University of California
Santa Barbara

The Importance of Eigenstructure in High-dimensional Statistics: Examples from Overparameterized Machine Learning and Graphical Models

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Ke Wang

Committee in charge:

Professor Alexander Franks, Co-Chair
Professor Sang-Yun Oh, Co-Chair
Professor Christos Thrampoulidis
Professor Yu-Xiang Wang

September 2022

The Dissertation of Ke Wang is approved.

Professor Christos Thrampoulidis

Professor Yu-Xiang Wang

Professor Alexander Franks, Committee Co-Chair

Professor Sang-Yun Oh, Committee Co-Chair

August 2022

The Importance of Eigenstructure in High-dimensional Statistics: Examples from Overparameterized Machine Learning and Graphical Models

by

Ke Wang

Acknowledgements

I would like to express my gratitude to all my PhD committee members: Prof. Alexander Franks, Prof. Sang-Yun Oh, Prof. Christos Thrampoulidis and Prof. Yu-Xiang Wang. I am very fortunate to work with Prof. Alexander Franks and Prof. Sang-Yun Oh on the graphical model project and work with Prof. Christos Thrampoulidis on the overparameterized machine learning project. Prof. Yu-Xiang Wang provided me with very helpful suggestions for these two projects.

I would extend my gratitude to all the collaborators, Dr. Vidya Muthukumar and Tina Behnia, for all the time that we spent thinking and learning together.

I would like to thank Prof. Andrew Carter and Prof. Tomoyuki Ichiba for their great support when they work as graduate student advisors.

I am grateful to all the support provided by the department staff, including Myranda Flores, Jamie Pillsbury-Fischler and Patrick Windmiller.

Many thanks go to all my friends and labmates for their help in my PhD program. I am particularly thankful to Joshua Bang, Megan Elcheikhali, Ganesh Kini, Alice Lepissier, Franky Meng, Orestis Paraskevas, Zachary Terner, Yuanbo Wang, Javier Zapata, Jiajing Zheng, Yi Zheng, Zhipu Zhou.

Curriculum Vitæ

Ke Wang

Education

2022	Ph.D. in Statistics, University of California, Santa Barbara.
2021	M.A. in Statistics, University of California, Santa Barbara.
2016	Master's degree in Applied Statistics, Cornell University.
2015	BSc in Actuarial Science, University of Kent.
2015	Bachelor in Economics in Actuarial Science and Risk Management, University of International Business and Economics.

Publications

- Ke Wang, Vidya Muthukumar and Christos Thrampoulidis. "Benign overfitting in multiclass classification: all roads lead to interpolation". NeurIPS 2021.
- Ke Wang and Christos Thrampoulidis. "Binary classification of Gaussian mixtures: abundance of support vectors, benign overfitting and regularization". SIAM Journal on Mathematics of Data Science, short version published at IEEE ICASSP 2021.
- Tina Behnia, Ke Wang and Christos Thrampoulidis. "On how to avoid exacerbating spurious correlations when models are overparameterized". 2022 IEEE International Symposium on Information Theory.
- Ke Wang, Alexander Franks and Sang-Yun Oh. "Learning Gaussian Graphical Models with Latent Confounders". Under rivision by Journal of Multivariate Analysis.

Abstract

The Importance of Eigenstructure in High-dimensional Statistics: Examples from Overparameterized Machine Learning and Graphical Models

by

Ke Wang

Modern data sets are large and complicated. The demand for understanding the nature of such big data has motivated the development of high-dimensional statistics. Understanding covariance matrices and their eigenstructure plays a central role in high-dimensional statistics. In this thesis, we examine the critical role of the eigenstructure of data covariance matrices in two popular high-dimensional problems. Understanding the covariance matrices and their eigenstructure is essential for both problems. First, we focus on overparameterized machine learning. This line of research is motivated by the empirical observation that deep neural networks can generalize well despite learning a number of parameters that far exceeds the size of the training set. Our work contributes to this field by proving that analogous behaviors can be observed in simpler, binary and multiclass linear classification models. We show the equivalence between support-vector machines (SVM) and the interpolating classifiers. We derive generalization bounds and characterize the role of regularization in the overparameterized regime. Our bounds reveal that the feature covariance matrix plays a central role in guaranteeing good generalization under overparameterization. Specifically, our analysis is the first to demonstrate this for multiclass rather than binary classification. The second topic is Gaussian graphical models (GGM). GGM are widely used to estimate network structures in several applications. Specifically, estimating graph structures accurately is challenging when latent confounders exist. In this work, we theoretically compare two commonly used methods

that can remove latent confounders when estimating GGM. The theory depends heavily on the analysis on feature covariance matrices and their inverse. Our results reveal that the eigenstructure of feature covariance matrices is crucial to determine the performance of different methods. Based on the theory, we propose a new method that combines the strengths of previous approaches. We demonstrate the effectiveness of our methodology with simulations in two real-world applications.

Contents

Chapter 1

Introduction

Many of the data sets arising in modern applications are very large, often with the feature dimension of the same order, or even larger than the sample size. The demand for understanding the nature of big data inspires the development of high-dimensional statistics. With techniques in high-dimensional statistics, we can theoretically characterize the performances of different machine learning and statistical estimation methods.

In high-dimensional statistics, analyzing covariance matrices is a very important and challenging topic. Covariance is the second-moment and characterizes linear relationships between the features. Specifically, let n be the number of observations, p be the number of features and matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the data matrix of features, the sample covariance matrix is defined as $(1/n)\boldsymbol{X}^T\boldsymbol{X}$. The terms $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{X}\boldsymbol{X}^T$ and their inverse appear in many machine learning and statistics problems. However, analyzing these terms can be challenging. The first challenge is due to the high dimensionality. When $p > n$, the sample covariance might not be a consistent estimator for the population covariance. More involved analysis is needed to understand the behavior of covariance matrices under high-dimensionality [163, Chapters 1 and 6]. The second challenge comes from the data distributions. In many analyses, the underlying distribution of data is

assumed to be multivariate Gaussian with zero mean. In many applications, however, it is more appropriate to use more complicated distributions and the data matrix $\boldsymbol{X}$ thus has a more complex form. For instance, in binary classification under Gaussian mixture models (GMM), the data distribution has a non-zero mean and thus the feature matrix $\boldsymbol{X}$ becomes

$$\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q}, \tag{1.1}$$

where $\boldsymbol{\eta}$ is the mean vector, $\boldsymbol{y}$ is the label vector and the noise matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times p}$ has independent zero-mean Gaussian rows. In multiclass classification under GMM with k classes, the feature matrix $\boldsymbol{X}$ includes more components:

$$\boldsymbol{X} = \sum_{j=1}^{k} \boldsymbol{v}_j\boldsymbol{\mu}_j^T + \boldsymbol{Q}, \tag{1.2}$$

where $\boldsymbol{\mu}_j$'s are mean vectors, $\boldsymbol{v}_j$ are label vectors and again the noise matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times p}$ has independent independent zero-mean Gaussian rows. In estimation Gaussian graphical models, if latent confounders exist, then one way to write the observed data matrix is

$$\boldsymbol{X} = \boldsymbol{X}_{true} + \boldsymbol{A}\boldsymbol{Z}, \tag{1.3}$$

where $\boldsymbol{X}_{true}$ is the uncorrupted data matrix and $\boldsymbol{A}\boldsymbol{Z}$ reflects the effect of latent confounding. Characterizing covariance matrices in these examples is challenging due to the extra components in data matrices, e.g., the mean vectors in (1.1) and (1.2) and the latent confounders in (1.3).

In this work, we focus on two important machine learning problems. Understanding the covariance matrices of feature matrix $\boldsymbol{X}$ and their eigenstructure is essential to both.

The first topic is overparameterized machine learning. The motivation of this line of research is to understand some surprising phenomenon observed in highly overparameterized problems, i.e., deep neural networks can still generalize well despite being highly overparameterized and being trained without explicit regularization [59, 169]. This curious phenomenon has inspired extensive research activity in establishing its statistical principles: Under what conditions is it observed? How do these depend on the data and on the training algorithm? When does regularization benefit generalization? While such questions remain wide open for deep neural nets, recent works have attempted gaining insights by studying simpler, often linear, models. We contribute to this growing line of work by examining linear classification. We study both binary and multiclass classification under Gaussian mixture models. The data matrix follows (1.1) in binary classification and follows (1.2) for the multiclass case. We will see in Chapters 2 and 3 that upper/lower bounding quadratic forms involving $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ plays central roles for these problems. One of our main contributions is to propose new methods that can effectively bound quadratic forms involving $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ even if the data matrix $\boldsymbol{X}$ includes mean and label vectors. Our results reveal that the bounds and conditions in these sections depend heavily on the eigenstructure of the covariance matrices of feature matrix $\boldsymbol{X}$.

The second topic is estimating Gaussian graphical models. Gaussian Graphical models are widely used to estimate network structure in domains ranging from biology to finance and one way to estimate the graph is to learn the inverse covariance matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ [92]. In practice, data is often corrupted by latent confounders which biases inference of the underlying true graphical structure. When latent confounders exist, the data matrix becomes (1.3). In our work, we theoretically understand $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ when $\boldsymbol{X}$ follows (1.3). Again, our theory shows the central role of eigenstructure of the covariance matrices of the data matrix. Based on our findings, we propose new methods that can

effectively remove the latent confounders. More details are introduced in Chapter 4.

We now provide a brief summary for each chapter.

1. Chapter 2.

   We study binary classification in the overparameterized regime under a generative Gaussian mixture model in which the feature vectors take the form $\boldsymbol{x} = \pm\boldsymbol{\eta} + \boldsymbol{q}$, where for a mean vector $\boldsymbol{\eta}$ and feature noise $\boldsymbol{q} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Motivated by recent results on the implicit bias of gradient descent, we study both max-margin SVM classifiers (corresponding to logistic loss) and min-norm interpolating classifiers (corresponding to least-squares loss). First, we leverage an idea introduced in [120] to relate the SVM solution to the min-norm interpolating solution. Second, we derive novel non-asymptotic bounds on the classification error of the latter. Combining the two, we present novel sufficient conditions on the covariance spectrum and on the signal-to-noise ratio (SNR) $SNR = \frac{\|\boldsymbol{\eta}\|_2^4}{\boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta}}$ under which interpolating estimators achieve asymptotically optimal performance as overparameterization increases. Interestingly, our results extend to a noisy model with constant probability noise flips. Contrary to previously studied discriminative data models, our results emphasize the crucial role of the SNR and its interplay with the data covariance. Finally, via a combination of analytical arguments and numerical demonstrations we identify conditions under which the interpolating estimator performs better than corresponding regularized estimates.

2. Chapter 3.

   Many modern machine learning applications operates in the multiclass setting. Motivated by this, we study benign overfitting in multiclass linear classification. Specifically, we consider the following training algorithms on separable data: (i) empirical risk minimization (ERM) with cross-entropy loss, which converges to the

multiclass support vector machine (SVM) solution; (ii) ERM with least-squares loss, which converges to the min-norm interpolating (MNI) solution; and, (iii) the one-vs-all SVM classifier. First, we provide a simple sufficient deterministic condition under which *all* three algorithms lead to classifiers that interpolate the training data and have equal accuracy. When the data is generated from Gaussian mixtures, this condition holds under high enough effective overparameterization. We also show that this sufficient condition is satisfied under "neural collapse", a phenomenon that is observed in training deep neural networks. Second, we derive novel bounds on the accuracy of the MNI classifier, thereby showing that all three training algorithms lead to benign overfitting under sufficient overparameterization. Ultimately, our analysis shows that good generalization is possible for SVM solutions beyond the realm in which typical margin-based bounds apply.

3. Chapter 4.

   In Chapter 4, we focus on estimating Gaussian graphical models (GMM) with latent confounders. We compare and contrast two strategies for inference in graphical models with latent confounders: Gaussian graphical models with latent variables (LVGGM) and PCA-based removal of confounding (PCA+GGM). While these two approaches have similar goals, they are motivated by different assumptions about confounding. In this work, we explore the connection between these two approaches and propose a new method, which combines the strengths of these two approaches. We prove the consistency and convergence rate for the PCA-based method and use these results to provide guidance about when to use each method. We demonstrate the effectiveness of our methodology using both simulations and two real-world applications.

Chapter 2

Benign overfitting in binary classification

2.1 Introduction

2.1.1 Motivation

Deep-learning models are increasingly more complex. They are designed with a huge number of parameters that far exceed the size of typical training data sets and training is often completed without any explicit regularization [89, 117, 132, 59]. As a consequence, after training, the models perfectly fit (or, so called interpolate) the data. Classical statistical wisdom suggests that such interpolating models overfit and as such they generalize poorly, e.g. [64]. But, the reality of modern deep-learning practice is very different: such overparameterized learning architectures achieve state-of-the-art generalization performance despite interpolating the data [169, 11, 121]. Interestingly, similar empirical findings, albeit in much simpler learning settings have been recorded in the literature even before the era of deep learning [158, 122, 43]; see discussion in [104].

Empirical observations like these raise a series of important questions [43, 169, 11, 13]: *Why and when are larger models better? What is the role of the training algorithm in this process? Can infinite overparameterization result in better generalization than any finite number of parameters or even training with explicit regularization?* Answering these questions is considered one of the main challenges in modern learning theory and has attracted significant research attention over the past couple of years or so, e.g., [11, 12, 110, 65, 100, 101, 39, 6, 119, 167, 32].

Among the earliest attempts towards analytically investigating the question "why do overparameterized models generalize well?" focused on linear-regression including both asymptotic and non-asymptotic analyses [65, 15, 118, 157]. While certainly a simplified model, this is a natural first step towards gaining insights about more complex models. Closest to our work, [10] derived non-asymptotic bounds on the squared prediction risk of the min-norm linear interpolator for a linear regression model with additive Gaussian noise and (sub)-Gaussian covariates. They subsequently used these bounds and identified conditions on the spectrum of the data covariance such that the risk asymptotically approaches the optimal Bayes error despite perfectly fitting to noisy data. This behavior was termed "*benign overfitting*" in their paper and the terminology has already been widely adapted in the literature.

A step further in the direction of understanding generalization in overparameterized regimes is the study of linear classification models, since arguably most deep learning success stories apply to classification settings. Classification is not only more relevant, but also typically harder to analyze. The challenge is that even in linear settings, the solution to logistic loss minimization is *not* given in closed form. This is to be contrasted to the solution to least-squares minimization typically used in regression (e.g. [10, 65]). As such, central questions have remained largely unexplored until very recently.

[149, 142, 116, 39, 84, 114, 80, 143] study overparaterized binary linear classification

in the proportional asymptotic regme, where the size n of the training set and the size p of the parameter vector grow large at a fixed rate. These works overcome the aforementioned challenge by relying on powerful tools from modern high-dimensional statistics [148, 153, 154] and yield asymptotic error predictions that are sharp, but remain limited to the proportional regime and are expressed in terms of complicated—and often hard to interpret and evaluate—systems of nonlinear equations.

A different approach, resulting in more general non-asymptotic, albeit non-sharp, bounds was initiated by [119] who studied a 'Signed' classification model with Gaussian features. Their key observation, that drives their analysis, is that the max-margin classifier linearly interpolates the data given sufficient overparameterization. This allowed the authors to establish a tight link between the (hard to directly analyze) SVM and the (amenable to analysis) LS solutions. In turn, this resulted in identifying sufficient conditions on the covariance spectrum needed for benign overfitting. While this paper was being prepared, a follow-up work [72] has extended their analysis to binary classification under generalized linear models (including the 'Signed' model as a special case) and to subGaussian/Haar-distributed features. Motivated by these works, we investigate the following related open questions: *Does the max-margin classifier interpolate data that are generated from generative (rather than discriminative) models? If so, under what conditions? How do optimally tuned regularized estimators compare to interpolating classifiers? Are there settings in which the latter perform better? How does label noise affect any interpolating properties of the max-margin classifier? What does this imply for benign overfitting?*

2.1.2 Contributions and novelty

We answer the questions above by focusing on the popular Gaussian mixture model (GMM). Unlike discriminative classification models, the GMM specifies the feature conditional distribution $\boldsymbol{x}|y$, setting it to be a multivariate Gaussian that is centered around a mean vector $y\boldsymbol{\eta}$ (of their respective class $y = \pm 1$) and has covariance matrix $\boldsymbol{\Sigma}$ (Section 2.2 for details). We outline our contributions below and then highlight the novelties compared to prior work.

(i) Abundance of support vectors (Section 2.3): We show for the first time that the max-margin classifier *linearly interpolates* GMM data given sufficient overparameterization. Notably, our analytic sufficient conditions for this to happen involve not only the covariance spectrum, but also the problem's signal-to-noise-ratio (SNR), which we define as $SNR = \|\boldsymbol{\eta}\|_2^4 / \boldsymbol{\eta}^T \boldsymbol{\Sigma} \boldsymbol{\eta}$. Thus, we uncover a key difference compared to discriminative data (e.g. Signed model [119, 72]). We complement our sufficient conditions with numerical results that suggest their tightness.

(ii) Non-asymptotic bounds for min-norm estimators (Section 2.4): We derive novel *non-asymptotic* error bounds for the min-norm linear interpolator. Our bounds explicitly capture the effect of the overparameterization ratio, of the covariance spectrum and of the SNR.

(iii) Interpolators' risk under high overparameterization (Section 2.5): Combining our findings above, we derive sufficient conditions on the spectrum of $\boldsymbol{\Sigma}$ and on the SNR that guarantee both the SVM and the LS solutions (a) perfectly interpolate the data, and, (b) achieve asymptotically optimal risk as overparameterization increases. Our conditions improve upon the state of the art [31] in the noiseless case (see discussion below).

(iv) The effect of regularization (Section 2.6): We study the effect of ridge-regularization on the risk. Interestingly, we identify regimes that the interpolating estimator (corre-

sponding to zero regularization) outperforms regularized estimates in the overparameterized regime.

(v) Interpolation and benign overfitting in noisy models (Section 2.7): We extend our findings to a *noisy* isotropic Gaussian mixture model, where labels are corrupted with constant probability. First, we find that the favorable interpolating property of SVM continues to hold, but under stronger conditions due to the label corruptions. Second, in the regime of interpolation, we upper bound the risk of the minimum-norm interpolator and use this result to identify regimes of benign overfitting, i.e. regimes where the SVM risk asymptotically approaches the Bayes risk despite perfectly fitting the data.

On the technical front, while our analysis uses tools similar to those in [10, 119], there are key differences in the GMM, which further complicate the analysis and impose new challenges. This can be illustrated at a high-level as described below (see also Section 2.8). We will show that at the heart of our analysis lies the challenge of upper/lower-bounding quadratic forms such as $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{y}$, where $\boldsymbol{y}$ is the label vector and $\boldsymbol{X}$ is the feature matrix of the training set. Under the GMM, and unlike in linear regression and discriminative classification models, the matrix $\boldsymbol{X}$ "includes" both the label vector $\boldsymbol{y}$ and the mean vector $\boldsymbol{\eta}$. Hence, considering $\boldsymbol{y}$ and $\boldsymbol{X}$ separately as in [10, 119] leads to sub-optimal bounds. Instead, we first show that it is possible to decompose the original quadratic form of interest into several more primitive quadratic forms on inverse-Wishart matrices (rather than on the original Gram matrix). This decomposition is central to our proof technique, but the technical challenge remains because: (a) the decomposition involves the new quadratic forms in a convoluted way requiring us to establish both lower and upper bounds for each one of them and then combine them carefully, and, (b) while more primitive, the desired bounds for the new quadratic forms do *not* follow from previous works. Besides, as mentioned above, a particular distinguishing feature of GMM

compared to previous works is that in the process of doing the above we need to carefully capture the impact of not only the covariance spectrum, but also of the model's SNR. Compared to previous works, we also complement our analysis with numerical results validating the tightness of our findings. Also, we study the effect of regularization and identify regimes in which interpolating estimators have optimal performance. Compared to [119, 72] we also extend our results to a noisy model with constant probability label corruptions.

The most closely related work in terms of problem setting and results is the recent paper by [31], which thus deserves its own discussion. [31] are the first to derive non-asymptotic risk bounds for overparameterized binary mixture models and use them to characterize benign-overfitting conditions. Notably, their bounds hold for sub-Gaussian features and for an adversarial noisy model that is more general than ours. On the other hand, in the special case of GMM, our results improve upon theirs as follows. In the noiseless case, we significantly relax the conditions under which interpolating estimators asymptotically attain Bayes optimal performance with increasing overparameterization. Also, our risk bounds capture the key role of the covariance structure unlike theirs. In the noisy case, our benign overfitting conditions are the same, but our risk bounds on the min-norm interpolator hold under relaxed scaling assumptions. It is worth mentioning that our proof strategy towards upper bounding the risk of SVM is also entirely different compared to [31]. In comparison to [31], we are also the first to establish interpolating conditions for the SVM solution under GMM data. Finally, our risk bounds also hold for regularized least-squares.

A more elaborate discussion on the above closely related works, as well as, a comparison to classical margin-based bounds is deferred to Section 2.9 due to space limitations. The Appendix includes detailed proofs of all our results.

Notation. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, let $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$, $\|\boldsymbol{v}\|_1 = \sum_{i=1}^p |v_i|$, $\|\boldsymbol{v}\|_{-1} =$

$\sum_{i=2}^{p} |v_i|$, $\|\boldsymbol{v}\|_\infty = \max_i\{|v_i|\}$ and $\boldsymbol{e}_i$ denotes the i-th standard basis vector. For a matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_2$ denotes its operator norm. $[n]$ denotes the set $\{1, 2, ..., n\}$. We also use standard "Big O" notations $\Theta(\cdot)$, $\omega(\cdot)$, e.g., see [34, Chapter 3]. Finally, we write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the (multivariate) Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and, $Q(x) = \mathbb{P}(Z > x)$, $Z \sim \mathcal{N}(0, 1)$ for the Q-function of a standard normal. Throughout, 'constants' refer to numbers that do *not* depend on the problem dimensions n or p.

2.2 Learning model

2.2.1 Data model

Consider the following supervised binary classification problem under a *Gaussian mixtures model* (GMM). Let $\boldsymbol{x} \in \mathbb{R}^p$ denote the feature vector and $y \in \{-1, +1\}$ its class label. The class label y takes one of the values $\{\pm 1\}$ with probabilities $\pi_{\pm 1}$ such that $\pi_{+1} + \pi_{-1} = 1$. The class-conditional probability $p(\boldsymbol{x}|y)$ follows Gaussian distribution. Specifically, conditional on $y = \pm 1$, the feature vector $\boldsymbol{x}$ is a Gaussian vector with mean vector $\pm\boldsymbol{\eta} \in \mathbb{R}^p$ and an invertible covariance matrix $\boldsymbol{\Sigma}$. Summarizing, the data pair $(\boldsymbol{x}, y)$ is generated such that

$$y = \begin{cases} 1, & \text{w.p. } \pi_{+1} \\ -1, & \text{w.p. } 1 - \pi_{+1} \end{cases} \qquad \text{and} \qquad \boldsymbol{x}|y \sim \mathcal{N}(y\boldsymbol{\eta}, \boldsymbol{\Sigma}). \tag{2.1}$$

We denote the eigenvalues of $\boldsymbol{\Sigma}$ by $\boldsymbol{\lambda} := [\lambda_1, \cdots, \lambda_p]$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, and write the eigendecomposition of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements are eigenvalues of $\boldsymbol{\Sigma}$ and the columns of matrix $\boldsymbol{V}$ are eigenvectors of $\boldsymbol{\Sigma}$. Using the eigenvecotrs of $\boldsymbol{\Sigma}$ as a basis, the mean vector $\boldsymbol{\eta}$ can be

expressed as $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$. Note that $\|\boldsymbol{\eta}\|_2 = \|\boldsymbol{\beta}\|_2$.

Consider training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ composed of n IID data pairs generated according to the GMM in (2.1). Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ denote the feature matrix and $\boldsymbol{y} = [y_1, \cdots, y_n]^T$ denote the class-label vector. Following (2.1), the data matrix $\boldsymbol{X}$ can be expressed as follows for a "noise matrix" $\boldsymbol{Q} \in \mathbb{R}^{n\times p}$ with independent $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows,

$$\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q}.$$

Data covariance structure

One of our contributions is demonstrating how the classification performance on data from the GMM depends crucially on the structure of the data covariance. To explicitly capture this dependency, we consider two ensembles for the spectrum of the data covariance $\boldsymbol{\Sigma}$.

Definition 2.2.1 (Balanced ensemble)**.** No eigenvalues of $\boldsymbol{\Sigma}$ are significantly larger than others. Specifically, there exists a constant $b > 1$ such that

$$bn\lambda_1 \leq \|\boldsymbol{\lambda}\|_{-1}, \tag{2.2}$$

where $\|\boldsymbol{\lambda}\|_{-1} = \sum_{i=2}^p \lambda_i$. An example of special interest is the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}$ with sufficient overparameterization, i.e., $p > Cn$, for some constant $C > 1$.

Definition 2.2.2 (Bi-level ensemble)**.** One eigenvalue of $\boldsymbol{\Sigma}$ is much larger than others. Specifically, there exist constants $b_1, b_2 > 1$ such that

$$b_1 n\lambda_1 \geq \|\boldsymbol{\lambda}\|_{-1} \quad \text{and} \quad b_2 n\lambda_2 \leq \sum_{i=3}^p \lambda_i. \tag{2.3}$$

The different nature of the two models leads to different conclusions on how the co-

variance structure affects our key results on abundance of support vectors and benign overfitting. Similar data covariance structures were considered in [119], but for the discriminative model $y_i = \mathrm{Sign}(\boldsymbol{x}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\eta}), \mathrm{i} \in [\mathrm{n}]$ with features $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. The two ensembles above are also related to the notions of effective ranks introduced by [10] in the study of benign overfitting for linear regression (see Section 2.9.2 for details).

Key summary quantities

As mentioned, our results naturally depend on the spectrum of $\boldsymbol{\Sigma}$. Specifically, we will identify $\|\boldsymbol{\lambda}\|_1$ and $\|\boldsymbol{\lambda}\|_2$ as two key relevant summary quantities. But as hinted by (2.1) the data covariance $\boldsymbol{\Sigma}$ is expected to interplay with the mean vector $\boldsymbol{\eta}$ in the results. We will show that this interplay is captured by the *the signal strength in the direction of* $\boldsymbol{\Sigma}$, which we denote

$$\sigma^2 := \|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}}^2 := \boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta} = \boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}.$$

Finally, the signal strength $\|\boldsymbol{\eta}\|_2$ will also be important. Note that the two quantities σ^2 and $\|\boldsymbol{\eta}\|_2$ define a natural notion of signal-to-noise ratio (SNR) for the GMM. To better see this, take inner products of both sides of (2.1) with $\boldsymbol{\eta}$ to express the label-feature relation as $\boldsymbol{x} = y\boldsymbol{\eta} + \boldsymbol{q} \implies y = \frac{\boldsymbol{\eta}^T\boldsymbol{x}}{\|\boldsymbol{\eta}\|_2^2} - \frac{\boldsymbol{\eta}^T\boldsymbol{q}}{\|\boldsymbol{\eta}\|_2^2}$, where $\boldsymbol{q} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Then, following the standard definition in random-design regression and noting that $\frac{\mathrm{Var}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x})}{\mathrm{Var}(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{q})} = \frac{c\|\boldsymbol{\eta}\|_2^4 + \boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta}} = \frac{c\|\boldsymbol{\eta}\|_2^4}{\boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta}} + 1$, for $0 \le c \le 1$ depending on π_{+1}, we let $\mathrm{SNR} := \frac{\|\boldsymbol{\eta}\|_2^4}{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}} = \frac{\|\boldsymbol{\beta}\|_2^4}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\beta}}$; Lemma 1 bounds the classification error in terms of the same quantity, which further validates its role as the SNR.

2.2.2 Training algorithm

Given access to the training set, we train a linear classifier $\hat{\boldsymbol{\eta}}$ by minimizing the empirical risk $\hat{\mathcal{R}}_{\mathrm{emp}}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \ell(y_i \cdot \boldsymbol{w}^T\boldsymbol{x}_i)$, where the loss function ℓ is chosen as: (i)

Least-squares (LS): $\ell(t) = (1-t)^2$, or, (ii) Logistic: $\ell(t) = \log\left(1+e^{-t}\right)$. Throughout, we focus on the *overparameterized* regime $p > n$. As is common, we run gradient descent (GD) on the empirical risk. The following results characterizing the *implicit bias* of GD for the square and logistic losses in the overparameterized regime are well-known. For one, when data can be linearly interpolated (i.e., $\exists \boldsymbol{\beta} \in \mathbb{R}^p$ such that $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta},\ \forall i \in [n]$), then GD on square loss with sufficiently small step size converges (as the number of iterations grow to infinity) to the solution of *min-norm interpolation*, e.g. [65]:

$$\hat{\boldsymbol{\eta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \ \text{ subject to } y_i = \boldsymbol{w}^T \boldsymbol{x}_i, \forall i \in [n]. \tag{2.4}$$

Second, when data are linearly separable (i.e., $\exists \boldsymbol{\beta} \in \mathbb{R}^p$ such that $y_i(\boldsymbol{x}_i^T \boldsymbol{\beta}) \geq 1,\ \forall i \in [n]$), then the normalized iterates of GD on logistic loss converge in direction [1] to the solution of *hard-margin SVM* [146, 77] (see also [138] for earlier similar results):

$$\hat{\boldsymbol{\eta}}_{\mathrm{SVM}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \ \text{ subject to } y_i \boldsymbol{w}^T \boldsymbol{x}_i \geq 1, \forall i \in [n]. \tag{2.5}$$

Now, specializing to data from the GMM, it can be shown that when $p > n+2$, then the data can be linearly interpolated with high probability (whp.). In turn, this easily implies that data are also linearly separable. See Appendix A.9 for a formal statement and proof of these claims. Combining those, in the overparameterized regime, whp., GD on data from the GMM converges to either (2.4) or (2.5) for a square and logistic loss, respectively.

The behavior above holds when no explicit regularization is used. To see the role of

[1] Precisely, convergence is in the sense of the normalized GD iterations $\boldsymbol{\eta}^t$, i.e. $\|\frac{\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}^t\|_2} - \frac{\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}}{\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}^t\|_2}\|_2 \overset{t\to\infty}{\to} 0$.

regularization, we also consider the ridge estimator given by

$$\hat{\boldsymbol{\eta}}_\tau = \arg\min_{\boldsymbol{w}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \tau\|\boldsymbol{w}\|_2^2\} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y}. \tag{2.6}$$

Note that $\hat{\boldsymbol{\eta}}_{\rm LS}$ can be obtained from (2.6) by setting $\tau = 0$ ($\boldsymbol{X}\boldsymbol{X}^T$ is non-singular whp. for $p > n$, e.g., [162]).

Henceforth, we focus on the classifiers in (2.5), (2.4), (2.6). With some abuse of terminology, we often refer to the minimum-norm interpolator in (2.4) as LS solution for brevity.

2.2.3 Classification error

For a new sample $(\boldsymbol{x}, y)$, the classifier $\hat{\boldsymbol{\eta}}$ classifies $\boldsymbol{x}$ as $\hat{y} = \text{sign}(\hat{\boldsymbol{\eta}}^{\rm T}\boldsymbol{x})$. Then, the classification error is measured by the expected 0-1 loss risk

$$\mathcal{R}(\hat{\boldsymbol{\eta}}) = \mathbb{E}[\mathbb{I}(\hat{y} \neq y)] = \mathbb{P}(\hat{\boldsymbol{\eta}}^T(y\boldsymbol{x}) < 0), \tag{2.7}$$

where the expectation is over the distribution of $(\boldsymbol{x}, y)$ generated as in (2.1). The following simple lemma gives an upper bound on $\mathcal{R}(\hat{\boldsymbol{\eta}})$.

Lemma 1. *Under the Gaussian-mixtures model, the classification error of a classifier $\hat{\boldsymbol{\eta}}$ satisfies, $\mathcal{R}(\hat{\boldsymbol{\eta}}) = Q(\frac{\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta}}{\sqrt{\hat{\boldsymbol{\eta}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}}})$. In particular, if $\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta} > 0$, then $\mathcal{R}(\hat{\boldsymbol{\eta}}) \leq \exp\big(-\frac{(\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta})^2}{2\hat{\boldsymbol{\eta}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}}\big)$.*

Proof. For a new draw $\boldsymbol{x}, y$, using $\boldsymbol{x} = y\boldsymbol{\eta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}, \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and symmetry of the Gaussian distribution, it can be easily checked that $\mathcal{R}(\hat{\boldsymbol{\eta}}) = \mathbb{P}(\hat{\boldsymbol{\eta}}^T(y\boldsymbol{q}) < -\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta}) = \mathbb{P}(\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\eta}}^T\boldsymbol{z} > \hat{\boldsymbol{\eta}}^T\boldsymbol{\eta})$. Now, $\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\eta}}^T\boldsymbol{z}$ is a zero-mean Gaussian random variable with variance $\hat{\boldsymbol{\eta}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}$. Thus, the advertised bounds follow directly: the first, by definition of the Q-function, and, the second, by the Chernoff bound for the Q-function, e.g., [163, Ch. 2]. □

Thanks to the lemma above, our goal of upper bounding the classification error, reduces to that of lower bounding the ratio $\frac{(\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta})^2}{2\hat{\boldsymbol{\eta}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}}$. We do this in Section 2.4 for the classifiers $\hat{\boldsymbol{\eta}}_\tau$ and $\hat{\boldsymbol{\eta}}_{\rm LS}$. In large, this is possible because these estimators can be conveniently written in closed forms (see (2.6)). In contrast, the SVM solution *cannot* be expressed in closed form. To get around this challenge, Section 2.3 establishes sufficient conditions under which the SVM-solution $\hat{\boldsymbol{\eta}}_{\rm SVM}$ linearly interpolates the data, thus, it coincides with the LS solution.

2.3 Link between SVM and linear-interpolation

This section establishes a link between the SVM solution in (2.5) and the LS solution in (2.4) for general $\boldsymbol{\Sigma}$. Specifically, Theorem 1 below identifies sufficient conditions under which all training data points become support vectors, i.e., $\hat{\boldsymbol{\eta}}_{\rm SVM}$ linearly interpolates the data: $\boldsymbol{x}_i^T\hat{\boldsymbol{\eta}}_{\rm SVM} = y_i,\ \forall i \in [n]$.

Theorem 1. *Assume n training samples following the GMM defined in Section 2.2. There exist constants $C_1, C_2 > 1$ such that, if the following conditions on the eigenvalues of $\boldsymbol{\Sigma}$ and on the signal strength in the direction of $\boldsymbol{\Sigma}$ defined as $\sigma^2 \triangleq \sum_{i=1}^p \lambda_i\beta_i^2$ hold:*

$$\|\boldsymbol{\lambda}\|_1 > 72\big(\|\boldsymbol{\lambda}\|_2 \cdot n\sqrt{\log n} + \|\boldsymbol{\lambda}\|_\infty \cdot n\sqrt{n}\log n + 1\big), \tag{2.8}$$

$$\|\boldsymbol{\lambda}\|_1 > C_1 n\sqrt{\log(2n)}\,\sigma, \tag{2.9}$$

then, the SVM-solution $\hat{\boldsymbol{\eta}}_{\rm SVM}$ satisfies the linear interpolation constraint in (2.4) *with probability at least* $(1 - \frac{C_2}{n})$.

For the isotropic case, condition (2.8) can be sharpened as shown in the following theorem.

Theorem 2. *Assume n training samples following the GMM with $\boldsymbol{\Sigma} = \mathbf{I}$. There exist constants $C_1, C_2 > 1$ such that, if the following conditions on the number of features p and the mean-vector $\boldsymbol{\eta}$ hold:*

$$p > 10n \log n + n - 1 \quad \textit{and} \quad p > C_1 n \sqrt{\log(2n)} \|\boldsymbol{\eta}\|_2, \tag{2.10}$$

then, the SVM-solution $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ satisfies the linear interpolation constraint in (2.4) *with probability at least* $(1 - \frac{C_2}{n})$.

The theorems establish two sufficient conditions each for all training samples to become support vectors. In the isotropic setting, the first condition requires that the number of features p is significantly larger than the number of observations n. For the anisotropic case, the corresponding condition is related the the effective ranks r_0 and R_0 [10, 119], i.e. $r_k := (\sum_{i>k}^{p} \lambda_i)/\lambda_{k+1}$ and $R_k = (\sum_{i>k}^{p} \lambda_i)^2/(\sum_{i>k}^{p} \lambda_i^2)$. The condition requires that the covariance spectrum has sufficiently slowly decaying eigenvalues (corresponding to sufficiently large R_0), and that it is not too "spiky" (corresponding to sufficiently large r_0). [119, Remark 4] provides a detailed discussion on how the effective ranks relate to different spectrum regimes. Specifically, the bi-level ensemble (Definition 2.2.2) does *not* satisfy (2.8). To see that, (2.8) implies $\|\boldsymbol{\lambda}\|_1 > 72n\sqrt{n}(\log n)\lambda_1$, meaning that $n\lambda_1$ should not be large compared to the sum of other eigenvalues. In contrast, the bi-level ensemble requires $b_1 n\lambda_1 > \|\boldsymbol{\lambda}\|_{-1}$. The second conditions in the two theorems above are the same to each other, since $\sigma = \|\boldsymbol{\eta}\|_2$ in the isotropic setting. These latter conditions relate to the SNR and constrain the signal strength in the direction of $\boldsymbol{\Sigma}$.

To better interpret the result of the two theorems we show corresponding numerical results in Figure 2.1. As explained, the figure also confirms the tightness of our theoretical prediction. In all our simulations throughout this paper, we fix $\pi_+ = 0.5$ and plot averages over 300 Monte-Carlo realizations. For simplicity, we choose diagonal $\boldsymbol{\Sigma}$; thus,

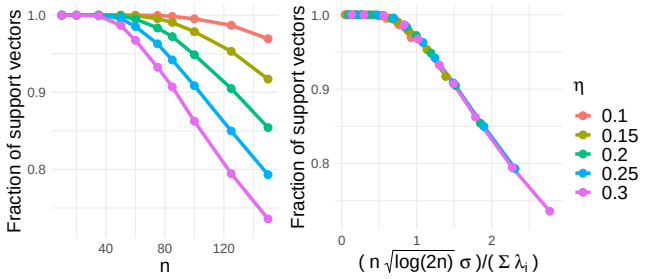


Figure 2.1: Proportion of support vectors for various values of σ^2. Note that the five curves nearly overlap when plotted versus $n\sqrt{\log(2n)}\sigma^2/\|\boldsymbol{\lambda}\|_1$ as predicted by (2.9) in our Theorem 1 confirming its tightness. See text for details on choices of $\boldsymbol{\eta}, \boldsymbol{\Sigma}$ and p.

$\boldsymbol{\eta} = \boldsymbol{\beta}$. In Fig. 2.1, we guarantee (2.8) by setting $p = 1500$ and varying n up to 150. For the eigenvalues of $\boldsymbol{\Sigma}$, we set $\lambda_1 = 7.5$, $\lambda_2 = \cdots = \lambda_{p-1} = 1$ and $\lambda_p = 0.2$. For $\boldsymbol{\eta}$, we chose $\eta_1 = \cdots = \eta_p = \eta$, where $\eta = 0.1, 0.15, 0.2, 0.25$ or 0.3. Fig. 2.1(Left) shows how the fraction of support vectors changes with n for different η. Smaller η results in higher proportion of support vectors. In order to verify the second condition in (2.9), Fig. 2.1(Right) plots the same curves over a re-scaled axis $n\sqrt{\log(2n)}\sigma/\|\boldsymbol{\lambda}\|_1$ (as suggested by (2.9)). Note that the 5 curves corresponding to different settings overlap in this new scaling, which agrees with the prediction of Theorem 1.

Next, we explain how Theorems 1 and 2 are useful for our purpose of studying the classification error of $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$. Suppose (2.8) and (2.9) (or (2.10) in the isotropic case) hold. Then $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}} = \hat{\boldsymbol{\eta}}_{\mathrm{LS}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$. Thus, under these conditions we can analyze the classification error of (2.5), by studying the simpler LS solution in (2.4). This observation was recently first exploited in [119] and sharpened in [72], but for a different data model. To see why the above statement is true, note that when (2.8) and (2.9) (or (2.10)) hold, then $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ satisfies the linear interpolation constraints; thus, it is feasible in (2.4). Consequently, $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ is in fact optimal in (2.4). To see the latter, assume for the sake of contradiction that $\|\hat{\boldsymbol{\eta}}_{\mathrm{LS}}\|_2 < \|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$. But, for all $i \in [n]$, $y_i(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T \boldsymbol{x}_i) = y_i^2 \geq 1$; thus, $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$

is feasible in (2.5), which contradicts our assumption. We will rely on this observation in Section 2.5 to study benign overfitting of SVM.

Finally, we compare our result to [119] that established similar conditions to Theorem 1, but for a 'Signed' model: $y_i = \text{sign}(\boldsymbol{x}_i^T \boldsymbol{\eta})$ with $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Interestingly, [119] obtained sufficient conditions that are identical to the first conditions in Theorems 1 and 2. More recently, [72] sharpened the overparameterization condition (2.8) to $\|\boldsymbol{\lambda}\|_1 \geq C_1\sqrt{n}\|\boldsymbol{\lambda}\|_2$ and $\|\boldsymbol{\lambda}\|_1 \geq C_2 n \log n \|\boldsymbol{\lambda}\|_\infty$ with large constants C_1 and C_2 for the anisotropic case. While their proof technique does not appear to be easily extended to the analysis of GMM, sharpening (2.8) can be an interesting future work. *The second conditions related to SNR are tailored to the GMM.* Intuitively, this is explained since in the 'Signed' model, the data are insensitive to the value of the signal strength $\|\boldsymbol{\eta}\|_2^2$; what matters is only the direction of $\boldsymbol{\eta}$. In contrast, both the direction and the scaling of the mean vector $\boldsymbol{\eta}$ are important in the GMM as apparent from (2.1). Our analysis captures this in a concrete way. Note that the first condition in Theorem 2 is sharper than in Theorem 1. This is because, in the isotropic case, we can leverage special properties of Wishart matrices; see Section 2.8.2 for more details.

2.4 Classification error

This section includes upper bounds on the classification error of the unregularized min-norm LS solution $\hat{\boldsymbol{\eta}}_{\text{LS}}$ and ℓ_2-regularized LS solution $\hat{\boldsymbol{\eta}}_\tau$ for the isotropic, balanced and bi-level ensembles. The implications of our bounds on $\hat{\boldsymbol{\eta}}_{\text{LS}}$ and $\hat{\boldsymbol{\eta}}_\tau$ are discussed later in Sections 2.5 and 2.6. The bounds that we provide can be achieved with probability $1-\delta$ over the randomness of the training set. We will assume throughout that $0 \leq \delta \leq 1/C$ for some universal constant C.

2.4.1 Balanced ensemble

Recall from Lemma 1 that $\hat{\boldsymbol{\eta}}^T\boldsymbol{\eta} > 0$ is needed to ensure that $\mathcal{R}(\hat{\boldsymbol{\eta}}) < 1/2$. The following lemma shows that this favorable event occurs with high probability provided sufficiently large overparameterization and high SNR.

Lemma 2. *Assume the balanced $\boldsymbol{\Sigma}$ ensemble (Definition 2.2.1). Fix $\delta \in (0,1)$ and suppose n is large enough such that $n > c\log(1/\delta)$ for some $c > 1$. Then, there exist constants $C_1, C_2 > 1$ such that with probability at least $1-\delta$, $\hat{\boldsymbol{\eta}}_\tau^T\boldsymbol{\eta} > 0$ provided that*

$$\|\boldsymbol{\eta}\|_2^2 > \frac{C_1 n\sigma^2}{\tau + \|\boldsymbol{\lambda}\|_1} + C_2\sigma. \tag{2.11}$$

We are now ready to state our main result for the balanced ensemble.

Theorem 3. *Assume the balanced $\boldsymbol{\Sigma}$ ensemble (Definition 2.2.1). Fix $\delta \in (0,1)$ and suppose large enough $n > c\log(1/\delta)$ for some $c > 1$. Further assume that (2.11) holds for constants C_1 and $C_2 > 1$. Then, there exists constants $C_3, C_4 > 1$ such that with probability at least $1-\delta$,*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_\tau) \le \exp\left(\frac{-\left(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\tau+\|\boldsymbol{\lambda}\|_1} - C_2\sigma\right)^2}{C_3\max\{1, \frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\}\|\boldsymbol{\lambda}\|_2^2 + C_4\sigma^2}\right). \tag{2.12}$$

The bound for the unregularized LS estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ can be obtained from (2.12) by setting $\tau = 0$. Thus, with probability at least $1-\delta$, $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})$ is upper bounded by

$$\exp\left(\frac{-\left(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\|\boldsymbol{\lambda}\|_1} - C_2\sigma\right)^2}{C_3\max\{1, \frac{n^2\sigma^2}{\|\boldsymbol{\lambda}\|_1^2}\}\|\boldsymbol{\lambda}\|_2^2 + C_4\sigma^2}\right). \tag{2.13}$$

By (2.13) we notice that the classification error depends on $\|\boldsymbol{\eta}\|_2^2$, $\|\boldsymbol{\lambda}\|_2^2$ and σ^2. Specifically, increasing $\|\boldsymbol{\eta}\|_2^2$ and/or decreasing either $\|\boldsymbol{\lambda}\|_2^2$ or σ^2 can make the bound smaller.

Increasing overparameterization can also help the bound decrease. To see that, consider for example the case $\lambda_1 = \lambda_2 = \cdots = \lambda_p$. Then, $\frac{n\sigma^2}{\|\boldsymbol{\lambda}\|_1} = \frac{n}{p}\|\boldsymbol{\eta}\|_2^2$ is directly related to the overparameterization ratio p/n and the numerator becomes $(\|\boldsymbol{\eta}\|_2^2(1 - C_1\frac{n}{p}) - C_2\sigma)^2$.

