mathbf{R}}\nu^k(\mathrm{d}w_0), \tag{A14}$$

where

$$\begin{aligned}
\mathcal{S}_k(\mathbf{Q}_\pm,\tilde{\mathbf{Q}}_\pm,\mathbf{R},\tilde{\mathbf{R}},\mathbf{w}_0) &= -\mathrm{i}\beta\left(\sum_{ab}Q_{ab}^+\tilde{Q}_{ab}^+ + \sum_{ab}Q_{ab}^-\tilde{Q}_{ab}^- + \sum_a R^a\tilde{R}^a\right)\\
&\quad - \frac{1}{p}\log\xi(\tilde{\mathbf{Q}}_\pm,\tilde{\mathbf{R}}) - \hat{\xi}(\mathbf{Q}_\pm,\mathbf{R},\mathbf{w}_0),\\
\xi(\tilde{\mathbf{Q}}_\pm,\tilde{\mathbf{R}}) &= \int \exp\left\{-\mathrm{i}\beta\sum_{ab}\tilde{Q}_{ab}^+(\mathbf{w}^a)^T\boldsymbol{\Sigma}_+\mathbf{w}^b - \mathrm{i}\beta\sum_{ab}\tilde{Q}_{ab}^-(\mathbf{w}^a)^T\boldsymbol{\Sigma}_-\mathbf{w}^b\right.
\end{aligned}$$

J. Stat. Mech. (2021) 113401

$$- \mathrm{i}\beta \sum_a \sqrt{p}\tilde{R}^a (\mathbf{w}^a)^T \bar{\boldsymbol{\mu}} \Bigg\} \nu^k(\mathrm{d}\mathbf{w}),$$

$$\hat{\xi}(\mathbf{Q}_\pm, \mathbf{R}, \mathbf{w}_0) = \alpha_+ \log \hat{I}_+ + \alpha_- \log \hat{I}_-, \tag{A15}$$

where

$$\hat{I}_\pm = \int \exp \Bigg\{ -\beta \sum_{a=1}^k V\left(u^a \pm w_0^a\right) + \mathrm{i}\sqrt{p} \sum_{a=1}^k u^a \tilde{u}^a$$

$$- \frac{p}{2} \sum_{ab} Q_{ab}^\pm \tilde{u}^a \tilde{u}^b - \mathrm{i}\sqrt{p} \sum_{a=1}^k R^a \mu \tilde{u}^a \Bigg\} \nu^k(\mathrm{d}u) \nu^k(\mathrm{d}\tilde{u}). \tag{A16}$$

Now we apply steepest descent method to the remaining integration. According to Varadhan's claim (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 (A14) to obtain

$$-\lim_{p\to\infty} \frac{1}{p} \Xi_k(\beta) = \mathcal{S}_k(\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbf{R}^\star, \tilde{\mathbf{R}}^\star, \mathbf{w}_0^\star),$$

where $\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbf{R}^\star, \tilde{\mathbf{R}}^\star, \mathbf{w}_0^\star$ is 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 $\mathcal{S}_k(\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbf{R}^\star, \tilde{\mathbf{R}}^\star, \mathbf{w}_0^\star)$ is indeed left unchanged by such change of variables. This is equivalent to postulating that $(w_0^a)^\star = w_0$, $R^a = R$, $\tilde{R}^a = \mathrm{i}\tilde{R}$,

$$(Q_{ab}^\pm)^\star = \begin{cases} q_1^\pm & \text{if } a = b \\ q_0^\pm & \text{otherwise} \end{cases}, \quad \text{and} \quad (\tilde{Q}_{ab}^\pm)^\star = \begin{cases} \mathrm{i}\dfrac{\beta\zeta_1^\pm}{2} & \text{if } a = b \\ \mathrm{i}\dfrac{\beta\zeta_0^\pm}{2} & \text{otherwise} \end{cases}, \tag{A17}$$