2.4.2 Isotropic ensemble

We have a slightly sharper bound on the classification error of the unregularized estimator in the isotropic regime, which is also easier to interpret. For simplicity, we only state the result for the min-norm interpolating solution (aka $\tau = 0$).

Theorem 4. *Assume $\boldsymbol{\Sigma} = \boldsymbol{I}$. Fix $\delta \in (0,1)$ and suppose large enough $n > c\log(1/\delta)$ for some $c > 1$. There exist constants $C, b > 1$ such that with probability at least $1 - \delta$, $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T\boldsymbol{\eta} > 0$ provided that $p > b \cdot n$ and $(1 - \frac{n}{p})\|\boldsymbol{\eta}\|_2 > C$. Further assume that these two conditions hold for $C, b > 1$. Then, there exist constants $C_1, C_2 > 1$ such that with probability at least $1 - \delta$:*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}) \leq \exp\Big(- \|\boldsymbol{\eta}\|_2^2 \frac{\big((1 - \frac{n}{p})\|\boldsymbol{\eta}\|_2 - C_1\big)^2}{C_2(\frac{p}{n} + \|\boldsymbol{\eta}\|_2^2)}\Big). \tag{2.14}$$

The bound depends on the overparameterization ratio p/n and the SNR $\|\boldsymbol{\eta}\|_2^2$ when $\boldsymbol{\Sigma} = \boldsymbol{I}$. To clarify the dependence, it is instructive to consider separately the following two regimes. (a) High-SNR regime: $\|\boldsymbol{\eta}\|_2^2 > \frac{p}{n}$. (b) Low-SNR regime: $\|\boldsymbol{\eta}\|_2^2 \leq \frac{p}{n}$.

The following is an immediate corollary of Theorem 4 specialized to the two regimes

Corollary 4.1. *Let the same assumptions of Theorem 4 hold. Then, there exists constants $C_1 > 1, C_2 > 0$ such that with probability at least $1 - \delta$, in the high-SNR regime:*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}) \leq \exp\Big(- C_2 \cdot \|\boldsymbol{\eta}\|_2^2 \cdot \big((1 - \frac{n}{p}) - C_1\frac{1}{\|\boldsymbol{\eta}\|_2}\big)^2\Big), \tag{2.15}$$

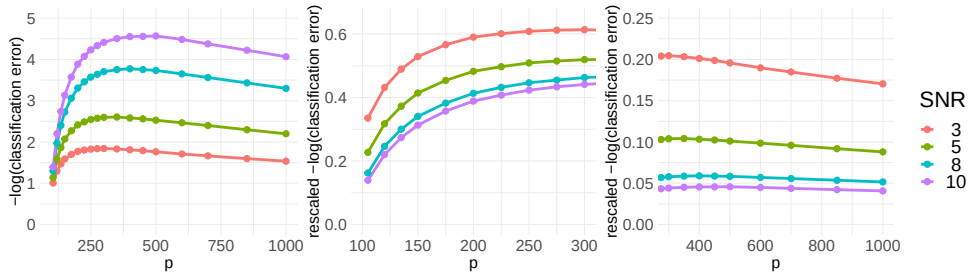


Figure 2.2: The left plot depicts $-\log($ classification error$)$ for $\|\boldsymbol{\eta}\|_2^2 = 3, 5, 8, 10$ as a function of p. The middle and right figures depict $-\log(\text{test error})/\|\boldsymbol{\eta}\|_2^2$ for small p (aka High-SNR regime) and $-\log(\text{test error})/\|\boldsymbol{\eta}\|_2^4$ for large p (aka Low-SNR regime), respectively. The rescalings are as suggested by our bounds (2.15) and (2.16), respectively. Note that, after rescaling, the error curves indeed become almost parallel as suggested by Corollary 4.1.

and, in the the low-SNR regime:

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}) \leq \exp\Big(- C_2 \cdot \|\boldsymbol{\eta}\|_2^4 \frac{((1-\frac{n}{p}) - C_1 \frac{1}{\|\boldsymbol{\eta}\|_2})^2}{p/n} \Big). \tag{2.16}$$

We use simulations to validate the above bounds. In Fig. 2.2(Left) we fix $n = 100$ and plot the classification error (in log-scale) as a function of p for four different SNR values $3, 5, 8$ and 10. Observe that $-\log \mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})$ initially increases until it reaches its maximum at some value of $p > n$ and then decreases as p gets even larger. This "increasing/decreasing" pattern is explained by the transition from the high-SNR to the low-SNR regime as per Corollary 4.1. On one hand, the negative of the exponent of the high-SNR bound (2.15) is increasing with p for $\|\boldsymbol{\eta}\|_2^2$. On the other hand, as p increases, and we move in the low-SNR regime, the negative of the exponent in (2.16) decreases with p when p is large enough. Additionally, in Figs. 2.2(Middle,Right), we plot re-normalized values $-\log \mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})/\|\boldsymbol{\eta}\|_2^2$ and $-\log \mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})/\|\boldsymbol{\eta}\|_2^4$. Notice that after appropriate normalization the curves become almost parallel to each other and almost overlap for large values of $\|\boldsymbol{\eta}\|_2^2$, as suggested by (2.15) and (2.16).

2.4.3 Bi-level ensemble

In this section we study the classification error under the bi-level ensemble in Definition 2.2.2, i.e. when one eigenvalue of $\boldsymbol{\Sigma}$ is much larger than the rest. Compared to the balanced ensemble, the analysis here depends on a more intricate way on the interaction between the mean vector and the spectrum of $\boldsymbol{\Sigma}$. To better understand this interaction we will assume $\boldsymbol{\beta}$ is one-sparse, i.e., the signal is concentrated in one direction. We will also assume, this time without loss of generality, [2] that $\boldsymbol{\Sigma}$ is diagonal; thus, $\boldsymbol{\beta} = \boldsymbol{\eta}$. Hence, taking $\boldsymbol{\beta}$ to be one-sparse with (say) the k-th element non-zero, the SNR becomes $\frac{\beta_k^2}{\lambda_k} = \frac{\eta_k^2}{\lambda_k}$. Specifically, $k = 1$ corresponds to the smallest SNR, for which we expect highest classification risk among all other choices of k. For better classification performance, large signal and noise components should *not* be in the same direction. This motivates the following assumption.

Assumption 1. The covariance matrix $\boldsymbol{\Sigma}$ is diagonal and its diagonal elements follow the bi-level structure in Definition 2.2.2. $\boldsymbol{\eta}$ is one-sparse with nonzero k-th element η_k and $k \neq 1$.

Under Assumption 1, the signal strength in the direction of $\boldsymbol{\Sigma}$ is $\sigma^2 = \lambda_k \eta_k^2$ and the ratio needed to be lower bounded $\frac{(\hat{\boldsymbol{\eta}}^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}}$ becomes $\frac{(\hat{\eta}_k \eta_k)^2}{\sum_{i=1}^p \lambda_i \hat{\eta}_k^2}$. The following theorem establishes an upper bound on the classification risk for this setting.

Theorem 5. *Let Assumption 1 hold. Fix $\delta \in (0,1)$ and large enough $n > c \log(1/\delta)$ for some $c > 1$. Let Assumption 1 hold. Then, there exist constants $c_1, c_2 > 1$ such that with probability at least $1-\delta$, $\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\eta} > 0$ provided that $\eta_k^2 > \frac{c_1 n \sigma^2}{\tau + \|\boldsymbol{\lambda}\|_{-1}} + c_2 \sigma$. Further assuming the above condition holds, there exist constants C_i's > 1 such that with probability at least*

[2] Recall $\boldsymbol{\Sigma} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^T$ and $\boldsymbol{\eta} = \boldsymbol{V} \boldsymbol{\beta}$. Thus, $\boldsymbol{X} = \boldsymbol{y} \boldsymbol{\eta}^T + \boldsymbol{Q} = (\boldsymbol{y} \boldsymbol{\beta}^T + \boldsymbol{Z} \boldsymbol{\Lambda}^{\frac{1}{2}}) \boldsymbol{V}^T =: \widetilde{\boldsymbol{X}} \boldsymbol{V}^T$, where $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ has IID standard normal entries. With this, it is not hard to check that $\frac{(\hat{\boldsymbol{\eta}}^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}} = \frac{(\boldsymbol{y}^T (\boldsymbol{X} \boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{X} \boldsymbol{\eta})^2}{\boldsymbol{y}^T (\boldsymbol{X} \boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{X} \boldsymbol{\Sigma} \boldsymbol{X}^T (\boldsymbol{X} \boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}} = \frac{(\boldsymbol{y}^T (\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T + \tau \boldsymbol{I})^{-1} \widetilde{\boldsymbol{X}} \boldsymbol{\beta})^2}{\boldsymbol{y}^T (\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T + \tau \boldsymbol{I})^{-1} \widetilde{\boldsymbol{X}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{X}}^T (\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}}$. Hence, after a change of basis, we can equivalently analyze the simplified model with diagonal covariance: $\widetilde{\boldsymbol{x}} = y \boldsymbol{\beta} + \widetilde{\boldsymbol{q}}$, $\widetilde{\boldsymbol{q}} \sim N(\boldsymbol{0}, \boldsymbol{\Lambda})$.

$1-\delta$,

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_\tau) \leq \exp\Big(\frac{-\Big(\eta_k^2(1-\frac{C_1 n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-C_2\sigma\Big)^2}{A+B+C_6(\lambda_k^2+\sigma^2)}\Big) \tag{2.17}$$

with $A = C_3\lambda_1^2\big(\frac{\tau+\|\boldsymbol{\lambda}\|_{-1}+C_4 n\sigma}{\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1}\big)^2$ *and* $B = C_5\big(\sum_{i\neq 1,k}\lambda_i^2\big)\big(1+\frac{C_4 n\sigma}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\big)^2$.

A bound for unregularized estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ can be obtained by setting $\tau = 0$. Recall the SNR under Assumption 1 is $\frac{\eta_k^4}{\sigma^2} = \frac{\eta_k^2}{\lambda_k}$. We observe that the bound above depends not only on the SNR, but also on λ_i, for $i \neq k$, i.e., the spectrum of $\boldsymbol{\Sigma}$ in *every* direction. Note that similar to previous sections, in (2.17), the term $\frac{n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}}$ on the numerator is related to the sufficiency of overparameterization. As we will see, the role of regularization in the bi-level ensemble is more subtle compared to the balanced ensemble and will be discussed in Section 2.6.

2.5 SVM generalization under high overparameterization

Now that we have captured the classification error of the min-norm LS estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ in (2.13) and (2.14), and we have established conditions ensuring $\hat{\boldsymbol{\eta}}_{\mathrm{LS}} = \hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ in Theorem 1 and Theorem 2, we establish sufficient conditions under which the classification error of hard-margin SVM vanishes as the overparameterization ratio p/n increases. Note that the bi-level ensemble will *not* satisfy the first condition in Theorem 1, hence we focus on the balanced and isotropic ensembles. For later use, define the term in (2.8) as $\lambda_* := 72\big(\|\boldsymbol{\lambda}\|_2 \cdot n\sqrt{\log n} + \|\boldsymbol{\lambda}\|_\infty \cdot n\sqrt{n}\log n + 1\big)$. We first focus on a special case where $\boldsymbol{\beta} = \begin{bmatrix}\beta & \beta & \dots & \beta\end{bmatrix}^T$ for simplicity.

Corollary 5.1. *Let the same assumption as in Theorem 3 hold with $\tau = 0$ and sufficiently large $n > C/\delta$ for some $C > 1$. Also let $\boldsymbol{\beta} = \begin{bmatrix} \beta & \beta & \dots & \beta \end{bmatrix}^T$. Then, for large C_i's > 1, with probability at least $(1-\delta)$, $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ linearly interpolates the data and the classification error $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}})$ approaches 0 as $p \to \infty$ provided the two following sets of conditions on $\|\boldsymbol{\lambda}\|_1$ hold:*

$$\|\boldsymbol{\lambda}\|_1 > \max\{\lambda_*, C_1\beta^2 n^2 \log(2n)\} \quad \textit{and} \quad \max\{\beta^{-2}\|\boldsymbol{\lambda}\|_2^2, \|\boldsymbol{\lambda}\|_1\} \leq C_2\beta^2 p^\alpha, \ \textit{for } \alpha < 2.$$

The first condition above requires sufficient overparameterization and the second one a large enough SNR. To see that, note for the setting of Corollary 5.1 that SNR $= p^2\beta^2/\|\boldsymbol{\lambda}\|_1$. Thus, the second condition imposes SNR $\geq cp^{2-\alpha}$ implying that SNR $\geq cp^\epsilon$ for some $\epsilon > 0$.

Corollary 5.1 assumes that $\boldsymbol{\beta}$ has equal elements. Now we allow the mean vector $\boldsymbol{\eta}$ to have different entry values but let $\boldsymbol{\Sigma} = \boldsymbol{I}$, then we have the following result.

Corollary 5.2. *Let the same assumptions as in Theorem 4 hold and n sufficiently large such that $n > C/\delta$ for some $C > 1$, thus $\boldsymbol{\Sigma} = \boldsymbol{I}$. Then, for large enough positive constant C_i's > 1, $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ linearly interpolates the data and the classification error $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}})$ approaches zero as $(p/n) \to \infty$ with probability at least $(1-\delta)$ provided either of the two following sets of conditions on the number of features p and mean-vector $\boldsymbol{\eta}$ hold:*
(1). High-SNR regime

$$\frac{1}{C_1} n\|\boldsymbol{\eta}\|_2^2 > p > \max\{10n\log n + n - 1, C_2 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2\}.$$

(2). Low-SNR regime

$$p > \max\{10n\log n + n - 1, C_3 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\}, \ \textit{and} \ \|\boldsymbol{\eta}\|_2^4 \geq C_4(\frac{p}{n})^\alpha, \ \textit{for } \alpha > 1.$$

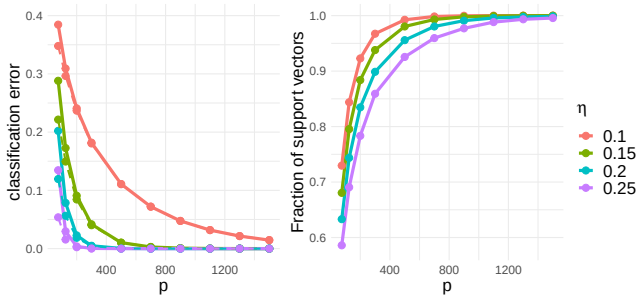


Figure 2.3: Numerical demonstration of benign overfitting for the GMM. The left plot shows the classification error with $n = 50$ and mean vector $\boldsymbol{\eta}$ with entries $\eta_1 = \ldots = \eta_p = \eta$. The solid lines correspond to the LS estimates and the (almost overlapping) dashed lines show the SVM solutions. The error vanishes with $p \to \infty$ indicating benign overfitting as predicted by Corollary 5.1. The right plot illustrates the proportion of support vectors in the same setting.

We first compare Corollaries 5.1 and 5.2 assuming both $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta & \beta & \ldots & \beta \end{bmatrix}^T$. Then $\|\boldsymbol{\lambda}\|_1 = \|\boldsymbol{\lambda}\|_2^2 = p$ and $\|\boldsymbol{\eta}\|_2^2 = \|\boldsymbol{\beta}\|_2^2 = p\beta$. It is not hard to check that under those assumptions, they both require $p > Cn^2 \log(2n)$, for sufficiently large constant C. One might expect that a sharper condition can be obtained by Corollary 5.2 when $\boldsymbol{\Sigma} = \boldsymbol{I}$. Unfortunately, that is not the case because although the first condition in Theorem 2 is sharper than that of Theorem 1, the second conditions become equivalent when $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta & \beta & \ldots & \beta \end{bmatrix}^T$ and are stronger than the first condition.

Remark 1 (Comparison of noiseless conditions to [31]). Using different tools to directly analyze $\hat{\boldsymbol{\eta}}_{\text{SVM}}$ (see Section 2.9.2), [31, Thm. 3.1] proved that for noisy mixtures with possibly adversarial corruptions and with subGaussian features

$$p > C_1 \max\{n\|\boldsymbol{\eta}\|_2^2, n^2 \log(n)\} \quad \text{and} \quad \|\boldsymbol{\eta}\|_2^4 \geq C_2 p^{\alpha}, \ \alpha > 1, \tag{2.18}$$

suffice for benign overfitting, i.e., for making the classification error asymptotically approach the noise level as $p/n \to \infty$. Our corollary 5.2 holds for the special case of Gaussian features and *noiseless* labels. Since labels are not corrupted, the noise floor

is zero. In this special case, our result relaxes significantly the sufficient conditions for which the risk approaches zero compared to a direct application of their result. To see this note that condition (2.18) is reminiscent of our 'low-SNR regime' condition (2) in Corollary 5.2. First, our condition relaxes the requirement on overparameterization from $p > Cn^2 \log(n)$ in (2.18) to $p > Cn\sqrt{\log(2n)}$. Second, our condition $\|\boldsymbol{\eta}\|_2^4 = \omega(p/n)$ on the SNR can be equivalent to theirs $\|\boldsymbol{\eta}\|_2^4 = \omega(p)$, for example in a setting of constant n. In order to better understand different conditions, consider a somewhat concrete setting in which n is fixed and only p and $\|\boldsymbol{\eta}\|_2$ grow large. Then for the classification error to go to 0 as $p \to \infty$, [31] requires (see (2.18)) that $\|\boldsymbol{\eta}\|_2 = \Theta(p^\beta)$ for $\beta \in (\frac{1}{4}, \frac{1}{2}]$. Instead, our Corollary 5.2 requires that $\|\boldsymbol{\eta}\|_2 = \Theta(p^\beta), \beta \in (1/4, 1/2]$ (low-SNR) or $\|\boldsymbol{\eta}\|_2 = \Theta(p^\beta)$ for $\beta \in (1/2, 1)$ (high-SNR). We repeat that this improvement is for zero label noise. In Section 2.7, where we study a noisy GMM, we show that our *sufficient* conditions can indeed change in the noisy case.

Finally, we present numerical illustrations validating Corollary 5.1. In Fig. 2.3, we let $\eta_1 = \cdots = \eta_p = \eta$ with $\eta = 0.1, 0.15, 0.2$ or 0.25. Thus, $\|\boldsymbol{\eta}\|_2^2 = \eta^2 p$. We also fix $n = 50$. The eigenvalues of $\boldsymbol{\Sigma}$ are generated as follows: $\lambda_1 = 0.005p$, $\lambda_p = 0.2 \cdot \frac{0.995p}{p-1}$ and $\lambda_2 = \cdots = \lambda_{p-1} = \frac{p-\lambda_1-\lambda_{p-1}}{p-2}$. This setting is different from the isotropic case, but ensures $\|\boldsymbol{\lambda}\|_1 \leq C_1 p$, $\|\boldsymbol{\lambda}\|_2 \leq C_2 p^{1/2}$ and conditions in Corollary 5.1 are satisfied. In Fig. 2.3(Left), we plot the classification error as a function of p for both LS estimates (solid lines) and SVM solutions (dashed lines). The solid and dashed curves almost overlap, so it can be hard to distinguish in the figure. We verify that as p increases, the classification error decreases towards zero. Similarly, Fig. 2.3(Right) reaffirms that all the data points become support vectors for sufficiently large p (cf. Theorem 2). In addition, Fig. 2.3(Left) shows that the classification error of SVM solutions is slightly better than that of LS estimates when p is small. The error becomes the same for large p, since then the SVM solutions are the same as LS solutions. Another observation is that

the classification error goes to zero very fast when SNR is high (e.g., purple curves), but the probability of interpolation increases at a slow rate. In contrast, when the SNR is low (e.g., red curves), the probability of interpolation increases fast, but the classification error decreases slowly. Intuitively, the harder the classification task (aka lower SNR), the larger the classification error and the more data points become support vectors.

2.6 On the role of regularization

In this section, we discuss how the ℓ_2-regularization affects the classification error of $\hat{\boldsymbol{\eta}}_\tau$ under the balanced and bi-level ensembles. For convenience, we start with a brief summary of our findings.

(a). Balanced ensemble:

1. The classification error is decreasing with τ. Thus, it is minimized as $\tau \to +\infty$.
2. Our bounds verify that in the limit $\tau \to +\infty$, $\hat{\boldsymbol{\eta}}_\tau$ has the same error as the so-called averaging estimator $\hat{\boldsymbol{\eta}}_{\text{Avg}} = \frac{1}{n}\sum_{i\in[n]} y_i \mathbf{x}_i$, where $\mathbf{x}_i^T$ is the i-th row of $\boldsymbol{X}$.
3. The averaging estimator is the best among the ridge-regularized estimator and the LS interpolating estimator.

(b). Bi-level ensemble:

1. Our upper bound on the classification error is *not* monotonically decreasing with τ. Hence regularization might *not* be helpful and the averaging estimator is *not* optimal.
2. There are regimes where $\tau = 0$ is optimal. Specifically, the interpolating estimator performs the best when λ_1 is large enough compared to other eigenvalues of $\boldsymbol{\Sigma}$ and overparameterization is sufficient.

These observations are illustrated in Figures 2.4, 2.5 and 2.6 which are discussed in detail in the next sections.

2.6.1 Balanced ensemble

We first analyze the bound in (2.12). Observe that both the terms $\frac{C_1 n\sigma^2}{\tau+\|\boldsymbol{\lambda}\|_1}$ in the numerator and $\frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}$ in the denominator decrease as the regularization parameter τ becomes larger. This suggests that, under the balanced ensemble, increasing regularization always helps decrease the error. The remaining terms, σ and $\|\boldsymbol{\lambda}\|_2^2$ in (2.12), that are not affected by changing τ reflect the intrinsic structure of the model and characterize the difficulty of the learning task. As $\tau \to +\infty$, the "regularization-sensitive" terms vanish and only those "regularization-insensitive" terms remain. Specifically, the upper bound on classification error becomes

$$\exp\Big(-\big(\frac{\|\boldsymbol{\eta}\|_2^2}{\sigma}-C_2\big)^2\big/\big(C_3\frac{\|\boldsymbol{\lambda}\|_2^2}{\sigma^2}+C_4\big)\Big). \tag{2.19}$$

In Appendix A.6 we show that the bound in (2.19) is the same as the bound for the so-called averaging estimator which simply returns

$$\hat{\boldsymbol{\eta}}_{\text{Avg}} = \boldsymbol{X}^T\boldsymbol{y}/n. \tag{2.20}$$

Therefore, under the balanced ensemble, the classification performance of the averaging estimator is superior to that of the ridge and interpolating estimators. A similar finding was recently reported in [114], but in an asymptotic setting and only for the isotropic case.

We now use numerical simulations to validate the above claims. In our simulations in Fig. 2.4, we fix $n = 100$ and vary p. To check (2.12), for each p, we set $\|\boldsymbol{\lambda}\|_2$ to be p

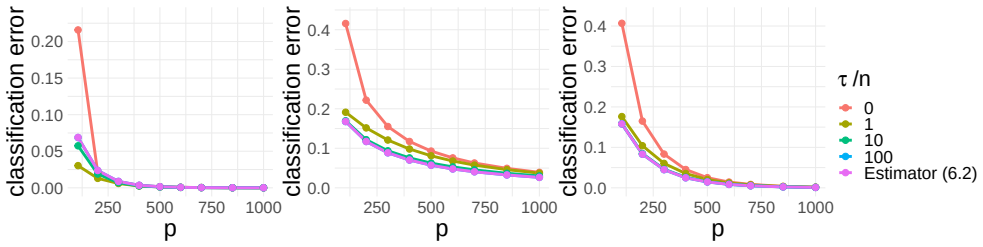


Figure 2.4: Classification error as a function of p under 3 model setups with different regularization parameter values τ. In the right plot all η_i's are the same. The middle/left ones correspond to the extreme cases of largest/smallest values σ^2; see text for details. Also plotted (in magenta) the averaging estimator defined in (2.20). As predicted by our theory, for fixed $\|\boldsymbol{\eta}\|_2$ and $\|\boldsymbol{\lambda}\|_2^2$, larger τ and smaller σ^2 lead to better performance and $\hat{\boldsymbol{\eta}}_\tau$ has the same performance as $\hat{\boldsymbol{\eta}}_{\text{Avg}}$ when τ is large.

and $\lambda_1 = \sqrt{0.0125p}$, $\lambda_p = \sqrt{0.000125p}$ and all the rest λ_i's are $\sqrt{(p-\lambda_1^2-\lambda_p^2)/(p-2)}$. This setup makes λ_1 slightly larger than other λ_i's and λ_p slightly smaller. For example, when $p = 1000$, then $\lambda_1 = 3.53$, $\lambda_p = 0.35$ and all other λ_i's are 0.99. Note that although λ_i's are not equal, those settings still satisfy the requirements of the balanced ensemble. Then, we look at different signals $\boldsymbol{\eta}$ with the same strength $\|\boldsymbol{\eta}\|_2^2 = (0.125^2)p$. To make σ^2 in (2.12) different, we consider 3 cases: all η_i's are the same, only η_1 nonzero and only η_p nonzero. The right plot in Fig. 2.4 shows the classification error of the same-η_i case, the middle one shows the nonzero-η_1 case and the left plot shows the nonzero-η_p case. To see the role of regularization, we look at the classification error with different τ values and also include the averaging estimator. We can see that among the three plots, the nonzero-η_p case (left) has the smallest classification error and the nonzero-η_1 case (middle) has the largest classification error. This is in agreement with the fact that the nonzero-η_p case has the smallest σ^2 and the the nonzero-η_1 case has the largest σ^2. For large p, regularization always helps reduce the classification error. When τ is large, the performance of $\hat{\boldsymbol{\eta}}_\tau$ becomes the same as that of $\hat{\boldsymbol{\eta}}_{\text{Avg}}$. All those observations are consistent with Theorem 3.

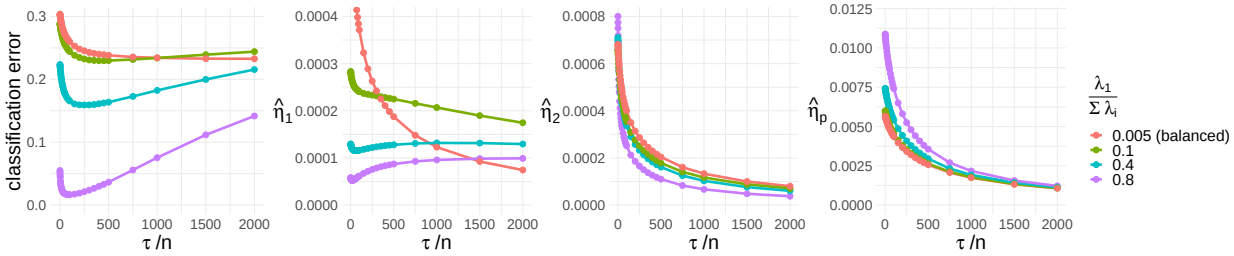


Figure 2.5: The left-most plot shows the classification error for different $\lambda_1/\|\boldsymbol{\lambda}\|_{-1}$ ratios with $n = 100$ and mean vector $\boldsymbol{\eta}$ with entries $\eta_1 = \cdots = \eta_{p-1} = 0$ and $\eta_p = \sqrt{200}$. Other plots show how $\hat{\eta}_i$'s vary with τ. As predicted by the theory, under the balanced ensemble, $\hat{\eta}_1$ and $\hat{\eta}_2$ decrease at similar rates, but have different behaviors when $\boldsymbol{\Sigma}$ has a highly spiky eigen-structure. See text for details.

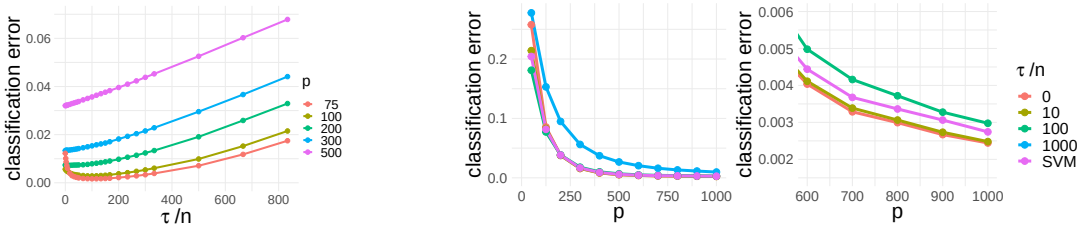
2.6.2 Bi-level ensemble

We have seen that regularization is always useful in reducing the classification risk in the balanced ensemble. For the bi-level ensemble, the story is quite different: the classification error is no longer monotonically decreasing as τ increases. Recall that under Assumption 1, with high probability, $\mathcal{R}(\hat{\boldsymbol{\eta}}_\tau)$ is upper bounded by (2.17). Moreover, when τ goes to infinity, it is not hard to check that this bound matches the corresponding for the averaging estimator (see Appendix A.6). Thus, in this case the averaging estimator is *not* optimal.

To see why (2.17) is no longer monotonically decreasing in τ, the term $\frac{\tau+\|\boldsymbol{\lambda}\|_{-1}+C_4 n\sigma}{\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1}$ in A is increasing in τ and thus A is increasing in τ when $\lambda_1 > C_4\sigma = C_4\sqrt{\lambda_k}\eta_k$, i.e., when λ_1 is large enough compared to λ_k and η_k. Note that (2.17) is obtained by lower bounding $\frac{(\eta_k\hat{\eta}_k)^2}{\sum_{i=1}^p \lambda_i\hat{\eta}_i^2}$ and A is related to the term $\lambda_1\hat{\eta}_1^2$, i.e., the estimate in the direction of λ_1. Since λ_1 is much larger than others, even if the regularization is useful in other directions, the performance won't keep improving as τ increases, because it won't help in the direction with the largest "noise". Term B in (2.17), on the other hand, is related to $\lambda_i\hat{\eta}_i^2$, for $i \neq 1$ or k, and it becomes smaller as τ becomes larger, thus regularization is useful in these directions. B becomes less important than A if λ_1 becomes larger than other λ_i's,

hence the regularization becomes less helpful in this case. Another observation is that in the numerator of (2.17), the term $\frac{n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}}$ decreases as τ increases. Note that when $\lambda_2 = \cdots = \lambda_p$, $\frac{n\lambda_k}{\|\boldsymbol{\lambda}\|_{-1}} = \frac{n}{p-1}$, hence this term measures the sufficiency of overparameterization. When the overparameterization is sufficient, i.e., p is much larger than n, $\frac{n\lambda_k}{\|\boldsymbol{\lambda}\|_{-1}}$ is already very small, hence $\frac{n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}}$ won't be much smaller than $\frac{n\lambda_k}{\|\boldsymbol{\lambda}\|_{-1}}$ even for large τ. In other words, strong regularization won't help very much. Summarizing all those observations, we conclude that the regularization becomes less useful in reducing the classification error when λ_1 is large enough relative to other eigenvalues and when overparameterization is sufficient. Under those conditions, $\tau = 0$ minimizes (2.17), therefore, the interpolating estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ has better performance than the regularized estimators. Since small or zero regularization can provide the best estimation in the bi-level setting with Assumption 1 in the overparameterization regime, it seems that the model structure itself provides the implicit regularization. This phenomenon is also discussed in [10, 118, 86, 119, 157].

The following numerical experiments validate our analysis. First in Fig. 2.5, we illustrate how the ratio $\lambda_1/\|\boldsymbol{\lambda}\|_{-1}$ affects the classification error and the role of regularization. In our simulation, $n = 100$, $p = 200$. $\boldsymbol{\eta} \in \mathbb{R}^{200}$ is one-sparse and only the last element is non-zero, i.e., $\boldsymbol{\eta}^T = [0, 0, \cdots, 0, \sqrt{200}]$. For the eigenvalues of $\boldsymbol{\Sigma}$, in the balanced ensemble, the diagonal elements are all equal, i.e., $\lambda_1 = \cdots = \lambda_{200} = 150$. In the bi-level ensemble, we fix $\|\boldsymbol{\lambda}\|_1 = 200 \cdot 150$ and let $\lambda_1 = \alpha\|\boldsymbol{\lambda}\|_1$, with $\alpha = 0.1, 0.4$ and 0.8. Then $\lambda_2 = \cdots = \lambda_p = (1-\alpha)\cdot\|\boldsymbol{\lambda}\|_1/(p-1)$. Note that larger α makes $\lambda_1/\|\boldsymbol{\lambda}\|_{-1}$ higher and that $\alpha = 0.005$ in the balanced ensemble. Fig. 2.5 illustrates how classification error and $\hat{\eta}_i$'s change with the regularization parameter τ. Based on previous analysis, we divide those $\hat{\eta}_i$'s into 3 groups, $\{\hat{\eta}_1; \hat{\eta}_2, \cdots, \hat{\eta}_{199}; \hat{\eta}_{200}\}$. $\hat{\eta}_1$ has true value 0 with large noise, $\hat{\eta}_2, \cdots, \hat{\eta}_{199}$ have true value 0 with small noise and $\hat{\eta}_{200}$ has non-zero true value with small noise. The figures show $\hat{\eta}_1$, $\hat{\eta}_2$ and $\hat{\eta}_{200}$. We can see that the classification error keeps decreasing as τ increases for the balanced ensemble (red curves). Part of the reason is that $\hat{\eta}_1$, $\hat{\eta}_2$



(a) (b)

Figure 2.6: For all the plots here, we fix $n = 30$, $\lambda_2 =, ..., = \lambda_p = 50$ and $\lambda_1/\|\boldsymbol{\lambda}\|_{-1} = 10$ (corresponding to the bi-level ensemble). (a) Classification error versus τ/n for different p and fixed $\eta_p = 25$. Observe that the classification error increases monotonically with τ for large p. (b) A regime where $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ performs the best and its classification error approaches 0 as $p \to \infty$. Specifically, we set here $\eta_p = 0.1\sqrt{50}p^{0.6}$. See text for details.

and $\hat{\eta}_{200}$ decrease at similar rates. In contrast, for the bi-level regime, as τ increases, the classification error decreases first, then increases. $\hat{\eta}_2$ decreases with τ, but $\hat{\eta}_1$ increases slowly with τ for large τ when $\lambda_1/\|\boldsymbol{\lambda}\|_{-1}$ is large. This is consistent with Theorem 5 in which A is increasing in τ when λ_1 is large enough. When λ_1 is not large enough, as the green curve shows, $\hat{\eta}_1$ decreases at a similar rate as $\hat{\eta}_1$ and all the curves are closer to those of the balanced ensemble.

Finally, we illustrate how the overparameterization ratio p/n affects the role of regularization in Fig. 2.6 (a). Here to guarantee p/n sufficiently large, we fix $n = 30$. We plot how the classification error changes with τ for $p = 75, 100, 200, 300$ and 500. $\boldsymbol{\eta}$ is one-sparse with $\eta_p = 25$. For eigenvalues of $\boldsymbol{\Sigma}$, to make $\lambda_1/\|\boldsymbol{\lambda}\|_{-1}$ sufficiently large, we set $\lambda_2 = \cdots = \lambda_p = 50$ and $\lambda_1 = 10\|\boldsymbol{\lambda}\|_{-1}$. We observe from Fig. 2.6 (a) that when p is large, the classification error increases with τ, thus $\tau = 0$ performs the best. The optimal choice of τ is larger than 0 when p is not large enough (e.g., $p = 75$ and 100). In Fig. 2.6 (b), we show a regime where $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ performs better than $\hat{\boldsymbol{\eta}}_\tau$ and $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ when p is large. Again we fix $n = 30$ to ensure sufficient overparameterization. Same as before, we set $\lambda_2 = \cdots = \lambda_p = 50$ and $\lambda_1 = 10\|\boldsymbol{\lambda}\|_{-1}$. To make the classification error approach

0 as $p \to \infty$, according to Corollary 14.1 in Appendix, we set $\eta_p = 0.1\sqrt{50}p^{0.6}$. Fig. 2.6 (b)(Left) shows the classification error over different p for various τ. We also added the curve for $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$. Fig. 2.6 (b)(Right) zooms in to $p \geq 600$. The classification error for the case with the largest τ is too large to be shown. We can see that the interpolating estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ performs better than the regularized estimators when p is sufficiently large.

2.7 Noisy GMM: Interpolation and benign overfitting

We extend our results to a probabilistic label-noise Gaussian mixture model.

2.7.1 Model and assumptions

We formally define the noisy model below; note that this is a special case of the adversarial noise model studied in [31].

Definition 2.7.1 (Noisy GMM)**.** A data pair $(\boldsymbol{x}, y_c) \in \mathbb{R}^p \times \{\pm 1\}$ is generated from the noisy Gaussian mixture model (GMM) with mean vector $\boldsymbol{\eta}$, covariance matrix $\boldsymbol{\Sigma}$ and corruption probability γ as follows. First, the clean data pair $(\boldsymbol{x}, y)$ is generated according to (2.1). Then the label y_c is generated by flipping the correct label y with probability γ. We assume that γ is independent of everything else (i.e., independent of the label y and the Gaussian noise term $\boldsymbol{q}$). Also, we assume that $0 \leq \gamma \leq 1/C$ for a large constant C.

We define the label vector with clean/corrupted labels as $\boldsymbol{y}/\boldsymbol{y}_c$. For brevity, we focus here on the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}$ and we derive analogues of Theorems 2, 4 and of Corollary 5.2. Throughout this section, we let $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}/\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ be the LS and SVM solutions

obtained by solving minimizations (2.4) and (2.5) but with the unobserved clean label vector $\boldsymbol{y}$ substituted by the observed corrupted vector $\boldsymbol{y}_c$.

2.7.2 Interpolation

Our first result establishes the equivalence between SVM and LS solutions for high enough effective overparameterization for noisy GMM data. As we will see, the required overparameterization conditions are now stronger compared to the noiseless case.

Theorem 6. *Assume n training samples following the noisy GMM with $\boldsymbol{\Sigma} = \mathbf{I}$. There exist large constants C_i's > 1 such that, if the following conditions on the number of features p and the mean-vector $\boldsymbol{\eta}$ hold:*

$$p > C_1 n \log n + n - 1 \quad \textit{and} \quad p > C_2 \max\{n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\}, \tag{2.21}$$

then, the SVM-solution $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ satisfies the linear interpolation constraint with probability at least $(1 - \frac{C_3}{n})$.

Note the extra term $n\|\boldsymbol{\eta}\|_2^2$ in the second condition above compared to Theorem 2. When $\mathrm{SNR} = \|\boldsymbol{\eta}\|_2^2 = \Omega(\log^{1/2}(\mathrm{n}))$, this new condition becomes dominant and the overparameterization ratio p/n should exceed SNR to guarantee interpolation. In Corollary 4.1, we called the regime $p/n \geq \mathrm{SNR}$ the low-noise regime. Hence, in the noisy case, we can guarantee equivalence of the SVM and LS solutions only in the low-SNR regime.

2.7.3 Error bounds

Our next result upper bounds the risk of the LS estimator. The bound holds in a regime where $\hat{\boldsymbol{\eta}}_{\mathrm{LS}} = \hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$, so it also applies to the risk of the SVM solution.

Theorem 7. *Assume that conditions in* (2.21) *hold for noisy GMM data with* $\boldsymbol{\Sigma} = \boldsymbol{I}$. *Fix* $\delta \in (0,1)$ *and suppose* n *is large enough such that* $n > c/\delta$ *for some* $c > 1$. *Then, there exist constants* $C, b > 1$ *such that with probability at least* $1-\delta$, $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T \boldsymbol{\eta} > 0$ *provided that* $p > b \cdot n$ *and* $(1 - \frac{n}{p})\|\boldsymbol{\eta}\|_2 > C$. *Further assume that these two conditions hold for* $C, b > 1$. *Then, there exist constants* $C_1, C_2 > 1$ *such that with probability at least* $1-\delta$*:*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}) \leq \gamma + \exp\Big(- C_2 \cdot \|\boldsymbol{\eta}\|_2^4 \frac{((1-\frac{n}{p}) - C_1 \frac{1}{\|\boldsymbol{\eta}\|_2})^2}{p/n} \Big). \tag{2.22}$$

Since the conditions in (2.21) hold, we operate here again in the low-SNR regime. The bound has two additive terms. The first term is the noise-level γ which we cannot beat due to the corruptions. The exponential term is the same as the bound for noiseless GMM in the low-SNR regime presented in Corollary 4.1.

Remark 2 (Comparison of risk bounds to [31]). For an adversarial noise model and subGaussian features [31] prove that

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}) \leq \gamma + \exp\Big(- C \frac{\|\boldsymbol{\eta}\|_2^4}{p} \Big), \tag{2.23}$$

in the following regime:

$$p > C \max\left\{ n^2 \log(n)\,,\, n\|\boldsymbol{\eta}\|_2^2 \right\}, \quad \|\boldsymbol{\eta}\|_2^2 \geq C \log(n) \quad \text{and} \quad n \geq C. \tag{2.24}$$

While our model is a special case of theirs, note that Theorem 7 holds under relaxed assumptions. Specifically, we relax (2.24) to

$$p > C \max\left\{ n \log(n)\,,\, n\|\boldsymbol{\eta}\|_2^2 \right\}, \quad \|\boldsymbol{\eta}\|_2^2 \geq C \quad \text{and} \quad n \geq C. \tag{2.25}$$

Also, assuming the special case (2.24) of [31] our bound in Theorem 7 reduces to the one

in (2.23).

2.7.4 Benign overfitting

Paralleling the exposition in Section 2.5, we use the results above to show that both the SVM and LS solutions approach the Bayes error as overparameterization increases. The requirements for this to happen are now stronger. However, the conclusion is somewhat more surprising in the noisy case: interpolating solutions nearly achieve optimal Bayes error despite perfectly fitting to corrupted labels. Borrowing the terminology introduced by [10], our result establishes "benign overfitting" for noisy GMM data.

Corollary 7.1. *Let the same assumptions as in Theorem 7 hold and n sufficiently large such that $n > C/\delta$ for some $C > 1$. Then, for large enough positive constant C_i's > 1, $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ linearly interpolates the data and the classification error $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}})$ approaches γ as $p/n \to \infty$ with probability at least $(1-\delta)$ provided the following sets of conditions on the number of features p and mean-vector $\boldsymbol{\eta}$ hold:*

$$p > \max\{C_2 n \log n + n - 1, C_3 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\}, \text{ and } \|\boldsymbol{\eta}\|_2^4 \geq C_4(\frac{p}{n})^{\alpha}, \text{ for } \alpha > 1.$$

Note that the benign overfitting condition above is identical to the condition of Corollary 5.2 for the Low-SNR regime in the noiseless case. However, instead of $\|\boldsymbol{\eta}\|_2 = \Theta(p^{\beta})$ with $\beta \in (\frac{1}{4}, 1)$ in the noiseless case, the conclusion of Corollary 7.1 holds under the stronger condition $\|\boldsymbol{\eta}\|_2 = \Theta(p^{\beta})$ for $\beta \in (\frac{1}{4}, \frac{1}{2}]$. We remark that (according also to the discussion in Remark 2), our conditions for benign overfitting of noisy GMM coincide with the conditions derived by [31].

2.8 Proofs outline

The complete proofs are given in the Appendix. Here, we provide an outline. For simplicity, we focus on the noiseless GMM in (2.1). At a high-level, the proofs for the noisy model remain the same with some more care needed to account for the mismatch between the clean and the corrupted labels (see Appendix A.8 for details).

2.8.1 Reductions to quadratic forms

We first show that the proofs of all theorems reduce to establishing lower/upper bounds on quadratic forms of the Gram matrix $(\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1}$.

Link between SVM solution and LS solution. We start with Theorems 1 and 2. As in [119, Theorem 1], it suffices to derive conditions under which the following complementary slackness condition of (2.5) is satisfied with high probability:

$$y_i \boldsymbol{e}_i^T (\boldsymbol{X}\boldsymbol{X}^T)^{-1} \boldsymbol{y} > 0, \text{ for all } i \in [n]. \tag{2.26}$$

Note that the LHS of (2.26) is a quadratic form involving $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$.

Classification error. When deriving upper bounds on the classification error, it suffices from Lemma 1 that we lower bound the ratio

$$\frac{(\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}_\tau} = \frac{(\boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{X} \boldsymbol{\eta})^2}{\boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{X} \boldsymbol{\Sigma} \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}}. \tag{2.27}$$

Specifically, when $\tau = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, we have

$$\frac{(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}_{\mathrm{LS}}} = \frac{(\boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T)^{-1} \boldsymbol{X} \boldsymbol{\eta})^2}{\boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T)^{-1} \boldsymbol{y}}. \tag{2.28}$$

Note that both the numerator and denominator above include terms such as $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T +$

$\tau \boldsymbol{I})^{-1}$ and $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$. Our key technical contribution is bounding those for GMM data.

Challenge. Bounding quadratic forms of $(\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1}$ is challenging for GMM data, since $\boldsymbol{X}\boldsymbol{X}^T = (\boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q})(\boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q})^T$, i.e. the Gram matrix "includes" both $\boldsymbol{y}$ and $\boldsymbol{\eta}$. Specifically, this is different to [119, 10, 157], since in their setting $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{Q}^T$ and their results on quadratic forms of inverse Wishart matrices do not directly apply here.