where the factor $i\beta/2$ is for future convenience. The next step consists in substituting the above expressions for $\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbf{R}^\star, \tilde{\mathbf{R}}^\star, \mathbf{w}_0^\star$ in $\mathcal{S}_k(\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbf{R}^\star, \tilde{\mathbf{R}}^\star, \mathbf{w}_0^\star)$ and then taking the limit $k \to 0$. We will consider separately each term of $\mathcal{S}_k(\mathbf{Q}_\pm^\star, \tilde{\mathbf{Q}}_\pm^\star, \mathbb{R}^\star, \tilde{R}^\star, \mathbf{w}_0^\star)$. Let us begin with the first term

$$- \mathrm{i}\beta \left(\sum_{ab} Q_{ab}^+ \tilde{Q}_{ab}^+ + \sum_{ab} Q_{ab}^- \tilde{Q}_{ab}^- + \sum_a R^a \tilde{R}^a \right)$$

$$= \frac{k\beta^2}{2}(\zeta_1^+ q_1^+ - \zeta_0^+ q_0^+) + \frac{k\beta^2}{2}(\zeta_1^- q_1^- - \zeta_0^- q_0^-) + k\beta R\tilde{R}. \tag{A18}$$

J. Stat. Mech. (2021) 113401

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

$$\begin{aligned}
\xi(\tilde{\mathbf{Q}}_\pm, \tilde{\mathbf{R}}) &= \int \exp\Bigg\{ \frac{\beta^2}{2}(\zeta_1^+ - \zeta_0^+) \sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}_+}^2 + \frac{\beta^2 \zeta_0^+}{2} \sum_{a,b=1}^k (\mathbf{w}^a)^T \boldsymbol{\Sigma}_+ \mathbf{w}^b \\
&\quad + \frac{\beta^2}{2}(\zeta_1^- - \zeta_0^-) \sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}_-}^2 + \frac{\beta^2 \zeta_0^-}{2} \sum_{a,b=1}^k (\mathbf{w}^a)^T \boldsymbol{\Sigma}_- \mathbf{w}^b \\
&\quad + \beta \sqrt{p} \sum_{a=1}^k \tilde{R} (\mathbf{w}^a)^T \bar{\boldsymbol{\mu}} \Bigg\} \nu^k(\mathrm{d}\mathbf{w}) \\
&= E \int \exp\Bigg\{ \frac{\beta^2}{2}(\zeta_1^+ - \zeta_0^+) \sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}_+}^2 + \beta \sqrt{\zeta_0^+} \sum_{a=1}^k (\mathbf{w}^a)^T \boldsymbol{\Sigma}_+^{1/2} \mathbf{z}_+ \\
&\quad + \frac{\beta^2}{2}(\zeta_1^- - \zeta_0^-) \sum_{a=1}^k \|\mathbf{w}^a\|_{\boldsymbol{\Sigma}_-}^2 + \beta \sqrt{\zeta_0^-} \sum_{a=1}^k (\mathbf{w}^a)^T \boldsymbol{\Sigma}_-^{1/2} \mathbf{z}_- \\
&\quad + \beta \sqrt{p} \sum_{a=1}^k \tilde{R} (\mathbf{w}^a)^T \bar{\boldsymbol{\mu}} \Bigg\} \nu^k(\mathrm{d}\mathbf{w}),
\end{aligned} \tag{A19}$$

where expectation is with respect to $\mathbf{z}_+, \mathbf{z}_- \sim N(0, I_{p\times p})$. Notice that, given $\mathbf{z}_+, \mathbf{z}_- \in \mathbb{R}^p$, the integrals over $\mathbf{w}^1, \ldots, \mathbf{w}^k$ factorize, whence