Our approach. For concreteness, consider the problem of bounding the quadratic form $T_1 := \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1}\boldsymbol{y}$. A possible approach is to start from bounds on the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I}$ and then obtain bounds for the eigenvalues of its inverse. Specifically, this turned out to be appropriate in the setting of [10, 119]. The situation is different here: the same eigenvalue approach fails to capture the dependence of $\boldsymbol{X}$ on $\boldsymbol{y}$ when bounding T_1 and results in suboptimal bounds. Instead of decoupling $\boldsymbol{y}$ and the inverse Gram matrix that appear in T_1, we consider both terms simultaneously. To make this possible we begin with the following decomposition of the Gram matrix:

$$\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I} = (\boldsymbol{Q}\boldsymbol{Q}^T + \tau \boldsymbol{I}) + \begin{bmatrix} \|\boldsymbol{\eta}\|_2 \boldsymbol{y} & \boldsymbol{Q}\boldsymbol{\eta} & \boldsymbol{y} \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\eta}\|_2 \boldsymbol{y}^T \\ \boldsymbol{y}^T \\ (\boldsymbol{Q}\boldsymbol{\eta})^T \end{bmatrix},$$

which already isolates the (translated) Wishart matrix $(\boldsymbol{Q}\boldsymbol{Q}^T + \tau \boldsymbol{I})$ from the terms $\boldsymbol{\eta}$ and $\boldsymbol{y}$. Once decomposed in this form, our observation is that with an appropriate application of the matrix inversion lemma we can now express quadratic forms of interest (such as T_1) in terms of five more primitive quadratic forms. This idea is materialized in the following key lemma.

Lemma 3. *Let* $\boldsymbol{U}_\tau := \boldsymbol{Q}\boldsymbol{Q}^T + \tau \boldsymbol{I}$ *(thus,* $\boldsymbol{U}_0 = \boldsymbol{Q}\boldsymbol{Q}^T$*) and* $\boldsymbol{d} := \boldsymbol{Q}\boldsymbol{\eta}$*. Further define the*

following five primitive quadratic forms

$$s := \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{y},\ t := \boldsymbol{d}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{d},\ h := \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{d},\ g_i := \boldsymbol{y}^T \boldsymbol{U}_0^{-1} \boldsymbol{e}_i,\ f_i := \boldsymbol{d}^T \boldsymbol{U}_0^{-1} \boldsymbol{e}_i,\ i \in [n], \tag{2.29}$$

and denote $D := s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2$. *With this notation, the following identity is true:*

$$\boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} = \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} - \frac{1}{D}\Big[\|\boldsymbol{\eta}\|_2 s, h^2 + h - st, s\Big] \begin{bmatrix} \|\boldsymbol{\eta}\|_2 \boldsymbol{y}^T \\ \boldsymbol{y}^T \\ \boldsymbol{d}^T \end{bmatrix} \boldsymbol{U}_\tau^{-1}. \tag{2.30}$$

The five quadratic forms defined in (2.29) involve now the inverse of the Wishart matrix $\boldsymbol{Q}\boldsymbol{Q}^T$ rather than of the original Gram matrix $\boldsymbol{X}\boldsymbol{X}^T$; this is why we call them "primitive". Despite that feature, bounding these terms still does *not* follow by mere application of results appearing in previous works [10, 157, 119]. Moreover, observe in identity (2.30) that the five primitive forms appear with mixed signs each and both in the numerator/denominator. Thus, it is critical to obtain both lower and upper bounds for them. We derive these in the two lemmas below, which together with Lemma 3 form key technical contributions of our work.

Lemma 4 (Balanced)**.** *Recall that* $\sigma^2 = \sum_{i=1}^p \lambda_i \beta_i^2$. *Assume that* $\boldsymbol{\Sigma}$ *follows the balanced ensemble defined in Definition 2.2.1. Fix* $\delta \in (0,1)$ *and suppose* n *is large enough such that* $n > c\log(1/\delta)$ *for some* $c > 1$. *Then, there exists constants* $C_1, C_2, C_3, C_6 > 1$, $C_5 > C_4 > 0$ *such that with probability at least* $1 - \delta$, *the following results hold:*

$$\frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_1)} \le s \le C_1 \frac{n}{(\tau + \|\boldsymbol{\lambda}\|_1)},\quad C_4 \frac{n\sigma^2}{(\tau + \|\boldsymbol{\lambda}\|_1)} \le t \le C_5 \frac{n\sigma^2}{(\tau + \|\boldsymbol{\lambda}\|_1)},$$
$$-C_2 \frac{n\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)} \le h \le C_2 \frac{n\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)},\quad \|\boldsymbol{d}\|_2^2 \le C_3 n\sigma^2,\quad \|\boldsymbol{y}^T \boldsymbol{U}_\tau^{-1}\|_2 \le C_6 \frac{\sqrt{n}}{(\tau + \|\boldsymbol{\lambda}\|_1)}.$$

We state our finding on $f_i, i \in [n]$ separately since it requires extra technical work to yield a bound that is uniform over $[n]$ and dimension independent. See Appendix A.7.3 for details.

Lemma 5. *Assume the condition in* (2.8) *is satisfied, Fix* $\delta \in (0,1)$ *and suppose large enough* $n > c/\delta, c > 1$. *There exists constant* $C > 1$ *such that with probability at least* $1 - \delta$,

$$\max_{i\in[n]} |f_i| \leq \frac{C\sqrt{\log(2n)}\sigma}{\|\boldsymbol{\lambda}\|_1}. \tag{2.31}$$

2.8.2 Proof sketch of Theorems 1 and 2

With the technical lemmas above, we are now ready to sketch the proof of Theorem 1. For simplicity here, consider the unregularized estimator ($\tau = 0$). As mentioned previously, it suffices to derive conditions under which (2.26) holds with high probability. Thanks to our Lemma 3, we derive the following decomposition in terms of the primitive terms defined in (2.29) (with $\tau = 0$ therein):

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{e}_i = \frac{g_i + hg_i - sf_i}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}. \tag{2.32}$$

The denominator above is positive with high probability. Thus, we only need to ensure that $y_i(g_i + hg_i - sf_i) > 0$. For this, we use Lemmas 4 and 5 (see also (A.6) for a lower bound on $y_i g_i$). Detailed proof is in Appendix A.2.1.

The proof of Theorem 2 is similar, except that the bounds on quadratic forms of the Wishart matrix are used when $\boldsymbol{\Sigma} = \boldsymbol{I}$, thus providing a sharper result. Specifically, when lower bounding $y_i g_i$, less overparameterization is needed, i.e., the first condition in (2.10) is sharper than (2.9).

2.8.3 Proof sketch of Theorems 3 and 4

As per Section 2.8.1, we will lower bound the ratio in (2.27). First, work with the denominator. Observe that $\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T = (\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})\boldsymbol{\Lambda}(\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})^T$. Further let $\boldsymbol{A} := (\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}$ and $\boldsymbol{z}_i$ denote the i-th column of $\boldsymbol{Z}$. Then, we show the following by applying the cyclic property of trace and the inequality $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \leq \frac{1}{2}(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} + \boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u})$, true for any PSD matrix $\boldsymbol{M}$:

$$\begin{aligned}
&\mathrm{Tr}\Big(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y}\Big) \\
=&\mathrm{Tr}\Big((\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})\boldsymbol{\Lambda}(\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})^T\boldsymbol{A}\Big) \\
\leq&2\Big(\sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2 + \sigma^2(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y})^2\Big),
\end{aligned}$$

Now, to upper bound $\sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2$, note $\|\boldsymbol{z}_i\|_2^2$'s are independent sub-exponentials; thus, for fixed $B > 0$, we can bound $\sum_{i=1}^{p}\lambda_i^2 B\|\boldsymbol{z}_i\|_2^2$ using the Bernstein's inequality. Specifically, we choose B as an upper bound on $\|\boldsymbol{A}\|_2 = \|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\|_2^2$, which we obtain thanks to Lemma 4 after the following decomposition as per Lemma 3: $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1} = \big((1+h)\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - s\boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\big)/D$. Similarly, we can upper bound $\sigma^2(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y})^2$ since again by Lemma 3 $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y} = s/D$.

Next, focus on the numerator in (2.27). Thanks to Lemma 3, we have the decomposition

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\eta} = \frac{s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h}{D}, \tag{2.33}$$

and the desired bound is obtained by a careful application of Lemma 4 that bounds the primitive quadratic appearing above. See Appendix A.3 for details and proof steps for Theorems 3 and 4.

2.8.4 Proof sketch of Theorem 5

We need to lower bound the ratio $\frac{(\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}_\tau} = \frac{(\eta_k \hat{\eta}_k)^2}{\sum_{i=1}^p \lambda_i \hat{\eta}_i^2}$. To do this, we divide $\hat{\eta}_i$'s into 3 groups ($\hat{\eta}_1$, $\hat{\eta}_k$ and the rest) and upper bound the following:

$$\frac{\lambda_1 \hat{\eta}_1^2}{(\eta_k \hat{\eta}_k)^2}, \quad \frac{\sum_{i \neq 1,k} \lambda_i \hat{\eta}_i^2}{(\eta_k \hat{\eta}_k)^2} \quad \text{and} \quad \frac{\lambda_k \hat{\eta}_k^2}{(\eta_k \hat{\eta}_k)^2},$$

where note from $\hat{\eta}_i = \boldsymbol{e}_i^T \hat{\boldsymbol{\eta}}$ that $\hat{\eta}_i = \sqrt{\lambda_i} \boldsymbol{z}_i^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}$, for $i \neq k$, and $\hat{\eta}_k = (\eta_k \boldsymbol{y}^T + \sqrt{\lambda_k} \boldsymbol{z}_k^T)(\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}$. As before, thanks to Lemma 3 this reduces to upper/lower bounding quadratic forms involving $\boldsymbol{U}_\tau^{-1} = (\boldsymbol{Q}\boldsymbol{Q}^T + \tau \boldsymbol{I})^{-1}$. However, because here λ_1 is much larger than other eigenvalues of $\boldsymbol{\Sigma}$, instead of directly bounding the eigenvalues of $\boldsymbol{U}_\tau$, we leverage the leave-one-out trick introduced in Bartlett et al. [10] and first separate λ_1 from the other eigenvalues. Specifically, by Woodbury's identity, $\boldsymbol{U}_\tau^{-1}$ is expressed as

$$\boldsymbol{U}_\tau^{-1} = (\tau \boldsymbol{I} + \sum_{i=2}^p \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T + \lambda_1 \boldsymbol{z}_1 \boldsymbol{z}_1^T)^{-1} = \boldsymbol{U}_{-1,\tau}^{-1} - \frac{\lambda_1 \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1}}{1 + \lambda_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1},$$

where $\boldsymbol{U}_{-1,\tau} = \tau \boldsymbol{I} + \sum_{i=2}^p \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T$. Now, we first bound the eigenvalues of $\boldsymbol{U}_{-1,\tau}$, and then use these results to bound the eigenvalues of $\boldsymbol{U}_\tau$ and $\boldsymbol{U}_\tau^{-1}$. See Appendix A.7.4 for details.

2.9 Discussion

Here, we include further details on how our results fit in the related literature.

2.9.1 Comparison to classical margin-based bounds

We start by arguing that classical bounds on the generalization of SVM are unimformative in the highly overparameterized settings of GMM data that we focus on. We do

this by quantitatively comparing our results with classical margin-based bounds applied to GMM data.

First, consider the following well-known bound.

Proposition 7.1. [145, Theorem 26.13]. *Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$ such that there exists some vector $\boldsymbol{\eta}^*$ with $\mathbf{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}(y \cdot \boldsymbol{\eta}^{*T}\boldsymbol{x} \geq 1) = 1$ and such that $\|\boldsymbol{x}\|_2 \leq R$ with probability 1. Let $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ be the SVM solution. Then with probability at least $1-\delta$, we have that*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}) \leq \frac{2R\|\boldsymbol{\eta}^*\|_2}{\sqrt{n}} + (1 + R\|\boldsymbol{\eta}^*\|_2)\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{2.34}$$

We apply Proposition 7.1 to the setting studied in Corollary 5.2. Specifically, we will apply the bound for $\boldsymbol{\eta}^* \leftarrow \boldsymbol{\eta}$. But, first we need to show that this choice satisfies the conditions of the proposition. To this end, by definition of $\boldsymbol{x}$, we have $y \cdot \boldsymbol{\eta}^T\boldsymbol{x} = \|\boldsymbol{\eta}\|_2^2 + \boldsymbol{\eta}^T(y\boldsymbol{q})$ with $\boldsymbol{q} \sim \mathcal{N}(0, \mathbf{I}_p)$. Therefore,

$$\begin{aligned}
\mathbf{P}(y \cdot \boldsymbol{\eta}^T\boldsymbol{x} \leq 1) &= \mathbf{P}(\boldsymbol{\eta}^T(y\boldsymbol{q}) \leq 1 - \|\boldsymbol{\eta}\|_2^2) = \mathbf{P}(\boldsymbol{\eta}^T(y\boldsymbol{q}) \geq \|\boldsymbol{\eta}\|_2^2 - 1) \\
&\leq \exp\left(-\frac{(\|\boldsymbol{\eta}\|_2^2 - 1)^2}{2\|\boldsymbol{\eta}\|_2^2}\right) \leq \exp\left(-\frac{\|\boldsymbol{\eta}\|_2^2}{2} + 1\right) \\
&\leq \exp\left(-C(p/n)^{2\alpha}\right) \overset{p/n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

The inequalities in the second line used Hoeffding's tail bound. In the third line, we used the conditions of Corollary 5.2 that $\|\boldsymbol{\eta}\|_2 \geq C_2(p/n)^\alpha$ for some $\alpha > 1/4$. Now, we compute the upper bound R. Bernstein's inequality gives with probability at least $1 - 2e^{-p/c}$,

$$\|\boldsymbol{\eta}\|_2^2 + (1 - (1/C))p \leq \|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{\eta}\|_2^2 + (1 + (1/C))p.$$

Thus, in our setting with probability 1, $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{\eta}\|_2 + C\sqrt{p} =: R$. Plugging this in (2.34) we see that

$$R\|\boldsymbol{\eta}^*\|_2/\sqrt{n} = \Theta\left(\|\boldsymbol{\eta}\|_2^2/\sqrt{n} + \sqrt{p/n}\|\boldsymbol{\eta}\|_2\right).$$

This bound becomes vacuous in the setting of Corollary 5.2. Indeed, by using $\|\boldsymbol{\eta}\|_2 \geq C_2(p/n)^\alpha$, we find that $\sqrt{p/n}\|\boldsymbol{\eta}\|_2 \to \infty$ as $p/n \to \infty$.

One might wonder if the conclusion would be different have we instead used a margin-based bound. We show that such bounds are also *not* able to explain why SVM nearly achieves Bayes optimal (aka zero) error in the highly overparameterized regime of Corollary 5.2.

Proposition 7.2. [145, Theorem 26.14]. *Assume the conditions of Proposition 7.1. Then, with probability at least* $1-\delta$*, we have that*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}) \leq \frac{4R\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2}{\sqrt{n}} + \sqrt{\frac{\log\left(4\log_2(\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2)/\delta\right)}{n}}. \tag{2.35}$$

In order to analytically evaluate the bound above, we need a means to control the inverse margin $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$. While it is not a-priori clear how to do this, our analysis establishes an upper bound on $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$ in the sufficiently overparameterized regime. Specifically, we do this as part of the proof of Theorem 3 in the process of upper bounding the correlation of the LS solution in Section A.3.2 (see Equation A.13). But in the setting of Corollary 5.2 $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2 = \|\hat{\boldsymbol{\eta}}_{\mathrm{LS}}\|_2$. Thus, (A.14) and (A.16) show that $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2^2 \leq \frac{C}{(1-n/p)\|\boldsymbol{\eta}\|_2^2}$. Recalling from above that $R = \|\boldsymbol{\eta}\|_2 + C\sqrt{p}$ and putting things together proves that

$$\frac{R\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2}{\sqrt{n}} = O\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{p}{n\|\boldsymbol{\eta}\|_2^2}}\right). \tag{2.36}$$

In the High-SNR regime of Corollary 5.2 recall that $p < n\|\boldsymbol{\eta}\|_2^2/C$, thus the value in (2.36) is $O(1 + 1/\sqrt{n})$. We see that (at least in the High-SNR regime) the bound we obtained by combining Proposition 7.2 with our upper bound of $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$ is indeed improved compared to that of Proposition 7.1. However, it still fails to predict the fact that the error goes to zero with increasing overparameterization (as predicted by Proposition 7.2). The bound is similarly inconclusive about the Low-SNR regime.

We end this section by noting that the fact that margin-based bounds are loose in the overparameterized regime has been previously also discussed in [116, 39] and [119]. Specifically, [116, 39] showed that Proposition 7.2 fails to predict the exact double-descent behavior of the risk in linear models even if the inverse margin $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$ in (2.35) is evaluated using sharp asymptotic formulas. Here, we have used our non-asymptotic bound for $\|\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}\|_2$ and showed that a margin-based argument is insufficient to yield the conclusions on Section 2.5. Finally, see also the discussion in [119, Sec. 6] where the authors demonstrate the deficiency of margin-based explanations in classification of signed data via numerical simulations. Here, we have arrived at the same conclusion, this time for GMM data, via an analytic study.

2.9.2 Comparison to previous works

We have already discussed how our results are motivated and how they differ from previous works in the Introduction. Here, we focus on the three most closely related papers [10, 119, 31] and provide a more detailed discussion.

Bartlett et al. [10]

As mentioned in the Introduction [10] is amongst the first to analytically study generalization principles under overparameterization. Our work is inspired by them, but

otherwise differs in four important aspects as outlined next.

(i) First, unlike linear regression, we study a linear classification model in which labels y are binary and covariates are of the form $\boldsymbol{x} = y\boldsymbol{\eta} + \boldsymbol{q}$. As discussed in Section 2.2 this implies that $y = \boldsymbol{x}^T\bar{\boldsymbol{\eta}} + z$ with $\bar{\boldsymbol{\eta}} := \boldsymbol{\eta}/\|\boldsymbol{\eta}\|_2^2$ and $z = \boldsymbol{q}^T\bar{\boldsymbol{\eta}}$. While this latter formulation resembles the linear regression model, where noise is additive, note here that the additive "noise" term z is highly signal dependent. The analysis of [10] makes heavy use of the assumption that noise is signal independent, hence their techniques *cannot* be directly applied to the GMM (see why in point (iii) below).

(ii) Second, our model is also different in that the feature vectors, although still Gaussian, are now signal dependent. Again, this does not allow a direct application of the technical results in [10] in our setting. Specifically, [10] show that in their setting bounding generalization can be mapped to a question about controlling the rate of decay of eigenvalues of inverse Wishart matrices. Instead, as explained in Section 2.8.1, in our setting we first express the generalization metric of interest as a non-trivial function of a number of simpler quadratic forms. While these quadratic forms involve inverse Wishart matrices, their statistics are not solely governed by the eigenstructure of the latter, but they also involve the mean vector $\boldsymbol{\eta}$.

(iii) Third, beyond the model itself what differs fundamentally in classification is the measure of generalization performance. Instead of the squared prediction risk studied by [10], relevant for us is the expected error as measured by the 0/1 loss. For Gaussian covariates, the former essentially reduces to the mean-squared error and the authors show that it suffices controlling a quantity $\boldsymbol{\epsilon}^T\mathbf{C}\boldsymbol{\epsilon}$, where $\mathbf{C} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ and $\boldsymbol{\epsilon}$ is the additive noise in the linear regression model [10, Lemma 7]. To do this, they exploit the assumption that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{X}$ and sub-Gaussian, which reduces the problem to upper bounding $\mathrm{Tr}(\mathbf{C})$ [10, Lemma 8]. Their subsequent analysis is tailored to this term. Instead, Lemma 1 shows that controlling the 0/1 risk requires bounding

the estimator's correlation. For the latter, we show that one needs to *both* upper bound $\boldsymbol{y}^T\mathbf{C}\boldsymbol{y}$ and lower bound $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta}$ (see (2.27)). Our goal is now more complicated compared to the situation faced in linear regression because: (a) In the first term $\boldsymbol{y}$ is not random (unlike $\boldsymbol{\epsilon}$). (b) The second quadratic form involves a matrix other than $\mathbf{C}$ and both vectors $\mathbf{y}$ and $\boldsymbol{\eta}$. (c) The feature matrix $\boldsymbol{X}$ is a non-centered Gaussian matrix whose mean involves both the response $\mathbf{y}$ and the mean vector $\boldsymbol{\eta}$.

(iv) The fourth difference is that in our setting, we are interested in the generalization performance of the SVM solution rather than the minimum-norm interpolator. The challenge is that the former is *not* given in closed form in terms of the label vector $\boldsymbol{y}$ and the feature matrix $\boldsymbol{X}$. The key innovation to circumvent this challenge is attributed to [119] who realized that under sufficient overparameterization SVM becomes equivalent to LS. We remark though that identifying the appropriate conditions for this to happen for GMM data is key contribution of our work (see Section 2.9.2). Following the above discussion emphasizing differences to the setting of [10] it should not be surprising the our error bounds in Section 2.4 are of different nature to those in [10]. The first key difference is that our bounds involve not only the eigenstructure of the covariance matrix, but also the mean vector of the GMM. Second, as a natural follow up, our conditions in Section 2.5 for which the classifier's error approaches the Bayes error are different to those in [10]. Despite the differences, it might be interesting to the reader noting that the two ensembles introduced in Definitions 2.2 and 2.3 can be expressed in terms of the notions of "effective ranks" defined by [10], i.e. $r_k := (\sum_{i>k}^{p} \lambda_i)/\lambda_{k+1}$.To see the relationship, let $\tilde{r}_k := (\sum_{i>k+1}^{p} \lambda_i)/\lambda_{k+1} = r_k - 1$. With this notation, in the balanced ensemble, $\tilde{r}_0 \geq bn$, which directly implies $r_0 \geq bn$. For large enough n, the reverse direction of implication is also true. In the bi-level ensemble, the first condition $\tilde{r}_0 \leq bn$ implies again $r_0 \leq b^{'}n$ for large enough n. Similarly, the second condition $\tilde{r}_1 \geq b_1 n$ implies $r_1 \geq b_1 n$.

Muthukumar et al. [119]

The paper by [119] is the most closely related to this work in terms of the approach that we follow. We complement the discussion in the Introduction regarding the different setting between the two works with a more detailed exposition of our key technical differences. For concreteness, we focus on the proof of equivalence between SVM and LS in Theorems 1 and 2, since the same differences apply to the error analysis in Theorems 3, 4 and 5.

There are two main steps in proving Theorems 1 and 2. The first step involves a deterministic sufficient condition guaranteeing that the constraints of the SVM optimization in (2.5) are active. The second step involves a probabilistic analysis of this deterministic condition using the generative statistical model at hand. The first part of our proof is same as in [119] and [72]. Specifically, we use their deterministic condition (2.26). On the other hand, the second part of our analysis is technically challenging. The reason is that unlike previous work where the covariates are zero mean Gaussians, in our case, $\boldsymbol{X} = \boldsymbol{Q} + \boldsymbol{y}\boldsymbol{\eta}^T$ for a zero-mean Gaussian matrix $\boldsymbol{Q}$. Note that the deterministic condition (2.26) to be checked involves the inverse Gram matrix. The key relevant technical argument in [119] (i.e., Lemma 1 therein) controls how far the inverse Wishart matrix $(\mathbf{Q}\mathbf{Q}^T)^{-1}$ is from $\left(\sum_{i\in[p]} |\lambda_i|\right) \mathbf{I}_d$. This results is clearly not sufficient in our case as $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ involves more terms. We repeat our strategy at circumventing this challenge as also sketched in Section 2.8.1. We start by expanding the terms in $(\boldsymbol{X}\boldsymbol{X}^T)$ and recognizing that after appropriate application of the matrix inversion lemma together with some algebra we can express the LHS of (2.26) as a function of five quadratic forms of either of two random matrices, $(\boldsymbol{Q}\boldsymbol{Q}^T)^{-1}$ or $\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{Q}^T)^{-1}$. It should be noted that this function involves the five quadratic forms in a convoluted way making it necessary to provide both upper and lower bounds for those forms (see Equation (2.32)). Besides

lower bounding one of the first two terms that involves $(\boldsymbol{Q}\boldsymbol{Q}^T)^{-1}$ using Lemma 1 in [119], *none* of the remaining quadratic forms appear in the analysis of [119]. Lemmas 9 and 10, where we obtain lower/upper bounds for them, form a main technical contribution of our work (see Appendix A.7 for details). Finally, the delicate piece of putting together those bounds to guarantee a positive quantity overall is also new compared to previous works (see Appendix A.2).

As we have highlighted in the previous sections, differences to [119] are not only technical. Most importantly, the differences extend to the conclusions regarding the conditions playing a key role for interpolation of the SVM solution and for the classification error of SVM to approach the Bayes error. See discussions in Sections 2.3 and 2.4. As a side technical note here, we have here relaxed the one-sparse assumption [119, Assumption 1] on the parameter vector $\boldsymbol{\eta}$ in the balanced ensemble. Finally, unlike [119], our bounds further apply to regularized LS and are extended to a model with label corruptions.

As a last note, we discuss the nice follow-up to [119] by [72], which involves two key contributions. The first concerns conditions for interpolation. The first step in their analysis (aka the deterministic condition (2.26) discussed above) is the same as in [119], but (2.26) is eventually expressed in a different equivalent form that allows tightening the probabilistic analysis that follows in the case of anisotropic convariance. Their second novelty involves relaxing the requirement for Gaussianity of the features to subGaussianity and Haar distribution. These improvements still only apply to the discriminative model, thus they are not directly applicable here.

Chatterji and Long [31]

We now compare our work to [31], who also derive non-asymptotic error bounds on the classification error of GMM data.

First, there are certain differences in the problem setting. On the one hand, [31] relaxes the assumption on Gaussianity by studying the case where $\boldsymbol{q}$ in (2.1) has sub-Gaussian entries [3]. On the other hand, while we require that $\boldsymbol{q}$ is Gaussian, our results capture explicitly the role of the data covariance matrix and its interplay with the mean vector via the key parameter $\sigma^2 = \boldsymbol{\eta}^T \boldsymbol{\Sigma} \boldsymbol{\eta}$. As we have seen in Sections 2.4, 2.5 and 2.6, the error behavior can differ substantially for different covariance structures (e.g., balanced vs bi-level ensembles). This phenomenon is *not* revealed by [31, Thm. 3.1] [4]. Another distinguishing feature of the results in [31] is that they apply to a noisy model that allows for (bounded number of) adversarial label corruptions. Our main focus is the noiseless GMM, but we also extended our results to a special case of their model in Section 2.7.

In terms of analysis, our techniques are very different. As mentioned we follow the high-level recipe of [119] (also adapted by [72]), that is first showing equivalence of SVM to LS and then deriving error bounds for the latter. Instead, [31] analyze the SVM solution by viewing it as the limit of gradient-descent updates on logistic loss minimization with sufficiently small step-size [146]. Specifically, they produce a recursive argument that at each iteration lower bounds the the expected margin of the current gradient-descent iterate on a clean point with respect to the margin of the previous iterate [31, Lem. 4.4]. We believe that both techniques are of interest. Via the connection to logistic loss minimization, their approach also yields insights on the degree to which one example (possibly a noisy one) can affect the quality of the learnt classifier [31, Lem. 4.8]. It also allows the study of subGaussian features (rather than Gaussian) rather naturally. On the other hand, the approach followed here leads to Theorems 1 and 2 on equivalence of

[3] This is interesting as for example it includes a Boolean noisy version of the rare-weak model by [78], for which our results do not directly apply.

[4] We note that the key role played by data covariance in double-decent and benign overfitting has been also emphasized in several related works, e.g., [65, 10, 119, 116, 30]

SVM to LS under sufficient effective overparameterization, which is a result of its own interest. Besides, as mentioned, our technique allows us to capture the effect of data covariance.

We already discussed in Sections 2.5 and 2.7 how our findings compare to those in [31]. In summary, for the noiseless case, we show that interpolating solutions asymptotically achieve the Bayes error under relaxed assumptions compared to the noisy model (see Remark 1. For the noisy model, our benign-overfitting conditions are identical, but our risk bounds hold under relaxed assumptions (see Remark 2). Finally, in addition to the risk bounds for SVM derived by [31], we also derive conditions for which SVM solution interpolates the data and we investigate regularized LS.

2.9.3 Contemporaneous and follow-up work

While the current version of our paper was undergoing review and after an earlier version of our paper [164], we became aware of contemporaneous independent work by [27]. Compared to our setting, [27] only requires sub-Gaussian features. Similar to us their results capture the key role of the spectrum of the data covariance. Their proofs for the correlated case build on ideas developed in our earlier version [164] for the isotropic case. Compared to them, we also derived bounds for regularized LS in our paper. A more detailed technical comparison between the two paper is as follows. First, [27] obtains a sharper first condition $\|\boldsymbol{\lambda}\|_1 \geq \max\{n\sqrt{n}\|\boldsymbol{\lambda}\|_\infty, n\|\boldsymbol{\lambda}\|_2\}$ for equivalence of SVM to LS in Theorem 1, by invoking stronger concentration arguments. Their second condition is the same as Theorem 1. For this, we further present insightful simulation results suggesting its tightness (see Figure 2.1). Regarding the classification error, [27] provides both upper and the lower bound for $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{SVM}})$. However, note that their results only apply to the balanced ensemble. For the anisotropic balanced setting, compared to Theorem 1, [27,

Theorem 3.1] proved that $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}}) \leq \exp\left(\frac{-C\|\boldsymbol{\eta}\|_2^4}{\|\boldsymbol{\lambda}\|_\infty + (\|\boldsymbol{\lambda}\|_1^2/n) + \sigma^2)}\right)$. Under the same assumptions in Theorem 1, [27, Theorem 3.1], the numerator of our corresponding bound in (2.13) can be simplified to the same as the result in [27]. However, the denominators are slightly different, where instead of $\|\boldsymbol{\lambda}\|_2^2/n$ in [27], we obtain $\|\boldsymbol{\lambda}\|_2^2$ and an additional $\|\boldsymbol{\lambda}\|_\infty$ term. For the isotropic setting, after some simplification, the bound on $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})$ in [27, Corollary 3.3] is the same as Theorem 4. Therefore, the benign overfitting condition ($\|\boldsymbol{\eta}\|_2 = \Theta(p^\beta), \beta \in (1/4, 1)$) is matching for finite n in the isotropic setting. As mentioned, we also investigate regularized LS in this paper. Additionally, in Section 2.7 we extend our results to a probabilistic label-noise model and derive conditions for benign overfitting that are not studied in [27].

More recently, [4] derived lower bounds for the conditions required to make SVM and LS solutions equivalent for discriminative models. For unconditional Gaussian covariates they show a sharp phase-transition characterizing the equivalence phenomenon. It is interesting to extend their analyses focusing on lower bounds to GMM data as studied in our paper. Finally, it is worth mentioning exciting related work [173, 161] that explores benign overfitting of *stochastic* gradient descent (SGD) (instead, note in Section 2.2 that our motivation for studying SVM or the minimum-norm interpolator comes from imiplicit bias of GD rather than SGD).

Chapter 3

Benign overfitting in multiclass classification

3.1 Introduction

Modern deep neural networks are overparameterized (high-dimensional) with respect to the amount of training data. Consequently, they achieve zero training error even on noisy training data, yet generalize well on test data [169]. Recent mathematical analysis has shown that fitting of noise in regression tasks can in fact be relatively benign for linear models that are sufficiently high-dimensional [10, 15, 65, 118, 86]. These analyses do not directly extend to classification, which requires separate treatment. In fact, recent progress on sharp analysis of interpolating binary classifiers [120, 31, 164, 27] revealed high-dimensional regimes in which binary classification generalizes well, but the corresponding regression task does *not* work and/or the success *cannot* be predicted by classical margin-based bounds [144, 9].

In an important separate development, these same high-dimensional regimes admit an equivalence of loss functions used for optimization at training time. The support

vector machine (SVM), which arises from minimizing the logistic loss using gradient descent [146, 77], was recently shown to satisfy a high-probability equivalence to interpolation, which arises from minimizing the squared loss [120, 72]. This equivalence suggests that interpolation is ubiquitous in very overparameterized settings, and can arise naturally as a consequence of the optimization procedure even when this is not explicitly encoded or intended. Moreover, this equivalence to interpolation and corresponding analysis implies that the SVM can generalize even in regimes where classical learning theory bounds are not predictive. In the logistic model case [120] and Gaussian binary mixture model case [31, 164, 27], it is shown that good generalization of the SVM is possible beyond the realm in which classical margin-based bounds apply. These analyses lend theoretical grounding to the surprising hypothesis that *squared loss can be equivalent to, or possibly even superior*, to the cross-entropy loss for classification tasks. Ryan Rifkin provided empirical support for this hypothesis on kernel machines [137, 136]; more recently, corresponding empirical evidence has been provided for state-of-the-art deep neural networks [74, 131].

These perspectives have thus far been limited to regression and *binary* classification settings. In contrast, most success stories and surprising new phenomena of modern machine learning have been recorded in *multiclass* classification settings, which appear naturally in a host of applications that demand the ability to automatically distinguish between large numbers of different classes. For example, the popular ImageNet dataset [141] contains on the order of 1000 classes. Whether a) good generalization beyond effectively low-dimensional regimes where margin-based bounds are predictive is possible, and b) equivalence of squared loss and cross-entropy loss holds in multiclass settings remained open problems.

This paper [165] makes significant progress towards a complete understanding of the optimization and generalization properties of high-dimensional linear multiclass classi-

fication, both for unconditional Gaussian covariates (where labels are generated via a multinomial logistic model), and Gaussian mixture models. Our contributions are listed in more detail below.

3.1.1 Our contributions

- We establish a *deterministic* sufficient condition under which the multiclass SVM solution has a very simple and symmetric structure: it is identical to the solution of a One-vs-All (OvA) SVM classifier that uses a *simplex-type* encoding for the labels (unlike the classical one-hot encoding). Moreover, the constraints at both solutions are active. Geometrically, this means that all data points are support vectors, and *they interpolate the simplex-encoding vector representation of the labels.* See Figure 3.2 for a numerical illustration confirming our finding.

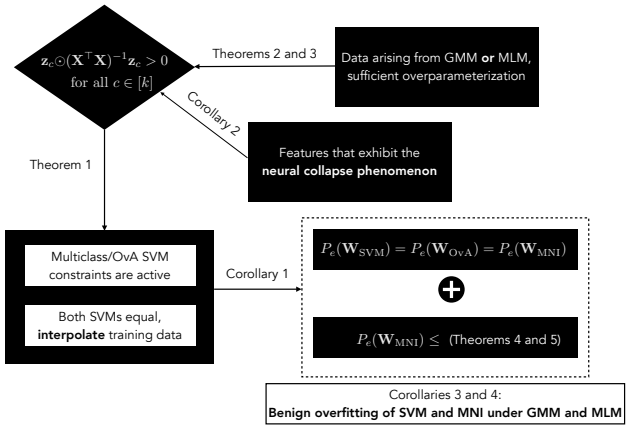


Figure 3.1: Contributions and organization.

- This implies a surprising equivalence between traditionally different formulations of multiclass SVM, which in turn are equivalent to the minimum-norm interpolating (MNI) classifier on the one-hot label vectors. Thus, we show that the outcomes of training with cross-entropy (CE) loss and squared loss are identical in terms of classification error.
- Next, for data following a Gaussian-mixtures model (GMM) or a Multinomial logistic model (MLM), we show that the above sufficient condition is satisfied with high-

probability under sufficient "effective" overparameterization. Our sufficient conditions are non-asymptotic and are characterized in terms of the data dimension, the number of classes, and functionals of the data covariance matrix. Our numerical results show excellent agreement with our theoretical findings. We also show that the sufficient condition of equivalence of CE and squared losses is satisfied when the "neural collapse" phenomenon occurs [125].

- Finally, we provide novel non-asymptotic bounds on the error of the MNI classifier for data generated either from the GMM and identify sufficient conditions under which benign overfitting occurs. A direct outcome of our results is that benign overfitting occurs under these conditions regardless of whether the cross-entropy loss or squared loss is used during training.

Figure 3.1 describes our contributions and their implications through a flowchart. To the best of our knowledge, these are the first results characterizing a) equivalence of loss functions, and b) generalization of interpolating solutions in the multiclass setting. The multiclass setting poses several challenges over and above the recently studied binary case. When presenting our results in later sections, we discuss in detail how our analysis circumvents these challenges.

3.1.2 Related work

Multiclass classification and the impact of training loss functions There is a classical body of work on algorithms for multiclass classification, e.g., [166, 21, 42, 36, 93] and several empirical studies of their comparative performance [136, 53, 3] (also see [71, 55, 90, 20, 38, 74, 131] for recent such studies in the context of deep nets). Many of these (e.g. [136, 74, 20]) have found that least-squares minimization yields competitive test classification performance to cross-entropy minimization. *Our proof of*

equivalence of the SVM and MNI solutions under sufficient overparameterization provides theoretical support for this line of work. This is a consequence of the implicit bias of gradient descent run on the CE and squared losses leading to the multiclass SVM [146, 77] and MNI [45] respectively. Numerous classical works investigated consistency [170, 93, 152, 129, 128] and finite-sample behavior, e.g., [88, 35, 96, 109, 97] of multiclass classification algorithms in the underparameterized regime. In contrast, our central focus is the highly overparameterized regime, where the typical uniform convergence techniques cannot apply.

Binary classification error analyses in overparameterized regime The recent wave of analyses of the minimum-ℓ_2-norm interpolator (MNI) in high-dimensional linear regression (beginning with [10, 15, 65, 118, 86]) prompted researchers to consider to what extent the phenomena of benign overfitting and double descent [14, 56] can be proven to occur in classification tasks. Even the binary classification setting turns out to be significantly more challenging to study owing to the discontinuity of the 0-1 test loss function. Sharp asymptotic formulas for the generalization error of binary classification algorithms in the linear high-dimensional regime have been derived in several recent works [73, 149, 107, 142, 150, 151, 39, 116, 81, 99, 143, 5, 103, 40]. These formulas are solutions to complicated nonlinear systems of equations that typically do not admit closed-form expressions. A separate line of work provides non-asymptotic error bounds for both the MNI classifier and the SVM classifier [31, 120, 164, 27]; in particular, [120] analyzed the SVM in a Gaussian covariates model by explicitly connecting its solution to the MNI solution. Subsequently, [164] also took this route to analyze the SVM and MNI in mixture models, and even more recently, [27] provided extensions of this result to sub-Gaussian mixtures. While these non-asymptotic analyses are only sharp in their dependences on the sample size n and the data dimension p, they provide closed-form

generalization expressions in terms of easily interpretable summary statistics. Interestingly, these results imply good generalization of the SVM beyond the regime in which margin-based bounds are predictive. Specifically, [120] identifies a separating regime for Gaussian covariates in which corresponding regression tasks would not generalize. In the Gaussian mixture model, margin-based bounds [144, 9] (as well as corresponding recently derived mistake bounds on interpolating classifiers [98]) would require the intrinsic signal-to-noise-ratio (SNR) to scale at least as $\omega(p^{1/2})$ for good generalization; however, the analyses of [164, 27] show that good generalization is possible for significantly lower SNR scaling as $\omega(p^{1/4})$. The above error analyses are specialized to the binary case, where closed-form error expressions are easy to derive [120]. The only related work applicable to the multiclass case is [155], which also highlights the numerous challenges of obtaining a sharp error analysis in multiclass settings. Specifically, [155] derived sharp generalization formulas for multiclass least-squares in underparameterized settings; extensions to the overparameterized regime and other losses beyond least-squares remained open. Finally, [85] recently derived sharp phase-transition thresholds for the feasibility of OvA-SVM on multiclass Gaussian mixture data in the linear high-dimensional regime. However, this does not address the more challenging multiclass-SVM that we investigate here.

Other SVM analyses The number of support vectors in the *binary* SVM has been characterized in low-dimensional separable and non-separable settings [41, 22, 108] and scenarios have been identified in which there is a vanishing fraction of support vectors, as this implies good generalization[1] via PAC-Bayes sample compression bounds [160, 61, 58]. In the highly overparameterized regime that we consider, perhaps surprisingly, the opposite behavior occurs: *all training points become support vectors with high probabil-*

[1]In this context, the fact that [120, 164, 27] provide good generalization bounds in the regime where support vectors proliferate is particularly surprising. In conventional wisdom, a proliferation of support vectors was associated with overfitting but this turns out to not be the case here.

ity [41, 22, 108, 120, 72]. In particular, [72] provided sharp non-asymptotic sufficient conditions for this phenomenon for both isotropic and anisotropic settings. The techniques in [120, 72] are highly specialized to the binary SVM and its dual, where a simple complementary slackness condition directly implies the property of interpolation. In contrast, the complementary slackness condition for the case of multiclass SVM *does not* directly imply interpolation; in fact, the operational meaning of "all training points becoming support vectors" is unclear in the multiclass SVM. *Our proof of deterministic equivalence goes beyond the complementary slackness condition and uncovers a surprising symmetric structure by showing equivalence of multiclass SVM to a simplex-type OvA classifier.* The simplex equiangular tight frame structure that we uncover is somewhat reminiscent of the recently observed neural collapse phenomenon in deep neural networks [125]; indeed, *Section 3.3.3 shows an explicit connection between our deterministic equivalence condition and the neural collapse phenomenon.* Further, [120, 72] focus on proving deterministic conditions for equivalence in the case of labels generated from covariates; the mixture model case (where covariates are generated from labels) turns out to be significantly more involved.

3.1.3 Organization

The paper is organized as follows. Section 3.2 describes the problem setting and sets up notation. Section 3.3 presents our main results on the equivalence between the multiclass SVM and MNI solutions for two data models: the Gaussian mixture model (GMM) and the multinomial logistic model (MLM). In the same section, we also show the equivalence under the Neural Collapse phenomenon. Section 3.4 presents our error analysis of the MNI solution (and, by our proved equivalence, the multiclass SVM) for the GMM and Section 3.5 presents consequent conditions for benign overfitting of multiclass

classification. Finally, Section 3.6 presents proofs of our main results; auxiliary proofs are deferred to the appendices. Please refer to the table of contents (before the appendices) for a more detailed decomposition of results and proofs.

Notation For a vector $\boldsymbol{v} \in \mathbb{R}^p$, let $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$, $\|\boldsymbol{v}\|_1 = \sum_{i=1}^p |v_i|$, $\|\boldsymbol{v}\|_\infty = \max_i\{|v_i|\}$. $\boldsymbol{v} > \mathbf{0}$ is interpreted elementwise. $\mathbf{1}_m$ / $\mathbf{0}_m$ denote the all-ones / all-zeros vectors of dimension m and $\mathbf{e}_i$ denotes the i-th standard basis vector. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|_2$ denotes its $2 \to 2$ operator norm and $\|\mathbf{M}\|_F$ denotes the Frobenius norm. $\odot$ denotes the Hadamard product. $[n]$ denotes the set $\{1, 2, ..., n\}$. We also use standard "Big O" notations $\Theta(\cdot)$, $\omega(\cdot)$, e.g., see [34, Chapter 3]. Finally, we write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the (multivariate) Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and, $Q(x) = \mathbb{P}(Z > x)$, $Z \sim \mathcal{N}(0, 1)$ for the Q-function of a standard normal. Throughout, constants refer to strictly positive numbers that do not depend on the problem dimensions n or p.

3.2 Problem setting

We consider the multiclass classification problem with k classes. Let $\boldsymbol{x} \in \mathbb{R}^p$ denote the feature vector and $y \in [k]$ represent the class label associated with one of the k classes. We assume that the training data has n feature/label pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$. We focus on the overparameterized regime, i.e., $p > Cn$, and we will frequently consider $p \gg n$. For convenience, we express the labels using the one-hot coding vector $\boldsymbol{y}_i \in \mathbb{R}^k$, where only the y_i-th entry of $\boldsymbol{y}_i$ is 1 and all other entries are zero, i.e., $\boldsymbol{y}_i = \boldsymbol{e}_{y_i}$. With this notation, the feature and label matrices are given in compact form as follows: $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{p \times n}$ and $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots \boldsymbol{v}_k \end{bmatrix}^T \in \mathbb{R}^{k \times n}$, where we have defined $\boldsymbol{v}_c \in \mathbb{R}^n, c \in [k]$ to denote the c-th row of the matrix $\boldsymbol{Y}$.

3.2.1 Data models

We assume that the data pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ are independently and identically distributed (IID). We will consider two models for the distribution of $(\boldsymbol{x}, y)$. For both models, we define the mean vectors $\{\boldsymbol{\mu}_j\}_{j=1}^k \in \mathbb{R}^p$, and the mean matrix is given by $\boldsymbol{M} := \begin{bmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \cdots & \boldsymbol{\mu}_k \end{bmatrix} \in \mathbb{R}^{p\times k}$.

Gaussian Mixture Model (GMM) In this model, the mean vector $\boldsymbol{\mu}_i$ represents the conditional mean vector for the i-th class. Specifically, each observation $(\boldsymbol{x}_i, y_i)$ belongs to to class $c \in [k]$ with probability π_c and conditional on the label y_i, $\boldsymbol{x}_i$ follows a multivariate Gaussian distribution. In summary, we have

$$\mathbb{P}(y = c) = \pi_c \text{ and } \boldsymbol{x} = \boldsymbol{\mu}_y + \boldsymbol{q},\ \boldsymbol{q} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3.1}$$

In this work, we focus on the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Our analysis can likely be extended to the more general anisotropic case, but we leave this to future work.

Multinomial Logit Model (MLM) In this model, the feature vector $\boldsymbol{x} \in \mathbb{R}^p$ is distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, and the conditional density of the class label y is given by the soft-max function. Specifically, we have

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \text{ and } \mathbb{P}(y = c|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\mu}_c^T \boldsymbol{x})}{\sum_{j\in[k]} \exp(\boldsymbol{\mu}_j^T \boldsymbol{x})}. \tag{3.2}$$

For this model, we analyze both the isotropic and anisotropic cases.

3.2.2 Data separability

We consider linear classifiers parameterized by $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_k \end{bmatrix}^T \in \mathbb{R}^{k \times p}$. Given input feature vector $\boldsymbol{x}$, the classifier is a function that maps $\boldsymbol{x}$ into an output of k via $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^k$ (for simplicity, we ignore the bias term throughout). We will operate in a regime where the training data are linearly separable. In multiclass settings, there exist multiple notions of separability. Here, we focus on (i) multiclass separability (also called k-class separability) (ii) one-vs-all (OvA) separability, and, recall their definitions below.

Definition 3.2.1 (multiclass and OvA separability)**.** The dataset $\{\boldsymbol{x}_i, y_i\}_{i \in [n]}$ is multiclass linearly separable when

$$\exists \boldsymbol{W} \ : \ (\boldsymbol{w}_{y_i} - \boldsymbol{w}_c)^T \boldsymbol{x}_i \geq 1, \ \forall c \neq y_i, c \in [k], \text{ and } \forall i \in [n]. \tag{3.3}$$

The dataset is one-vs-all (OvA) separable when

$$\exists \boldsymbol{W} \ : \ \boldsymbol{w}_c^T \boldsymbol{x}_i \begin{cases} \geq 1 \text{ if } y_i = c \\ \leq -1 \text{ if } y_i \neq c \end{cases}, \forall c \in [k], \text{ and } \forall i \in [n]. \tag{3.4}$$

Under both data models of the previous section (i.e., GMM and MLM), we have $\text{rank}(\boldsymbol{X}) = \text{n}$ almost surely in the overparameterized regime $p > n$. This directly implies OvA separability. It turns out that OvA separability implies multiclass separability, but not vice versa (see [16] for a counterexample).

3.2.3 Classification error

Consider a linear classifier $\widehat{\boldsymbol{W}}$ and a fresh sample $(\boldsymbol{x}, y)$ generated following the same distribution as the training data. As is standard, we predict $\hat{y}$ by a "winner takes it all

strategy", i.e., $\hat{y} = \arg\max_{j\in[k]} \widehat{\boldsymbol{w}}_j^T \boldsymbol{x}$. Then, the classification error conditioned on the true label being c, which we refer to as the *class-wise classification error*, is defined as

$$\mathbb{P}_{e|c} := \mathbb{P}(\hat{y} \neq y | y = c) = \mathbb{P}(\widehat{\boldsymbol{w}}_c^T \boldsymbol{x} \leq \max_{j\neq c} \widehat{\boldsymbol{w}}_j^T \boldsymbol{x}). \tag{3.5}$$

In turn, the *total classification error* is defined as

$$\mathbb{P}_e := \mathbb{P}(\hat{y} \neq y) = \mathbb{P}(\arg\max_{j\in[k]} \widehat{\boldsymbol{w}}_j^T \boldsymbol{x} \neq y) = \mathbb{P}(\widehat{\boldsymbol{w}}_y^T \boldsymbol{x} \leq \max_{j\neq y} \widehat{\boldsymbol{w}}_j^T \boldsymbol{x}). \tag{3.6}$$

3.2.4 Classification algorithms

Next, we review several different training strategies for which we characterize the total/class-wise classification error in this paper.

Multiclass SVM Consider training $\boldsymbol{W}$ by minimizing the cross-entropy (CE) loss

$$\mathcal{L}(\boldsymbol{W}) := -\log\left(\frac{e^{\boldsymbol{w}_{y_i}^T \boldsymbol{x}_i}}{\sum_{c\in[k]} e^{\boldsymbol{w}_c^T \boldsymbol{x}_i}}\right)$$

with the gradient descent algorithm (with constant step size η). In the separable regime, the CE loss $\mathcal{L}(\boldsymbol{W})$ can be driven to zero. Moreover, [146, Thm. 7] showed that the normalized iterates $\{\boldsymbol{W}^t\}_{t\geq 1}$ converge as

$$\lim_{t\to\infty} \Big\| \frac{\boldsymbol{W}^t}{\log t} - \boldsymbol{W}_{\mathrm{SVM}} \Big\|_F = 0,$$

where $\boldsymbol{W}_{\mathrm{SVM}}$ is the solution of the *multiclass SVM* [166] given by

$$\boldsymbol{W}_{\mathrm{SVM}} := \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{sub. to} \quad (\boldsymbol{w}_{y_i} - \boldsymbol{w}_c)^T \boldsymbol{x}_i \geq 1, \ \forall i \in [n], c \in [k] \text{ s.t. } c \neq y_i. \tag{3.7}$$

It is important to note that the normalizing factor $\log t$ here does *not* depend on the class label; hence, in the limit of GD iterations, the solution $\boldsymbol{W}^t$ decides the same label as multiclass SVM for any test sample.

One-vs-all SVM In contrast to Equation (3.7), which optimizes the hyperplanes $\{\boldsymbol{w}_c\}_{c\in[k]}$ jointly, the one-vs-all (OvA)-SVM classifier solves k separable optimization problems that maximize the margin of each class with respect to all the rest. Concretely, the OvA-SVM solves the following optimization problem for all $c \in [k]$:

$$\boldsymbol{w}_{\mathrm{OvA},c} := \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \quad \text{sub. to} \quad \boldsymbol{w}^T \boldsymbol{x}_i \begin{cases} \geq 1, & \text{if } \boldsymbol{y}_i = c, \\ \leq -1, & \text{if } \boldsymbol{y}_i \neq c, \end{cases} \quad \forall i \in [n]. \tag{3.8}$$

In general, the solutions to Equations (3.7) and (3.8) are different. While the OvA-SVM does not have an obvious connection to any training loss function, its relevance will become clear in Section 3.3. Perhaps surprisingly, we will prove that in the highly overparameterized regime the multiclass SVM solution is identical to a slight variant of (3.8).

Min-norm interpolating (MNI) classifier An alternative to the CE loss is the square loss $\mathcal{L}(\boldsymbol{W}) := \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}\|_2^2 = \frac{1}{2n}\sum_{i=1}^n \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{y}_i\|_2^2$. Since the square loss is tailored to regression, it might appear that the CE loss is more appropriate for classification. Perhaps surprisingly, one of the main messages of this paper is that under sufficient

effective overparameterization the two losses actually have equivalent performance. Our results lend theoretical support to empirical observations of competitive classification accuracy between the square loss and CE loss in practice [137, 74, 131].

Towards showing this, we note that when the linear model is overparameterized (i.e. $p > n$) and assuming rank($\boldsymbol{X}$) = n (e.g this holds almost surely under both the GMM and MLM), the data can be linearly interpolated, i.e. the square-loss can be driven to zero. Then, it is well-known [45] that gradient descent with sufficiently small step size and appropriate initialization converges to the minimum-norm -interpolating (MNI) solution, given by:

$$\boldsymbol{W}_{\text{MNI}} := \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F, \quad \text{sub. to} \quad \boldsymbol{X}^T\boldsymbol{w}_c = \boldsymbol{v}_c, \forall c \in [k]. \tag{3.9}$$

Since $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the MNI solution is given in closed form as

$$\boldsymbol{W}_{\text{MNI}}^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Y}^T.$$

From here on, we refer to (3.9) as the MNI classifier.

3.3 Equivalence of solutions and geometry of support vectors

In this section, we show the equivalence of the solutions of the three classifiers defined above in certain high-dimensional regimes.

3.3.1 A key deterministic condition

We first establish a key deterministic property of SVM that holds for *generic* multiclass datasets $(\boldsymbol{X}, \boldsymbol{Y})$ (i.e. not necessarily generated by either the GMM or MLM). Specifically, Theorem 8 below derives a sufficient condition (cf. (3.12)) under which the multiclass SVM solution has a surprisingly simple structure. First, the constraints are *all* active at the optima (cf. (3.13)). Second, and perhaps more interestingly, this happens in a very specific way; the feature vectors interpolate a simplex representation of the multiclass labels, as specified below:

$$\hat{\boldsymbol{w}}_c^T \boldsymbol{x}_i = z_{ci} := \begin{cases} \frac{k-1}{k} & , \ c = y_i \\ -\frac{1}{k} & , \ c \neq y_i \end{cases} \quad \text{for all } i \in [n], c \in [k]. \tag{3.10}$$

To interpret this, define an adjusted k-dimensional label vector $\widetilde{\boldsymbol{y}}_i := [z_{1i}, z_{2i}, \ldots, z_{ki}]^T$ for each training sample $i \in [n]$. This can be understood as a k-dimensional vector encoding of the original label y_i that is different from the classical one-hot encoding representation $\boldsymbol{y}_i$; in particular, it has entries either $-1/k$ or $1-1/k$ (rather than 0 or 1). We call this new representation a *simplex representation*, based on the following observation. Consider k data points that each belong to a different class $1, \ldots, k$, and their corresponding vector representations $\widetilde{\boldsymbol{y}}_1, \ldots, \widetilde{\boldsymbol{y}}_k$. Then, it is easy to verify that the vectors $\{\boldsymbol{0}, \widetilde{\boldsymbol{y}}_1, \ldots, \widetilde{\boldsymbol{y}}_k\}$ are affinely independent; hence, they form the vectices of a k-simplex.

Theorem 8. *For a multiclass separable dataset with feature matrix* $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ *and label matrix* $\boldsymbol{Y} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k]^T \in \mathbb{R}^{k \times n}$*, denote by* $\boldsymbol{W}_{SVM} = [\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, \ldots, \hat{\boldsymbol{w}}_k]^T$ *the multiclass SVM solution of* (3.7)*. For each class* $c \in [k]$ *define vectors* $\boldsymbol{z}_c \in \mathbb{R}^n$ *such*

that

$$\boldsymbol{z}_c = \boldsymbol{v}_c - \frac{1}{k}\mathbf{1}_n,\ c \in [k]. \tag{3.11}$$

Let $(\boldsymbol{X}^T\boldsymbol{X})^+$ *be the Moore-Penrose generalized inverse*[2] *of the Gram matrix* $\boldsymbol{X}^T\boldsymbol{X}$ *and assume that the following condition holds*

$$\boldsymbol{z}_c \odot (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{z}_c > \mathbf{0}, \quad \forall c \in [k]. \tag{3.12}$$

Then, the SVM solution $\boldsymbol{W}_{SVM}$ *is such that all the constraints in* (3.7) *are active. That is,*

$$(\hat{\boldsymbol{w}}_{y_i} - \hat{\boldsymbol{w}}_c)^T\boldsymbol{x}_i = 1,\ \forall c \neq y_i, c \in [k],\ \textit{and}\ \ \forall i \in [n]. \tag{3.13}$$

Moreover, the features interpolate the simplex representation. That is,

$$\boldsymbol{X}^T\hat{\boldsymbol{w}}_c = \boldsymbol{z}_c,\ \forall c \in [k]. \tag{3.14}$$

For $k = 2$ classes, it can be easily verified that Equation (3.12) reduces to the condition in Equation (22) of [120] for the binary SVM. Compared to the binary setting, the conclusion for the multiclass case is richer: provided that Equation (3.12) holds, we show that not only are all data points support vectors, but also, they satisfy a set of simplex OvA-type constraints as elaborated above. The proof of Equation (3.14) is particularly subtle and involved: unlike in the binary case, it does *not* follow directly from a complementary slackness condition on the dual of the multiclass SVM. A key technical contribution that we provide to remedy this issue is a novel reparameterization of the

[2]Most of the regimes that we study are ultra-high-dimensional (i.e. $p \gg n$), and so $\boldsymbol{X}^T\boldsymbol{X}$ is invertible with high probability. Consequently, $(\boldsymbol{X}^T\boldsymbol{X})^+$ can be replaced by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ in these cases.

SVM dual. The complete proof of Theorem 8 and this reparameterization is provided in Section 3.6.1.

We make a few additional remarks on the interpretation of Equation (3.14).

First, our proof shows a somewhat stronger conclusion: when Equation (3.12) holds, the multiclass SVM solutions $\hat{\boldsymbol{w}}_c, c \in [k]$ are same as the solutions to the following *simplex OvA-type classifier* (cf. Equation (3.8)):

$$\min_{\boldsymbol{w}_c} \ \frac{1}{2}\|\boldsymbol{w}_c\|_2^2 \qquad \text{sub. to} \quad \boldsymbol{x}_i^T \boldsymbol{w}_c \begin{cases} \geq \frac{k-1}{k} & , y_i = c, \\ \leq -\frac{1}{k} & , y_i \neq c, \end{cases} \quad \forall i \in [n], \tag{3.15}$$

for all $c \in [k]$. We note that the OvA-type classifier above can also be interpreted as a binary cost-sensitive SVM classifier [75] that enforces the margin corresponding to all other classes to be $(k-1)$ times smaller compared to the margin for the labeled class of the training data point. This simplex structure is illustrated in Figure 3.2, which evaluates the solution of the multiclass SVM on a 4-class Gaussian mixture model with isotropic noise covariance. The mean vectors are set to be mutually orthogonal and equal in norm, with SNR $\|\boldsymbol{\mu}\|_2 = 0.2\sqrt{p}$. We also set $n = 50$, $p = 1000$ to ensure sufficient effective overparameterization (in a sense that will be formally defined in subsequent sections). Figure 3.2 shows the inner product $\widehat{\boldsymbol{w}}_c^T \boldsymbol{x}$ drawn from 8 samples. These inner products are consistent with the simplex OvA structure defined in Equation (3.14), i.e., $\widehat{\boldsymbol{w}}_c^T \boldsymbol{x}_i = 3/4$ if $y_i = c$ and $\widehat{\boldsymbol{w}}_c^T \boldsymbol{x}_i = -1/4$ if $y_i \neq c$.

Second, Equation (3.14) shows that when Equation (3.12) holds, then the multiclass SVM solution $\boldsymbol{W}_{\mathrm{SVM}}$ has the same classification error as that of the minimum-norm interpolating solution. In other words, we can show that the minimum-norm classifiers that interpolate the data with respect to *either* the one-hot representations $\boldsymbol{y}_i$ or the simplex representations $\widetilde{\boldsymbol{y}}_i$ of (3.10) have identical classification performance. This conclusion,

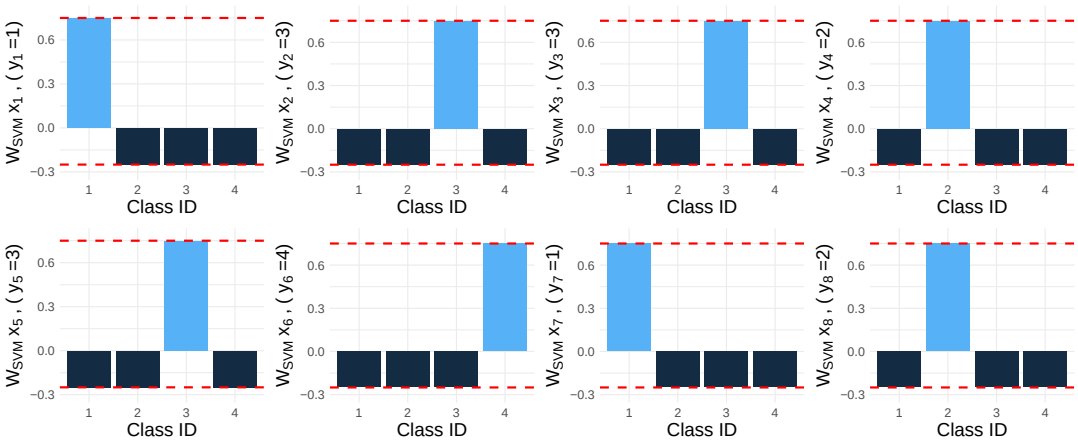


Figure 3.2: Inner products $\boldsymbol{W}_{\mathrm{SVM}}\boldsymbol{x}_c \in \mathbb{R}^4$ for features $\boldsymbol{x}_i$ that each belongs to the c-th class for $c \in [k]$ and $k = 4$ total classes. The red lines correspond to the values $(k-1)/k = 3/4$ and $-1/k = -1/4$ of the simplex encoding described in Theorem 8. Observe that the inner products $\boldsymbol{W}_{\mathrm{SVM}}\boldsymbol{x}_c$ match with these values, that is, Equation (3.10) holds.

stated as a corollary below, drives our classification error analysis in Section 3.4.

Corollary 8.1 (SVM=MNI)**.** *Under the same assumptions as in Theorem 8, and provided that the inequality in Equation* (3.12) *holds, it holds that* $\mathbb{P}_{e|c}(\boldsymbol{W}_{SVM}) = \mathbb{P}_{e|c}(\boldsymbol{W}_{MNI})$ *for all* $c \in [k]$*. Thus, the total classification errors of both solutions are equal:* $\mathbb{P}_e(\boldsymbol{W}_{SVM}) = \mathbb{P}_e(\boldsymbol{W}_{MNI})$.

The corollary follows directly by combining Theorem 8 with the following lemma applied with the choice $\alpha = 1, \beta = -1/k$. We include a detailed proof below for completeness.

Lemma 6. *For constants* $\alpha > 0, \beta$*, consider the MNI-solution* $\boldsymbol{w}_c^{\alpha,\beta} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^+(\alpha\boldsymbol{v}_c + \beta\mathbf{1}), c \in [k]$ *corresponding to a target vector of labels* $\alpha\boldsymbol{v}_c + \beta\mathbf{1}_n$*. Let* $\mathbb{P}_{e|c}^{\alpha,\beta}, c \in [k]$ *be the class-conditional classification errors of the classifier* $\boldsymbol{w}^{\alpha,\beta}$*. Then, for any different set of constants* $\alpha' > 0, \beta'$*, it holds that* $\mathbb{P}_{e|c}^{\alpha,\beta} = \mathbb{P}_{e|c}^{\alpha',\beta'}, \forall c \in [k]$.

Proof. Note that $\boldsymbol{w}_c^{\alpha=1,\beta=0} = \mathbf{w}_{\mathrm{MNI},c}, c \in [k]$ and for arbitrary $\alpha > 0, \beta$, we have: $\boldsymbol{w}_c^{\alpha,\beta} = \alpha\mathbf{w}_{\mathrm{MNI},c} + \beta\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^+\mathbf{1}$. Moreover, it is not hard to check that $\mathbf{w}_{\mathrm{MNI},c}^\top\boldsymbol{x} \leq$

$\max_{j\neq c} \mathbf{w}_{\mathrm{MNI},j}^\top \boldsymbol{x}$ if and only if $(\alpha \boldsymbol{w}_{\mathrm{MNI}c}+\mathbf{b})^\top \boldsymbol{x} \leq \max_{j\neq c}(\alpha \mathbf{w}_{\mathrm{MNI},j}+\mathbf{b})^\top \boldsymbol{x}$, for any $\mathbf{b} \in \mathbb{R}^p$. The claim then follows by choosing $\mathbf{b} = \beta \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^+ \mathbf{1}$ and noting that $\alpha > 0, \beta$ were chosen arbitrarily. □

3.3.2 Connection to effective overparameterization

Theorem 8 establishes a *deterministic* condition that applies to any multiclass separable dataset as long as the data matrix $\boldsymbol{X}$ is full-rank. In this subsection, we show that the inequality in Equation (3.12) occurs with high-probability under both the GMM and MLM data models provided that there is sufficient effective overparameterization.

Gaussian mixture model

We assume an equal-energy, equal-prior setting as detailed below.

Assumption 2 (Equal energy/prior). The mean vectors have equal energy and the priors are equal, i.e. we have $\|\boldsymbol{\mu}\|_2 := \|\boldsymbol{\mu}_c\|_2$ and $\pi_c = \pi = 1/k$, for all $c \in [k]$.

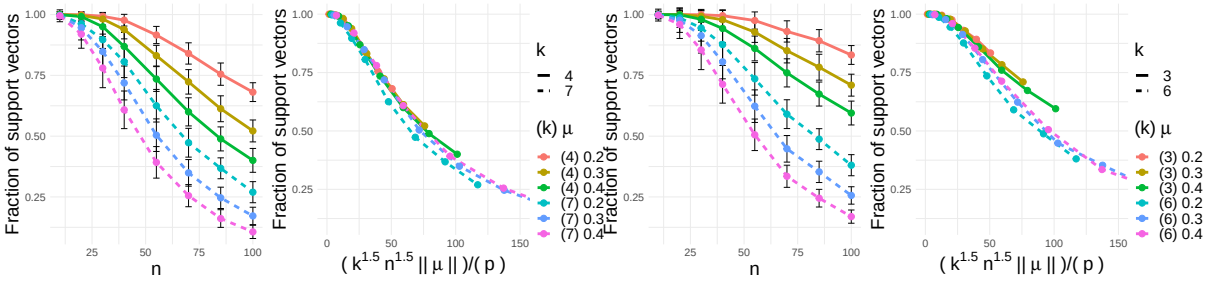
Theorem 9. *Assume that the training set follows a multiclass GMM with* $\boldsymbol{\Sigma} = \mathbf{I}_p$*, Assumption 2 holds, and the number of training samples* n *is large enough. There exist constants* $c_1, c_2, c_3 > 1$ *and* $C_1, C_2 > 1$ *such that Equation* (3.12) *holds with probability at least* $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$*, provided that*

$$p > C_1 k^3 n \log(kn) + n - 1 \quad \text{and} \quad p > C_2 k^{1.5} n \sqrt{n} \|\boldsymbol{\mu}\|_2. \tag{3.16}$$

Theorem 9 establishes a set of two conditions under which Equation (3.12) and the conclusions of Theorem 8 hold, i.e. $\boldsymbol{W}_{\mathrm{SVM}} = \boldsymbol{W}_{\mathrm{MNI}}$. The first condition requires sufficient overparameterization $p = \Omega(k^3 n \log(kn))$, while the second one requires that the signal strength is not too large. Intuitively, we can understand these conditions as fol-

lows. Note that Equation (3.12) is satisfied provided that the inverse Gram matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is "close" to identity, or any other positive-definite diagonal matrix. Recall from Equation (3.1) that $\boldsymbol{X} = \boldsymbol{M}\boldsymbol{Y} + \boldsymbol{Q} = \sum_{j=1}^{k} \boldsymbol{\mu}_j \boldsymbol{v}_j^T + \boldsymbol{Q}$ where $\boldsymbol{Q}$ is a $p \times n$ standard Gaussian matrix. The first inequality in Equation (3.16) (i.e. a lower bound on the data dimension p) is sufficient for $(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}$ to have the desired property; the major technical challenge is that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ involves additional terms that intricately depend on the label matrix $\boldsymbol{Y}$ itself. Our key technical contribution is showing that these extra terms do *not* drastically change the desired behavior, provided that the norms of the mean vectors (i.e. signal strength) are sufficiently small. At a high-level we accomplish this with a recursive argument as follows. Denote $\boldsymbol{X}_0 = \boldsymbol{Q}$ and $\boldsymbol{X}_i = \sum_{j=1}^{i} \boldsymbol{\mu}_j \boldsymbol{v}_j^T + \boldsymbol{Q}$ for $i \in [k]$. Then, at each stage i of the recursion, we show how to bound quadratic forms involving $\left(\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1}$ using bounds established previously at stage $i-1$ on quadratic forms involving $\left(\boldsymbol{X}_{i-1}^T\boldsymbol{X}_{i-1}\right)^{-1}$. A critical property for the success of our proof strategy is the observation that the rows of $\boldsymbol{Y}$ are always orthogonal, that is, $\boldsymbol{v}_i^T\boldsymbol{v}_j = 0$, for $i \neq j$. The complete proof of the theorem is given in Section 3.6.2.

We first present numerical results that support the conclusions of Theorem 9. (In all our figures, we show averages over 100 Monte-Carlo realizations, and the error bars show the standard deviation at each point.) Figure 3.3(a) plots the fraction of support vectors satisfying Equation (3.14) as a function of training size n. We fix dimension $p = 1000$ and class priors $\pi = \frac{1}{k}$. To study how the outcome depends on the number of classes k and signal strength $\|\boldsymbol{\mu}\|_2$, we consider $k = 4, 7$ and three equal-energy scenarios where $\forall c \in [k] : \|\boldsymbol{\mu}_c\|_2 = \|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$ with $\mu = 0.2, 0.3, 0.4$. Observe that smaller μ results in larger proportion of support vectors for the same value of n. To verify our theorem's second condition (on the signal strength) in Equation (3.16), Figure 3.3(a) also plots the same set of curves over a re-scaled axis $k^{1.5}n^{1.5}\|\boldsymbol{\mu}\|_2/p$. The six curves corresponding to different settings nearly overlap in this new scaling, showing that the



(a) $k = 4$ and 7 (b) $k = 3$ and 6

Figure 3.3: Fraction of training examples satisfying Equation (3.14) (also called "support vectors") in the GMM case. The error bars show the standard deviation. Figure (a) considers $k = 4$ and 7, and Figure (b) considers $k = 3$ and 6. On the legend, "(4) 0.3" corresponds to $k = 4$ and $\|\boldsymbol{\mu}\|_2/\sqrt{p} = 0.2$. Observe that the curves nearly overlap when plotted versus $k^{1.5}n^{1.5}\|\boldsymbol{\mu}\|_2/p$ as predicted by the second condition in Equation (3.16) of Theorem 9.

condition is order-wise tight. In Figure 3.3(b), we repeat the experiment in Figure 3.3(a) for different values of $k = 3$ and $k = 6$. Again, these curves nearly overlap when the x-axis is scaled according to the second condition on signal strength in Equation (3.16). We conjecture that our second condition on the signal strength is tight up to an extra $\sqrt{n}$ factor, which we believe is an artifact of the analysis[3]. We also believe that the k^3 factor in the first condition can be relaxed slightly to k^2 (as in the MLM case depicted in Figure 3.4, which considers a rescaled x-axis and shows *exact* overlap of the curves for all values of k). Sharpening these dependences on both k and n is an interesting direction for future work.

Multinomial logistic model

We now consider the MLM data model and anisotropic data covariance. Explicitly, the eigendecomposition of the covariance matrix is given by $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$, where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_p]$. We also define the effective dimensions $d_2 := \|\boldsymbol{\lambda}\|_1^2/\|\boldsymbol{\lambda}\|_2^2$ and $d_\infty :=$

[3]Support for this belief comes from the fact that [164] shows that $p > C_2\|\boldsymbol{\mu}\|_2 n$ is sufficient for the SVM = interpolation phenomenon to occur in the case of GMM and *binary* classification.

$\|\boldsymbol{\lambda}\|_1/\|\boldsymbol{\lambda}\|_\infty$. The following result contains sufficient conditions for the SVM and MNI solutions to coincide.

Theorem 10. *Assume n training samples following the MLM defined in* (3.2)*. There exist constants c and $C_1, C_2 > 1$ such that Equation* (3.12) *holds with probability at least $(1 - \frac{c}{n})$ provided that*

$$d_\infty > C_1 k^2 n \log(kn) \text{ and } d_2 > C_2(\log(kn) + n). \tag{3.17}$$

In fact, the only conditions we require on the generated labels is conditional independence.

For the isotropic case $\boldsymbol{\Sigma} = \mathbf{I}_p$, this implies that Equation (3.12) *holds with probability at least $(1 - \frac{c}{n})$ provided that*

$$p > C_1 k^2 n \log(kn). \tag{3.18}$$

The sufficient conditions in Theorem 10 require that the spectral structure in the covariance matrix $\boldsymbol{\Sigma}$ has sufficiently slowly decaying eigenvalues (corresponding to sufficiently large d_2), and that it is not too "spiky" (corresponding to sufficiently large d_∞). When $\boldsymbol{\Sigma} = \mathbf{I}_p$, the conditions reduce to sufficient overparameterization. For the special case of $k = 2$ classes, our conditions reduce to those in [72] for binary classification. The dominant dependence on k, given by k^2, is a byproduct of the "unequal" margin in Equation (3.10). Figure 3.4 empirically verifies the sharpness of this factor.

The proof of Theorem 10 is provided in Appendix B.2. We now numerically validate our results in Theorem 10 in Figure 3.4, focusing on the isotropic case. We fix $p = 1000$, vary n from 10 to 100 and the numbers of classes from $k = 3$ to $k = 6$. We choose orthogonal mean vectors for each class with equal energy $\|\boldsymbol{\mu}\|_2^2 = p$. The left-most plot in Figure 3.4 shows the fraction of support vectors satisfying Equation (3.14) as a

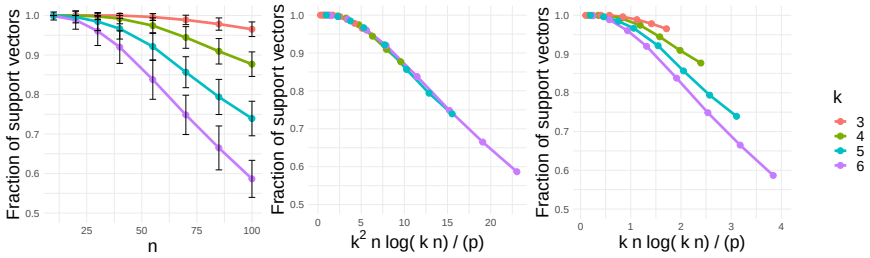


Figure 3.4: Fraction of training examples satisfying equality in the simplex label representation in Equation (3.14) in the MLM case with $\mathbf{\Sigma} = \mathbf{I}_p$. The middle plot shows that the curves overlap when plotted versus $k^2 n \log(kn)/p$ as predicted by Equation (3.18).

function of n. Clearly, smaller number of classes k results in higher proportion of support vectors with the desired property for the same number of measurements n. To verify the condition in Equation (3.18), the middle plot in Figure 3.4 plots the same curves over a re-scaled axis $k^2 n \log(kn)/p$ (as suggested by Equation (3.18)). We additionally draw the same curves over $kn \log(kn)/p$ in the right-most plot of Figure 3.3. Note the overlap of the curves in the middle plot. We now numerically validate our results in Theorem 10 in Figure 3.4, focusing on the isotropic case. We fix $p = 1000$, vary n from 10 to 100 and the numbers of classes from $k = 3$ to $k = 6$. We choose orthogonal mean vectors for each class with equal energy $\|\boldsymbol{\mu}\|_2^2 = p$. The left-most plot in Figure 3.4 shows the fraction of support vectors satisfying Equation (3.14) as a function of n. Clearly, smaller number of classes k results in higher proportion of support vectors with the desired property for the same number of measurements n. To verify the condition in Equation (3.18), the middle plot in Figure 3.4 plots the same curves over a re-scaled axis $k^2 n \log(kn)/p$ (as suggested by Equation (3.18)). We additionally draw the same curves over $kn \log(kn)/p$ in the right-most plot of Figure 3.3. Note the overlap of the curves in the middle plot.

3.3.3 Connection to Neural Collapse

In this section, we provide a distinct set of sufficient conditions on the feature vectors that guarantee Equation (3.12), and hence the conclusions of Theorem 8 hold. Interestingly, these sufficient conditions relate to the recently discovered, so called *neural-collapse* phenomenon that is empirically observed in the training process of overparameterized deep nets [125] (see also e.g. [172, 115, 63, 105, 48, 49, 130, 62] for several recent follow-ups).

Corollary 10.1. *Recall the notation in Theorem 8. Assume exactly balanced data, that is* $|\{i : y_i = c\}| = n/k$ *for all* $c \in [k]$. *Also, assume that the following two conditions hold:*

- ***Feature collapse (NC1):*** *For each* $c \in [k]$ *and all* $i \in [n] : y_i = c$, *it holds that* $\boldsymbol{x}_i = \boldsymbol{\mu}_c$, *where* $\boldsymbol{\mu}_c \triangleq \frac{k}{n} \sum_{i:y_i=c} \boldsymbol{x}_i$ *is the "mean" vector of the corresponding class.*

- ***Simplex ETF structure (NC2)****: The matrix of mean vectors,*

$$\boldsymbol{M} := [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k]_{p\times k},$$

is the matrix of a simplex Equiangular Tight Frame (ETF), i.e., for some orthogonal matrix $\mathbf{U}_{p\times k}$ *(with* $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k$*) and* $\alpha \in \mathbb{R}$, *it holds that*

$$\boldsymbol{M} = \alpha\sqrt{\frac{k}{n}}\mathbf{U}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T\right). \tag{3.19}$$

Then, the sufficient condition (3.12) *of Theorem 8 holds for the Gram matrix* $\boldsymbol{X}^T\boldsymbol{X}$.

Proof. For simplicity, denote the sample size of each class as $m := n/k$. Without loss of generality under the corollary's assumptions, let the columns of the feature matrix $\boldsymbol{X}$ be ordered such that $\boldsymbol{X} = [\boldsymbol{M}, \boldsymbol{M}, \ldots, \boldsymbol{M}] = \boldsymbol{M} \otimes \mathbf{1}_m^T$. Accordingly, we have $\boldsymbol{z}_c =$

$(\mathbf{e}_c \otimes \mathbf{1}_m) - \frac{1}{k}(\mathbf{1}_k \otimes \mathbf{1}_m)$ where $\mathbf{e}_c$ is the c-th basis vector in $\mathbb{R}^k$. Then, the feature Gram matrix is computed as

$$\boldsymbol{X}^T\boldsymbol{X} = \left(\boldsymbol{M}^T\boldsymbol{M}\right) \otimes \left(\mathbf{1}_m\mathbf{1}_m^T\right) = \frac{\alpha^2}{m}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) \otimes \left(\mathbf{1}_m\mathbf{1}_m^T\right). \tag{3.20}$$

Observe here that we can write $\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) = \boldsymbol{V}\boldsymbol{V}^T$ for $\boldsymbol{V} \in \mathbb{R}^{k\times(k-1)}$ having orthogonal columns (i.e. $\boldsymbol{V}^T\boldsymbol{V} = \mathbf{I}_{k-1}$) and $\boldsymbol{V}^T\mathbf{1}_k = \mathbf{0}_k$. Using this and the fact that $(\boldsymbol{V}\boldsymbol{V}^T)^+ = (\boldsymbol{V}\boldsymbol{V}^T)$, it can be checked from (3.20) that

$$(\boldsymbol{X}^T\boldsymbol{X})^+ = \frac{1}{\alpha^2 m}\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) \otimes \left(\mathbf{1}_m\mathbf{1}_m^T\right). \tag{3.21}$$

Putting things together, we get, for any $c \in [k]$, that

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{z}_c =& \frac{1}{\alpha^2 m}\left(\left(\mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T\right) \otimes \left(\mathbf{1}_m\mathbf{1}_m^T\right)\right)(\mathbf{e}_c \otimes \mathbf{1}_m)\\ =& \frac{1}{\alpha^2}\left(\mathbf{e}_c - \frac{1}{k}\mathbf{1}_k\right) \otimes \mathbf{1}_m = \frac{1}{\alpha^2}\boldsymbol{z}_c.\end{aligned}$$

Therefore, it follows immediately that

$$\boldsymbol{z}_c \odot \boldsymbol{M}^+\boldsymbol{z}_c = \frac{1}{\alpha^2}\boldsymbol{z}_c \odot \boldsymbol{z}_c > \mathbf{0},$$

as desired. This completes the proof. □

It might initially appear that the structure of the feature vectors imposed by the properties NC1 and NC2 is too specific to be relevant in practice. To the contrary, [125] showed via a principled experimental study that these properties occur at the last layer of overparameterized deep nets across several different data sets and DNN architectures. Specifically, the experiments conducted in [125] suggest that training overparameterized

deep nets on classification tasks with CE loss in the absence of weight decay (i.e., without explicit regularization) results in learned feature representations in the final layer that converge[4] to the ETF structure described by NC1 and NC2. Furthermore, it was recently shown in [62] that the neural collapse phenomenon continues to occur when the last-layer features of a deep net are trained with the recently proposed *supervised contrastive loss* (SCL) function [83] and a linear model is independently trained on these learned last-layer features. (In fact, [62, 83] showed that this self-supervised procedure can yield superior generalization performance compared to CE loss.)

To interpret Corollary 10.1 in view of these findings, consider the following two-stage classification training process:

- First, train (without weight-decay and continuing training beyond the interpolation regime) the last-layer feature representations of an overparameterized deep-net with either CE or SCL losses.

- Second, taking as inputs those learned feature representations of the first stage, train a linear multiclass classifier (often called the "head" of the deep-net) with CE loss.

Then, from Corollary 10.1, the resulting classifier from this two-stage process interpolates the simplex label representation, and the classification accuracy is the same as if we had used the square loss in the second stage of the above training process. Thus, our results lend strong theoretical justification to the empirical observation that *square-loss and CE loss yield near-identical performance in large-scale classification tasks [137, 136, 74, 131].*

[4]Here, "convergence" is with respect to an increasing number of training epochs. Since the architecture is overparameterized, it can perfectly separate the data. Hence, the training 0-1 error can be driven to zero. Nevertheless, training continues despite having achieved zero 0-1 training error, since the CE loss continues to drop. [125] refers to this regime as the terminal phase of training (TPT). In sum, [125] show that neural collapse is observed in TPT.

3.4 Generalization bounds

In this section, we derive non-asymptotic bounds on the error of the MNI classifier for data generated from GMM.

3.4.1 Gaussian mixture model

We present classification error bounds under the additional assumption of orthogonal means for ease of exposition — this can be relaxed with some additional work as described in Appendix B.3.1.

Assumption 3 (Orthogonal means). In addition to Assumption 2, we assume that the means are orthogonal, that is $\boldsymbol{\mu}_c^T \boldsymbol{\mu}_j = 0$, for all $c \neq j \in [k]$.

Theorem 11. *Let Assumption 3 and the condition in Equation* (3.16) *hold. Further assume constants* $C_1, C_2, C_3 > 1$ *such that* $\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right) \|\boldsymbol{\mu}\|_2 > C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\}$. *Then, there exist additional constants* c_1, c_2, c_3 *and* $C_4 > 1$ *such that both the MNI solution* $\boldsymbol{W}_{MNI}$ *and the multiclass SVM solution* $\boldsymbol{W}_{SVM}$ *satisfy*

$$\mathbb{P}_{e|c} \leq (k-1) \exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right) \|\boldsymbol{\mu}\|_2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\}\right)^2}{C_4 \left(1 + \frac{kp}{n\|\boldsymbol{\mu}\|_2^2}\right)} \right) \tag{3.22}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, *for every* $c \in [k]$. *Moreover, the same bound holds for the total classification error* $\mathbb{P}_e$.

For large enough n, Theorem 11 reduces to the results in [164] when $k = 2$ (with slightly different constants). There are two major challenges in the proof of Theorem 11, which is presented in Appendix B.3.1. First, in contrast to the binary case the classification error does *not* simply reduce to bounding correlations between vector means $\boldsymbol{\mu}_c$

and their estimators $\hat{\boldsymbol{w}}_c$. Second, just as in the proof of Theorem 9, technical complications arise from the multiple mean components in the training data matrix $\boldsymbol{X}$. We use a variant of the recursion-based argument described in Section 3.6.2 to obtain our final bound.

3.5 Conditions for benign overfitting

Thus far, we have studied the classification error of the MNI classifier under the GMM data model (Theorem 11), and shown equivalence of the multiclass SVM and MNI solutions (Theorems 8, 9 and Corollary 8.1). Combining these results, we now provide sufficient conditions under which the classification error of the multiclass SVM solution (also of the MNI) approaches 0 as the number of parameters p increases. First, we state our sufficient conditions for harmless interpolation under the GMM model — these arise as a consequence of Theorem 11, and the proof is provided in Appendix B.3.2.

Corollary 11.1. *Let the same assumptions as in Theorem 11 hold. Then, for finite number of classes k and sufficiently large sample size n, there exist positive constants c_i's and C_i's > 1, such that the multiclass SVM classifier $\boldsymbol{W}_{SVM}$ in* (3.7) *satisfies the simplex interpolation constraint in* (3.14) *and its total classification error approaches* 0 *as $\left(\frac{p}{n}\right) \to \infty$ with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, provided that the following conditions hold:*

(1). When $\|\boldsymbol{\mu}\|_2^2 > \frac{kp}{n}$,

$$\frac{n}{C_1 k}\|\boldsymbol{\mu}\|_2^2 > p > \max\{C_2 k^3 n \log(kn) + n - 1, C_3 k^{1.5} n^{1.5} \|\boldsymbol{\mu}\|_2\}.$$

(2). When $\|\boldsymbol{\mu}\|_2^2 \leq \frac{kp}{n}$,

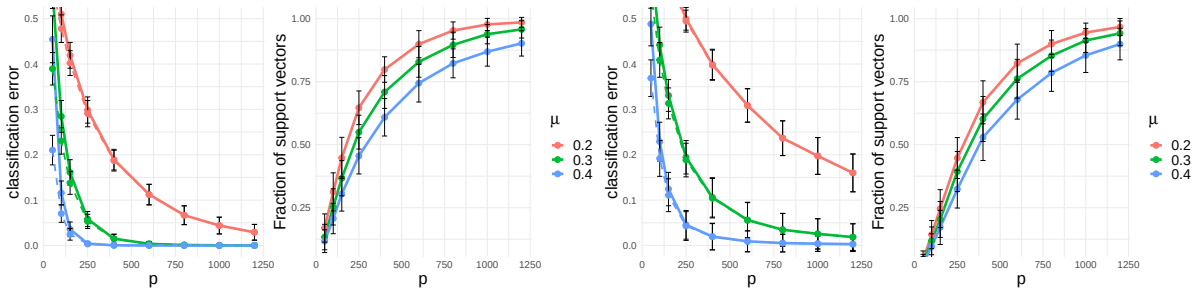
$$p > \max\{C_2 k^3 n \log(kn) + n - 1, C_3 k^{1.5} n^{1.5} \|\boldsymbol{\mu}\|_2, \frac{n\|\boldsymbol{\mu}\|_2^2}{k}\},$$
$$\textit{and } \|\boldsymbol{\mu}\|_2^4 \geq C_4 \left(\frac{p}{n}\right)^{\alpha}, \quad \textit{for } \alpha > 1.$$

When n is fixed, the conditions for benign overfitting for $\boldsymbol{W}_{SVM}$ become

$$\|\boldsymbol{\mu}\|_2 = \Theta(p^{\beta}), \ \textit{for } \beta \in (1/4, 1).$$

Note that the upper bound on $\|\boldsymbol{\mu}\|_2$ comes from the conditions that make SVM=MNI in Theorem 9; indeed, a distinct corollary of Theorem 11 is that $\boldsymbol{W}_{\mathrm{MNI}}$ overfits benignly with sufficient signal strength $\|\boldsymbol{\mu}\|_2 = \Omega(p^{1/4})$. We can compare our result with the binary case [164]. When k and n are both finite, the condition $\|\boldsymbol{\mu}\|_2 = \Theta(p^{\beta})$ for $\beta \in (1/4, 1)$ is the same as the binary result.

We particularly note that, like in the binary case, Corollaries 11.1 imply benign overfitting in regimes that cannot be explained by classical *training-data-dependent* bounds based on the margin [144]. While the shortcomings of such margin-based bounds in the highly overparameterized regime are well-documented, e.g. [44], we provide a brief description here for completeness. For the GMM, we verify here that the margin-based bounds could only predict benign overfitting if we had the significantly stronger condition $\beta \in (1/2, 1)$ (see also [164, Section 9.1]): in the regime where SVM = MNI, the margin is exactly equal to 1. The margin-based bounds (as given in, e.g. [9]), can be verified to scale as $\mathcal{O}\left(\sqrt{\frac{\mathrm{trace}(\boldsymbol{\Sigma}_{\mathrm{un}})}{n\|\boldsymbol{\Sigma}_{\mathrm{un}}\|_2}}\right)$ with high probability, where $\boldsymbol{\Sigma}_{\mathrm{un}} := \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\top}\right]$ denotes the *unconditional* covariance matrix under the GMM. In the case of the binary GMM and isotropic noise covariance, an elementary calculation shows that the spectrum of $\boldsymbol{\Sigma}_{\mathrm{un}}$ is given by $\begin{bmatrix}\|\boldsymbol{\mu}\|_2^2 + 1 & 1 & \dots & 1\end{bmatrix}$; plugging this into the above bound requires $\|\boldsymbol{\mu}\|_2^2 \gg \frac{p}{n}$



(a) $k = 4$ (b) $k = 6$

Figure 3.5: Evolution of total classification error and fraction of support vectors as a function of p in the GMM case. Figure (a) considers $k = 4$ and Figure (b) considers $k = 6$. We consider the energy of all class means to be $\|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$, where $\mu = 0.2, 0.3$ and 0.4. Observe that the total classification error approaches 0 and the fraction of support vectors approaches 1 as p gets larger.

for the margin-based upper bound to scale as $o(1)$. This clearly does not explain benign overfitting when SVM = MNI, which we showed requires $\|\boldsymbol{\mu}\|_2^2 \leq \frac{p}{n}$.

Finally, we present numerical illustrations validating our benign overfitting results in Corollary 11.1. In Figure 3.5(a), we set the number of classes $k = 4$. To guarantee sufficient overparameterization, we fix $n = 40$ and vary p from 50 to 1200. We simulate 3 different settings for the mean matrices: each has orthogonal and equal-norm mean vectors $\|\boldsymbol{\mu}\|_2 = \mu\sqrt{p}$, with $\mu = 0.2, 0.3$ and 0.4. Figure 3.5 plots the classification error as a function of p for both MNI estimates (solid lines) and multiclass SVM solutions (dashed lines). Different colors correspond to different mean norms. The solid and dashed curves almost overlap as predicted from our results in Section 3.3. We verify that as p increases, the classification error decreases towards zero. Observe that the fraction of support vectors approaches 1 as p gets larger. Further, the classification error goes to zero very fast when μ is large, but then the proportion of support vectors increases at a slow rate. In contrast, when μ is small, the proportion of support vectors increases fast, but the classification error decreases slowly. Figure 3.5(b) uses the same setting as in Figure 3.5(a) except for setting $k = 6$ and $n = 30$. Observe that the classification error

continues to go to zero and the proportion of support vectors continues to increase, but both become slower as the number of classes is now greater.

3.6 Proofs of main results

In this section, we provide the proofs of Theorems 8 and 9. The proof techniques we developed for these results convey novel technical ideas that also form the core of the rest of the proofs, which we defer to the Appendix B.

3.6.1 Proof of Theorem 8

Argument sketch. We split the proof of the theorem in three steps. To better convey the main ideas, we first outline the three steps in this paragraph before discussing their details in the remaining of this section.