$$\begin{aligned}
\xi(\tilde{\mathbf{Q}}_\pm, \tilde{\mathbf{R}}) &= E \Bigg\{ \Bigg[\int \exp\Bigg\{ \frac{\beta^2}{2}(\zeta_1^+ - \zeta_0^+) \|\mathbf{w}\|_{\boldsymbol{\Sigma}_+}^2 + \beta \sqrt{\zeta_0^+} \mathbf{w}^T \boldsymbol{\Sigma}_+^{1/2} \mathbf{z}_+ \\
&\quad + \frac{\beta^2}{2}(\zeta_1^- - \zeta_0^-) \|\mathbf{w}\|_{\boldsymbol{\Sigma}_-}^2 + \beta \sqrt{\zeta_0^-} \mathbf{w}^T \boldsymbol{\Sigma}_-^{1/2} \mathbf{z}_- \\
&\quad + \beta \sqrt{p} \tilde{R} \mathbf{w}^T \bar{\boldsymbol{\mu}} \Bigg\} \nu^k(d\mathbf{w}) \Bigg]^k \Bigg\}.
\end{aligned}$$

Finally, after integration over $\nu^k(\mathrm{d}\tilde{u})$, (A16) becomes

$$\begin{aligned}
\hat{I}_\pm = \int \exp\Bigg\{ &-\beta \sum_{a=1}^k V(u^a \pm w_0) - \frac{1}{2} \sum_{ab} (u^a - R\mu)(\mathbf{Q}_\pm^{-1})_{ab} (u^b - R\mu) \\
&- \frac{1}{2} \log \det \mathbf{Q}_\pm \Bigg\} \nu^k(\mathrm{d}u).
\end{aligned} \tag{A20}$$

We can next take the limit $\beta \to \infty$. The analysis of the saddle point parameters $q_0^\pm, q_1^\pm, \zeta_0^\pm, \zeta_1^\pm$ shows that $q_0^\pm, q_1^\pm$ have the same limit with $q_1^\pm - q_0^\pm = (q^\pm/\beta) + o(\beta^{-1})$ and $\zeta_0^\pm, \zeta_1^\pm$ have the same limit with $\zeta_1^\pm - \zeta_0^\pm = (-\zeta^\pm/\beta) + o(\beta^{-1})$. Substituting the above

J. Stat. Mech. (2021) 113401

expression in (A18) and (A19), in the limit of $k \to 0$, we then obtain

$$-\mathrm{i}\beta\left(\sum_{ab} Q^+_{ab}\tilde{Q}^+_{ab} + \sum_{ab} Q^-_{ab}\tilde{Q}^-_{ab} + \sum_a R^a\tilde{R}^a\right)$$
$$= \frac{k\beta}{2}(\zeta_0^+ q^+ - \zeta^+ q_0^+) + \frac{k\beta}{2}(\zeta_0^- q^- - \zeta^- q_0^-) + k\beta R\tilde{R}, \tag{A21}$$

and

$$\xi(\tilde{\mathbf{Q}}_\pm, \tilde{\mathbf{R}}) = E\left\{\left[\int \exp\left\{-\frac{\beta\zeta^+}{2}\|\mathbf{w}\|^2_{\mathbf{\Sigma}_+} + \beta\sqrt{\zeta_0^+}\mathbf{w}^T\mathbf{\Sigma}_+^{1/2}\mathbf{z}_+ - \frac{\beta\zeta^-}{2}\|\mathbf{w}\|^2_{\mathbf{\Sigma}_-}\right.\right.\right.$$
$$\left.\left.\left. + \beta\sqrt{\zeta_0^-}\mathbf{w}^T\mathbf{\Sigma}_-^{1/2}\mathbf{z}_- + \beta\sqrt{p}\tilde{R}\mathbf{w}^T\bar{\boldsymbol{\mu}}\right\}\nu^k(\mathrm{d}\mathbf{w})\right]^k\right\}. \tag{A22}$$