Step 1: The first key step to prove Theorem 8 is constructing a new parameterization of the dual of the multiclass SVM, which we show takes the following form:

$$\max_{\boldsymbol{\beta}_c \in \mathbb{R}^n, c \in [k]} \quad \sum_{c \in [k]} \boldsymbol{\beta}_c^T \boldsymbol{z}_c - \frac{1}{2} \|\boldsymbol{X} \boldsymbol{\beta}_c\|_2^2 \tag{3.23}$$

$$\text{sub. to} \qquad \beta_{y_i,i} = -\sum_{c \neq y_i} \beta_{c,i}, \ \forall i \in [n] \quad \text{and} \quad \boldsymbol{\beta}_c \odot \boldsymbol{z}_c \geq \boldsymbol{0}, \forall c \in [k].$$

Here, for each $c \in [k]$ we let $\boldsymbol{\beta}_c = [\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,n}] \in \mathbb{R}^n$. We also show by complementary slackness the following implication for any *optimal* $\beta^*_{c,i}$ in (3.23):

$$z_{c,i} \beta^*_{c,i} > 0 \implies (\hat{\boldsymbol{w}}_{y_i} - \hat{\boldsymbol{w}}_c)^T \boldsymbol{x}_i = 1. \tag{3.24}$$

Thus, to prove Equation (3.13), it will suffice showing that $z_{c,i} \beta^*_{c,i} > 0, \forall i \in [n], c \in [k]$ provided that Equation (3.12) holds.

Step 2: To do this, we prove that the *unconstrained* maximizer in (3.23), that is $\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{z}_c,\ \forall c \in [k]$ is feasible, and therefore optimal, in (3.23). Now, note that Equation (3.12) is equivalent to $\boldsymbol{z}_c \odot \hat{\boldsymbol{\beta}}_c > 0$; thus, we have found that $\hat{\boldsymbol{\beta}}_c, c \in [k]$ further satisfies the n *strict* inequality constraints in (3.24) which completes the proof of the first part of the theorem (Equation (3.13)).

Step 3: Next, we outline the proof of Equation (3.14). We consider the simplex-type OvA-classifier in (3.15). The proof has two steps. First, using similar arguments to what was done above, we show that when Equation (3.12) holds, then all the inequality constraints in (3.15) are active at the optimal. That is, the minimizers $\boldsymbol{w}_{\mathrm{OvA},c}$ of (3.15) satisfy Equation (3.14). Second, to prove that Equation (3.14) is satisfied by the minimizers $\hat{\boldsymbol{w}}_c$ of the multiclass SVM in (3.7), we need to show that $\boldsymbol{w}_{\mathrm{OvA},c} = \hat{\boldsymbol{w}}_c$ for all $c \in [k]$. We do this by showing that, under Equation (3.12), the duals of (3.7) and (3.15) are equivalent. By strong duality, the optimal costs of the primal problems are also the same. Then, because a) the objective is the same for the two primals, b) $\boldsymbol{w}_{\mathrm{OvA},c}$ is feasible in (3.15) and c) (3.7) is strongly convex, we can conclude with the desired.

Step 1: Key alternative parameterization of the dual. We start by writing the dual of the multiclass SVM, repeated here for convenience:

$$\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{W}\|_F^2 \quad \text{sub. to} \ \ (\boldsymbol{w}_{y_i} - \boldsymbol{w}_c)^\top \boldsymbol{x}_i \geq 1,\ \forall i \in [n], c \in [k] : c \neq y_i. \tag{3.25}$$

We have dual variables $\{\lambda_{c,i}\}$ for every $i \in [n], c \in [k] : c \neq y_i$ corresponding to the constraints on the primal form above. Then, the dual of the multiclass SVM takes the form

$$\max_{\lambda_{c,i} \geq 0} \sum_{i \in [n]} \Big(\sum_{\substack{c \in [k] \\ c \neq y_i}} \lambda_{c,i} \Big) - \frac{1}{2} \sum_{c \in [k]} \Big\| \sum_{i \in [n]: y_i = c} \Big(\sum_{\substack{c' \in [k] \\ c' \neq y_i}} \lambda_{c',i} \Big) \boldsymbol{x}_i - \sum_{i \in [n]: y_i \neq c} \lambda_{c,i} \boldsymbol{x}_i \Big\|_2^2. \tag{3.26}$$

Let $\hat{\lambda}_{c,i}, i \in [n], c \in [k] : c \neq y_i$ be maximizers in Equation (3.26). By complementary slackness, we have

$$\hat{\lambda}_{c,i} > 0 \implies (\hat{\boldsymbol{w}}_{y_i} - \hat{\boldsymbol{w}}_c)^\top \boldsymbol{x}_i = 1. \tag{3.27}$$

Thus, it will suffice to prove that $\hat{\lambda}_{c,i} > 0, \forall i \in [n], c \in [k] : c \neq y_i$ provided that (3.12) holds.

It is challenging to work directly with Equation (3.26) because the variables $\lambda_{c,i}$ are coupled in the objective function. Our main idea is to re-parameterize the dual objective in terms of new variables $\{\beta_{c,i}\}$, which we define as follows for all $c \in [k]$ and $i \in [n]$:

$$\beta_{c,i} = \begin{cases} \sum_{c' \neq y_i} \lambda_{c',i} & , y_i = c, \\ -\lambda_{c,i} & , y_i \neq c. \end{cases} \tag{3.28}$$

For each $c \in [k]$, we denote $\boldsymbol{\beta}_c = [\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,n}] \in \mathbb{R}^n$. With these, we show that the dual objective becomes

$$\sum_{c \in [k]} \boldsymbol{\beta}_c^\top \boldsymbol{z}_c - \frac{1}{2} \sum_{c \in [k]} \Big\| \sum_{i \in [n]} \beta_{c,i} \boldsymbol{x}_i \Big\|_2^2 = \sum_{c \in [k]} \boldsymbol{\beta}_c^\top \boldsymbol{z}_c - \frac{1}{2} \|\boldsymbol{X} \boldsymbol{\beta}_c\|_2^2. \tag{3.29}$$

The equivalence of the quadratic term in $\boldsymbol{\beta}$ is straightforward. To show the equivalence of the linear term in $\boldsymbol{\beta}$, we denote $A := \sum_{i \in [n]} \Big(\sum_{c \in [k], c \neq y_i} \lambda_{c,i} \Big)$, and simultaneously get

$$A = \sum_{i \in [n]} \beta_{y_i, i} \qquad \text{and} \qquad A = \sum_{i \in [n]} \sum_{c \neq y_i} (-\beta_{c,i}),$$

by the definition of variables $\{\beta_{c,i}\}$ in Equation (3.28). Then, we have

$$\begin{aligned} A = \frac{k-1}{k} \cdot A + \frac{1}{k} \cdot A &= \frac{k-1}{k} \sum_{i \in [n]} \beta_{y_i,i} + \frac{1}{k} \sum_{i \in [n]} \sum_{c \neq y_i} (-\beta_{c,i}) \\ &\overset{\text{(i)}}{=} \sum_{i \in [n]} \boldsymbol{z}_{y_i,i} \beta_{y_i,i} + \sum_{i \in [n]} \sum_{c \neq y_i} \boldsymbol{z}_{c,i} \beta_{c,i} \\ &= \sum_{i \in [n]} \sum_{c \in [k]} \boldsymbol{z}_{c,i} \beta_{c,i} = \sum_{c \in [k]} \boldsymbol{\beta}_c^\top \boldsymbol{z}_c. \end{aligned}$$

Above, inequality (i) follows from the definition of $\boldsymbol{z}_c$ in Equation (3.11), rewritten coordinate-wise as:

$$z_{c,i} = \begin{cases} \frac{k-1}{k}, & y_i = c, \\ -\frac{1}{k}, & y_i \neq c. \end{cases}$$

Thus, we have shown that the objective of the dual can be rewritten in terms of variables $\{\beta_{c,i}\}$. After rewriting the constraints in terms of $\{\beta_{c,i}\}$, we have shown that the dual of the SVM (Equation (3.7)) can be equivalently written as in Equation (3.23). Note that the first constraint in (3.23) ensures consistency with the definition of $\boldsymbol{\beta}_c$ in Equation (3.28). The second constraint guarantees the non-negativity constraint of the original dual variables in (3.26), because we have

$$\beta_{c,i} z_{c,i} = \frac{\lambda_{c,i}}{k} \text{ for all } i \in [n], c \in [k] : c \neq y_i.$$

Consequently, we have

$$\beta_{c,i} z_{c,i} \geq 0 \iff \lambda_{c,i} \geq 0 \tag{3.30}$$

for all $c \in [k]$ and $i \in [n] : y_i \neq c$. In fact, the equivalence above also holds with

the inequalities replaced by strict inequalities. Also note that the second constraint for $c = y_i$ yields $\frac{k-1}{k} \sum_{c' \neq y_i} \lambda_{c',i} \geq 0$, which is automatically satisfied when Equation (3.30) is satisfied. Thus, these constraints are redundant.

Step 2: Proof of Equation (3.13)**.** Define

$$\hat{\boldsymbol{\beta}}_c := (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{z}_c, \ \forall c \in [k].$$

This specifies an *unconstrained* maximizer in (3.23). We will show that this unconstrained maximizer $\hat{\boldsymbol{\beta}}_c, c \in [k]$ is feasible in the constrained program in (3.23). Thus, it is in fact an optimal solution in (3.23).

To prove this, we will first prove that $\hat{\boldsymbol{\beta}}_c, c \in [k]$ satisfies the n equality constraints in (3.23). For convenience, let $\boldsymbol{g}_i \in \mathbb{R}^n, i \in [n]$ denote the i-th row of $(\boldsymbol{X}^\top \boldsymbol{X})^+$. Then, for the i-th element $\hat{\beta}_{c,i}$ of $\hat{\boldsymbol{\beta}}_c$, it holds that $\hat{\beta}_{c,i} = \boldsymbol{g}_i^\top \boldsymbol{z}_c$. Thus, for all $i \in [n]$, we have

$$\hat{\beta}_{y_i,i} + \sum_{c \neq y_i} \hat{\beta}_{c,i} = \boldsymbol{g}_i^\top \Big(\boldsymbol{z}_{y_i} + \sum_{c \neq y_i} \boldsymbol{z}_c \Big) = \boldsymbol{g}_i^\top \Big(\sum_{c \in [k]} \boldsymbol{z}_c \Big) = 0,$$

where in the last equality we used the definition of $\boldsymbol{z}_c$ in (3.11) and the fact that $\sum_{c \in [k]} \boldsymbol{v}_c = \mathbf{1}_n$, since each column of the label matrix $\boldsymbol{Y}$ has exactly one non-zero element equal to 1. Second, since Equation (3.12) holds, $\hat{\boldsymbol{\beta}}_c, c \in [k]$ further satisfies the n *strict* inequality constraints in (3.23).

We have shown that the unconstrained maximizer is feasible in the constrained program (3.23). Thus, we can conclude that it is also a global solution to the latter. By Equation (3.30), we note that the original dual variables $\{\lambda_{c,i}\}$ are all strictly positive. This completes the proof of the first part of the theorem, i.e. the proof of Equation (3.13).

Step 3: Proof of Equation (3.14)**.** To prove Equation (3.14), consider the following

OvA-type classifier: for all $c \in [k]$,

$$\min_{\boldsymbol{w}_c} \ \frac{1}{2}\|\boldsymbol{w}_c\|_2^2 \qquad \text{sub. to} \quad \boldsymbol{x}_i^\top \boldsymbol{w}_c \begin{cases} \geq \frac{k-1}{k}, & y_i = c, \\ \leq -\frac{1}{k}, & y_i \neq c, \end{cases} \quad \forall i \in [n]. \tag{3.31}$$

To see the connection with Equation (3.14), note the condition for the constraints in (3.31) to be active is exactly Equation (3.14). Thus, it suffices to prove that the constraints of (3.31) are active under the theorem's assumptions. We work again with the dual of (3.31):

$$\max_{\boldsymbol{\nu}_c \in \mathbb{R}^k} \ -\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\nu}_c\|_2^2 + \boldsymbol{z}_c^\top \boldsymbol{\nu}_c \qquad \text{sub. to} \quad \boldsymbol{z}_c \odot \boldsymbol{\nu}_c \geq \boldsymbol{0}. \tag{3.32}$$

Again by complementary slackness, the desired Equation (3.14) holds provided that all dual constraints in (3.32) are strict at the optimal.

We now observe two critical similarities between (3.32) and (3.23): (i) the two dual problems have the same objectives (indeed the objective in (3.23) is separable over $c \in [k]$); (ii) they share the constraint $\boldsymbol{z}_c \odot \boldsymbol{\nu}_c \geq \boldsymbol{0}$ / $\boldsymbol{z}_c \odot \boldsymbol{\beta}_c \geq \boldsymbol{0}$. From this observation, we can use the same argument as for (3.23) to show that when Equation (3.12) holds, $\hat{\boldsymbol{\beta}}_c$ is optimal in (3.32).

Now, let $\mathrm{OPT}_{(3.25)}$ and $\mathrm{OPT}^{\mathrm{c}}_{(3.31)}$ be the optimal costs of the multiclass SVM in (3.25) and of the simplex-type OvA-SVM in (3.31) parameterized by $c \in [k]$. Also, denote $\mathrm{OPT}_{(3.23)}$ and $\mathrm{OPT}^{\mathrm{c}}_{(3.32)}, \mathrm{c} \in [\mathrm{k}]$ the optimal costs of their respective duals in (3.23) and (3.32), respectively. We proved above that

$$\mathrm{OPT}_{(3.23)} = \sum_{\mathrm{c} \in [\mathrm{k}]} \mathrm{OPT}^{\mathrm{c}}_{(3.32)}. \tag{3.33}$$

Further let $\boldsymbol{W}_{\mathrm{OvA}} = [\boldsymbol{w}_{\mathrm{OvA},1}, \ldots, \boldsymbol{w}_{\mathrm{OvA},k}]$ be the optimal solution in the simplex-type

OvA-SVM in (3.32). We have proved that under Equation (3.12) $\boldsymbol{w}_{\mathrm{OvA},c}$ satisfies the constraints in (3.31) with equality, that is $\boldsymbol{X}^\top \boldsymbol{w}_{\mathrm{OvA},c} = \boldsymbol{z}_c,\ \forall c \in [k]$. Thus, it suffices to prove that $\boldsymbol{W}_{\mathrm{OvA}} = \boldsymbol{W}_{\mathrm{SVM}}$. By strong duality (which holds trivially for (3.31) by Slater's conditions), we get

$$\begin{aligned}
\mathrm{OPT}^{\mathrm{c}}_{(3.31)} = \mathrm{OPT}^{\mathrm{c}}_{(3.32)},\ \mathrm{c} \in [\mathrm{k}] \implies & \sum_{c\in[k]} \mathrm{OPT}^{\mathrm{c}}_{(3.31)} = \sum_{\mathrm{c}\in[\mathrm{k}]} \mathrm{OPT}^{\mathrm{c}}_{(3.32)} \\
\overset{(3.33)}{\implies} & \sum_{c\in[k]} \mathrm{OPT}^{\mathrm{c}}_{(3.31)} = \mathrm{OPT}_{(3.23)} \\
\overset{(3.31)}{\implies} & \sum_{c\in[k]} \frac{1}{2}\|\boldsymbol{w}_{\mathrm{OvA},c}\|_2^2 = \mathrm{OPT}_{(3.23)}. \qquad (3.34)
\end{aligned}$$

Again, by strong duality we get $\mathrm{OPT}_{(3.23)} = \mathrm{OPT}_{(3.25)}$. Thus, we have

$$\sum_{c\in[k]} \frac{1}{2}\|\boldsymbol{w}_{\mathrm{OvA},c}\|_2^2 = \mathrm{OPT}_{(3.25)}.$$

Note also that $\boldsymbol{W}_{\mathrm{OvA}}$ is feasible in (3.25) since

$$\boldsymbol{X}^\top \boldsymbol{w}_{\mathrm{OvA},c} = \boldsymbol{z}_c,\ \forall c \in [k] \implies (\boldsymbol{w}_{\mathrm{OvA},y_i} - \boldsymbol{w}_{\mathrm{OvA},c})^\top \boldsymbol{x}_i = 1,\ \forall c \neq y_i, c \in [k],\ \text{and } \forall i \in [n].$$

Therefore, $\boldsymbol{W}_{\mathrm{OvA}}$ is optimal in (3.25). Finally, note that the optimization objective in (3.25) is strongly convex. Thus, it has a unique minimum and therefore $\boldsymbol{W}_{\mathrm{SVM}} = \boldsymbol{W}_{\mathrm{OvA}}$ as desired.

3.6.2 Proof of Theorem 9

In this section, we provide the proof of Theorem 9. First, we remind the reader of the prescribed approach outlined in Section 3.3.2 and introduce some necessary notation. Second, we present the key Lemma 7, which forms the backbone of our proof. The proof

of the lemma is rather technical and is deferred to Appendix B.1.1 along with a series of auxiliary lemmas. Finally, we end this section by showing how to prove Theorem 9 using Lemma 7.

Argument sketch and notation. We begin by presenting high-level ideas and defining notation that is specific to this proof. For $c \in [k]$, we define

$$\boldsymbol{A}_c := (\mathbf{Q} + \sum_{j=1}^{c} \boldsymbol{\mu}_j \boldsymbol{v}_j^T)^T (\mathbf{Q} + \sum_{j=1}^{c} \boldsymbol{\mu}_j \boldsymbol{v}_j^T).$$

Recall that in the above, $\boldsymbol{\mu}_j$ denotes the j^{th} class mean of dimension p, and $\boldsymbol{v}_j$ denotes the n-dimensional indicator that each training example is labeled as class j. Since we have made an equal-energy assumption on the class means (Assumption 3), we will denote $\|\boldsymbol{\mu}\|_2 := \|\boldsymbol{\mu}_c\|_2$ throughout this proof as shorthand. Further, recall from Equation (3.1) that the feature matrix can be expressed as $\boldsymbol{X} = \boldsymbol{M}\boldsymbol{Y} + \boldsymbol{Q}$, where $\boldsymbol{Q} \in \mathbb{R}^{p \times n}$ is a standard Gaussian matrix. Thus, we have

$$\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{A}_k \qquad \text{and} \qquad \mathbf{Q}^T \mathbf{Q} = \boldsymbol{A}_0.$$

As discussed in Section 3.3.2, our goal is to show that the inverse Gram matrix $\boldsymbol{A}_k^{-1}$ is "close" to a positive definite diagonal matrix. Indeed, in our new notation, the desired inequality in Equation (3.12) becomes

$$z_{ci} \mathbf{e}_i^T \boldsymbol{A}_k^{-1} \boldsymbol{z}_c > 0, \quad \text{for all} \quad c \in [k] \quad \text{and } i \in [n]. \tag{3.35}$$

The major challenge in showing inequality (3.35) is that $\boldsymbol{A}_k = (\mathbf{Q} + \sum_{j=1}^{k} \boldsymbol{\mu}_j \boldsymbol{v}_j^T)^T (\mathbf{Q} + \sum_{j=1}^{k} \boldsymbol{\mu}_j \boldsymbol{v}_j^T)$ involves multiple mean components through the sum $\sum_{j=1}^{c} \boldsymbol{\mu}_j \boldsymbol{v}_j^T$. This makes it challenging to bound quadratic forms involving the Gram matrix $\boldsymbol{A}_k^{-1}$ directly.

Instead, our idea is to work recursively starting from bounding quadratic forms involving $\boldsymbol{A}_0^{-1}$. Specifically, we denote $\mathbf{P}_1 = \boldsymbol{Q} + \boldsymbol{\mu}_1 \boldsymbol{v}_1^T$ and derive the following recursion on the $\boldsymbol{A}_0, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ matrices:

$$\boldsymbol{A}_1 = \mathbf{P}_1^T \mathbf{P}_1 = \boldsymbol{A}_0 + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{Q}^T \boldsymbol{\mu}_1 & \boldsymbol{v}_1 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{\mu}_1^T \boldsymbol{Q} \end{bmatrix},$$

$$\boldsymbol{A}_2 = (\mathbf{P}_1 + \boldsymbol{\mu}_2 \boldsymbol{v}_2^T)^T (\mathbf{P}_1 + \boldsymbol{\mu}_2 \boldsymbol{v}_2^T) = \boldsymbol{A}_1 + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_2 & \mathbf{P}_1^T \boldsymbol{\mu}_2 & \boldsymbol{v}_2 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_2^T \\ \boldsymbol{v}_2^T \\ \boldsymbol{\mu}_2^T \mathbf{P}_1 \end{bmatrix}, \quad (3.36)$$

and so on, until $\boldsymbol{A}_k$ (see Appendix B.4.1 for the complete expressions for the recursion). Using this trick, we can exploit bounds on quadratic forms involving $\boldsymbol{A}_0^{-1}$ to obtain bounds for quadratic forms involving $\boldsymbol{A}_1^{-1}$, and so on until $\boldsymbol{A}_k^{-1}$.

There are two key ideas behind this approach. First, we will show how to use a leave-one-out argument and the Matrix Inversion Lemma to express (recursively) the quadratic form $\mathbf{e}_i^T \boldsymbol{A}_k^{-1} \boldsymbol{z}_c$ in (3.35) in terms of simpler quadratic forms, which are more accessible to bound directly. For later reference, we define these auxiliary forms here. Let $\mathbf{d}_c := \mathbf{Q}^T \boldsymbol{\mu}_c$, for $c \in [k]$ and define the following quadratic forms involving $\boldsymbol{A}_c^{-1}$ for $c, j, m \in [k]$ and $i \in [n]$:

$$\begin{aligned} s_{mj}^{(c)} &:= \boldsymbol{v}_m^T \boldsymbol{A}_c^{-1} \boldsymbol{v}_j, \\ t_{mj}^{(c)} &:= \mathbf{d}_m^T \boldsymbol{A}_c^{-1} \mathbf{d}_j, \\ h_{mj}^{(c)} &:= \mathbf{v}_m^T \boldsymbol{A}_c^{-1} \mathbf{d}_j, \\ g_{ji}^{(c)} &:= \mathbf{v}_j^T \boldsymbol{A}_c^{-1} \mathbf{e}_i, \\ f_{ji}^{(c)} &:= \mathbf{d}_j^T \boldsymbol{A}_c^{-1} \mathbf{e}_i. \end{aligned} \quad (3.37)$$

For convenience, we refer to terms above as *quadratic forms of order c* or *the c-th order quadratic forms*, where c indicates the corresponding superscript. A complementary useful observation facilitating our approach is the observation that the class label indicators are orthogonal by definition, i.e. $\boldsymbol{v}_i^T \boldsymbol{v}_j = 0$, for $i, j \in [k]$. (This is a consequence of the fact that any training data point has a unique label and we are using here one-hot encoding.) Thus, the newly added mean component $\boldsymbol{\mu}_{c+1}\boldsymbol{v}_{c+1}^T$ is orthogonal to the already existing mean components included in the matrix $\boldsymbol{A}_c$ (see Equation (3.36)). Consequently, we will see that adding new mean components will only slightly change the magnitude of these these quadratic forms as c ranges from 0 to k.

Identifying and bounding quadratic forms of high orders. Recall the desired inequality (3.35). We can equivalently write the definition of $\boldsymbol{z}_c$ in Equation (3.11) as

$$\boldsymbol{z}_c = \frac{k-1}{k}\boldsymbol{v}_c + \sum_{j \neq c}\left(-\frac{1}{k}\right)\boldsymbol{v}_j = \tilde{z}_{c(c)}\boldsymbol{v}_c + \sum_{j \neq c}\tilde{z}_{j(c)}\boldsymbol{v}_j, \tag{3.38}$$

where we denote

$$\tilde{z}_{j(c)} = \begin{cases} -\frac{1}{k}, & \text{if } \ j \neq c \\ \frac{k-1}{k}, & \text{if } \ j = c \end{cases}.$$

Note that by this definition, we have $\tilde{z}_{y_i(c)} := z_{ci}$. This gives us

$$\begin{aligned} z_{ci}\mathbf{e}_i^T \boldsymbol{A}_k^{-1}\boldsymbol{z}_c &= z_{ci}^2 \mathbf{e}_i^T \boldsymbol{A}_k^{-1}\boldsymbol{v}_{y_i} + \sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)}\mathbf{e}_i^T \boldsymbol{A}_k^{-1}\boldsymbol{v}_j, \\ &= z_{ci}^2 g_{y_i i}^{(k)} + \sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(k)}. \end{aligned} \tag{3.39}$$

Note that this expression (Equation (3.39)) involves the k-th order quadratic forms $g_{ji}^{(k)} = \mathbf{e}_i^T \boldsymbol{A}_k^{-1}\boldsymbol{v}_j$. For each such form, we use the matrix inversion lemma to leave the

j-th mean component in $\boldsymbol{A}_k$ out and express it in terms of the *leave-one-out* versions of quadratic forms that we defined in (3.37), as below (see Appendix B.4.1 for a detailed derivation):

$$g_{ji}^{(k)} = \mathbf{e}_i^T \boldsymbol{A}_k^{-1} \boldsymbol{v}_j = \frac{(1+h_{jj}^{(-j)})g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}}{s_{jj}^{(-j)}(\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(-j)}) + (1+h_{jj}^{(-j)})^2}\,. \tag{3.40}$$

Specifically, above we defined $s_{jj}^{(-j)} := \boldsymbol{v}_j^T \boldsymbol{A}_{-j}^{-1} \boldsymbol{v}_j$, where $\boldsymbol{A}_{-j}$ denotes the version of the Gram matrix $\boldsymbol{A}_k$ with the j-th mean component left out. The quadratic forms $h_{jj}^{(-j)}$, $f_{ji}^{(-j)}$, $g_{ji}^{(-j)}$ and $t_{jj}^{(-j)}$ are defined similarly in view of Equation (3.37).

Specifically, to see how these "leave-one-out" quadratic forms relate directly to the forms in Equation (3.37), note that it suffices in (3.40) to consider the case where $j = k$. Indeed, observe that when $j \neq k$ we can simply change the order of adding mean components, described in Equation (3.36), so that the j-th mean component is added last. On the other hand, when $j = k$ the leave-one-out quadratic terms in (3.40) involve the Gram matrix $\boldsymbol{A}_{k-1}$. Thus, they are equal to the quadratic forms of order $k-1$, given by $s_{kk}^{(k-1)}, t_{kk}^{(k-1)}$, $h_{kk}^{(k-1)}$, $g_{ki}^{(k-1)}$ and $f_{ki}^{(k-1)}$.

The following technical lemma bounds all of these quantities and its use is essential in the proof of Theorem 9. Its proof, which is deferred to Appendix B.1, relies on the recursive argument outlined above: We start from the quadratic forms of order 0 building up all the way to the quadratic forms of order $k-1$.

Lemma 7 (Quadratic forms of high orders)**.** *Let Assumption 2 hold and further assume that $p > Ck^3 n \log(kn) + n - 1$ for large enough constant $C > 1$ and large n. There exist constants c_i's and C_i's > 1 such that the following bounds hold for every $i \in [n]$ and*

$j \in [k]$ with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$,

$$\begin{aligned}
\frac{C_1 - 1}{C_1} \cdot \frac{n}{kp} \leq & s_{jj}^{(-j)} \leq \frac{C_1 + 1}{C_1} \cdot \frac{n}{kp}, \\
t_{jj}^{(-j)} \leq & \frac{C_2 n \|\boldsymbol{\mu}\|_2^2}{p}, \\
-\tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{k}p} \leq & h_{jj}^{(-j)} \leq \tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{k}p}, \\
|f_{ji}^{(-j)}| \leq & \frac{C_4 \sqrt{n} \|\boldsymbol{\mu}\|_2}{p}, \\
g_{ji}^{(-j)} \geq & \left(1 - \frac{1}{C_5}\right) \frac{1}{p}, \quad \textit{for } j = y_i, \\
|g_{ji}^{(-j)}| \leq & \frac{1}{C_6 k^2 p}, \quad \textit{for } j \neq y_i,
\end{aligned}$$

where $\tilde{\rho}_{n,k} = \min\{1, \sqrt{\log(2n)/k}\}$. Observe that the bounds stated in the lemma hold for any $j \in [k]$ and the bounds themselves are independent of j.

Completing the proof of Theorem 9. We now show how to use Lemma 7 to complete the proof of the theorem. Following the second condition in the statement of Theorem 9, we define

$$\epsilon_n := \frac{k^{1.5} n \sqrt{n} \|\boldsymbol{\mu}\|_2}{p} \leq \tau, \tag{3.41}$$

where τ is a sufficiently small positive constant, the value of which will be specified later in the proof. First, we will show that the denominator of Equation (3.40) is strictly positive on the event where Lemma 7 holds. We define

$$\det_{-j} := s_{jj}^{(-j)} (\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(-j)}) + (1 + h_{jj}^{(-j)})^2.$$

By Lemma 7, the quadratic forms $s_{jj}^{(-j)}$ are of the same order $\Theta\left(\frac{n}{kp}\right)$ for every $j \in [k]$.

Similarly, we have $t_{jj}^{(-j)} = \mathcal{O}\left(\frac{n}{p}\|\boldsymbol{\mu}\|_2^2\right)$ and $|h_{jj}^{(-j)}| = \tilde{\rho}_{n,k}\mathcal{O}\left(\frac{\epsilon_n}{k^2\sqrt{n}}\right)$ for $j \in [k]$. Thus, we have

$$\frac{n\|\boldsymbol{\mu}\|_2^2}{C_1 kp}\left(1 - \frac{C_2 n}{p}\right) + \left(1 - \frac{C_3\epsilon_n}{k^2\sqrt{n}}\right)^2 \leq \det{}_{-j} \leq \frac{C_1 n\|\boldsymbol{\mu}\|_2^2}{kp} + \left(1 + \frac{C_3\epsilon_n}{k^2\sqrt{n}}\right)^2, \quad (3.42)$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$, for every $j \in [k]$. Here, we use the fact that $t_{jj}^{-j} \geq 0$ by the positive semidefinite property of the leave-one-out Gram matrix $\boldsymbol{A}_{-j}^{-1}$. Next, we choose τ in Equation (3.41) to be sufficiently small so that $C_3\tau \leq 1/2$. Provided that p is sufficiently large compared to n, there then exist constants $C_1', C_2' > 0$ such that we have

$$C_1' \leq \frac{\det_{-m}}{\det_{-j}} \leq C_2', \quad \text{for all } j, m \in [k],$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$. Now, assume without loss of generality that $y_i = k$. Equation (3.42) shows that there exists constant $c > 0$ such that $\det_{-j} > c$ for all $j \in [k]$ with high probability provided that p/n is large enough (guaranteed by the first condition of the theorem). Hence, to make the right-hand-side of Equation (3.39) positive, it suffices to show that the numerator will be positive. Accordingly, we will show that

$$z_{ci}^2\big((1 + h_{kk}^{(-k)})g_{ki}^{(-k)} - s_{kk}^{(-k)} f_{ki}^{(-k)}\big) + C z_{ci} \sum_{j \neq k} \tilde{z}_j \big((1 + h_{jj}^{(-j)})g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}\big) > 0, \quad (3.43)$$

for some $C > 1$.

We can show by simple algebra that it suffices to consider the worst case of $z_{ci} = -1/k$. To see why this is true, we consider the simpler term $z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}|$.

Clearly, Equation (3.43) is positive only if the above quantity is also positive. Lemma 7 shows that when $z_{ci} = -1/k$, then $z_{ci}^2 g_{y_i i}^{(-y_i)} \geq \left(1 - \frac{1}{C_1}\right) \frac{1}{k^2 p}$ and $|z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}| \leq \frac{1}{C_2 k^3 p}$, for $j \neq y_i$. Hence

$$z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}| \geq \left(1 - \frac{1}{C_3}\right) \frac{1}{k^2 p}.$$

Here, $z_{ci} = -1/k$ minimizes the lower bound $z_{ci}^2 g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} z_{ci}\tilde{z}_{j(c)} g_{ji}^{(-j)}|$. To see this, we first drop the positive common factor $|z_{ci}|$ in the equation above and get $|z_{ci}| g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} \tilde{z}_{j(c)} g_{ji}^{(-j)}|$. If we had $z_{ci} = -1/k$, then $|\tilde{z}_{j(c)}|$ is either $(k-1)/k$ or $1/k$. In contrast, if we consider $z_{ci} = (k-1)/k$, then we have $|\tilde{z}_{j(c)}| = 1/k$ for all $j \neq y_i$ and so the term $|z_{ci}| g_{y_i i}^{(-y_i)} - |\sum_{j \neq y_i} \tilde{z}_{j(c)} g_{ji}^{(-j)}|$ is strictly larger.

Using this worst case, i.e. $z_{ci} = -1/k$, and the trivial inequality $|\tilde{z}_{j(c)}| < 1$ for $j \neq y_i$ together with the bounds for the terms $s_{jj}^{(-j)}, t_{jj}^{(-j)}, h_{jj}^{(-j)}$ and $f_{ji}^{(-j)}$ derived in Lemma 7 gives us

$$\begin{aligned}
&(3.43) \\
\geq &\frac{1}{k^2}\left(\left(1 - \frac{C_1\epsilon_n}{k^2\sqrt{n}}\right)\left(1 - \frac{1}{C_2}\right)\frac{1}{p} - \frac{C_3\epsilon_n}{k^{1.5}n} \cdot \frac{n}{kp}\right) \\
&- k \cdot \frac{1}{C_4 k}\left(\left(1 + \frac{C_5\epsilon_n}{k^2\sqrt{n}}\right)\frac{1}{k^2 p} - \frac{C_6\epsilon_n}{k^{1.5}n}\frac{n}{kp}\right) \\
\geq &\frac{1}{k^2}\left(1 - \frac{1}{C_9} - \frac{C_{10}\epsilon_n}{k^2\sqrt{n}} - \frac{C_{11}\epsilon_n}{k^2} - C_{12}\epsilon_n\right)\frac{1}{p} \\
\geq &\frac{1}{k^2 p}\left(1 - \frac{1}{C_9} - C_{10}\tau\right), \qquad (3.44)
\end{aligned}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ for some constants C_i's > 1. Above, we recalled the definition of ϵ_n and used from Lemma 7 that $h_{jj}^{(-j)} \leq \tilde{\rho}_{n,k} \frac{C_{11}\epsilon_n}{k^2\sqrt{n}}$ and $|f_{ji}^{(-j)}| \leq \frac{C_{12}\epsilon_n}{k^{1.5}n}$ with high probability. To complete the proof, we choose τ to be a small enough constant to guarantee $C_{10}\tau < 1 - 1/C_9$, and substitute this in Equation (3.44) to get the desired condition of Equation (3.43).

Chapter 4

Learning Gaussian graphical models with latent confounders

4.1 Introduction

In many domains, it is useful to characterize relationships between features using network models. For example, networks have been used to identify transcriptional patterns and regulatory relationships in genetic networks and applied as a way to characterize functional brain connectivity and cognitive disorders [50, 159, 7, 133]. One of the most common methods for inferring a network from observations is the Gaussian graphical model (GGM). A GGM is defined with respect to a graph, in which the nodes correspond to joint Gaussian random variables and the edges correspond to the conditional dependencies among pairs of variables. A key property of the GGM is that the presence or absence of edges can be obtained from the precision matrix for multivariate Gaussian random variables [92]. Similar to LASSO regression [156], we can infer the sparse graph structure via sparse precision matrix estimation with l_1-regularized maximum likelihood estimation. This family of approaches is called graphical lasso (Glasso) [52, 168].

In practice, however, network inference may be complicated due to the presence of latent confounders. For example, when characterizing relationships between the stock prices of publicly trade companies, the existence of overall market and sector factors induces extra correlation between stocks [33], which can obscure the underlying network structure between companies.

We focus on estimating $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, the precision matrix encoding the graph structure of interest [168, 52, 24, 126]. When latent confounders are present, the covariance matrix for the observed data, $\mathbf{\Sigma}_{obs}$ can be expressed as

$$\mathbf{\Sigma}_{obs} = \mathbf{\Sigma} + L_{\mathbf{\Sigma}}, \tag{4.1}$$

where the positive semidefinite matrix $L_{\mathbf{\Sigma}}$ reflects the effect of latent confounders. One approach, which we will call PCA+GGM, is motivated by confounders that affect the marginal correlation between observed variables [126] and uses principal component analysis (PCA) as a preprocessing step to remove the effect of these confounders [79, 8]. PCA removes the leading eigencomponents from $\mathbf{\Sigma}_{obs}$ which are assumed to be $L_{\mathbf{\Sigma}}$, then a second stage of standard GGM inference follows. PCA+GGM has shown to be useful in estimating gene co-expression networks, where correlated measurement noise and batch effects induce large extraneous marginal correlations between observed variables [57, 94, 147, 54, 51, 76].

Alternatively, equation (4.1) can be reparametrized as the observed precision matrix by applying the Sherman-Morrison identity [70] as,

$$\mathbf{\Omega}_{obs} = \mathbf{\Sigma}_{obs}^{-1} = \mathbf{\Omega} - L_{\mathbf{\Omega}}, \tag{4.2}$$

where $L_{\mathbf{\Omega}}$ again reflects the effect of unobserved confounding, i.e. unobserved nodes in

a graph [28]. One such approach, known as latent variable Gaussian Graphical Models (LVGGM), uses parameterization (4.2) and involves joint inference for $\mathbf{\Omega}$ and $L_{\mathbf{\Omega}}$. The motivation behind LVGGM is to address the effect of unobserved variables in the complete data graph, which affect the partial correlations of the variables in the observed precision matrix $\mathbf{\Omega}$. This perspective can be particularly useful when the unobserved variables would have been included in the graph, had they been observed.

In previous work, either parameterization (4.1), e.g. PCA+GGM, or (4.2), e.g. LVGGM, has been used, depending on the source of confounding and motivations as described earlier. Typically, LVGGM is appropriate when confounding is induced by unobserved nodes in a complete data graph of interest, whereas PCA+GGM is more appropriate when confounding corresponds to nuisance variables, e.g. from batch effects.

In practice, the selection between these two methods will depend on user's belief about the type of confounding present in the observed data. In this paper, our goal is to explore a way to address the effect of confounders in order to obtain the graph structure encoded in $\mathbf{\Omega}$ without making such selection.

To achieve this goal, we generalize two seemingly different methods, PCA+GGM and LVGGM, into a common framework for addressing the effect of $L_{\mathbf{\Sigma}}$ in order to obtain the graph structure encoded in $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. Based on the generalization, We propose a new method, PCA+LVGGM, to address two different sources of confounding. The combined approach is more general, since PCA+LVGGM contains both LVGGM and PCA+GGM as special cases. To our knowledge, the two methods of addressing confounding have not been discussed together in the literature.

In summary, in this paper,

- we carefully compare PCA+GGM and LVGGM, and illustrate the connection and difference between these two methods. We first theoretically characterize the per-

formance of PCA+GGM. Different from [126] who derives asymptotic results, we provide a non-asymptotic convergence result for the performance of PCA+GGM. We observe that the performance of PCA+GGM are largely determined by the spectral structure of $\boldsymbol{\Sigma}$ and $L_{\boldsymbol{\Sigma}}$.

- we propose PCA+LVGGM, which combines elements of PCA+GGM and LVGGM. In simulation, PCA+LVGGM can outperform PCA+GGM or LVGGM when the data is corrupted by multiple confounders. We perform extensive numerical experiments to validate the theory, compare the performance of the three methods, and demonstrate the utility of our approach in two applications.

The remainder of this paper is organized as follows: In section 4.2, we introduce the problem definition for GGM, LVGGM and PCA+GGM followed by a brief literature review. Next, we introduce our hybrid method, PCA+LVGGM, and present a novel theoretical results for PCA+GGM in section 4.3. We use these result to analyze the similarities and differences between LVGGM and PCA+GGM. In section 4.4, we compare the utility of the various approaches in the simulation setting. Finally, in section 4.5 we apply the methods on two real world data sets. We also extend our analysis to joint estimation of multiple graphs when latent confounders exist. The analysis is in Appendix C.5.

Notation: For a vector $\boldsymbol{v} = [v_1, ..., v_p]^T$, define $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$, $\|\boldsymbol{v}\|_1 = \sum_{i=1}^p |v_i|$ and $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$. For a matrix $\boldsymbol{M}$, let M_{ij} be its (i,j)-th entry. Define the Frobenius norm $\|\boldsymbol{M}\|_F = \sqrt{\sum_i \sum_j M_{ij}^2}$, the element-wise ℓ_1-norm $\|\boldsymbol{M}\|_1 = \sum_i \sum_j |M_{ij}|$ and $\|\boldsymbol{M}\|_\infty = \max_{(i,j)} |M_{ij}|$. We also define the spectral norm $\|\boldsymbol{M}\|_2 = \sup_{\|\boldsymbol{v}\|_2 \leq 1} \|\boldsymbol{M}\boldsymbol{v}\|_2$ and $\|\boldsymbol{M}\|_{L_1} = \max_j \sum_i |M_{ij}|$. The nuclear norm $\|\boldsymbol{M}\|_*$ is defined as the sum of the singular values of $\boldsymbol{M}$. When $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ is symmetric, its eigendecomposition is $\boldsymbol{M} = \sum_{i=1}^p \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$, where λ_i is the i-th eigenvalue of $\boldsymbol{M}$, and $\boldsymbol{v}_i$ is the i-th eigenvector. We

assume that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. We call $\lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$ the i-th eigencomponent of $\boldsymbol{M}$.

4.2 Problem setup and review

4.2.1 Gaussian graphical models

Consider a p-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ with covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{\Omega}$. Let $G = (V, E)$ be the graph associated with $\boldsymbol{X}$, where V is the set of nodes (or vertices) corresponding to the elements of $\boldsymbol{X}$, and E is the set of edges connecting nodes. The graph shows the conditional independence relations between elements of $\boldsymbol{X}$. For any pair of connected nodes, the corresponding pairs of variables in $\boldsymbol{X}$ are conditionally independent given the rest variables, *i.e.*, $X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$, for all $(i, j) \notin E$. If $\boldsymbol{X}$ is multivariate Gaussian, then X_i and X_j are conditionally independent given other variables if and only if $\boldsymbol{\Omega}_{ij} = 0$, and thus the graph structure can be recovered from the precision matrix of $\boldsymbol{X}$.

Without loss of generality, we assume the variable $\boldsymbol{X}$ has mean zero in this paper. Assuming that the graph is sparse, given a random sample $\{\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(n)}\}$ following the distribution of $\boldsymbol{X}$, the Glasso estimate $\hat{\boldsymbol{\Omega}}_{Glasso}$[168, 52] is obtained by solving the following log-likelihood based ℓ_1-regularized function:

$$\underset{\Omega \succ 0}{\text{minimize}} \quad \text{Tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}_n) - \log\det(\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1, \tag{4.3}$$

where Tr denotes the trace of a matrix and $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}^{(k)} \boldsymbol{X}^{(k)^T}$ is the sample covariance matrix. Many alternative objective functions for sparse precision matrix estimation have been proposed [24, 112, 127, 82]. The behavior and convergence rates of these approaches are well studied [19, 18, 139, 91, 134, 25].

In presence of latent confounders, Glasso and other GGM methods would likely re-

cover a more dense precision matrix owing to spurious partial correlations introduced between observed variables. In other words, even when the underlying graph is sparse conditioned on the latent variables, the observed graph is dense marginally.

4.2.2 Latent variable Gaussian graphical models

One method for controlling the effects of confounders is the Latent Variable Gaussian Graphical Model (LVGGM) approach first proposed by Chandrasekaran et al. [29]. They assume that the number of latent factors is small compared to the number of observed variables, and that the conditional dependencies among the observed variables conditional on the latent factors is sparse. Consider a $(p+r)$ dimensional mean-zero normal random variable $\boldsymbol{X} = (\boldsymbol{X}_O, \boldsymbol{X}_H)^T$, where $\boldsymbol{X}_O \in \mathbb{R}^p$ is observed and $\boldsymbol{X}_H \in \mathbb{R}^r$ is latent. Let $\boldsymbol{X}$ have precision matrix $\boldsymbol{\Omega} \in \mathbb{R}^{(p+r)\times(p+r)}$, and the submatrices $\boldsymbol{\Omega}_O \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Omega}_H \in \mathbb{R}^{r\times r}$ and $\boldsymbol{\Omega}_{O,H} \in \mathbb{R}^{p\times r}$ specify the dependencies between observed variables, between latent variables and between the observed and latent variables respectively. By Schur complement, the inverse of the observed covariance matrix satisfies:

$$\boldsymbol{\Omega}_{obs} = \boldsymbol{\Sigma}_{obs}^{-1} = \boldsymbol{\Omega}_O - \boldsymbol{\Omega}_{O,H}\boldsymbol{\Omega}_H^{-1}\boldsymbol{\Omega}_{O,H}^T = \boldsymbol{\Omega} - L_{\boldsymbol{\Omega}}. \tag{4.4}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Omega}_O$ encodes the conditional independence relations of interest and is sparse by assumption. $L_{\boldsymbol{\Omega}} = \boldsymbol{\Omega}_{O,H}\boldsymbol{\Omega}_H^{-1}\boldsymbol{\Omega}_{O,H}^T$ reflects the low-rank effect of latent variables $\boldsymbol{X}_H$. Based on this sparse plus low-rank decomposition [29] proposed the following problem:

$$\begin{aligned} \underset{\boldsymbol{\Omega},\, L_{\boldsymbol{\Omega}}}{\text{minimize}} \quad & -\ell(\boldsymbol{\Omega} - L_{\boldsymbol{\Omega}}; \boldsymbol{\Sigma}_n) + \lambda\|\boldsymbol{\Omega}\|_1 + \gamma\|L_{\boldsymbol{\Omega}}\|_* \\ \text{subject to} \quad & L_{\boldsymbol{\Omega}} \succeq 0, \\ & \boldsymbol{\Omega} - L_{\boldsymbol{\Omega}} \succ 0, \end{aligned} \tag{4.5}$$

where $\mathbf{\Sigma}_n$ is the observed sample covariance matrix and $\ell(\mathbf{\Omega}, \mathbf{\Sigma}) = \log(\det(\mathbf{\Omega})) - \mathrm{Tr}(\mathbf{\Omega}\mathbf{\Sigma})$ is the Gaussian log-likelihood function. The ℓ_1-norm encourages sparsity on $\mathbf{\Omega}$ and the nuclear norm encourages low-rank structure on $L_{\mathbf{\Omega}}$.