Similarly, using (A17), we obtain

$$\sum_{ab}(u^a - R\mu)(\mathbf{Q}_\pm^{-1})_{ab}(u^b - R\mu) = \frac{\beta\sum_a(u^a - R\mu)^2}{q^\pm} - \frac{\beta^2 q_0^\pm\{\sum_a(u^a - R\mu)\}^2}{(q^\pm)^2},$$

$$\log\det\mathbf{Q}_\pm = \log\left[(q_1^\pm - q_0^\pm)^k\left(1 + \frac{kq_0^\pm}{q_1^\pm - q_0^\pm}\right)\right] = \frac{k\beta q_0^\pm}{q^\pm},$$

where we retain only the leading order terms. Therefore, (A20) becomes

$$\hat{I}_\pm = \exp\left(-\frac{k\beta q_0^\pm}{2q^\pm}\right)\int Dz_\pm\left(\int\exp\left\{-\beta V(u \pm w_0) - \frac{\beta(u - R\mu - \sqrt{q_0^\pm}z_\pm)^2}{2q^\pm}\right.\right.$$
$$\left.\left. + \frac{\beta q_0^\pm z_\pm^2}{2q^\pm}\right\}\mathrm{d}u\right)^k,$$

where the expectation $D_z = \int\frac{\mathrm{d}z}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$. Substituting this expression in (A15), we obtain

$$\hat{\xi}(\mathbf{Q}_\pm, \mathbb{R}, \mathbf{w}_0) = -k\beta E\left\{\alpha_+\min_u\left[V(u + w_0) + \frac{(u - R\mu - \sqrt{q_0^+}z_+)^2}{2q^+}\right]\right.$$
$$\left. + \alpha_-\min_u\left[V(u - w_0) + \frac{(u - R\mu - \sqrt{q_0^-}z_-)^2}{2q^-}\right]\right\}, \tag{A23}$$

where the expectation is with respect to $z_+, z_- \sim N(0, 1)$. Putting (A21)–(A23) together into (A14) and then into (A6), we obtain

J. Stat. Mech. (2021) 113401

$$\mathcal{F} = \frac{1}{2}(\zeta_0^+ q^+ - \zeta^+ q_0^+) + \frac{1}{2}(\zeta_0^- q^- - \zeta^- q_0^-) + R\tilde{R}$$
$$+ \alpha_+ \underset{u\in R}{E\min} \left\{ V(u+w_0) + \frac{\left(u - R\mu - \sqrt{q_0^+} z_+\right)^2}{2q^+} \right\}$$
$$+ \alpha_- \underset{u\in R}{E\min} \left\{ V(u-w_0) + \frac{\left(u - R\mu - \sqrt{q_0^-} z_-\right)^2}{2q^-} \right\}$$
$$+ \frac{1}{p} \underset{\mathbf{w}\in R^p}{E\min} \left\{ \frac{\zeta^+}{2}\|\mathbf{w}\|^2_{\mathbf{\Sigma}_+} + \frac{\zeta^-}{2}\|\mathbf{w}\|^2_{\mathbf{\Sigma}_-} - \left\langle \sqrt{\zeta_0^+}\mathbf{\Sigma}_+^{1/2}\mathbf{z}_+ \right.\right.$$
$$\left.\left. + \sqrt{\zeta_0^-}\mathbf{\Sigma}_-^{1/2}\mathbf{z}_- + \sqrt{p}\tilde{R}\bar{\boldsymbol{\mu}}, \mathbf{w} \right\rangle + \sum_{j=1}^{p} J_\lambda(w_j) \right\}, \tag{A24}$$

where the expectations are with respect to $z_+, z_- \sim N(0,1)$, and $\mathbf{z}_+, \mathbf{z}_- \sim N(0, \mathbf{I}_{p\times p})$, with z_+, z_- and $\mathbf{z}_+, \mathbf{z}_-$ independent from each other. Here $\zeta^\pm, \zeta_0^\pm, q^\pm, q_0^\pm, R, \tilde{R}$ are order parameters that can be determined from the saddle point equations of $\mathcal{F}$. Define the functions $F_\pm$, $G_\pm$, and $H_\pm$ as

$$F_\pm = E_z\left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z_\pm\right),$$

$$G_\pm = E_z\left\{\left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z_\pm\right) z\right\},$$

$$H_\pm = E_z\left\{\left(\hat{u}_\pm - R\mu \mp w_0 - \sqrt{q_0^\pm} z_\pm\right)^2\right\},$$

where

$$\hat{u}_\pm = \mathrm{argmin}_{u\in R}\left\{V(u \pm w_0) + \frac{(u - R\mu \mp w_0 - \sqrt{q_0^\pm} z_\pm)^2}{2q^\pm}\right\}.$$