The sparse plus low-rank decomposition is ill-posed if $L_{\mathbf{\Omega}}$ is not dense. If $\boldsymbol{L}_{\mathbf{\Omega}}$ is sparse, then it is indistinguishable from $\mathbf{\Omega}$, that is, the sparse plus low-rank decomposition works well only when the sparse component is not low-rank and the low-rank component is not sparse [28]. In practice, $\boldsymbol{L}_{\mathbf{\Omega}}$ is dense if the latent variables have widespread effects.. Identifiability of $\mathbf{\Omega}$ coincides with the incoherence condition in the matrix completion problem [26] which requires that $|\boldsymbol{v}_k^T \boldsymbol{e}_i|$ is small for all $k \in \{1, ...r\}$ and $i \in \{1, ..., p\}$ where $\boldsymbol{v}_k$ is the k-th eigenvector of $L_{\mathbf{\Omega}}$ and $\boldsymbol{e}_i$ is the i-th standard basis vector. More analysis on LVGGM can be found in [1] and [113].

Finally, [135] shows that the standard GGM approaches can still recover $\mathbf{\Omega}$ in the presence of latent confounding as long as the spectral norm of the low-rank component is sufficiently small compared to that of $\mathbf{\Sigma}$. This is also verified in our simulations.

4.2.3 PCA+GGM

Unlike LVGGM, which involves a decomposition of the observed data precision matrix, PCA+GGM involves a decomposition of the observed data covariance matrix:

$$\mathbf{\Sigma}_{obs} = \mathbf{\Omega}_{obs}^{-1} = (\mathbf{\Omega} - L_{\mathbf{\Omega}})^{-1} = \mathbf{\Omega}^{-1} + L_{\mathbf{\Sigma}}. \tag{4.6}$$

Motivated by confounding from measurement error and batch effects, [126] proposed the principal components correction (PC-correction) for removing $L_{\mathbf{\Sigma}}$. Consider observed data $\boldsymbol{X}_{obs}$, such that

$$\boldsymbol{X}_{obs} = \boldsymbol{X} + \boldsymbol{A}\boldsymbol{Z}, \tag{4.7}$$

Procedure 1 PCA+GGM

Input: Sample covariance matrix, $\hat{\boldsymbol{\Sigma}}_{obs} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_{obs}^{(k)} \boldsymbol{X}_{obs}^{(k)^T}$; rank of $\hat{L}_{\boldsymbol{\Sigma}}$, r

Output: Precision matrix estimate, $\hat{\boldsymbol{\Omega}}$

1: Estimate $\hat{L}_{\boldsymbol{\Sigma}}$ from eigencomponents:

$$\hat{\boldsymbol{\Sigma}}_{obs} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T, \ \hat{L}_{\boldsymbol{\Sigma}} = \sum_{i=1}^{r} \hat{\lambda}_i \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T$$

2: Remove $\hat{L}_{\boldsymbol{\Sigma}}$:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{obs} - \hat{L}_{\boldsymbol{\Sigma}}.$$

3: Using $\hat{\boldsymbol{\Sigma}}$, compute $\hat{\boldsymbol{\Omega}}$ as solution to (4.3)

Procedure 2 PCA+LVGGM

Input: Sample covariance matrix, $\hat{\boldsymbol{\Sigma}}_{obs}$; rank of $\hat{L}_{\boldsymbol{\Sigma}}$, r_P; rank of $\hat{L}_{\boldsymbol{\Omega}}$, r_L

Output: Precision matrix estimate, $\hat{\boldsymbol{\Omega}}$

1: Estimate $\hat{L}_{\boldsymbol{\Sigma}}$ from eigencomponents:

$$\hat{\boldsymbol{\Sigma}}_{obs} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T, \ \hat{L}_{\boldsymbol{\Sigma}} = \sum_{i=1}^{r_P} \hat{\lambda}_i \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T$$

2: Remove $\hat{L}_{\boldsymbol{\Sigma}}$:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{obs} - \hat{L}_{\boldsymbol{\Sigma}}.$$

3: Using $\hat{\boldsymbol{\Sigma}}$, compute $\hat{\boldsymbol{\Omega}}$ as solution to (4.5) with γ such that $\text{rank}(\hat{L}_{\boldsymbol{\Omega}}) = r_L$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_r)$. Matrix $\boldsymbol{A} \in \mathbb{R}^{p \times r}$ is non-random so that $L_{\boldsymbol{\Sigma}} = \boldsymbol{A}\boldsymbol{A}^T$. In general, additional structural assumptions are needed to distinguish $L_{\boldsymbol{\Sigma}}$ from $\boldsymbol{\Sigma}$. As we will discuss in section 4.3, one of our contributions is to show that if the spectral norm of $L_{\boldsymbol{\Sigma}}$ is large relative to that of $\boldsymbol{\Sigma}$, then under mild conditions, $L_{\boldsymbol{\Sigma}}$ is close to the sum of the first few eigencomponents of $\boldsymbol{\Sigma}_{obs}$. Therefore, one can remove the first r eigencomponents from $\boldsymbol{\Sigma}_{obs}$ [126]. This PCA+GGM method is described in Procedure 1. Note that the number of principal components needs to be determined a priori, which we discuss in subsequent sections.

4.2.4 Combining PCA+GGM and LVGGM

As previously mentioned, while LVGGM and PCA+GGM solve the same problem, they are motivated by different sources of confounding. In applications, the observed data may be corrupted by multiple sources of confounding, and thus elements from both methods are needed. For example, in the biological application discussed in section 4.5.1, both batch effects and unmeasured biological variables likely confound estimates of graph

structure between observed variables. This motivates us to propose the PCA+LVGGM strategy described below.

As (4.4) illustrated, the observed precision matrix $\mathbf{\Omega}^{'}$ may have been corrupted by a latent factor $L_{\mathbf{\Omega}}$:

$$\mathbf{\Omega}^{'} = \mathbf{\Omega} - L_{\mathbf{\Omega}}. \tag{4.8}$$

Now, rewriting (4.8) in terms of $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ and $\mathbf{\Sigma}^{'} = \mathbf{\Omega}^{'-1}$, applying the Sherman-Morrison identity on $\mathbf{\Omega}^{'}$ gives,

$$\mathbf{\Sigma}^{'} = \mathbf{\Sigma} + L^{'}_{\mathbf{\Omega}}, \tag{4.9}$$

where $L^{'}_{\mathbf{\Omega}}$ is still a low-rank matrix. If $\mathbf{\Sigma}^{'}$ is further corrupted by an additive latent factor represented by $L_{\mathbf{\Sigma}}$, the following equation described the observed matrix $\mathbf{\Sigma}_{obs}$:

$$\mathbf{\Sigma}_{obs} = \mathbf{\Sigma}^{'} + L_{\mathbf{\Sigma}} = \mathbf{\Sigma} + L^{'}_{\mathbf{\Omega}} + L_{\mathbf{\Sigma}} \tag{4.10}$$

In the above example, following our theoretical analysis in section 4.3, if the spectral norm of $L_{\mathbf{\Sigma}}$ is much larger than that of $\mathbf{\Sigma}$ and $L^{'}_{\mathbf{\Omega}}$, then removing $L_{\mathbf{\Sigma}}$ using the PC-correction is likely to be effective. If the spectral norm of $L^{'}_{\mathbf{\Omega}}$ is not much larger than that of $\mathbf{\Sigma}$, then PC-correction is not a good choice to remove $L^{'}_{\mathbf{\Omega}}$. If $L^{'}_{\mathbf{\Omega}}$ is dense, then $\mathbf{\Omega}$ and $L_{\mathbf{\Omega}}$ can be well estimated by LVGGM. In (4.10), the overall confounding $L^{'}_{\mathbf{\Omega}} + L_{\mathbf{\Sigma}}$ is the sum of two low-rank components with different norms, we can consider using both methods: first remove $L_{\mathbf{\Sigma}}$ via eigendecomposition, then apply LVGGM to estimate $\mathbf{\Omega}$ and $L_{\mathbf{\Omega}}$. We call this procedure PCA+LVGGM and it is shown in Procedure 2. We discuss methods for setting the ranks for $L_{\mathbf{\Sigma}}$ (defined as r_P) and $L^{'}_{\mathbf{\Omega}}$ (defined as r_L) in section 4.3.5.

4.3 Theoretical analysis and model comparisons

In this section, we investigate the theoretical properties of PCA+GGM. Our results reveal precisely how the eigenstructure of the observed covariance matrix affects the the performance of PCA+GGM. The theoretical analysis provides practical insights into when each graph estimation method should (or should not) be applied. Specifically, we derive the convergence rate of PCA+GGM and compare it to that of LVGGM. As shown in theoretical analysis by [126], the low-rank confounder can be well estimated by PC-correction if the number of features $p \to \infty$ with the number of observations n fixed. We provide a non-asymptotic analysis depending on p and n and our result shows that the graph can be recovered exactly when $n \to \infty$ with fixed p. When additional assumptions are satisfied, e.g. spiky covariance structure and incoherent eigenvectors, the convergence rate can be improved to $O(\sqrt{\frac{\log p}{n}})$.

4.3.1 Convergence analysis on PCA+GGM

Without loss of generality, we consider the case of a rank-one confounder. Assume that we have a random sample of p-dimensional random vectors:

$$\boldsymbol{X}_{obs}^{(i)} = \boldsymbol{X}^{(i)} + \sigma \boldsymbol{v} \boldsymbol{Z}^{(i)},\ i = 1..., n, \tag{4.11}$$

where $Cov(\boldsymbol{X}^{(i)}) = \boldsymbol{\Sigma}$ and $\boldsymbol{Z}^{(i)}$ is a univariate standard normal random variable. $\boldsymbol{v} \in \mathbb{R}^p$ is a non-random vector with unit norm, and σ is a non-negative scalar constant. Without loss of generality, we assume that $\boldsymbol{X}^{(i)} \perp\!\!\!\perp \boldsymbol{Z}^{(i)}$. To see how $\boldsymbol{v}$ affects estimation, we assume that $\boldsymbol{v}$ is the k-th eigenvector of $\boldsymbol{\Sigma}$. The discussion on general $\boldsymbol{v}$ is deferred to section

4.3.2. Therefore, the covariance matrix of $\boldsymbol{X}_{obs}^{(i)}$ is:

$$\boldsymbol{\Sigma}_{obs} = \boldsymbol{\Sigma} + \sigma^2 \boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{\Sigma}_{-k} + (\lambda_k(\boldsymbol{\Sigma}) + \sigma^2)\boldsymbol{v}\boldsymbol{v}^T, \tag{4.12}$$

where $\boldsymbol{\Sigma}_{-k}$ is the matrix $\boldsymbol{\Sigma}$ without the k-th eigencomponent and $\lambda_k(\boldsymbol{\Sigma})$ is the k-th eigenvalue of $\boldsymbol{\Sigma}$. When $\sigma^2 > \lambda_1(\boldsymbol{\Sigma})$, $\lambda_k(\boldsymbol{\Sigma}) + \sigma^2$ becomes the first eigenvalue of $\boldsymbol{\Sigma}_{obs}$, and $\boldsymbol{v}$ is the corresponding first eigenvector. We remove the first principal component from the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{obs}$:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{obs} - \hat{\lambda}_1 \hat{\boldsymbol{\theta}}_1 \hat{\boldsymbol{\theta}}_1^T, \tag{4.13}$$

where $\hat{\lambda}_1$ is the first eigenvalue of $\hat{\boldsymbol{\Sigma}}_{obs}$ and $\hat{\boldsymbol{\theta}}_1$ is the first eigenvector of $\hat{\boldsymbol{\Sigma}}_{obs}$. Then we use $\hat{\boldsymbol{\Sigma}}$ to estimate $\boldsymbol{\Omega}$. We first show that under mild conditions, $\hat{\boldsymbol{\Sigma}}$ is close to $\boldsymbol{\Sigma}$. Following [19, 3.1], we assume that there exists a constant M such that:

$$\lambda_1(\boldsymbol{\Sigma}_{obs}) \leq M \quad \textit{and} \quad \lambda_p(\boldsymbol{\Sigma}_{obs}) \geq \frac{1}{M}. \tag{4.14}$$

Theorem 12. *Let λ_i be the i-th eigenvalue of $\boldsymbol{\Sigma}_{obs}$ and $\nu = \lambda_1 - \lambda_2$ be the eigengap of $\boldsymbol{\Sigma}_{obs}$. Suppose $\boldsymbol{\Sigma}_{obs}$ satisfies condition (4.14) and $\boldsymbol{X}_{obs}^{(i)}$ is generated as (4.11). Further assume that $\sigma^2 > \lambda_1(\boldsymbol{\Sigma})$. Suppose $n \geq p$ and $\|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{p}{n}} \leq \frac{1}{128}$, then:*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \leq C_1 \sqrt{\frac{\log p}{n}} + C_2 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{p}{n}} + C_3 \sqrt{\frac{p}{n}} + \lambda_k(\boldsymbol{\Sigma}) \|\boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T\|_\infty,$$

with probability greater than $1 - C_4/p$ for constants C_i's > 1.

Proof. By (4.12) and (4.13),

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} &= (\hat{\boldsymbol{\Sigma}}_{obs} - \hat{\lambda}_1 \hat{\boldsymbol{\theta}}_1 \hat{\boldsymbol{\theta}}_1^T) - (\boldsymbol{\Sigma}_{obs} - \lambda_1 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T + \lambda_k \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T) \\ &= (\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}) + (\lambda_1 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T - \hat{\lambda}_1 \hat{\boldsymbol{\theta}}_1 \hat{\boldsymbol{\theta}}_1^T) - \lambda_k \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T,\end{aligned}$$

where λ_k is the k-th eigenvalue of $\boldsymbol{\Sigma}_{obs}$, and $\boldsymbol{\theta}_k$ is the k-th eigenvector of $\boldsymbol{\Sigma}_{obs}$. At a high level, we bound $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty$ by bounding the norms of $\boldsymbol{\Sigma}_{obs} - \hat{\boldsymbol{\Sigma}}_{obs}$, $\lambda_1 - \hat{\lambda}_1$ and $\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1$. The details of the complete proof is in Appendices C.1.1 and C.1.2. □

The bound in Theorem 12 can be further simplified as $C_s\sqrt{\frac{p}{n}} + \lambda_k(\boldsymbol{\Sigma})\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty$ for some large constant C_s. We express it in the above form because it provides more insight on how each term affects the result. Now we analyze the bound in Theorem 12 in detail.

The error bound in Theorem 12 depends on the largest eigenvalue of $\boldsymbol{\Sigma}_{obs}$, the eigengap $\nu = \lambda_1(\boldsymbol{\Sigma}_{obs}) - \lambda_2(\boldsymbol{\Sigma}_{obs})$, the eigenvector of the confounder and n and p. The term $\sqrt{\frac{\nu+1}{\nu^2}}$ shows that if the eigengap ν is larger, the estimation error bound will be smaller. Recall that when $\sigma^2 > \lambda_1(\boldsymbol{\Sigma})$, $\lambda_k(\boldsymbol{\Sigma}) + \sigma^2$ becomes the first eigenvalue of $\boldsymbol{\Sigma}_{obs}$. Hence if $\sigma^2 \gg \lambda_1(\boldsymbol{\Sigma})$, then the eigengap ν is large. The fact that a larger eigengap leads to a better convergence rate is closely related to the concept of "effective dimension" (also known as "effective rank"). The effective rank, $r(\boldsymbol{M})$, of any positive semidefinite matrix $\boldsymbol{M} \in \mathbb{R}^{p\times p}$, is defined as:

$$r(\boldsymbol{M}) := \frac{\mathrm{Tr}(\boldsymbol{M})}{\lambda_1(\boldsymbol{M})} = \frac{\sum_{i=1}^p \lambda_i(\boldsymbol{M})}{\lambda_1(\boldsymbol{M})} \leq C, \tag{4.15}$$

where $C \geq 1$ can be viewed as the effective dimension of $\boldsymbol{M}$ ([113, 87, 163]). M is approximately low-rank if the first few eigenvalues are much larger than the rest, and $r(\boldsymbol{M})$ will be much smaller than the observed dimension p. In this case, we can significantly reduce the magnitude of the dependence on $O(\sqrt{\frac{p}{n}})$ by replacing p with effective dimension C,

in (4.15). We provide a sharper bound for matrices with small effective rank in Theorem 13.

Next, we reason about the last term in the error bound, $\lambda_k(\mathbf{\Sigma})\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty$. In practice, in many sparse graphs inferred from real world data, the first few eigenvalues of $\mathbf{\Sigma}$ are much larger than the rest, i.e., $\lambda_1(\mathbf{\Sigma}) \gg \lambda_k(\mathbf{\Sigma})$ for large enough $k > 1$. This is also true for many common graph data generating models (see Appendix C.3). This means that if the eigenvector of the low-rank component is one of the first few eigenvectors of $\mathbf{\Sigma}$, then the error bound will be much larger. This result shows that the first few eigencomponents play a more important role in determining the structure of $\mathbf{\Sigma}$ and its inverse. Thus, the error of the PCA+GGM estimator will be large if those first few eigencomponents are removed by PC-correction.

Note that $\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty$ is upper bounded by 1, since $\boldsymbol{\theta}_1$ is the eigenvector of some matrix, and thus has unit Euclidean norm; however, $\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty$ can be much smaller than 1 when $\boldsymbol{\theta}_1$ is incoherent with standard basis, e.g. dense. One extreme case is when all the elements of $\boldsymbol{\theta}_1$ are $\frac{1}{\sqrt{p}}$, in which case $\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty = \frac{1}{p}$. This setup corresponds to a scenario in which the confounder has a widespread effect over all the p variables in the signal, which is in accordance with one requirement in LVGGM. LVGGM requires the low-rank component to be dense. For both PCA+GGM and LVGGM, more "widespread" confounding implies smaller estimation error. Based on these observations, we provide a tighter bound under small effective rank and incoherent $\boldsymbol{\theta}_1$.

Theorem 13. *Following the same notations and assumptions for $\mathbf{\Sigma}_{obs}$ in Theorem 12 and again assume that $\sigma^2 > \lambda_1(\mathbf{\Sigma})$. Further assume that there exist constants C_i's > 1 such that the effective rank of $\mathbf{\Sigma}_{obs}$ (defined in (4.15)) $r(\mathbf{\Sigma}_{obs}) \leq C_1 n$, the eigengap ν*

satisfies $\sqrt{p}\nu \geq C_2(p\lambda_1(\mathbf{\Sigma}) \vee \sigma^2)$ *and* $\boldsymbol{\theta}_1$ *is incoherent, i.e.* $\|\boldsymbol{\theta}\|_\infty \leq C_3/\sqrt{p}$. *Then:*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq C_4\sqrt{\frac{\log p}{n}},$$

with probability greater than $1 - C_5/p$ *for some* C_i*'s* > 1.

Proof. The complete proof is in Appendices C.1.1 and C.1.2. □

After obtaining $\hat{\mathbf{\Sigma}}$, we can use Glasso, CLIME [24] or any sparse GGM estimation approach to estimate $\mathbf{\Omega}$. We can have a good estimate of $\mathbf{\Omega}$ when $\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_\infty$ is small. With the same input $\hat{\mathbf{\Sigma}}$, the theoretical convergence rate of the estimate obtained from CLIME is of the same order as the Glasso estimate. The derivation of the error bound of Glasso requires the irrepresentability condition and restricted eigenvalue conditions (see [134]). Due to the length of the article, we only show the proof of the edge selection consistency for CLIME, meaning that for the theoretical analysis, we apply CLIME method after obtaining $\hat{\mathbf{\Sigma}}$.

The CLIME estimator $\hat{\mathbf{\Omega}}_1$ is obtained by solving:

$$\begin{aligned} \underset{\mathbf{\Omega}}{\text{minimize}} \quad & \|\mathbf{\Omega}\|_1 \\ \text{subject to} \quad & \|\hat{\mathbf{\Sigma}}\mathbf{\Omega} - \boldsymbol{I}\|_\infty \leq \lambda_n. \end{aligned} \tag{4.16}$$

Since $\hat{\mathbf{\Omega}}_1$ might not be symmetric, we need the symmetrization step to obtain $\hat{\mathbf{\Omega}}$.

Following [24], we assume that $\mathbf{\Omega}$ is in the following class:

$$U(s_0, M_0) = \{\mathbf{\Omega} = \omega_{ij} : \ \mathbf{\Omega} \succ \mathbf{0}, \|\mathbf{\Omega}\|_{L_1} < M_0, \max_{1\leq i\leq p} \sum_{i=1}^{p} I_{\{\omega_{ij}\neq 0\}} \leq s_0(p)\}, \tag{4.17}$$

where we allow s_0 and M_0 to grow as p and n increase. With $\hat{\mathbf{\Sigma}}$ obtained from equation (4.13) as the input of (4.16), we have the following result.

Theorem 14. *Suppose that assumptions in Theorem 12 hold,* $\mathbf{\Omega} \in U(s_0, M_0)$*, and* λ_n *is chosen as* $M_0(C_1\sqrt{\frac{\log p}{n}} + C_2\sqrt{\frac{p}{n}} + C_3\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{p}{n}} + \lambda_k(\mathbf{\Sigma})\|\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty)$*, then:*

$$\|\mathbf{\Omega} - \hat{\mathbf{\Omega}}\|_\infty \leq 2M_0\lambda_n,$$

with probability greater than $1 - C_4/p$*.* C_i*'s are defined the same as in Theorem 12. When assumptions in Theorem 13 hold,* $\mathbf{\Omega} \in U(s_0, M_0)$ *and* $\lambda_n^{'}$ *is chosen as* $M_0(C_5\sqrt{\frac{\log p}{n}})$*, then:*

$$\|\mathbf{\Omega} - \hat{\mathbf{\Omega}}\|_\infty \leq 2M_1\lambda_n^{'},$$

with probability greater than $1 - C_6/p$ *for* C_i*'s* > 1*.*

Proof. The main steps follow the proof of Theorem 6 in [24]. The complete proof is in Appendix C.1.3. □

Therefore, if the minimum magnitude of $\mathbf{\Omega}$ is larger than the error bounds above, then we can have exact edge selection with high probability.

4.3.2 Generalizations

The analysis in previous sections assumes that the low-rank confounder has rank 1, is independent of $\boldsymbol{X}$ and the eigenvector of the covariance of the $L_{\mathbf{\Sigma}}$ is one of the eigenvectors of $\mathbf{\Sigma}$. We now comment on more general settings.

- **Higher rank:** For ease of interpretation, we assume that the confounder can be expressed as $\sum_{i=1}^{r} \sigma_i \boldsymbol{v}_i$. If $\min_i\{\sigma_i^2\} > \lambda_1(\mathbf{\Sigma})$, then when running PCA+GGM, the low-rank component can be removed due to its larger norm compared to that of

$\boldsymbol{\Sigma}$. According to Theorem 12, PCA+GGM can still perform well if $\boldsymbol{v}_i$'s are not the top eigenvectors of $\boldsymbol{\Sigma}$.

- **General $\boldsymbol{v}$:** When the eigenvector of the low-rank component is not one of the eigenvectors of $\boldsymbol{\Sigma}$, we can express that vector using the eigenvectors of $\boldsymbol{\Sigma}$ as basis. For example, we assume that in (4.11), $\boldsymbol{v} = \sum_{i=1}^{p} a_i \boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_i$ means the i-th eigenvector of $\boldsymbol{\Sigma}$. We say $\boldsymbol{v}$ is closely aligned with $\boldsymbol{\theta}_1$ if $|a_1|$ is significantly large compared with other $|a_i|$'s. Equivalently, $|\boldsymbol{v}^T\boldsymbol{\theta}_1| \gg |\boldsymbol{v}^T\boldsymbol{\theta}_i|$ for $i \neq 1$, if $\boldsymbol{v}$ is closely aligned with $\boldsymbol{\theta}_1$. In this case, the first eigencomponent of $\boldsymbol{\Sigma}$ will be removed, thus leading to a poor estimate of $\boldsymbol{\Sigma}$ using PC-correction. If the eigenvector of the low-rank component is not closely aligned with the first few eigenvectors of $\boldsymbol{\Sigma}$, then we won't lose too much useful information when removing the top principal components and PCA+GGM can still perform well.

4.3.3 Comparison with LVGGM

Now we compare LVGGM to PCA+GGM in more detail. We observe that PCA+GGM can be viewed as a supplement to LVGGM. The assumptions of PCA+GGM can be well satisfied when the assumptions of LVGGM cannot be satisfied. In (4.12), now let $\boldsymbol{v}$ be the k-th eigenvector of $\boldsymbol{\Omega}$ (thus the $(p-k+1)$-th eigenvector of $\boldsymbol{\Sigma}$), the Sherman-Morrison identity gives

$$\boldsymbol{\Sigma}_{obs}^{-1} = \boldsymbol{\Omega} - \frac{\lambda_k(\boldsymbol{\Omega})^2}{\lambda_k(\boldsymbol{\Omega}) + (1/\sigma^2)} \boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{\Omega} - L_{\boldsymbol{\Omega}}. \tag{4.18}$$

We can see that as σ increases, $\frac{\lambda_k(\boldsymbol{\Omega})^2}{\lambda_k(\boldsymbol{\Omega})+(1/\sigma^2)}$ increases. In the simulations in section 4.4, we observe that LVGGM performs poorly when $\boldsymbol{v}$ is closely aligned with the first few eigenvectors of $\boldsymbol{\Omega}$ (thus the last few eigenvectors of $\boldsymbol{\Sigma}$). One way to interpret why LVGGM does not work well under this setting is because the nuclear norm penalty in LVGGM will shrink large eigenvalues. Specifically, when k is small and σ^2 is large, $\frac{\lambda_k(\boldsymbol{\Omega})^2}{\lambda_k(\boldsymbol{\Omega})+(1/\sigma^2)}$

is large. Therefore, the nuclear norm regularization in LVGGM introduces larger bias. Additionally, when k is small, $\boldsymbol{v}$ is one of the top eigenvectors of $\boldsymbol{\Omega}$. We empirically observe that the top eigenvectors of $\boldsymbol{\Omega}$ can be coherent with standard basis and this will lead to the identifiability issue of LVGGM, thus increasing the error of LVGGM estimator.

This observation is consistent with the conclusion in [1]. They impose a spikiness condition, which is a weaker condition than the incoherence condition in [29]. The spikiness condition requires that $\|L_{\boldsymbol{\Omega}}\|_\infty$ is not too large. (4.18) shows that $L_{\boldsymbol{\Omega}}$ tends to have a larger spectral norm when $\boldsymbol{v}$ is aligned with the first few eigenvectors of $\boldsymbol{\Omega}$ and σ is large, since in this case, $\frac{\lambda_k(\boldsymbol{\Omega})^2}{\lambda_k(\boldsymbol{\Omega})+(1/\sigma^2)}$ is close to $\lambda_1(\boldsymbol{\Omega})$. The large norm of $L_{\boldsymbol{\Omega}}$ implies that the spikiness condition is not well satisfied, thus the error bound of LVGGM is large. Note, however, that the first few eigenvectors of $\boldsymbol{\Omega}$ are the last few eigenvectors of $\boldsymbol{\Sigma}$. Our analysis shows that the error bound of the estimate of PCA+GGM is small when $\boldsymbol{v}$ is aligned with the first few eigenvectors of $\boldsymbol{\Omega}$ and σ is large.

4.3.4 PCA+LVGGM

In this section we discuss the PCA+LVGGM method briefly. We use the same formulation as (4.8) to (4.10). We claim that PCA+LVGGM outperforms using PCA+GGM or LVGGM individually when $L_{\boldsymbol{\Sigma}}$'s spectral norm is large compared to that of $L'_{\boldsymbol{\Omega}}$ and $\boldsymbol{\Sigma}$, $L_{\boldsymbol{\Sigma}}$'s vectors are not aligned with the first few eigenvectors of $\boldsymbol{\Sigma}$, and the spectral norm of $L'_{\boldsymbol{\Omega}}$ is not significantly larger than that of $\boldsymbol{\Sigma}$. This is because based on Theorem 12 and 14, PCA+GGM is effective only when the spectral norm of the low-rank confounding is larger than that of the signal. PC-correction, however, can only effectively remove $L_{\boldsymbol{\Sigma}}$ but not $L'_{\boldsymbol{\Omega}}$ because the norm of $L'_{\boldsymbol{\Omega}}$ is not significantly larger than that of $\boldsymbol{\Sigma}$. In contrast, LVGGM can estimate $L'_{\boldsymbol{\Omega}}$ well, but not $L_{\boldsymbol{\Sigma}}$ because it has a larger spectral norm and its

eigenvectors might be aligned with the first few eigenvectors of $\mathbf{\Omega}$.

4.3.5 Tuning parameter selection

In both LVGGM and PCA+GGM there are crucial tuning parameters to select. For LVGGM, recall that λ controls the sparsity of $\mathbf{\Omega}$ and γ controls the rank of $L_{\mathbf{\Omega}}$. Chandrasekaran et al. [29] argues that λ should be proportional to $\sqrt{\frac{p}{n}}$, the rate in the convergence analysis, and choose γ among a range of values that makes the graph structure of $\hat{\mathbf{\Omega}}$ stable [see 28, for more detail].

When using PCA+GGM, we need to determine the rank first (i.e. how many principal components should be removed). [94] and [95] suggest using the `sva` function from `Bioconductor`, which is based on parallel analysis [69, 23, 102]. Parallel analysis compares the eigenvalues of the sample correlation matrix to the eigenvalues of a random correlation matrix for which no factors are assumed. Given the number of principal components to remove, we can use model selection tools such as AIC, BIC or cross-validation to choose the sparsity parameter in Glasso. One may also decide how many principal components to remove by considering the number of top eigenvalues of the observed covariance matrix (see section 4.3.1 and section 4.3.2). Note that these rank selection approaches perform well when the low-rank confounding has large enough spectral norm compared to the norm of signal (more details on these conditions are discussed in [95, 102]). We will see in later sections that these conditions can be satisfied in many real-world applications. When the spectral norm of the latent confounder is small, approaches which do not account for confounding, such as Glasso and CLIME, are actually robust enough to perform well even when confounding exists. This is theoretically proved by [135] and our simulations in next section also confirm this.

The PCA+LVGGM method has three tuning parameters: the rank of $L_{\mathbf{\Sigma}}$, γ and

λ. To start, we first look at eigenvalues or use the `sva` package to determine the total rank of the low-rank component, $L_{\mathbf{\Sigma}} + L'_{\mathbf{\Omega}}$. We think it is natural to determine the rank of confounder first because we will see in later applications, we can have some domain knowledge on the ranks of coufounders, e.g. in finance applications, some financial theory suggests the number of latent variables in the market. We then need to partition the total rank between $L_{\mathbf{\Sigma}}$ and $L'_{\mathbf{\Omega}}$. If we determine that $\text{rank}(L_{\mathbf{\Sigma}} + L'_{\mathbf{\Omega}}) = k$, we look for an eigengap in the first k eigenvalues and allocate the largest $m < k$ eigenvalues for PC-removal. Our experiments in section 4.5 show that domain knowledge can be used to motivate the number of components for PC-removal. After removing the principal components, we choose γ in LVGGM so tha $L'_{\mathbf{\Omega}}$ is approximately rank $k - m$. We observe that when running LVGGM, the rank won't change for a range of λ values when using a fixed γ. Thus, it suffices to fix γ first to control the rank, then determine λ to control the sparsity.

Practically, network estimation is often used to help exploratory data analysis and hypothesis generation. For these purposes, model selection methods such as AIC, BIC or cross-validation may tend to choose models that are too dense [37]. This fact can also be observed by our experiments. Therefore, we recommend that model selection should be based on prior knowledge and practical purposes, such as network interpretability and stability, or identification of important edges with low false discovery rate [111]. Thus, we recommend that the selection of tuning parameters should be driven by applications. For example, for biological applications, the model should be biologically plausible, sufficiently complex to include important information and sparse enough to be interpretable. In this context, a robustness analysis can be used to explore how edges change over a range of tuning parameters.

4.4 Simulations

In this section, numerical experiments illustrate the utility of each sparse plus low rank method. In section 4.4.1 we illustrate the behavior of Glasso, LVGGM and PCA+GGM under different assumptions about rank-one confounding. In section 4.4.2, we show the efficacy of PCA+LVGGM in a variety of simulation scenarios. In all experiments, we set $p = 100$ and use the scale-free and random networks from `huge.generator` function in `R` package `huge` [171]. To generate random networks, each pair of off-diagonal elements are randomly set, while the graph is generated using B-A algorithm under scale-free structures [2]. Due to space limit, we only include results on the scale-free structure.

4.4.1 The efficacy of LVGGM and PCA+GGM

We compare the relative performance of PCA+GGM, LVGGM and Glasso in the presence of a rank-one confounder, L. Guided by our analysis in section 4.3, we show that the relationship between L and the eigenstructure of $\mathbf{\Sigma}$ determines the performance of these three methods. We first generate the data with

$$\begin{aligned} \boldsymbol{X}_{obs}^{(i)} &= \boldsymbol{X}^{(i)} + L^{(i)},\ i = 1, ..., n, \\ L^{(i)} &= \sigma \boldsymbol{V} \boldsymbol{Z}^{(i)}, \end{aligned}$$

where $\boldsymbol{X}^{(i)} \in \mathbb{R}^p$ is normally distributed with mean zero and covariance matrix $\mathbf{\Sigma}$. $\boldsymbol{Z}^{(i)} \in \mathbb{R}^r$, the low-rank confounder, follows a normal distribution with mean zero and identity covariance matrix. $\boldsymbol{V} \in \mathbb{R}^{p \times r}$ is a non-random semi-orthogonal matrix satisfying $\boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{I}$, and $\sigma \in \mathbb{R}$ represents the magnitude of the confounder. Without loss of generality, we assume that $\boldsymbol{X}^{(i)}$ and $\boldsymbol{Z}^{(i)}$ are independent. We illustrate the performance of different methods under various choices for, $\boldsymbol{V}$, the eigencomponents of L.

We first set $r = 1$, $p = 100$ and $n = 200$. The largest eigenvalue of $\mathbf{\Sigma}$ is around 5. We use $\boldsymbol{v}_i$ to denote the i-th eigenvector of $\mathbf{\Sigma}$. When examining the effect of σ, we choose the 95-th eigenvector of $\mathbf{\Sigma}$ as $\boldsymbol{V}$ to ensure that $\boldsymbol{V}$ is not closely aligned with the first few eigenvectors of $\mathbf{\Sigma}$. We then compare the cases with $\sigma^2 = 20$ and 3. Next, we examine the effect of eigenvectors. We fix σ^2 as 20, and use the i-th eigenvector of $\mathbf{\Sigma}$ as $\boldsymbol{V}$, where $i \in \{1, 60, 95\}$. Following previous notation, we use $\boldsymbol{v}_1$, $\boldsymbol{v}_{60}$ and $\boldsymbol{v}_{95}$ as $\boldsymbol{V}$. 1 is chosen as the rank for PC-correction and LVGGM. We generate ROC curves [64, 9.2.5] for each method based on 50 simulated samples and use the average to draw the ROC curves (Figure 4.1). We truncate the ROC curves at FPR=0.2, since the estimates with large FPR are typically less useful in practice.

From Figure 4.1, we observe that when the confounder has large norm and its eigenvectors are not closely aligned with the first few eigenvectors of $\mathbf{\Sigma}$, PCA+GGM performs better than LVGGM and Glasso. LVGGM preforms the best when the confounder has large norm and its eigenvectors are not aligned with the last few eigenvectors of $\mathbf{\Sigma}$ (also the first few eigenvectors of $\mathbf{\Sigma}^{-1}$). When the low-rank component does not have a large norm, Glasso also performs well. This reaffirms the fact that Glasso can be robust enough to address the low-rank confounding with small norm.

4.4.2 The efficacy of PCA+LVGGM

In this section, we use examples to demonstrate the efficacy of PCA+LVGGM. We introduce corruption of the signal with two low-rank confounders. The data is generated as follows:

$$\boldsymbol{X}_{obs}^{(i)} = \boldsymbol{X}^{(i)} + \boldsymbol{V}_1\boldsymbol{D}_1\boldsymbol{Z}_1^{(i)} + \boldsymbol{V}_2\boldsymbol{D}_2\boldsymbol{Z}_i^{(i)},\ i = 1, ..., n,$$

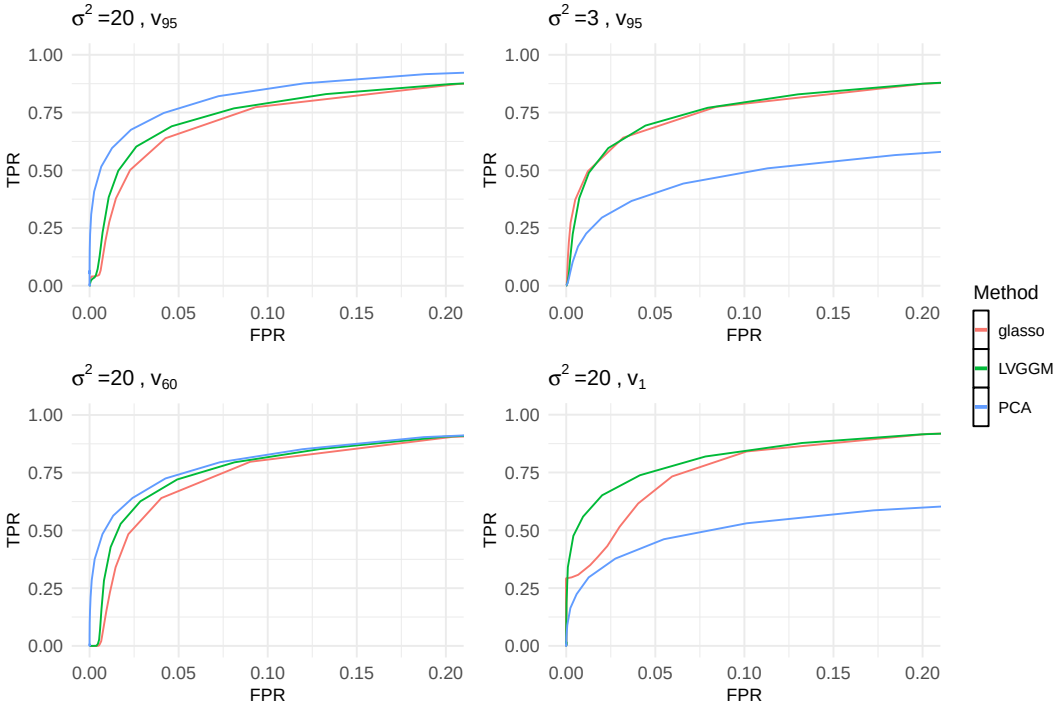


Figure 4.1: $n = 200$. The low-rank component has rank 1. σ^2 is the magnitude of the low-rank component, $\boldsymbol{v}_i$ is the i-th eigenvector of $\boldsymbol{\Sigma}$. The first row illustrates the effect of σ: we use the 95-th eigenvector of $\boldsymbol{\Sigma}$ as $\boldsymbol{V}$, and set σ^2 to 20 and 3 from left to right. The second row illustrates the effect of $\boldsymbol{V}$: we fixed σ^2 as 20, and use the 60-th and first eigenvector of $\boldsymbol{\Sigma}$ as $\boldsymbol{V}$ from left to right. PCA+GGM works the best when σ^2 is large and $\boldsymbol{V}$ is not aligned with the first eigenvector of $\boldsymbol{\Sigma}$. LVGGM works the best when σ^2 is large and $\boldsymbol{V}$ is not aligned with the last eigenvector of $\boldsymbol{\Sigma}$. When σ^2 is small, Glasso works as well as the other two.

where $\boldsymbol{X}_i \in \mathbb{R}^p$ is normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{Z}_1^{(i)} \in \mathbb{R}^{d_1}$, corresponding to the first source of low-rank confounder, has a normal distribution with mean zero and covariance matrix $\boldsymbol{I}_{d_1}$. $\boldsymbol{V}_1 \in \mathbb{R}^{p\times d_1}$ is a non-random, semi-orthogonal matrix satisfying $\boldsymbol{V}_1^T\boldsymbol{V}_1 = \boldsymbol{I}$, and $\boldsymbol{D}_1 \in \mathbb{R}^{d_1\times d_1}$ is a diagonal matrix, measuring the magnitude of the first confounder. Similarly, $\boldsymbol{Z}_2^{(i)} \in \mathbb{R}^{d_2}$, corresponding to the second source of low-rank confounder, has normal distribution with mean zero and covariance matrix $\boldsymbol{I}_{d_2}$. $\boldsymbol{V}_2 \in \mathbb{R}^{p\times d_2}$ is a semi-orthogonal matrix satisfying $\boldsymbol{V}_2^T\boldsymbol{V}_2 = \boldsymbol{I}$, and $\boldsymbol{D}_2 \in \mathbb{R}^{d_2\times d_2}$ is a diagonal matrix, measuring the magnitude of the second low-rank confounder. Without loss of generality, we assume that $\boldsymbol{X}^{(i)}$, $\boldsymbol{Z}_1^{(i)}$ and $\boldsymbol{Z}_2^{(i)}$ are three

pairwise independent vectors. Hence the observed covariance matrix is

$$Cov(\boldsymbol{X}_{obs}) = \boldsymbol{\Sigma}_{obs} = \boldsymbol{\Sigma} + \boldsymbol{V}_1\boldsymbol{D}_1^2\boldsymbol{V}_1^T + \boldsymbol{V}_2\boldsymbol{D}_2^2\boldsymbol{V}_2^T = \boldsymbol{\Sigma} + L_1 + L_2. \tag{4.19}$$

Our first simulation setup (case 1) shows an ideal case for PCA+LVGGM, meaning that PCA+LVGGM method performs much better than using PCA+GGM, LVGGM, or Glasso. Let $d_1 = d_2 = 3$. We set $p = 100$ and $n = 100$. In our first example, the columns of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ come from the eigenvectors of $\boldsymbol{\Sigma}$. We expect that PC-correction removes L_2, so we set the diagonal elements of $\boldsymbol{D}_2^2$ all 50, and use the last 3 eigenvectors of $\boldsymbol{\Sigma}$ as $\boldsymbol{V}_2$. This can guarantee that PC-correction performs much better than LVGGM and Glasso when removing L_2. Then we use LVGGM to estimate L_1, so we need a moderately large magnitude. We set all diagonal elements of $\boldsymbol{D}_1^2$ to 20, and use the first 3 eigenvectors of $\boldsymbol{\Sigma}$ as $\boldsymbol{V}_1$. This ensures that LVGGM performs better than PC-correction and Glasso when estimating L_1.

Using the `sva` package, we estimate the rank of $L_1 + L_2$ to be 6. Then we look at the eigenvalues of the observed sample covariance matrix and we can see the first 3 eigenvalues are much larger than the 4-th to 6-th eigenvalues (shown in the top row of Figure 4.2). We therefore allocate 3 to PC-correction, and $6-3 = 3$ to LVGGM. We also try allocating 1 to PC-correction and 5 to LVGGM. Then we compare more approaches, including using PC-correction individually by removing only 3 principal components or 6 principal components, using LVGGM with rank 6 for the low-rank component as well as the uncorrected approach Glasso. We still use 50 datasets and draw the ROC curve for the averages with varying sparsity parameters λ. The ROC for the scale-free example is in the bottom row of Figure 4.2. We also include the AUC (area under the ROC curve) for each method. We compare PCA+LVGGM with rank 3 in PC-correction with other methods. For each data set, we calculate (AUC of PCA+LVGGM)/(AUC of one

other method), then compute the sample mean and sample standard deviation of that ratio over 50 data sets to compare the average performance and the variance of different methods. The results for the scale-free graph are shown in the first column of Table 4.1. We can see that PCA+LVGGM with rank 3 for PC-correction and 3 for LVGGM do perform much better than other methods for both graph structures, indicating that if the assumptions are satisfied, our method and parameter tuning procedure are useful.

Finally, we try setups that are more similar to real world data. We still use (4.19) to generate the data and set $p = 100$ and $n = 100$. Differently from previous settings, we now use some randomly generated eigenvectors as columns of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$. We look at the distribution of eigenvalues of gene co-expression and stock return data covariance matrices, and try to make simulation settings similar to those examples. We run two setups - the first is called a large-magnitude case (case 2), with $\boldsymbol{D}_1^2$ a diagonal matrix with diagonal elements $(7, 6, 6)$ and $\boldsymbol{D}_2^2$ a diagonal matrix with diagonal elements $(20, 10, 10)$. The second setup is referred to a moderately large magnitude case (case 3), in which the low-rank component has the same eigenvectors as the large-magnitude case, but the elements of $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ become smaller, with diagonal elements of $\boldsymbol{D}_1^2$ $(3, 3, 3)$ and diagonal elements of $\boldsymbol{D}_2^2$ $(10, 8, 6)$.

Using the `sva` package, we estimate the rank of $L_1 + L_2$ to be 6 for both case 2 and case 3. We observe that the first 3 eigenvalues are larger than the rest, so we allocate 3 to PC-correction and use $6 - 3 = 3$ as the rank for the low-rank component for LVGGM. We also try allocating 1 to PC-correction and 5 to LVGGM, using PC-correction by removing only 3 PC-components and 6 PC-components, using LVGGM with rank 6 and using Glasso. Again, we run over 50 datasets and include ROC curves and AUC tables.