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

$$\hat{\mathbf{w}} = (\xi^+\mathbf{\Sigma}_+ + \xi^-\mathbf{\Sigma}_- + \lambda\mathbf{I}_p)^{-1}\left(\sqrt{\xi_0^+}\mathbf{\Sigma}_+^{1/2}\mathbf{z}_+ + \sqrt{\xi_0^-}\mathbf{\Sigma}_-^{1/2}\mathbf{z}_- + \sqrt{p}\tilde{R}\bar{\boldsymbol{\mu}}\right). \tag{A25}$$

J. Stat. Mech. (2021) 113401

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

$$\xi_0^{\pm} = \frac{\alpha_{\pm}}{(q^{\pm})^2} H_{\pm}, \tag{A26}$$

$$\xi^{\pm} = \frac{\alpha_{\pm} G_{\pm}}{\sqrt{q_0^{\mp} q^{\pm}}}, \tag{A27}$$

$$q_0^{\pm} = \frac{1}{p} E_z \|\mathbf{w}\|_{\mathbf{\Sigma}_{\pm}}^2, \tag{A28}$$

$$q^{\pm} = \frac{1}{p\sqrt{\zeta_0^{\pm}}} E \left\langle \mathbf{\Sigma}_{\pm}^{1/2} \mathbf{z}_{\pm}, \hat{\mathbf{w}} \right\rangle \tag{A29}$$

$$R = \frac{1}{\sqrt{p}} E_z \langle \bar{\boldsymbol{\mu}}, \hat{\mathbf{w}} \rangle, \tag{A30}$$

$$\tilde{R} = \frac{\alpha_+ \mu}{q^+} F_+ + \frac{\alpha_- \mu}{q^-} F_-, \tag{A31}$$

$$\frac{\alpha_+}{q^+} F_+ = \frac{\alpha_-}{q^-} F_-. \tag{A32}$$

The above formulas are for general positive definite covariance matrix $\mathbf{\Sigma}_{\pm}$. Then after applying the spiked population assumption (2) and integrating over $\mathbf{z}_{\pm}$, we obtain the explicit nonlinear equations for determining six parameters $q_0^{\pm}, q^{\pm}, R, w_0$ as

$$q_0^{\pm} = \alpha_+ H_+ + \alpha_- H_- + \left(\frac{\alpha_{\pm}\mu}{\sigma_{\pm}} F_{\pm} + \frac{\alpha_{\mp}\mu\sigma_{\pm}}{\sigma_{\mp}^2} F_{\mp} \right) \sum_{k=1}^{2K+1} \frac{(1+\lambda_k^{\pm}) R_k^2}{\left\{ 1 - \frac{\alpha_+ \lambda_k^+ G_+}{\sqrt{q_0^+}} - \frac{\alpha_- \lambda_k^- G_-}{\sqrt{q_0^-}} \right\}^2},$$

$$R = \left(\frac{\alpha_+ \mu}{\sigma_+} F_+ + \frac{\alpha_- \mu \sigma_+}{\sigma_-^2} F_- \right) \sum_{k=1}^{K+1} \frac{R_k^2}{\left\{ 1 - \frac{\alpha_+ \lambda_k^+ G_+}{\sqrt{q_0^+}} - \frac{\alpha_- \lambda_k^- G_-}{\sqrt{q_0^-}} \right\}},$$

$$\frac{\alpha_+}{q^+} F_+ = \frac{\alpha_-}{q^-} F_-,$$

$$\frac{q^+}{\sigma_+^2} = \frac{q^-}{\sigma_-^2},$$

$$\frac{q^+ \lambda}{\sigma_+^2} = 1 + \alpha_+ G_+ + \alpha_- G_-.$$

Then, the other five parameters $\zeta_0^{\pm}$, $\zeta^{\pm}$, and $\tilde{R}$ can be obtained using equations (A26), (A27), and (A31).