From Figure 4.2 and Table 4.1, we can see that other approaches considered hardly outperform PCA+LVGGM. Actually, using PCA+GGM or LVGGM can be viewed as a special case of the PCA+LVGMM methods. To see that, we can have LVGGM from

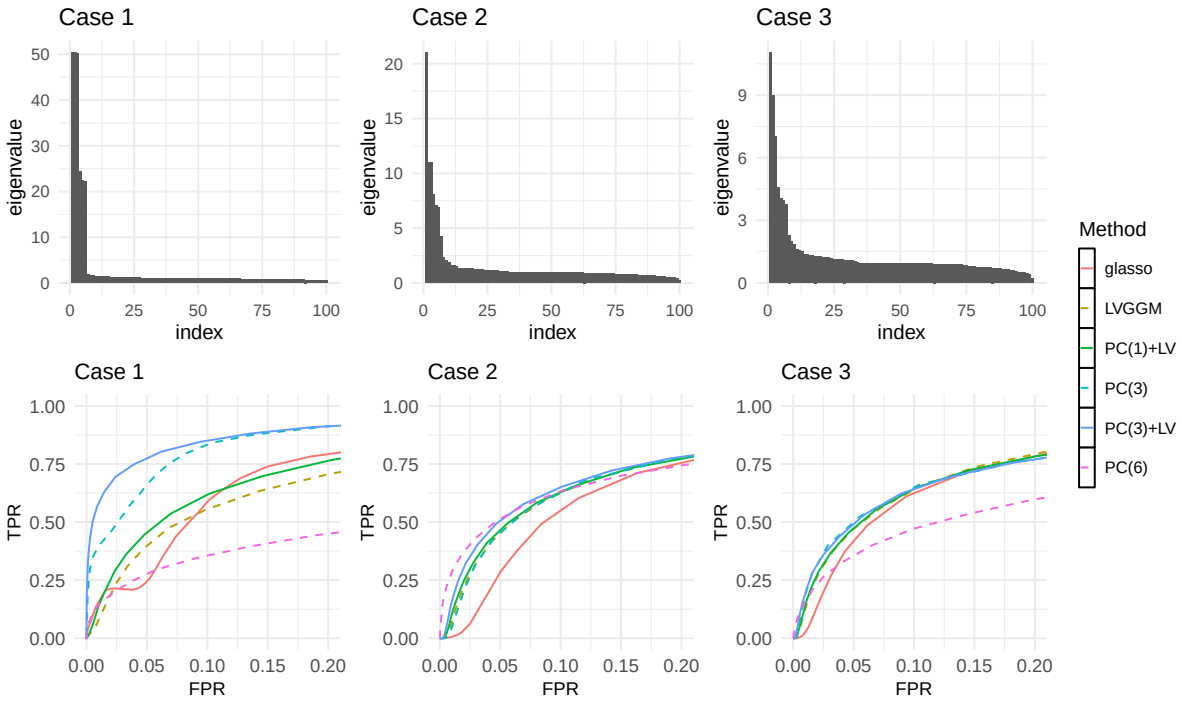


Figure 4.2: We use the scale-free structure when generating graphs. The first row shows the eigenvalues of $\mathbf{\Sigma}_{obs}$ under 3 setups, and the second row shows the corresponding ROC curves with different methods. PC(k) means that we use k as the rank in PC-correction and PC(k)+LV means that we use PCA+LVGGM with k as the rank for PC-correction.

Method	Case 1	Case 2	Case 3
PCA(3)+LVGGM	**1**	**1**	1
Glasso	1.58(0.073)	1.25(0.080)	1.07(0.048)
LVGGM	1.64(0.088)	1.06(0.044)	1.01(0.036)
PCA(Full)	2.46(0.22)	1.01(0.12)	1.36(0.14)
PCA(3)	1.08(0.017)	1.05(0.027)	**0.99(0.018)**
PCA(1)+LVGGM	1.47(0.11)	1.04(0.038)	1.01(0.035)

Table 4.1: We use the scale-free structure when generating graphs. We compute the ratio of AUC between PCA+LVGGM with rank 3 in PC-correction and other methods, using PCA+LVGGM as the numerator. The table shows the sample mean and sample standard deviations of that ratio (in the parenthesis) over 50 data sets. In case 3, the magnitude of the confounding is not as large as other cases, so PC-correction with rank 3 has the best performance.

PCA+LVGGM by allocating a rank of 0 to PC-correction. From the simulation and real data examples, we observe that using PCA+GGM with higher ranks often removes some useful information, resulting in more false negatives. On the other hand, if the effect of multiple confounders exists in the data that are not well represented by the first few principal components, using PCA+GGM alone might not be enough to remove the additional sources of noise. Note that LVGGM may not be enough to remove the confounding with large norm, leading to spurious connections between nodes. In this case, we would suggest PCA+LVGGM as a default setting and a starting point for problems with low-rank confounding. We can adjust different rank allocations based on the specific problems and goals of interest.

4.5 Applications

4.5.1 Gene co-expression networks

Our first application is to reanalyze the gene co-expression networks originally analyzed by [126]. The goal of gene co-expression network analysis is to identify transcriptional patterns indicating functional and regulatory relationships between genes. In biology, it is of great interest to infer the structure of these networks; however, the construction of such networks from data is challenging, since the data is usually corrupted by technical and unwanted biological variability known to confound expression data. The influence of such artifacts can often introduce spurious correlation between genes; if we apply sparse precision matrix inference directly without addressing confounding, we may obtain a graph including many false positive edges. [126] uses PCA+GGM to estimate this network and shows that PC-correction can be an effective way to control the false discovery rate of the network. In practice, however, some effects of confounding may not

be represented in the top few principal components. This motivates the more flexible PCA+LVGGM approach. The PC-correction effectively removes high variance confounding, and then LVGGM subsequently accounts for any remaining low-rank confounding. We consider gene expression data from 3 diverse tissues: blood, lung and tibial nerve, with sample sizes between 300 to 400 each. 1000 genes are chosen from each tissue. More detail about the source of the data and pre-processing steps are introduced in Appendix C.4.

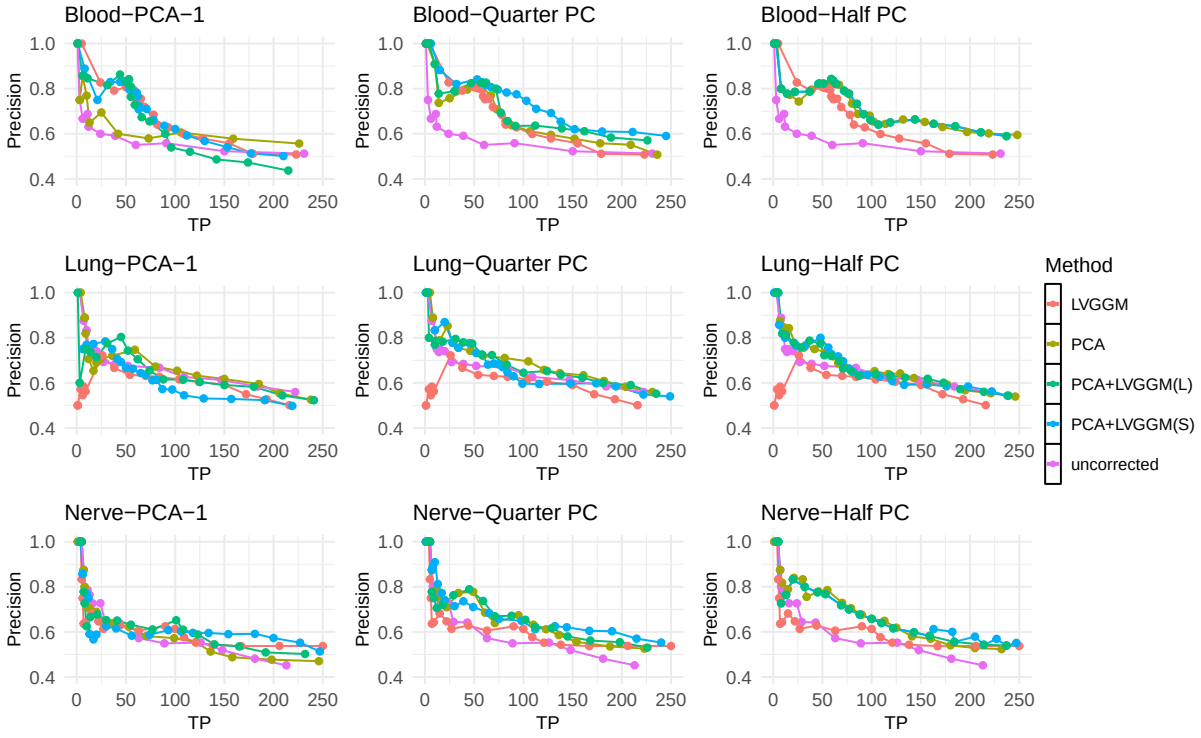


Figure 4.3: Precision-recall plots for gene expression data. TP represents number of true positives. PCA+LVGGM(L) means larger γ in LVGGM and PCA+LVGGM(S) means small γ in LVGGM. We can see that PCA+LVGGM performs the best or equivalently well compared to other approaches for almost all 3 tissues.

We observe that all of the sample covariance matrices are approximately low-rank by looking at the eigenvalues of the covariance matrices of genes, indicating the potential existence of high variance confounding. Then we use `sva` package to estimate the rank

for PC-correction and call this the full `sva` rank correction. [126] suggests that the rank estimated by `sva` might be so large that some useful network signal is removed. To reduce the effect of over-correction, we apply the PC-correction with half and one quarter of the `sva` rank, which we refer to as half `sva` rank correction and quarter `sva` rank correction, respectively. For many tissues, the first eigenvalue is much larger than the rest, this motivates us to try rank-1 PC-correction. We include the results with half `sva` rank, quarter `sva` rank and rank 1 PC-corrections in Figure 4.3. After running the above PC-corrections to remove high-variance confounding, we run LVGGM as an additional step to further estimate and remove the low-rank noise with moderate variance. We use two different values as the γ parameters in LVGGM. Larger γ leads to removing lower-rank confounding and smaller γ leads to remove higher-rank confounding. We show the results for both choices of γ. We use different λ[1] to control sparsity of the estimated graph and draw Figure 4.3 similar to the precision recall plot. The y-axis represents the precision (True Positives/(True Positives + False Positives)), and the x-axis is the number of true positives. We can see that PCA+LVGGM can yield better or equivalently good results compared to other methods, indicating that it can be useful to run LVGGM after the PC-correction when estimating gene co-expression networks.

4.5.2 Stock return data

In finance, the Capital Asset Pricing Model (CAPM) states that there is a widespread market factor which dominates the movement of all stock prices. Empirical evidence for the market trend can be found in the first principal component of the stock data, which is dense and has approximately equal loadings across all stocks (Figure 4.5, left). In fact, the first few eigenvalues of the stock correlation matrix are significantly larger than

[1]Specifically, [126] provides a range of proper λ and based on this, we use 50 values of λ between 0.3 and 1.

the rest [46], which suggests that only a few latent factors are mainly driving stock correlations.

In this section, we posit that the conditional dependence structure after accounting for these latent effects is more likely to reflect direct relationships between companies aside from the market and, perhaps, sector trends. Our interest is in recovering the undirected graphical model (conditional dependence) structure between stock returns after controlling for potential low rank confounders.

We compare networks inferred by PCA+LVGGM, PCA+GGM, LVGGM and Glasso by analyzing monthly returns of component stocks in S&P 100 index between 2008 and 2019 [66]. The 49 chosen companies are in 6 sectors: technology (10 companies), finance (11), energy (7), health (8), capital goods (7) and non-cyclical stocks (6). For PCA+GGM, we remove the first eigenvector which corresponds to the overall market trend. For the other latent variable methods we use the `sva` package to identify a plausible rank. For PCA+LVGGM, we remove the first principal component corresponding to the overall market trend and use LVGGM to estimate remaining latent confounders and the graph. Figure 4.4 shows the networks obtained by each approach.

For each method, the sparsity-inducing tuning parameter was chosen to minimize negative log-likelihood using a 6-fold cross-validation procedure, and the number of low rank components are chosen manually. Specifically, in cross-validation, we use negative log-likelihood to measure the out-of-sample error and choose the parameters that minimize the average out-of sample error over 6 validation sets. We observe that when using LVGGM, allocating rank 1 or 2 to the low-rank component won't make the estimates very different from Glasso, while allocating ranks higher than 6 to the low-rank component leads to higher out-of-sample error, so 5, the rank picked by sva, is among the best choices. For PCA-based methods, removing more than 1 principal components leads to higher out-of-sample error. As expected, the Glasso result is denser than the networks

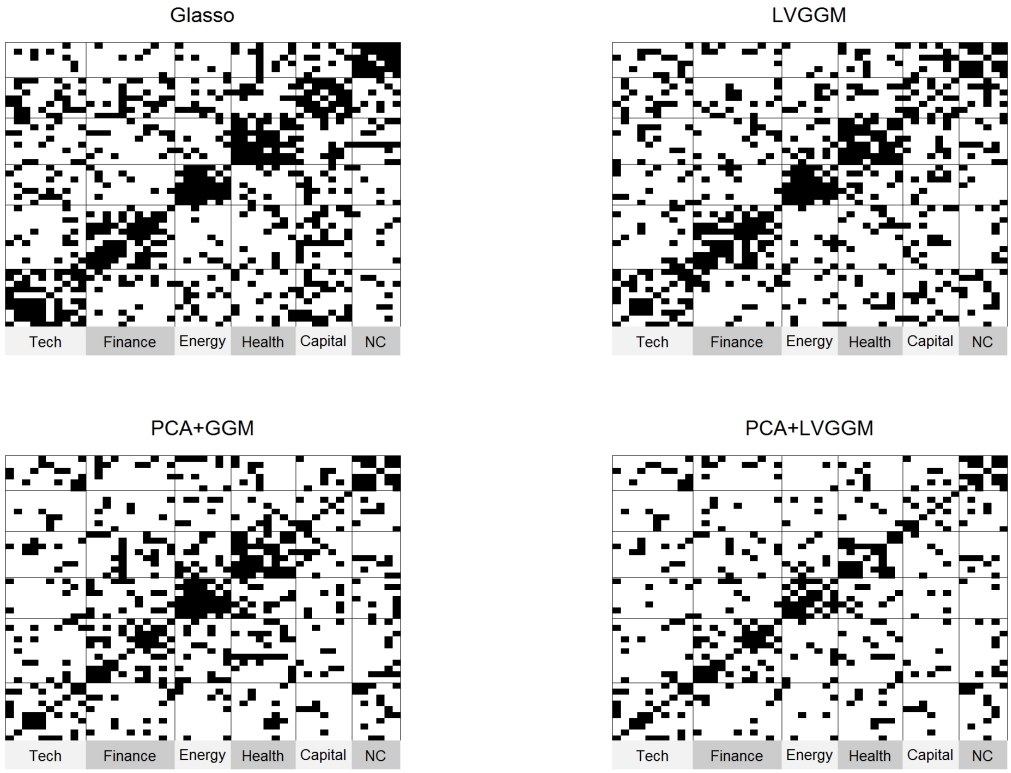


Figure 4.4: Stock connections between 2008 and 2019 learned by different methods. The following sectors are included: Tech, Finance, Energy, Health, Capital goods and Non-cyclical (NC) from left to right.

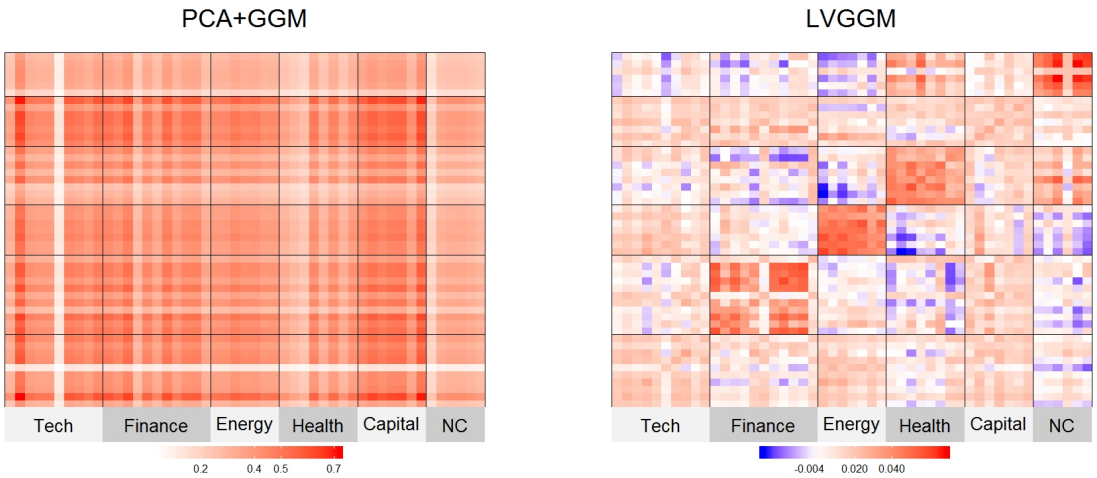
learned with sparse plus low rank methodology with PCA+LVGGM yielding the sparsest network.

For LVGGM, we note that the method effectively controls for sector effect but is less effective in controlling for the effect of the overall market trend. Let $\hat{\mathbf{\Sigma}}_{obs}$ be the empirical observed covariance matrix and $\hat{\boldsymbol{v}}_i$ be its i-th eigenvector. We have the following observations: first, the first principal component is closely aligned with the overall market trend, because the absolute value of the inner product between the first eigenvector of $\hat{\mathbf{\Sigma}}_{obs}$ and the normalized "all ones" vector is 0.98. Second, the observed empirical covariance matrix has an approximately low-rank structure, because the first eigenvalue of $\hat{\mathbf{\Sigma}}_{obs}$ is

18.25 and the second is 3.5 and all other eigenvalues are close or smaller than 1. Third, LVGGM does not capture the full effect of the market trend. Let $\hat{L}'_{\mathbf{\Omega}}$ be the estimate of $L_{\mathbf{\Omega}}{}'$ in (4.9). When we apply LVGGM on $\hat{\mathbf{\Sigma}}_{obs}$, the inner product between the first eigenvector of $\hat{L}'_{\mathbf{\Omega}}$ and $\hat{\boldsymbol{v}}_1$ is close to 1 but the first eigenvalue of $\hat{L}'_{\mathbf{\Omega}}$ is only 0.55, much smaller than the first eigenvalue of $\hat{\mathbf{\Sigma}}_{obs}$.

We argue that PCA+LVGGM is the most appropriate method for this application because it appropriately controls for both market and sector effects. Let $\hat{L}_{\mathbf{\Sigma}}$ and $\hat{L}'_{\mathbf{\Omega}}$ be the estimates of the low-rank components defined in (4.10). For PCA+LVGGM, we remove $L_{\mathbf{\Sigma}}$ by removing the first eigencomponent of $\mathbf{\Sigma}_{obs}$, then run LVGGM to estimate $L_{\mathbf{\Omega}}$ and $\mathbf{\Omega}$. We claim that PCA+LVGGM can remove the confounding effect fully in the market trend direction, as well as the remaining confounding effect in other directions. To see that, first, $\hat{\boldsymbol{v}}_1$ is removed in PC-correction. Second, the inner product between the first eigenvector of $\hat{L}'_{\mathbf{\Omega}}$ and $\hat{\boldsymbol{v}}_2$, the second eigenvector of $\hat{\mathbf{\Sigma}}_{obs}$, is 0.99. The first eigenvalue of $\hat{L}'_{\mathbf{\Omega}}$ is 0.4 and the second eigenvalue of $\hat{\mathbf{\Sigma}}_{obs}$ is 3.5. This shows that when applying LVGGM, only part of the information in the direction of $\hat{\boldsymbol{v}}_2$ has been removed. We know that the direction of $\hat{\boldsymbol{v}}_1$ reflects the market trend, but $\hat{\boldsymbol{v}}_2$ might include both true graph information and some latent confounding effect, hence using LVGGM might be a good choice for capturing the confounding effect in the direction of $\hat{\boldsymbol{v}}_2$. Overall PCA+LVGGM, therefore, might be a better choice than LVGGM and the PCA-based method.

Figure 4.5 shows heat maps of $\hat{L}_{\mathbf{\Sigma}}$ obtained via PCA and $\hat{L}'_{\mathbf{\Omega}}$ obtained with LVGGM (rank 5). As expected, the elements of $\hat{L}_{\mathbf{\Sigma}}$ are roughly equal in magnitude, reflect the market trend and the large first eigenvalue of $\hat{\mathbf{\Sigma}}_{obs}$. In contrast, $\hat{L}'_{\mathbf{\Omega}}$ shows a block-diagonal structure and its elements have smaller magnitudes, which suggests that LVGGM does not adequately account for the overall the market trend. On the other hand, the block diagonal structure of $\hat{L}'_{\mathbf{\Omega}}$ reflects inferred sector effects. PCA+GGM is most effective at



(a) (b)

Figure 4.5: (a) Rank one approximation to $\hat{L}_{\mathbf{\Sigma}}$ obtained with PCA. (b) $\hat{L}'_{\mathbf{\Omega}}$ obtained with LVGGM with rank 5. The rank-one approximation to $\hat{L}_{\mathbf{\Sigma}}$ is close to a constant matrix. In contrast, $\hat{L}'_{\mathbf{\Omega}}$ reflects sector effects but does not reflect the strong effect due to overall market trends.

reducing confounding from overall market trends and LVGGM is more effective at accounting for remaining confounding, such as the sector effect. Therefore, PCA+LVGGM, which combines the benefits of PCA and LVGGM is arguably the most appropriate choice for addressing the latent confounding in this context.

Appendix A

Appendix for Chapter 2

A.1 Key Technical Lemmas

For the reader's convenience, we repeat here some definitions and lemmas that were previously stated in Section 2.8.2. Define $\boldsymbol{U}_\tau := \boldsymbol{Q}\boldsymbol{Q}^T + \tau\boldsymbol{I}$ and $\boldsymbol{d} := \boldsymbol{Q}\boldsymbol{\eta}$; thus $\boldsymbol{U}_0 = \boldsymbol{Q}\boldsymbol{Q}^T$. The lemma below expresses $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}$ in terms of the following quadratic forms:

$$
\begin{aligned}
s &= \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{y}, \\
t &= \boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{d}, \\
h &= \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{d}, \\
g_i &= \boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i, \quad i \in [n], \\
f_i &= \boldsymbol{d}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i, \quad i \in [n].
\end{aligned}
$$

Lemma 8. *Define* $D := s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2$, *then*

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1} = \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - \frac{1}{D}\Big[\|\boldsymbol{\eta}\|_2 s, h^2 + h - st, s\Big]\begin{bmatrix} \|\boldsymbol{\eta}\|_2\boldsymbol{y}^T \\ \boldsymbol{y}^T \\ \boldsymbol{d}^T \end{bmatrix}\boldsymbol{U}_\tau^{-1}. \tag{A.1}$$

The lemma below derives upper/lower bounds for those quadratic forms involving the inverse Gram matrix $\boldsymbol{U}_\tau^{-1}$.

Lemma 9 (Balanced)**.** *Recall that* $\sigma^2 = \sum_{i=1}^p \lambda_i\beta_i^2$. *Assume the* $\boldsymbol{\Sigma}$ *follows the balanced ensemble defined in Definition 2.2.1. Fix* $\delta \in (0,1)$ *and suppose* n *is large enough such that* $n > c\log(1/\delta)$ *for some* $c > 1$. *Then, there exists constants* $C_1, C_2, C_3, C_6, C_7 > 1$, $C_5 > C_4 > 0$ *such that with probability at least* $1-\delta$, *the following results hold:*

$$\begin{aligned}
\frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_1)} \leq & s \leq C_1\frac{n}{(\tau + \|\boldsymbol{\lambda}\|_1)}, \\
C_4\frac{n\sigma^2}{(\tau + \|\boldsymbol{\lambda}\|_1)} \leq & t \leq C_5\frac{n\sigma^2}{(\tau + \|\boldsymbol{\lambda}\|_1)}, \\
-C_2\frac{n\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)} \leq & h \leq C_2\frac{n\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)}, \\
\|\boldsymbol{d}\|_2^2 \leq & C_3 n\sigma^2, \\
\|\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\|_2 \leq & C_6\frac{\sqrt{n}}{(\tau + \|\boldsymbol{\lambda}\|_1)}, \\
\|\boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\|_2 \leq & C_7\frac{\sqrt{n}\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)}.
\end{aligned}$$

To bound the term f_i, we need some additional work, which leads to the following result.

Lemma 10. *Assume that the condition in* (2.8) *is satisfied, Fix* $\delta \in (0,1)$ *and suppose* n *is large enough such that* $n > c/\delta$ *for some* $c > 1$. *Then, there exists a constant* $C > 1$

such that with probability at least $1-\delta$,

$$\max_{i\in[n]}|f_i| \leq \frac{C\sqrt{\log(2n)}\sigma}{\|\boldsymbol{\lambda}\|_1}. \tag{A.2}$$

The proofs of Lemmas 8, 9 and 10 are given in Section A.7. We will also need the following lemmas adapted from [119, Proof of Theorem 1].

Lemma 11. *Let* $\boldsymbol{E} = \boldsymbol{Q}\boldsymbol{Q}^T - \|\boldsymbol{\lambda}\|_1 \cdot \boldsymbol{I}$ *and* $\boldsymbol{E}' = \frac{1}{\|\boldsymbol{\lambda}\|_1} \cdot (\boldsymbol{Q}\boldsymbol{Q}^T)^{-1}\boldsymbol{E}$. *Assume that the condition in* (2.8) *is satisfied, then there exists a constant* $C > 1$ *such that with probability at least* $(1-\frac{C}{n})$,

$$\|\boldsymbol{E}'\|_2 \leq \frac{1}{2\sqrt{n}\|\boldsymbol{\lambda}\|_1}. \tag{A.3}$$

Lemma 12. *Let* $d'(n) := (p-n+1)$. *With probability at least* $(1-\frac{2}{n^2})$,

$$y_i g_i = y_i(\boldsymbol{e}_i^T \boldsymbol{U}_0^{-1}\boldsymbol{y}) \geq \frac{1}{4\sqrt{n}} \frac{2\sqrt{n}d'(n) - 2n\sqrt{4\log(n)d'(n)} - 4n\log(n)}{(d'(n)+\sqrt{4\log(n)d'(n)})(d'(n)-\sqrt{4\log(n)d'(n)})}, \text{ for } i \in [n].$$

A.2 Proof of Theorem 1 and Theorem 2

A.2.1 Proof of Theorem 1

Now we are ready to prove Theorem 1. In this section, we only consider the unregularized estimator, i.e., $\tau = 0$. Define $\gamma^* := (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$. Using duality (see [119, Appendix C.1]), all the constraints in (2.5) hold with equality provided that

$$y_i\gamma_i^* > 0, \text{ for all } i \in [n]. \tag{A.4}$$

Hence it suffices to derive conditions under which (A.4) holds with high probability. Note that $\gamma_i^* = \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{e}_i$, for all $i \in [n]$. Using (A.1) and some algebra steps, it can be checked that:

$$\begin{aligned}
\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{e}_i &= \boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i - \frac{1}{D}\begin{bmatrix}\|\boldsymbol{\eta}\|_2 s & h^2+h-st & s\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i \\ \boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i \\ \boldsymbol{d}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i\end{bmatrix} \\
&= g_i - \frac{1}{D}\begin{bmatrix}\|\boldsymbol{\eta}\|_2 s & h^2+h-st & s\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\eta}\|_2 g_i \\ g_i \\ f_i\end{bmatrix} \\
&= \frac{g_i(s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2) - \|\boldsymbol{\eta}\|_2^2 s g_i - (h^2+h-st)g_i - s f_i}{D} \\
&= \frac{g_i + h g_i - s f_i}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}. \qquad \text{(A.5)}
\end{aligned}$$

Here, s, h, t, g_i and f_i are as defined in Section A.1 with $\tau = 0$. The denominator of (A.5) is non-negative, thus to make $\gamma_i > 0$, we only need to study the numerator:

$$y_i(g_i + h g_i - s f_i) = (1 + \boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{d})y_i(\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{y}) - y_i(\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{d})\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{y}.$$

First, consider the term $y_i(\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{y})$. By the proof of [119, Theorem 1], if (2.8) is satisfied, then with probability at least $(1 - \frac{C}{n})$,

$$y_i g_i \geq \frac{1}{2\|\boldsymbol{\lambda}\|_1}. \qquad \text{(A.6)}$$

We know that $\boldsymbol{U}_0^{-1} = \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} - \boldsymbol{E}'$. Thus for $\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{d}$, by Lemma 9 and 11, with probability at least $(1 - \frac{C}{n})$,

$$\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{d} = \boldsymbol{y}^T(\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} - \boldsymbol{E}')\boldsymbol{d} \geq -\frac{C_1 n\sigma}{\|\boldsymbol{\lambda}\|_1} - \frac{C_2\sqrt{n}\sigma}{\|\boldsymbol{\lambda}\|_1} \geq -\frac{C_3 n\sigma}{\|\boldsymbol{\lambda}\|_1},$$

where the first inequality above follows from the fact $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \geq -\|\boldsymbol{v}\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{M}\|_2$. Lemma 10 gives for every $i \in [n]$, with the same high probability,

$$y_i\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{d} = y_i f_i \geq -\max_{i\in[n]}|f_i| \geq -\frac{C_4\sqrt{\log(2n)}\sigma}{\|\boldsymbol{\lambda}\|_1}.$$

Similarly, the fact $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \leq \|\boldsymbol{v}\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{M}\|_2$ gives with probability at least $1 - \delta$,

$$\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{y} = \boldsymbol{y}^T(\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} - \boldsymbol{E}')\boldsymbol{y} \leq \boldsymbol{y}^T\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{y} \leq \frac{C_5 n}{\|\boldsymbol{\lambda}\|_1}.$$

Combining the results above gives

$$\begin{aligned} y_i(g_i + hg_i - sf_i) &\geq \Big(\frac{\|\boldsymbol{\lambda}\|_1 - C_1 n\sigma}{\|\boldsymbol{\lambda}\|_1}\Big)\frac{1}{2\|\boldsymbol{\lambda}\|_1} - \frac{C_2 n\sqrt{\log(2n)}\sigma}{\|\boldsymbol{\lambda}\|_1^2} \\ &\geq \frac{\|\boldsymbol{\lambda}\|_1 - C_1 n\sigma - 2C_2 n\sqrt{\log(2n)}\sigma}{2\|\boldsymbol{\lambda}\|_1^2}. \end{aligned}$$

To make the expression above positive, it suffices to have $\|\boldsymbol{\lambda}\|_1 \geq Cn\sqrt{\log(2n)}\sigma$. This completes the proof.

A.2.2 Proof of Theorem 2

According to section A.2.1, we need to study:

$$y_i(g_i + hg_i - sf_i) = (1 + \boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{d})y_i(\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{y}) - y_i(\boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{d})\boldsymbol{y}^T\boldsymbol{U}_0^{-1}\boldsymbol{y}.$$

First, consider the term $y_i(\boldsymbol{e}_i^T \boldsymbol{U}_0^{-1}\boldsymbol{y})$. By Lemma 12, if $d^{'}(n) = p - n + 1 > 9n\log(n)$, then, $4n\log(n) < \frac{4}{9}d^{'}(n)$ gives

$$\begin{aligned} y_i g_i = y_i(\boldsymbol{e}_i^T \boldsymbol{U}_0^{-1}\boldsymbol{y}) &> \frac{1}{4\sqrt{n}} \frac{2\sqrt{n}d^{'}(n) - \frac{4}{3}\sqrt{n}d^{'}(n) - \frac{4}{9}d^{'}(n)}{(d^{'}(n) + \sqrt{4/(9n)}d^{'}(n))d^{'}(n)} \\ &> \frac{1}{4\sqrt{n}} \frac{2\sqrt{n}d^{'}(n) - \frac{4}{3}\sqrt{n}d^{'}(n) - \frac{4}{9}\sqrt{n}d^{'}(n)}{2d^{'}(n)^2} \\ &> \frac{1}{4\sqrt{n}} \frac{(2 - \frac{4}{3} - \frac{4}{9})\sqrt{n}}{2p} \\ &> \frac{1}{36p}. \end{aligned} \tag{A.7}$$

Second, by Lemmas 9, 10 and (A.7), we find that for large enough constants C_i's > 1, with probability at least $1 - \frac{C_1}{n^2}$,

$$\begin{aligned} y_i(g_i + hg_i - sf_i) &= (1+h)y_i g_i - y_i s f_i \\ &\geq (1 - |h|)\frac{1}{36p} - \max_{i\in[n]} |f_i| s \\ &\geq (1 - \frac{C_2 n}{p}\|\boldsymbol{\eta}\|_2)\frac{1}{36p} - \frac{C_3\sqrt{\log(2n)}}{p}\|\boldsymbol{\eta}\|_2 \frac{n}{p} \\ &\geq \frac{p - C_2 n\|\boldsymbol{\eta}\|_2 - 36C_3\sqrt{2\log(2n)}n\|\boldsymbol{\eta}\|_2}{36p^2} \\ &\geq \frac{p - 36C_4\sqrt{2\log(2n)}n\|\boldsymbol{\eta}\|_2}{36p^2}. \end{aligned}$$

To make the expression above positive, it suffices to have $p > C_5 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2$, for large enough $C_5 > 1$. The result above holds for every γ_i^*, $i \in [n]$ with probability $1 - \frac{C_1}{n^2}$ each (by Lemma 11). Applying union bound over all n training data points, we conclude that $y_i\gamma_i^* > 0$ for all i with probability at least $1 - \frac{C_1}{n}$. This completes the proof.

A.3 Proof of Theorem 3 and Theorem 4

A.3.1 Proof of Theorem 3

From Section 2.8, we need to lower bound the ratio

$$\frac{(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\eta})^2}{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}}. \tag{A.8}$$

We will upper bound the denominator and lower bound the numerator. We first look at the denominator. We know $\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T = (\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})\boldsymbol{\Lambda}(\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})^T$. Define $\boldsymbol{A} := (\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}$. Then by the cyclic property of trace, the denominator of (A.8) can be expressed as

$$\begin{aligned}
&\mathrm{Tr}\Big(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}\Big)\\
&= \mathrm{Tr}\Big((\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})\boldsymbol{\Lambda}(\boldsymbol{y}\boldsymbol{\beta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}})^T\boldsymbol{A}\Big)\\
&= \sum_{i=1}^p \lambda_i^2\boldsymbol{z}_i^T\boldsymbol{A}\boldsymbol{z}_i + \sum_{i=1}^p \lambda_i\beta_i^2(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) + 2\sum_{i=1}^p \lambda_i^{1.5}\beta_i\boldsymbol{z}_i^T\boldsymbol{A}\boldsymbol{y}\\
&\le 2\Big(\sum_{i=1}^p \lambda_i^2\boldsymbol{z}_i^T\boldsymbol{A}\boldsymbol{z}_i + \sigma^2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}\Big)\\
&\le 2\Big(\sum_{i=1}^p \lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2 + \sigma^2(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y})^2\Big),
\end{aligned}$$

where the first inequality follows from the inequality $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \le \frac{1}{2}(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} + \boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u})$ for positive semidefinite matrix $\boldsymbol{M}$ and $\boldsymbol{z}_i$ is the i-th column of matrix $\boldsymbol{Z}$. Thus, we need to upper bound $\sum_{i=1}^p \lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2$ and $\sigma^2(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y})^2$. For $\sum_{i=1}^p \lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2$, note that $\|\boldsymbol{z}_i\|_2^2$'s are independent sub-exponential random variables [162, Chapter 2], thus for a fixed number $B > 0$, $\sum_{i=1}^p \lambda_i^2 B\|\boldsymbol{z}_i\|_2^2$ is the weighted sum of sub-exponential random variables, with the weights given by $B\lambda_i^2$ in blocks of size n [10, Lemma 7 and

Corollary 1]. By Lemma 15.1, with probability at least $1 - 2e^{-x}$,

$$\begin{aligned}\sum_{i=1}^{p} \lambda_i^2 B\|\boldsymbol{z}_i\|_2^2 &\le Bn\sum_{i=1}^{p}\lambda_i^2 + Bc\max\Big(\lambda_1^2 x, \sqrt{xn\sum_{i}^{p}\lambda_i^4}\Big)\\ &\le Bn\sum_{i=1}^{p}\lambda_i^2 + Bc\max\Big(x\sum_{i=1}^{p}\lambda_i^2, \sqrt{xn}\sum_{i}^{p}\lambda_i^2\Big)\\ &\le CnB\sum_{i=1}^{p}\lambda_i^2,\end{aligned}$$

for $x < n/c_0$. The number B above should be replaced by the upper bound of $\|\boldsymbol{A}\|_2$. Recall $\boldsymbol{A} := (\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}$, thus $\|\boldsymbol{A}\|_2 = \|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\|_2^2$. Further recalling $D := s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2$, by Lemma 8,

$$\begin{aligned}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1} &= \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - \frac{1}{D}\begin{bmatrix}\|\boldsymbol{\eta}\|_2 s & h^2 + h - st & s\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_\tau^{-1}\\ &= \frac{1}{D}\Big((1+h)\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - s\boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\Big).\end{aligned}$$

Therefore, by Lemma 9, with probability at least $1 - \delta$,

$$\begin{aligned}\|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\|_2 &\le \frac{1}{D}\bigg(\Big(1 + |h|\Big)\|\boldsymbol{y}\|_2\|\boldsymbol{U}_\tau^{-1}\|_2 + s\|\boldsymbol{d}\|_2\|\boldsymbol{U}_\tau^{-1}\|_2\bigg)\\ &\le \frac{1}{D}\bigg(\Big(1 + \frac{C_1 n\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)}\Big)\frac{C_2\sqrt{n}}{(\tau + \|\boldsymbol{\lambda}\|_1)} + \frac{C_3 n\sqrt{n}\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)^2}\bigg).\end{aligned}$$

The above result can be further simplified. If $1 \le \frac{n\sigma}{\tau + \|\boldsymbol{\lambda}\|_1}$, then,

$$\|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1}\|_2 \le \frac{1}{D}\frac{C_4 n\sqrt{n}\sigma}{(\tau + \|\boldsymbol{\lambda}\|_1)^2}.$$

If $1 > \frac{n\sigma}{\tau+\|\boldsymbol{\lambda}\|_1}$, then,

$$\|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\|_2 \leq \frac{1}{D}\frac{C_5\sqrt{n}}{(\tau+\|\boldsymbol{\lambda}\|_1)}.$$

Combining the above gives, with probability at least $1-\delta$,

$$\sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{A}\|_2\|\boldsymbol{z}_i\|_2^2 \leq \frac{C}{D^2}\frac{n^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\max\{1,\frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\}\|\boldsymbol{\lambda}\|_2^2. \tag{A.9}$$

Now we look at $\sigma^2(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y})^2$. We need to upper bound $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}$. Using Lemma 9 gives with probability at least $1-\delta$,

$$\begin{aligned}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y} &= s - \frac{\|\boldsymbol{\eta}\|_2^2 s^2 + sh^2 + 2sh - s^2 t}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2} \\ &= \frac{s}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2} \\ &= \frac{s}{D} \leq \frac{1}{D}\frac{Cn}{(\tau+\|\boldsymbol{\lambda}\|_1)}.\end{aligned}$$

Therefore, $\sigma^2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} \leq \frac{C}{D^2}\frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}$. Hence the denominator of (A.8) is upper bounded by

$$\frac{1}{D^2}\frac{n^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\Big(C_1\max\{1,\frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\}\|\boldsymbol{\lambda}\|_2 + C_2\sigma^2\Big). \tag{A.10}$$

Now we look at the numerator of (A.8). By Lemma 8,

$$\begin{aligned}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{\eta} &= \|\boldsymbol{\eta}\|_2^2\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y} + \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{Q}\boldsymbol{\eta} \\ &= \frac{s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h}{D}.\end{aligned}$$

The numerator needs to be lower bounded and Lemma 9 gives with probability at least $1-\delta$,

$$s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h \geq s(\|\boldsymbol{\eta}\|_2^2 - t) + h \geq \frac{n}{C(\tau + \|\boldsymbol{\lambda}\|_1)}\Big(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\tau + \|\boldsymbol{\lambda}\|_1} - C_2\sigma\Big). \tag{A.11}$$

Combining (A.10) and (A.11) gives with probability at least $1-\delta$, (A.8) is lower bounded by

$$\frac{\Big(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\tau+\|\boldsymbol{\lambda}\|_1} - C_2\sigma\Big)^2}{C_3 \max\{1, \frac{n^2\sigma^2}{(\tau+\|\boldsymbol{\lambda}\|_1)^2}\}\|\boldsymbol{\lambda}\|_2^2 + C_4\sigma^2}. \tag{A.12}$$

This completes the proof of the theorem.

A.3.2 Proof of Theorem 4

We need to lower bound the ratio

$$\frac{(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta})^2}{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}}. \tag{A.13}$$

Here we will lower bound $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta}$ and upper bound $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$. By Lemma 1, we know that the bound is not useful if $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta} < 0$, hence we need the conditions that ensure $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta} \geq 0$ with high probability. Using (A.1) and some algebra steps, it can be checked that:

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y} = s - \frac{\|\boldsymbol{\eta}\|_2^2 s^2 + sh^2 + 2sh - s^2 t}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2} = \frac{s}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}. \tag{A.14}$$

Similarly,

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y} + \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Q}\boldsymbol{\eta} = \frac{s\|\boldsymbol{\eta}\|_2^2 - st + h^2 + h}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}.$$

Combining the above gives

$$\frac{(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta})^2}{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}} = \frac{\Big(s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h\Big)^2}{s\Big(s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2\Big)}. \tag{A.15}$$

The numerator needs to be lower bounded and Lemma 9 gives with probability at least $1-\delta$,

$$\begin{aligned} s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h &\geq s(\|\boldsymbol{\eta}\|_2^2 - t) + h \\ &\geq \frac{n}{C_1 p}(1 - \frac{n}{p})\|\boldsymbol{\eta}\|_2^2 - C_2\frac{n}{p}\|\boldsymbol{\eta}\|_2 \\ &\geq \frac{n}{C_1 p}\Big((1 - \frac{n}{p})\|\boldsymbol{\eta}\|_2^2 - C_3\|\boldsymbol{\eta}\|_2\Big). \end{aligned} \tag{A.16}$$

Similarly, the denominator is upper bounded by:

$$\begin{aligned} s\Big(s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2\Big) &\leq s\Big(s\|\boldsymbol{\eta}\|_2^2 + (1+|h|)^2\Big) \\ &\leq C_1\frac{n}{p}\Big(C_1\frac{n}{p}\|\boldsymbol{\eta}\|_2^2 + (1 + C_2\frac{n}{p}\|\boldsymbol{\eta}\|_2)^2\Big) \\ &\leq C_1\frac{n}{p}\Big(C_3\frac{n}{p}\|\boldsymbol{\eta}\|_2^2 + C_4\Big) \\ &\leq C_5\frac{n^2}{p^2}\Big(\|\boldsymbol{\eta}\|_2^2 + \frac{p}{n}\Big), \end{aligned}$$

where we also use the fact $(a+b)^2 \leq 2(a^2+b^2)$ and $n < p$. Combining the above results gives with probability at least $1-\delta$,

$$\frac{(\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta})^2}{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}} \geq \|\boldsymbol{\eta}\|_2^2 \frac{\left((1-\frac{n}{p})\|\boldsymbol{\eta}\|_2 - C_3\right)^2}{C_6(\frac{p}{n}+\|\boldsymbol{\eta}\|_2^2)}. \tag{A.17}$$

To ensure the classification error is smaller than 0.5, we need $p > b{\cdot}n$ and $(1-\frac{n}{p})\|\boldsymbol{\eta}\|_2 > C_3$ for $b > 1$ to make $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}^T\boldsymbol{\eta} > 0$ with high probability. This completes the proof of the theorem.

A.4 Proof of Theorem 5 and benign overfitting for the bi-level ensemble

A.4.1 Proof of Theorem 5

We first introduce some new notations. Following Assumption 1, we assume that the covariance matrix $\boldsymbol{\Sigma}$ is diagonal and the mean vector $\boldsymbol{\eta}$ is one-sparse ($\eta_k \neq 0$ and $k \neq 1$). Hence the data matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q} = \boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Z}\boldsymbol{\Lambda}^{\frac{1}{2}}.$$

Let $\boldsymbol{z}_i$ be the i-th column of the matrix $\boldsymbol{Z}$ above whose elements are IID standard Gaussian. Recall $\boldsymbol{U}_\tau := \boldsymbol{Q}\boldsymbol{Q}^T + \tau \boldsymbol{I}$, define

$$
\begin{aligned}
s &:= \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{y}, \\
t_k &:= \boldsymbol{z}_k^T \boldsymbol{U}_\tau^{-1} \boldsymbol{z}_k, \\
f_1 &:= \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{z}_1, \\
f_k &:= \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{z}_k, \\
g_1 &:= \boldsymbol{z}_1^T \boldsymbol{U}_\tau^{-1} \boldsymbol{z}_k.
\end{aligned}
$$

Lemma 13 (Bi-level)**.** *Assume that $\boldsymbol{\Sigma}$ follows the bi-level ensemble defined in Definition 2.2. Fix $\delta \in (0,1)$ and suppose n is large enough such that $n > c\log(1/\delta)$ for some $c > 1$. Then, there exists constants C_i's > 1 such that with probability at least $1-\delta$, the following results hold:*

$$
\begin{aligned}
\frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_{-1})} \leq & s \leq \frac{C_2 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}, \\
-\frac{C_3 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1})} \leq & f_k \leq \frac{C_3 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}, \\
\frac{n}{C_4(\tau + \|\boldsymbol{\lambda}\|_{-1})} \leq & t_k \leq \frac{C_5 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}, \\
-\frac{C_6 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1} + n\lambda_1)} \leq & f_1 \leq \frac{C_6 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1} + n\lambda_1)}, \\
-\frac{C_7 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1} + n\lambda_1)} \leq & g_1 \leq \frac{C_7 n}{(\tau + \|\boldsymbol{\lambda}\|_{-1} + n\lambda_1)}, \\
\|\boldsymbol{y}^T \boldsymbol{U}_\tau^{-1}\|_2 \leq & \frac{C_8 \sqrt{n}}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}, \\
\|\boldsymbol{z}_k^T \boldsymbol{U}_\tau^{-1}\|_2 \leq & \frac{C_9 \sqrt{n}}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}.
\end{aligned}
$$

Now we are ready to prove Theorem 5. We know from the proof outline that we need

to lower bound

$$\frac{(\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}_\tau^T \boldsymbol{\Sigma} \hat{\boldsymbol{\eta}}_\tau} = \frac{(\eta_k \hat{\eta}_k)^2}{\sum_{i=1}^p \lambda_i \hat{\eta}_i^2}. \tag{A.18}$$

We divide $\hat{\eta}_i$'s into 3 groups: $\hat{\eta}_1$, $\hat{\eta}_k$ and the rest. Rather than lower bounding (A.18), we will upper bound its reciprocal. Specifically, we will upper bound

$$\frac{\lambda_1 \hat{\eta}_1^2}{(\eta_k \hat{\eta}_k)^2}, \quad \frac{\sum_{i \neq 1,k} \lambda_i \hat{\eta}_i^2}{(\eta_k \hat{\eta}_k)^2} \quad \text{and} \quad \frac{\lambda_k \hat{\eta}_k^2}{(\eta_k \hat{\eta}_k)^2}, \tag{A.19}$$

then reverse the sum of the upper bounds of the three ratios above to obtain the lower bound of (A.18).