A1. Derivation of claim 4

Let $\lambda = 0$, from equations (15)–(17) in claim 3, we obtain

J. Stat. Mech. (2021) 113401

$$q_0 - \frac{R^2}{\gamma^2} = \alpha E\{(\hat{u} - a)^2\},$$

$$\frac{R}{\gamma^2} = \alpha\mu E(\hat{u} - a),$$

$$1 = -\frac{\alpha}{\sqrt{q_0}} E\{(\hat{u} - a)z\},$$

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

$$1 - \frac{r^2}{\gamma^2} = \alpha \left\{ \int_{z_c - x}^{z_c} (z_c - z)^2 Dz + x^2 \int_{-\infty}^{z_c - x} Dz \right\} \tag{A33}$$

$$\frac{r}{\gamma^2} = \alpha\mu \left\{ \int_{z_c - x}^{z_c} (z_c - z) Dz + x \int_{-\infty}^{z_c - x} Dz \right\} \tag{A34}$$

$$1 = \alpha \int_{z_c - x}^{z_c} Dz. \tag{A35}$$

(A33) and (A34) lead to

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

For fixed α, μ has upper bound in order for the above equation to have a solution. Because of (A35), the biggest value for μ we can achieve is when $x \to \infty$. Therefore, the phase transition is determined by

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

where $\Phi(z_c) = 1/\alpha$. Note that $\kappa = 1/\alpha$, substituting the spike covariance matrix (12) and (20), we obtain (21). ■

References

Bai Z and Yao J 2012 On sample eigenvalues in a generalized spiked population model *J. Multivariate Anal.* **106** 167–77

Baik J and Silverstein J W 2006 Eigenvalues of large sample covariance matrices of spiked population models *J. Multivariate Anal.* **97** 1382–408

Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 Optimal errors and phase transitions in high-dimensional generalized linear models *Proc. Natl Acad. Sci. USA* **116** 5451–60

Bayati M and Montanari A 2011 The LASSO risk for Gaussian matrices *IEEE Trans. Inf. Theory* **58** 1997–2017

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

Berthier R, Montanari A and Nguyen P-M 2020 State evolution for approximate message passing with non-separable functions *Inf. Inference* **9** 33–79

J. Stat. Mech. (2021) 113401

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
Celentano M, Montanari A and Wei Y 2020 The LASSO with general Gaussian designs with applications to hypothesis testing (arXiv:2007.13716)
Deng Z, Kammoun A and Thrampoulidis C 2019 A model of double descent for high-dimensional binary linear classification *Inf. Inference* iaab002
Dobriban E and Wager S 2018 High-dimensional asymptotics of prediction: Ridge regression and classification *Ann. Stat.* **46** 247–79
Guo D and Verdu S 2005 Randomly spread CDMA: asymptotics via statistical physics *IEEE Trans. Inf. Theory* **51** 1983–2010
Fernández-Delgado M, Cernadas E, Barro S and Amorim D 2014 Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–81
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 *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 *Int. Conf. Machine Learning* pp 3452–62
Gerbelot C, Abbara A and Krzakala F 2020 Asymptotic errors for teacher-student convex generalized linear models (or: how to prove Kabashima's replica formula) (arXiv:2006.06581)
Hastie T, Buja A and Tibshirani R 1995 Penalized discriminant analysis *Ann. Stat.* **23** 73–102
Hastie T, Montanari A, Rosset S and Tibshirani R J 2019 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)
Huang H 2017 Asymptotic behavior of support vector machine for spiked population model *J. Mach. Learn. Res.* **18** 1–45
Huang H 2018 Asymptotic behavior of margin-based classification methods *2018 IEEE Statistical Signal Processing Workshop (SSP)* pp 463–7
Huang H, Liu Y, Yuan M and Marron J S 2015 Statistical significance of clustering using soft thresholding *J. Comput. Graph. Stat.* **24** 975–93
Johnstone I M 2001 On the distribution of the largest eigenvalue in principal components analysis *Ann. Stat.* **29** 295–327
Laloux L, Cizeau P, Potters M and Bouchaud J-P 2000 Random matrix theory and financial correlations *Int. J. Theor. Appl. Finance* **3** 391–7
Lam X Y, Marron J S, Sun D and Toh K-C 2018 Fast algorithms for large-scale generalized distance weighted discrimination *J. Comput. Graph. Stat.* **27** 368–79
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 X, Parker J, Fan C, Perou C M and Marron J S 2009 *Visualization of Cross-Platform Microarray Normalization* (New York: Wiley) ch 14 pp 167–81
Liu Y, Zhang H H and Wu Y 2011 Soft or hard classification? Large margin unified machines *J. Am. Stat. Assoc.* **106** 166–77
Loureiro B, Gerbelot C, Cui H, Goldt S, Krzakala F, Mézard M and Zdeborová L 2021a Learning curves of generic features maps for realistic datasets with a teacher-student model (arXiv:2102.08127)
Loureiro B, Sicuro G, Gerbelot C, Pacco A, Krzakala F and Zdeborová L 2021b Learning Gaussian mixtures with generalised linear models: precise asymptotics in high-dimensions (arXiv:2106.03791)
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 and Liao Z 2019 High dimensional classification via empirical risk minimization: improvements and optimality (arXiv:1905.13742)
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. Acoustics, Speech and Signal Processing (ICASSP)* pp 3357–61
Marcenko V A and Pastur L A 1967 Distribution of eigenvalues for some sets of random matrices *Math. USSR Sbornik* **1** 457–83
Marron J S, Todd M J and Ahn J 2007 Distance-weighted discrimination *J. Am. Stat. Assoc.* **102** 1267–71
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* (Singapore: World Scientific)
Mignacco F, Krzakala F, Lu Y, Urbani P and Zdeborova L 2020 The role of regularization in classification of high-dimensional noisy Gaussian mixture *Int. Conf. Machine Learning* pp 6874–83

J. Stat. Mech. (2021) 113401

Montanari A, Ruan F, Sohn Y and Yan J 2019 The generalization error of maxmargin linear classifiers: high-dimensional asymptotics in the overparametrized regime (arXiv:1911.01544)
Qiao X, Zhang H H, Liu Y, Todd M J and Marron J S 2010 Asymptotic properties of distance-weighted discrimination *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
Sear R P and Cuesta J A 2003 Instabilities in complex mixtures with a large number of components *Phys. Rev. Lett.* **91** 245701
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 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
TCGA 2010 The Cancer Genome Atlas Research Network (http://cancergenome.nih.gov/wwd/pilot_program/research_network/cgcc.asp)
Telatar E 1999 Capacity of multi-antenna Gaussian channels *Eur. Trans. Telecommun.* **10** 585–95
Vapnik V N 1995 *The Nature of Statistical Learning Theory* (New York: Springer)
Wang B and Zou H 2016 Sparse distance weighted discrimination *J. Comput. Graph. Stat.* **25** 826–38
Wang B and Zou H 2018 Another look at distance-weighted discrimination *J. R. Stat. Soc.* B **80** 177–98
Wang K and Thrampoulidis C 2021 Binary classification of Gaussian mixtures: abundance of support vectors, benign overfitting and regularization (arXiv:2011.09148)
Wu Y and Verdu S 2012 Optimal phase transitions in compressed sensing *IEEE Trans. Inf. Theory* **58** 6241–63
Zhu J and Hastie T 2005 Kernel logistic regression and the import vector machine *J. Comput. Graph. Stat.* **14** 185–205

J. Stat. Mech. (2021) 113401