Following the fact that $\hat{\eta}_i = \boldsymbol{e}_i^T \hat{\boldsymbol{\eta}}$, we have

$$\hat{\eta}_i = \sqrt{\lambda_i} \boldsymbol{z}_i^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}, \quad \text{for} \;\; i \neq k, \tag{A.20}$$

$$\hat{\eta}_k = \eta_k \boldsymbol{y}^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y} + \sqrt{\lambda_k} \boldsymbol{z}_k^T (\boldsymbol{X}\boldsymbol{X}^T + \tau \boldsymbol{I})^{-1} \boldsymbol{y}. \tag{A.21}$$

To upper bound the 3 terms in (A.19), we need to lower bound $(\eta_k \hat{\eta}_k)^2$. Recall that under Assumption 1, σ^2 is $\lambda_k \eta_k^2$. By Lemma 8 and using our newly defined notations, we have

$$\begin{aligned}
\eta_k \hat{\eta}_k &= \frac{\eta_k^2 s}{D} + \sqrt{\lambda_k} \eta_k \Big(f_k - \frac{\|\boldsymbol{\eta}\|_2^2 s f_k + ((\sqrt{\lambda_k} \eta_k f_k)^2 + \sqrt{\lambda_k} \eta_k f_k - s(\lambda_k \eta_k^2 t_k)) f_k + \sqrt{\lambda_k} \eta_k t_k s}{D} \Big) \\
&= \frac{1}{D} \Big(\eta_k^2 s (1 - \lambda_k t_k) + \sigma f_k + \sigma^2 f_k^2 \Big),
\end{aligned}$$

where D becomes $(\sigma f_k + 1)^2 + s(\eta_k^2 - \sigma^2 t_k)$. Lemma 13 gives with probability at least $1 - \delta$,

$$\eta_k \hat{\eta}_k \geq \frac{1}{D} \frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_{-1})} \Big(\eta_k^2 (1 - \frac{C_2 n \lambda_k}{\tau + \|\boldsymbol{\lambda}\|_{-1}}) - \sigma \Big). \tag{A.22}$$

Now we upper bound $\sum_{i\neq 1,k}\lambda_i\hat{\eta}_i^2$. From the proof of Theorem 3, we know

$$\sum_{i\neq 1,k}\lambda_i\hat{\eta}_i^2=\sum_{i\neq 1,k}\lambda_i^2\boldsymbol{z}_i^T\boldsymbol{A}\boldsymbol{z}_i\leq\sum_{i\neq 1,k}\lambda_i^2\|\boldsymbol{z}_i\|_2^2\|\boldsymbol{A}\|_2,$$

where $\boldsymbol{A}=(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}$. Then we need to upper bound $\|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\|_2$. Following Lemmas 8 and 13, we have

$$\begin{aligned}\|\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\|_2&=\|\frac{1}{D}\Big(\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}(1+\sigma f_k)-\boldsymbol{z}_k^T\boldsymbol{U}_\tau^{-1}(\sigma s)\Big)\|_2\\&\leq\frac{1}{D}\Big((1+\sigma f_k)\|\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\|_2+(\sigma s)\|\boldsymbol{z}_k^T\boldsymbol{U}_\tau^{-1}\|_2\Big)\\&\leq\frac{1}{D}\frac{C_1\sqrt{n}}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big(1+\frac{C_2n\sigma}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big).\end{aligned}$$

Note for a fixed number $B>0$, $\sum_{i\neq 1,k}\lambda_i^2\|\boldsymbol{z}_i\|_2^2B$ is the weighted sum of sub-exponential random variables. By [10, Lemma 7], with probability at least $1-2e^{-x}$,

$$\sum_{i=1}^p\lambda_i^2B\|\boldsymbol{z}_i\|_2^2\leq Cn\sum_{i=1}^p\lambda_i^2B,$$

for $x<n/c_0$. Combining (A.22) and bounds above with B replaced by the upper bound of $\boldsymbol{A}$ gives with probability at least $1-\delta$,

$$\frac{\sum_{i\neq 1,k}\lambda_i\hat{\eta}_i^2}{(\eta_k\hat{\eta}_k)^2}\leq\frac{C_1\sum_{i\neq 1,k}\lambda_i^2}{\Big(\eta_k^2(1-\frac{C_2n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-\sigma\Big)^2}\Big(1+\frac{C_3n\sigma}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big)^2.\tag{A.23}$$

Next we upper bound $\lambda_1\hat{\eta}_1^2$. (A.20) gives

$$\begin{aligned}\hat{\eta}_1 &= \sqrt{\lambda_1}\boldsymbol{z}_1^T(\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I})^{-1}\boldsymbol{y}\\ &= \frac{\sqrt{\lambda_1}}{D}(f_1+\sigma f_1 f_k-\sigma g_1 s)\\ &\le \frac{\sqrt{\lambda_1}}{D}\frac{C_6 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1)}\Big(1+\frac{C_3 n\sigma}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big).\end{aligned}$$

Combining the result above and (A.22) gives with probability at least $1-\delta$,

$$\frac{\lambda_1\hat{\eta}_1^2}{(\eta_k\hat{\eta}_k)^2} \le \frac{C_1\lambda_1^2}{\Big(\eta_k^2(1-\frac{C_2 n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-\sigma\Big)^2}\frac{(\tau+\|\boldsymbol{\lambda}\|_{-1})^2}{(\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1)^2}\Big(1+\frac{C_3 n\sigma}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big)^2. \quad (A.24)$$

In addition, $\frac{\lambda_k\hat{\eta}_k^2}{(\eta_k\hat{\eta}_k)^2}$ in (A.19) is $\frac{\lambda_k}{\eta_k^2}$. Then the sum of (A.23), (A.24) and $\frac{\lambda_k}{\eta_k^2}$ is

$$\frac{A+B+\lambda_k\Big(\eta_k(1-\frac{C_2 n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-\sqrt{\lambda_k}\Big)^2}{\Big(\eta_k^2(1-\frac{C_2 n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-\sigma\Big)^2} \le \frac{A+B+\lambda_k\Big(\eta_k^2+\lambda_k\Big)}{\Big(\eta_k^2(1-\frac{C_2 n\lambda_k}{\tau+\|\boldsymbol{\lambda}\|_{-1}})-\sigma\Big)^2},$$

where we use $(a-b)^2\le a^2+b^2$ for $a,b>0$ and with

$$\begin{aligned}A &= C_3\lambda_1^2\Big(\frac{\tau+\|\boldsymbol{\lambda}\|_{-1})}{\tau+n\lambda_1+\|\boldsymbol{\lambda}\|_{-1}}\Big)^2\Big(1+\frac{C_4 n\sigma}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)^2,\\ B &= C_5\Big(\sum_{i\neq 1,k}\lambda_i^2\Big)\Big(1+\frac{C_4 n\sigma}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)^2,\end{aligned}$$

for large constants C_i's. The inverse of the upper bound of $\frac{\lambda_k}{\eta_k^2}$ gives the lower bound of (A.18). To ensure $\hat{\boldsymbol{\eta}}_\tau^T\boldsymbol{\eta}>0$ with high probability, we need $\eta_k^2 > \frac{c_1 n\sigma^2}{\tau+\|\boldsymbol{\lambda}\|_{-1}}+c_2\sigma$ for $c_1, c_2>1$.

A.4.2 Benign overfitting for the bi-level ensemble

For the bi-level ensemble, the first condition in Theorem 1 is not satisfied, hence we can no longer analyze $\hat{\boldsymbol{\eta}}_{\mathrm{SVM}}$ by studying $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$. We show a regime that suffices to make the classification error of $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ vanish as p increase to $+\infty$. Consider the setting:

$$\lambda_2 = \cdots = \lambda_p = \lambda \;\text{ and }\; \lambda_1 = \alpha p\lambda, \;\text{ for } \alpha > 1. \tag{A.25}$$

For large enough p, the setting above can ensure the bi-level ensemble condition in (2.3) is satisfied.

Corollary 14.1. *Assume that the data generating process follows Assumption 1 and* (A.25). *Fix $\delta \in (0,1)$ and suppose n is finite but large enough such that $n > c\log(1/\delta)$ for some $c > 1$. Then for large enough $C > 1$, with probability at least $1-\delta$, $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})$, the expected 0-1 loss of the least squares estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$, approaches 0 as $p \to \infty$, provided that $\eta_k > C\sqrt{\lambda}p^r$, for $r > \frac{1}{2}$.*

Proof. First, the bound on the unregularized estimator $\hat{\boldsymbol{\eta}}_{\mathrm{LS}}$ can be obtained by setting $\tau = 0$ in (2.17). Thus, with probability at least $1-\delta$, $\mathcal{R}(\hat{\boldsymbol{\eta}}_{\mathrm{LS}})$ is upper bounded by

$$\exp\left(\frac{-\Big(\eta_k^2(1-\frac{C_1 n\lambda_k}{\|\boldsymbol{\lambda}\|_{-1}}) - C_2\sigma\Big)^2}{A+B+C_6(\lambda_k^2+\sigma^2)}\right), \tag{A.26}$$
$$\text{with } A = C_3\lambda_1^2\left(\frac{\|\boldsymbol{\lambda}\|_{-1}+C_4 n\sigma}{\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1}\right)^2, B = C_5\Big(\sum_{i\neq 1,k}\lambda_i^2\Big)\Big(1+\frac{C_4 n\sigma}{\|\boldsymbol{\lambda}\|_{-1}}\Big)^2.$$

Note from (A.25) that $\|\boldsymbol{\lambda}\|_{-1} = (p-1)\lambda$. We first look at the denominator of the exponent

of (A.26). Assuming $n\eta_k \le p\sqrt{\lambda}$, we have

$$\begin{aligned} A &= C_1\lambda_1^2\Big(\frac{\|\boldsymbol{\lambda}\|_{-1} + C_2 n\sigma)}{n\lambda_1 + \|\boldsymbol{\lambda}\|_{-1}}\Big)^2 \\ &= C_1\alpha^2 p^2\lambda^2\Big(\frac{(p-1)\lambda + C_2 n\sigma}{n\alpha p\lambda + (p-1)\lambda}\Big)^2 \\ &\le C_3\alpha^2 p^2\lambda^2\Big(\frac{p\lambda}{n\alpha p\lambda + (p-1)\lambda}\Big)^2 \\ &\le C_4\alpha^2 p^2\lambda^2\Big(\frac{p\lambda}{n\alpha p\lambda}\Big)^2, \quad \text{by } n\alpha p \gg 1 \\ &\le C_4\frac{p^2\lambda^2}{n^2}. \end{aligned}$$

Moreover,

$$\begin{aligned} B &= C_5\Big(\sum_{i\neq 1,k}\lambda_i^2\Big)\Big(1 + \frac{C_6 n\sigma}{\|\boldsymbol{\lambda}\|_{-1}}\Big)^2 \\ &= C_5(p-2)\lambda^2\Big(1 + \frac{C_6 n\sigma}{\|\boldsymbol{\lambda}\|_{-1}}\Big)^2 \\ &\le C_7(p-2)\Big(\lambda^2 + \frac{n^2\sigma^2}{(p-1)^2}\Big) \\ &\le C_8(p-2)\lambda^2, \end{aligned}$$

where the last inequality comes from $n\eta_k \le p\sqrt{\lambda}$. Combining the results above, we have the denominator of the exponent of (A.26) is upper bounded by

$$C_4\frac{p^2\lambda^2}{n^2} + C_8(p-2)\lambda^2 + C_9(\lambda^2 + \sigma^2) \le C_{10}(\frac{p^2}{n^2}\lambda^2 + p\lambda^2). \tag{A.27}$$

Now we look at the numerator.

$$\begin{aligned}\left(\eta_k^2(1-\frac{C_0 n\lambda_k}{\|\boldsymbol{\lambda}\|_{-1}})-C_2\sigma\right)^2 &\geq \eta_k^4(1-C\frac{n}{p})^2-C_9\sqrt{\lambda}\eta_k^3\\ &\geq \eta_k^4-2C\eta_k^4\frac{n}{p}-C_9\sqrt{\lambda}\eta_k^3. \end{aligned} \tag{A.28}$$

Let $\eta_k = C\sqrt{\lambda}p^r$, for some $C>1$. Combining (A.27) and (A.28), if $p>n^2$, in (A.27), $\frac{p^2}{n^2}>p$, then the negative exponent of (A.26) is lower bounded by

$$\frac{\eta_k^4 n^2}{\lambda^2 p^2}-C_{10}\frac{\eta_k^4 n^3}{\lambda^2 p^3}-C_{11}\frac{\eta_k^3 n^2\sqrt{\lambda}}{\lambda^2 p^2}. \tag{A.29}$$

Then

$$\begin{aligned}\text{(A.29)} &\geq n^2p^{4r-2}-C_{10}n^3p^{4r-3}-C_{11}n^2p^{3r-2}\\ &\geq n^2p^{4r-2}-C_{10}np^{4r-2}-C_{11}n^2p^{3r-2},\end{aligned}$$

where we use $\frac{p^2}{n^2}>p$ in the last inequality. Thus to make the bound above approach $+\infty$ as $p\to\infty$, it suffices to have $r>\frac{1}{2}$.

If $p\leq n^2$, then the negative exponent of (A.26) is lower bounded by

$$\frac{\eta_k^4}{\lambda^2 p}-C_{10}\frac{\eta_k^4 n}{\lambda^2 p^2}-C_{11}\frac{\eta_k^3\sqrt{\lambda}}{\lambda^2 p}\geq p^{4r-1}-C_{10}\frac{n}{p}p^{4r-1}-C_{11}p^{3r-1}.$$

It suffices to have $r>\frac{1}{4}$ to make the bound above approach $+\infty$ as $p\to\infty$. Combing previous results, it suffices to have $\eta_k = C\sqrt{\lambda}p^r$, for $r>\frac{1}{2}$ and some $C>1$. Recall that we assume $n\eta_k\leq p\sqrt{\lambda}$, and actually $n\eta_k > p\sqrt{\lambda}$ is stronger than the condition $\eta_k > C\sqrt{\lambda}p^r$, for $r>\frac{1}{2}$ and some $C>1$ for finite n, hence the later condition is sufficient to make the classification error approach 0 as $p\to\infty$. □

A.5 Proof of Corollaries 5.1 and 5.2

A.5.1 Proof of Corollary 5.1

We need to find the conditions that make the negative exponent of (2.13) vanish as p increases when conditions in Theorem 1 hold. Recall Theorem 1 requires

$$\|\boldsymbol{\lambda}\|_1 > \max\{\lambda_*, C_1 n\sqrt{\log(2n)}\sigma\}, \tag{A.30}$$

for some $C_1, C_2 > 1$ and $\lambda_* = 72\Big(\|\boldsymbol{\lambda}\|_2 \cdot n\sqrt{\log n} + \|\boldsymbol{\lambda}\|_\infty \cdot n\sqrt{n}\log n + 1\Big)$. It is not hard to check that (A.30) implies that $\frac{n^2\sigma^2}{\|\boldsymbol{\lambda}\|_1^2} < 1$. Then the negative exponent of (2.13) is lower bounded by

$$\begin{aligned}
\frac{\Big(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\|\boldsymbol{\lambda}\|_1} - C_2\sigma\Big)^2}{C_3\max\{1, \frac{n^2\sigma^2}{\|\boldsymbol{\lambda}\|_1^2}\}\sum_{i=1}^p \lambda_i^2 + C_4\sigma^2} \geq & \frac{\Big(\|\boldsymbol{\eta}\|_2^2 - \frac{C_1 n\sigma^2}{\|\boldsymbol{\lambda}\|_1} - C_2\sigma\Big)^2}{C_5\Big(\sum_{i=1}^p \lambda_i^2 + \sigma^2\Big)} \\
\geq & \frac{\Big(\|\boldsymbol{\eta}\|_2^2 - C_1\frac{\|\boldsymbol{\lambda}\|_1}{n\sqrt{\log(2n)}}\frac{n\sigma}{\|\boldsymbol{\lambda}\|_1} - C_2\sigma\Big)^2}{C_5\Big(\sum_{i=1}^p \lambda_i^2 + \sigma^2\Big)} \\
\geq & \frac{\Big(\|\boldsymbol{\eta}\|_2^2 - C_1\frac{\sigma}{\sqrt{\log(2n)}} - C_2\sigma\Big)^2}{C_5\Big(\sum_{i=1}^p \lambda_i^2 + \sigma^2\Big)} \\
\geq & \frac{\Big(\|\boldsymbol{\eta}\|_2^2 - C_6\sigma\Big)^2}{C_5\Big(\sum_{i=1}^p \lambda_i^2 + \sigma^2\Big)} \\
\geq & \frac{\|\boldsymbol{\eta}\|_2^4 - C_7\|\boldsymbol{\eta}\|_2^2\sigma}{C_5\Big(\sum_{i=1}^p \lambda_i^2 + \sigma^2\Big)}.
\end{aligned} \tag{A.31}$$

Note that $\boldsymbol{\beta} = \begin{bmatrix} \beta & \beta & ... & \beta \end{bmatrix}^T$, hence $\sigma = \beta\sqrt{\|\boldsymbol{\lambda}\|_1}$. Looking at the denominator of (A.31), when $\sum_{i=1}^p \lambda_i^2 \leq \sigma^2$, i.e. $\|\boldsymbol{\lambda}\|_2^2 \leq \beta^2\|\boldsymbol{\lambda}\|_1$,

$$\begin{aligned}
\text{(A.31)} &\geq \frac{\|\boldsymbol{\eta}\|_2^4 - C_7\|\boldsymbol{\eta}\|_2^2\sigma}{C_8\sigma^2} \\
&\geq \frac{(p\beta^2)^2 - C_7(p\beta^2)\sqrt{\beta^2\|\boldsymbol{\lambda}\|_1}}{C_8\Big(\beta^2\|\boldsymbol{\lambda}\|_1\Big)} \\
&\geq \Big(\frac{p\beta}{\sqrt{\|\boldsymbol{\lambda}\|_1}}\Big)^2 - \frac{C_9 p\beta}{\sqrt{\|\boldsymbol{\lambda}\|_1}}. \qquad \text{(A.32)}
\end{aligned}$$

To guarantee (A.32) $\to \infty$ as $p \to \infty$, it suffices to have $\|\boldsymbol{\lambda}\|_1 \leq C\beta^2 p^\alpha$, for $\alpha < 2$. Note that the second condition in Theorem 1 becomes

$$\|\boldsymbol{\lambda}\|_1 > Cn\sqrt{\log(2n)}\beta\sqrt{\|\boldsymbol{\lambda}\|_1} \iff \|\boldsymbol{\lambda}\|_1 > C^2 n^2 \log(2n)\beta^2.$$

Combing the conditions above, the SVM solution goes to 0 with $p \to \infty$ provided the assumptions of Theorem 3 with $\tau = 0$ hold and

$$\max\{\lambda_*, C_1\beta^2 n^2 \log(2n)\} < \|\boldsymbol{\lambda}\|_1 \leq C_2\beta^2 p^\alpha, \quad \text{for } \alpha < 2. \qquad \text{(A.33)}$$

When $\sum_{i=1}^p \lambda_i^2 > \sigma^2$, i.e. $\|\boldsymbol{\lambda}\|_2^2 > \beta^2\|\boldsymbol{\lambda}\|_1$,

$$\begin{aligned}
\text{(A.31)} &\geq \frac{\|\boldsymbol{\eta}\|_2^4 - C_7\|\boldsymbol{\eta}\|_2^2\sigma}{C_8\Big(\sum_{i=1}^p \lambda_i^2\Big)} \\
&\geq \Big(\frac{p\beta^2}{\sqrt{\sum_{i=1}^p \lambda_i^2}}\Big)^2 - \frac{C_9 p\beta^2}{\sqrt{\sum_{i=1}^p \lambda_i^2}}. \qquad \text{(A.34)}
\end{aligned}$$

To guarantee (A.34) $\to \infty$ as $p \to \infty$, it suffices to have $\sum_{i=1}^p \lambda_i^2 \leq C\beta^4 p^\alpha$, for $\alpha < 2$, which is equivalent to $\|\boldsymbol{\lambda}\|_2 \leq C\beta^2 p^\alpha$, for $\alpha < 1$. Combing the conditions in Theorem 1, the SVM solution goes to 0 with $p \to \infty$ provided the assumptions of Theorem 3 with

$\tau = 0$ hold and

$$\|\boldsymbol{\lambda}\|_1 > \max\{\lambda_*, C_1\beta^2 n^2 \log(2n)\} \quad \text{and} \quad \|\boldsymbol{\lambda}\|_2 \leq C\beta^2 p^\alpha, \quad \text{for} \quad \alpha < 1. \tag{A.35}$$

Combining (A.33) and (A.35) completes the proof.

A.5.2 Proof of Corollary 5.2

We start from the high-SNR regime. In fact, we can assume a bit stronger that

$$\|\boldsymbol{\eta}\|_2^2 > C\frac{p}{n}, \quad \text{for some large} \quad C > 1. \tag{A.36}$$

Then the exponent in (2.15) becomes:

$$\begin{aligned}\|\boldsymbol{\eta}\|_2^2(1-\frac{n}{p})^2 + C_1 - 2C_1(1-\frac{n}{p})\|\boldsymbol{\eta}\|_2 &> \|\boldsymbol{\eta}\|_2^2 - 2\|\boldsymbol{\eta}\|_2^2\frac{n}{p} - 2C_1\|\boldsymbol{\eta}\|_2 \\ &> C\frac{p}{n} - 2\|\boldsymbol{\eta}\|_2^2\frac{n}{p} - 2C_1\|\boldsymbol{\eta}\|_2,\end{aligned} \tag{A.37}$$

where the last inequality comes from (A.36). Following Theorem 2, we further assume that

$$p > 10n\log n + n - 1 \quad \text{and} \quad p > C_2 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, \tag{A.38}$$

for some constant $C_2 > 1$. Then combining the relationships above gives

$$\begin{aligned}(\text{A.37}) &> C\frac{p}{n} - 2\Big(\frac{p}{C_2 n\sqrt{\log(2n)}}\Big)^2\frac{n}{p} - 2C_1\frac{p}{n\sqrt{\log(2n)}} \\ &= C\frac{p}{n} - \frac{2p}{C_2 n\log(2n)} - \frac{2C_1 p}{n\sqrt{\log(2n)}} \\ &= \frac{p}{n}\Big(C - \frac{2C_1}{\sqrt{\log(2n)}} - \frac{2}{C_2\log(2n)}\Big). \end{aligned} \tag{A.39}$$

Notice that (A.39) $\to \infty$ as $(p/n) \to \infty$ for sufficiently large C and n. Thus we have proved that in the high-SNR regime, the error of SVM solution goes to 0 with $(p/n) \to \infty$ provided that the assumptions of Theorem 4 hold and

$$\frac{1}{C}n\|\boldsymbol{\eta}\|_2^2 > p > \max\{10n\log n + n - 1, C_1 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2\},$$

for sufficiently large constants $C, C_1 > 1$.

For the low-SNR regime, assume that

$$p > 10n\log n + n - 1,\ p > C_1 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, \tag{A.40}$$

$$\text{and } \|\boldsymbol{\eta}\|_2^2 \le \frac{p}{n},\ \|\boldsymbol{\eta}\|_2^4 = C_2(\frac{p}{n})^\alpha,\ \text{ for } \alpha > 1. \tag{A.41}$$

Then the exponent in (2.16) becomes:

$$\begin{aligned}\frac{n}{p}\|\boldsymbol{\eta}\|_2^4\Big((1-\frac{n}{p}) - C_3\frac{1}{\|\boldsymbol{\eta}\|_2}\Big)^2 &> \frac{n}{p}\|\boldsymbol{\eta}\|_2^4 - 2\frac{n^2}{p^2}\|\boldsymbol{\eta}\|_2^4 - 2\frac{n}{p}C_3\|\boldsymbol{\eta}\|_2^3 \\ &\ge C_2\Big(\frac{p}{n}\Big)^{\alpha-1} - 2C_2\Big(\frac{p}{n}\Big)^{\alpha-2} - 2C_3C_2\Big(\frac{p}{n}\Big)^{0.75\alpha-1}, \end{aligned} \tag{A.42}$$

where the last inequality comes from (A.40) and (A.41). (A.42) will go to $+\infty$ as $(p/n) \to \infty$ provided that $\alpha > 1$. Overall in the low-SNR regime, we need the assumptions of

Theorem 3 plus

$$p > \max\{10n\log n + n - 1, C_1 n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\},$$
$$\text{and } \|\boldsymbol{\eta}\|_2^4 \geq C_2(\frac{p}{n})^{\alpha}, \quad \text{for } \alpha \in (1,2].$$

A.6 Results for the averaging estimator

The theorem below shows an upper bound on the classification error for the averaging estimator $\hat{\boldsymbol{\eta}}_{\text{Avg}}$. Note the result below is for general $\boldsymbol{\Sigma}$, i.e. no balanced or bi-level structure is required.

Theorem 15. *Assume that the data are generated with the GMM model. Fix $\delta \in (0,1)$ and suppose n is large enough such that $n > c\log(1/\delta)$ for some $c > 1$. Then, there exist a constant $c_1 > 1$ such that with probability at least $1-\delta$, $\hat{\boldsymbol{\eta}}_{Avg}^T\boldsymbol{\eta} > 0$ provided that $\|\boldsymbol{\eta}\|_2^2 > c_1\sigma$. Then, there exists constants C_i's > 1 such that with probability at least $1-\delta$,*

$$\mathcal{R}(\hat{\boldsymbol{\eta}}_{Avg}) \leq \exp\left(\frac{-\Big(\|\boldsymbol{\eta}\|_2^2 - C_1\sigma\Big)^2}{C_2\|\boldsymbol{\lambda}\|_2^2 + C_3\sigma^2}\right). \tag{A.43}$$

The bound above is the same as (2.19).

Proof. We need to lower bound $\frac{(\hat{\boldsymbol{\eta}}_{\text{Avg}}^T\boldsymbol{\eta})^2}{\hat{\boldsymbol{\eta}}_{\text{Avg}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}_{\text{Avg}}}$. Recall that $\hat{\boldsymbol{\eta}}_{\text{Avg}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{y}$. For the denominator,

$$\hat{\boldsymbol{\eta}}_{\text{Avg}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\eta}}_{\text{Avg}} \leq \frac{2}{n^2}\Big(n^2\boldsymbol{\eta}^T\boldsymbol{\Sigma}\boldsymbol{\eta} + \boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T\boldsymbol{y}\Big),$$

where we use the fact $\boldsymbol{v}^T\boldsymbol{u} \leq \frac{1}{2}(\boldsymbol{v}^T\boldsymbol{v} + \boldsymbol{u}^T\boldsymbol{u})$. Then we need to upper bound $\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T\boldsymbol{y}$.

Following what we show in the proof of Theorem 3, with probability at least $1-\delta$,

$$\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T\boldsymbol{y} = \mathrm{Tr}\Big(\sum_{i=1}^{p}\lambda_i^2\boldsymbol{z}_i^T(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{z}_i\Big) \le \sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{y}\boldsymbol{y}^T\|_2\|\boldsymbol{z}_i\|_2^2 \le Cn\sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{z}_i\|_2^2 \le Cn^2\|\boldsymbol{\lambda}\|_2^2,$$

where the last inequality follows the fact that $\sum_{i=1}^{p}\lambda_i^2\|\boldsymbol{z}_i\|_2^2$ is the weighted sum of sub-exponential variables. Next we lower bound the numerator $\hat{\boldsymbol{\eta}}_{\mathrm{Avg}}^T\boldsymbol{\eta}$, Lemma 9 gives with probability at least $1-\delta$,

$$\hat{\boldsymbol{\eta}}_{\mathrm{Avg}}^T\boldsymbol{\eta} = \frac{1}{n}\boldsymbol{y}^T\boldsymbol{y}\|\boldsymbol{\eta}\|_2^2 + \frac{1}{n}\boldsymbol{y}^T\boldsymbol{d} \ge \|\boldsymbol{\eta}\|_2^2 - C\sigma.$$

We need $\|\boldsymbol{\eta}\|_2^2 - C\sigma > 0$ to guarantee $\hat{\boldsymbol{\eta}}_{\mathrm{Avg}}^T\boldsymbol{\eta} > 0$ with high probability. Combining results above completes the proof. □

A.7 Proof of Lemmas

A.7.1 Proof of Lemmas 2

For Lemma 2, the proof of Theorem 3 gives

$$\hat{\boldsymbol{\eta}}_{\tau}^T\boldsymbol{\eta} = \frac{s(\|\boldsymbol{\eta}\|_2^2 - t) + h^2 + h}{D},$$

for $D > 0$. Then we proceed by directly applying (A.11).

A.7.2 Proof of Lemma 8

Recall

$$\boldsymbol{X}\boldsymbol{X}^T+\tau\boldsymbol{I}=\boldsymbol{Q}\boldsymbol{Q}^T+\tau\boldsymbol{I}+\|\boldsymbol{\eta}\|_2^2\boldsymbol{y}\boldsymbol{y}^T+\boldsymbol{Q}\boldsymbol{\eta}\boldsymbol{y}^T+\left(\boldsymbol{Q}\boldsymbol{\eta}\boldsymbol{y}^T\right)^T=\boldsymbol{U}_\tau+\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y} & \boldsymbol{d} & \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}.$$

Thus, by Woodbury identity [70], $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ can be expressed as:

$$\boldsymbol{U}_\tau^{-1}-\boldsymbol{U}_\tau^{-1}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y} & \boldsymbol{d} & \boldsymbol{y}\end{bmatrix}\left[\boldsymbol{I}+\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_\tau^{-1}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y} & \boldsymbol{d} & \boldsymbol{y}\end{bmatrix}\right]^{-1}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_\tau^{-1}. \tag{A.44}$$

We first compute the inverse of the 3×3 matrix $\boldsymbol{A}:=\left[\boldsymbol{I}+\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_\tau^{-1}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y} & \boldsymbol{d} & \boldsymbol{y}\end{bmatrix}\right]$.

By our definitions of s, h and t in Section A.1:

$$\boldsymbol{A}=\begin{bmatrix}1+\|\boldsymbol{\eta}\|_2^2 s & \|\boldsymbol{\eta}\|_2 h & \|\boldsymbol{\eta}\|_2 s\\ \|\boldsymbol{\eta}\|_2 s & 1+h & s\\ \|\boldsymbol{\eta}\|_2 h & t & 1+h\end{bmatrix}.$$

Recalling $\boldsymbol{A}^{-1}=\frac{1}{\det(\boldsymbol{A})}\mathrm{adj}(\boldsymbol{A})$, where $\det(\boldsymbol{A})$ is the determinant of $\boldsymbol{A}$ and $\mathrm{adj}(\boldsymbol{A})$ is the adjoint of $\boldsymbol{A}$, it can be checked that:

$$\det(\boldsymbol{A})=D=s(\|\boldsymbol{\eta}\|_2^2-t)+(h+1)^2,$$

and

$$\mathrm{adj}(\boldsymbol{A}) = \begin{bmatrix} (h+1)^2 - st & \|\boldsymbol{\eta}\|_2(st - h - h^2) & -\|\boldsymbol{\eta}\|_2 s \\ -\|\boldsymbol{\eta}\|_2 s & h + 1 + \|\boldsymbol{\eta}\|_2^2 s & -s \\ \|\boldsymbol{\eta}\|_2(st - h - h^2) & \|\boldsymbol{\eta}\|_2^2 h^2 - t(1 + \|\boldsymbol{\eta}\|_2^2 s) & h + 1 + \|\boldsymbol{\eta}\|_2^2 s \end{bmatrix}.$$

Combining the above gives

$$\begin{aligned} \boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{X}^T + \tau\boldsymbol{I})^{-1} &= \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - \begin{bmatrix} \|\boldsymbol{\eta}\|_2 s & h & s \end{bmatrix} \boldsymbol{A}^{-1} \begin{bmatrix} \|\boldsymbol{\eta}\|_2 \boldsymbol{y}^T \\ \boldsymbol{y}^T \\ \boldsymbol{d}^T \end{bmatrix} \boldsymbol{U}_\tau^{-1} \\ &= \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1} - \frac{1}{D} \begin{bmatrix} \|\boldsymbol{\eta}\|_2 s & h^2 + h - st & s \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\eta}\|_2 \boldsymbol{y}^T \\ \boldsymbol{y}^T \\ \boldsymbol{d}^T \end{bmatrix} \boldsymbol{U}_\tau^{-1}. \end{aligned}$$

This completes the proof of the lemma.

A.7.3 Proof of Lemma 9 and Lemma 10

To prove Lemma 9, we need to bound the eigenvalues of $\boldsymbol{U}_\tau$. Recall $\boldsymbol{U}_0 = \boldsymbol{Q}\boldsymbol{Q}^T = \sum_{i=1}^p \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T$, where $\boldsymbol{z}_i \in \mathbb{R}^n$ are independent vectors with IID standard normal elements. Let $\lambda_k(\boldsymbol{M})$ represent the k-th eigenvalue of matrix $\boldsymbol{M}$. We start from Bartlett et al. [10, Lemma 5 (3)]:

Lemma 14. *There are constants $b, c \geq 1$ such that, for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$, if $\frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}} \geq bn$, then*

$$\frac{1}{c}\sum_{i>k}\lambda_i \leq \lambda_n\Big(\sum_{i>k}^p \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T\Big) \leq \lambda_1\Big(\sum_{i>k}^p \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T\Big) \leq c\sum_{i>k}\lambda_i.$$

First note that the balanced ensemble requirement $bn\lambda_1 \leq \|\boldsymbol{\lambda}\|_{-1}$ implies $bn\lambda_1 \leq \|\boldsymbol{\lambda}\|_1$. We can then obtain the bounds for eigenvalues of $\boldsymbol{U}_0$ by letting $k = 0$ in Lemma 14. Then the eigenvalues of $\boldsymbol{U}_\tau$ are bounded as follows.

Lemma 15. *Assume the balanced $\boldsymbol{\Sigma}$ assumption is satisfied. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$ for some $c > 1$. There is a constant $C > 1$ such that with probability at least $1 - \delta$, the largest and smallest eigenvalues of $\boldsymbol{U}_\tau$ satisfy:*

$$\frac{1}{C}(\tau+\sum_{i=1}^{p}\lambda_i) \leq \tau+\frac{1}{C}\sum_{i=1}^{p}\lambda_i \leq \lambda_n(\boldsymbol{U}_\tau) \leq \lambda_1(\boldsymbol{U}_\tau) \leq \tau+C\sum_{i=1}^{p}\lambda_i \leq C(\tau+\sum_{i=1}^{p}\lambda_i). \quad \text{(A.45)}$$

Now we are ready to prove Lemma 9.

Bounds for s

For $s = \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{y}$, from (A.45) and $\|\boldsymbol{y}\|_2^2 = n$, the variational characterization of eigenvalues gives:

$$s = \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{y} \leq \|\boldsymbol{y}\|_2^2\lambda_1(\boldsymbol{U}_\tau^{-1}) \leq n\frac{1}{\lambda_n(\boldsymbol{U}_\tau)} \leq C_1\frac{n}{\tau + \|\boldsymbol{\lambda}\|_1}.$$

The lower bound can be derived in a similar way and is omitted for brevity.

Bounds for t and h

We begin by presenting the definitions of sub-Gaussian and sub-exponential norms. For a detailed discussion of sub-Gaussian and sub-exponential variables, we refer the readers to [162, Chapter 2].

Definition A.7.1. For a sub-Gaussian variable X defined in [162, 2.5], the sub-Gaussian

norm of X, denoted by $\|X\|_{\psi_2}$, is defined as

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[e^{X^2/t^2}] < 2\}.$$

Then [162, Example 2.5.8 (a)] states that if $X \sim \mathcal{N}(0, \sigma^2)$, then X is sub-Gaussian with $\|X\|_{\psi_2} < C\sigma$, where C is an absolute constant.

Definition A.7.2. For a sub-exponential variable X defined in [162, 2.7], the sub-exponential norm of X, denoted by $\|X\|_{\psi_1}$, is defined as

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[e^{|X|/t}] < 2\}.$$

[162, Lemma 2.7.6] shows that sub-exponential is sub-Gaussian squared.

Lemma 16. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

We now look at $\|\boldsymbol{d}\|_2$. Recall $\boldsymbol{d} = \boldsymbol{Q\eta} = \boldsymbol{Z\Lambda}^{\frac{1}{2}}\boldsymbol{\beta}$. $\|\boldsymbol{d}\|_2^2 = \sum_{j=1}^n {d_j}^2$, where $d_j = \sum_{i=1}^p \sqrt{\lambda_i}\beta_i z_{ji}$ and z_{ji}'s are IID standard Gaussian variable. Hence d_j is Gaussian with mean zero and variance $\sum_{i=1}^p \lambda_i\beta_i^2$ and ${d_j}^2$ is sub-exponential with $\|d_j^2\|_{\psi_1} < c\sum_{i=1}^p \lambda_i\beta_i^2$ and mean $\sum_{i=1}^p \lambda_i\beta_i^2$. To bound $\|\boldsymbol{d}\|_2$, we need the Bernstein's inequality [162, Theorem 2.8.2]:

Lemma 17. *Let $\xi_1, ..., \xi_n$ be independent, mean zero, sub-exponential random variables with sub-exponential norm $\|\xi\|_{\psi_1}$, and $a = (a_1, ..., a_n) \in \mathbb{R}^n$. Then for every $t \geq 0$, we*

have

$$\mathbb{P}\Big(|\sum_{i=1}^{n} a_i\xi_i| \geq t\Big) \leq 2\exp\Big\{-c\min\Big(\frac{t^2}{\|\xi\|_{\psi_1}^2 \cdot \sum_{i=1}^n a_i^2}, \frac{t}{\|\xi\|_{\psi_1} \cdot \max_{i\in[n]}|a_i|}\Big)\Big\}.$$

Corollary 15.1. *Suppose $\{a_i\}$ is a non-increasing sequence of non-negative numbers such that $\sum_i a_i < \infty$. Then there is a constant c such that for any sequence of independent, zero-mean sub-exponential random variables $\{\xi_i\}$ with sub-exponential norm $\|\xi\|_{\psi_1}$, and any $x > 0$, with probability at least $1 - 2e^{-x}$,*

$$|\sum_{i=1} a_i\xi_i| \leq c\|\xi\|_{\psi_1} \cdot \max\Big(a_1 x, \sqrt{x\sum_i a_i^2}\Big).$$

Let fix the length of the sequence as n and let $a_i = 1$, for $i \in [n]$. Then combing the inequality above with $x = n/c$ and the fact that ${d_j}^2$'s are sub-exponential gives with probability at least $1 - 2e^{-\frac{n}{c}}$,

$$\|\boldsymbol{d}\|_2 \leq C\sqrt{n\sum_{i=1}^{p}\lambda_i\beta_i^2} = C\sqrt{n}\sigma. \tag{A.46}$$

Recall $t = \boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{d}$ and $h = \boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\boldsymbol{d}$, we can obtain the upper and lower bounds of t by the variational characterization of eigenvalues. The bounds of h can be derived from the fact $-\|\boldsymbol{d}\|_2\|\boldsymbol{y}\|_2\|\boldsymbol{U}_\tau^{-1}\|_2 \leq h \leq \|\boldsymbol{d}\|_2\|\boldsymbol{y}\|_2\|\boldsymbol{U}_\tau^{-1}\|_2$. The bounds for $\|\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\|_2$ and $\|\boldsymbol{d}^T\boldsymbol{U}_\tau^{-1}\|_2$ can be obtained from Cauchy-Schwarz for matrices.

Proof of Lemma 10

Now we prove Lemma 10. Recall

$$f_i = \boldsymbol{e}_i^T\boldsymbol{U}_0^{-1}\boldsymbol{d} = {\boldsymbol{e}_i}^T(\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} - \boldsymbol{E}')\boldsymbol{d},$$

thus,

$$\begin{aligned}\max_{i\in[n]}|f_i| &\leq \frac{1}{\|\boldsymbol{\lambda}\|_1}\|\boldsymbol{d}\|_\infty + \|\boldsymbol{e}_i^T\boldsymbol{E}'\boldsymbol{d}\|_\infty \\ &\leq \frac{1}{\|\boldsymbol{\lambda}\|_1}\|\boldsymbol{d}\|_\infty + \|\boldsymbol{e}_i^T\boldsymbol{E}'\boldsymbol{d}\|_2,\end{aligned}$$

where the last equality comes from the fact that the ℓ_2 norm of a vector won't be smaller than its infinity norm. By Markov's inequality [163, 2.1.1], for sufficiently large constant $C > 1$,

$$\mathbb{P}(\max_{i\in[n]}|d_i| \geq C\mathbb{E}[\max_{i\in[n]}|d_i|]) \leq \delta.$$

Thus it suffices to bound $\mathbb{E}[\max_{i\in[n]}|d_i|]$. We know that the elements of $\boldsymbol{d}$ are IID zero-mean Gaussian variables with variance $\sum_{i=1}^p \lambda_i\beta_i^2$. By [163, Exercise 2.11],

$$\mathbb{E}[\max_{i\in[n]}|d_i|] \leq \sqrt{\sum_{i=1}^p \lambda_i\beta_i^2}\sqrt{2\log(2n)} = \sqrt{2\log(2n)}\sigma.$$

Thus with probability at least $1-\delta$,

$$\|\boldsymbol{d}\|_\infty \leq C\sqrt{2\log(2n)}\sigma.$$

To bound $\|\boldsymbol{e}_i^T\boldsymbol{E}'\boldsymbol{d}\|_2$, using $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \leq \|\boldsymbol{v}\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{M}\|_2$ and the bound on $\|\boldsymbol{d}\|_2$ in (A.46) and the bound on $\|\boldsymbol{E}'\|_2$ in (A.3) give, for $n > c/\delta$ and for every $i \in [n]$,

$$\begin{aligned}\|\boldsymbol{e}_i^T\boldsymbol{E}'\boldsymbol{d}\|_2 &\leq \|\boldsymbol{e}_i\|_2\|\boldsymbol{d}\|_2\|\boldsymbol{E}'\|_2 \\ &\leq \frac{C_1\sigma}{\|\boldsymbol{\lambda}\|_1}.\end{aligned}$$

Combining results above completes the proof.

A.7.4 Proof of Lemma 13

To prove Lemma 13, the first step is to separate the largest eigenvalue from others. Specifically, by Woodbury identity, $\boldsymbol{U}_\tau^{-1}$ can be expressed as

$$\boldsymbol{U}_\tau^{-1} = (\tau \boldsymbol{I} + \sum_{i=2}^{p} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T + \lambda_1 \boldsymbol{z}_1 \boldsymbol{z}_1^T)^{-1} \tag{A.47}$$

$$= \boldsymbol{U}_{-1,\tau}^{-1} - \frac{\lambda_1 \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1}}{1 + \lambda_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1}, \tag{A.48}$$

where $\boldsymbol{U}_{-1,\tau} = \tau \boldsymbol{I} + \sum_{i=2}^{p} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^T$. By Lemma 15 above, with probability at least $1 - \delta$,

$$\frac{1}{C}(\tau + \sum_{i=2}^{p} \lambda_i) \leq \lambda_n(\boldsymbol{U}_{-1,\tau}) \leq \lambda_1(\boldsymbol{U}_{-1,\tau}) \leq C(\tau + \sum_{i=2}^{p} \lambda_i). \tag{A.49}$$

Then we need to bound $\|\boldsymbol{z}_1\|_2$ and $\|\boldsymbol{z}_k\|_2$. In Lemma 15.1, let $x < \frac{n}{c_0}$ with sufficiently large c_0, if $n > C_0 \log(1/\delta)$ for some $C_0 > 1$, then there exist $C_1, C_2 > 1$ such that with probability at least $1 - \delta$,

$$\frac{1}{C_1} n \leq \|\boldsymbol{z}_i\|_2^2 \leq C_2 n, \quad i \in [p].$$

Now we are ready to derive the bounds in Lemma 13.

For $s = \boldsymbol{y}^T \boldsymbol{U}_\tau^{-1} \boldsymbol{y}$, by (A.48) and (A.49) and using the variational characterization of

eigenvalues and $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{u} \leq \|\boldsymbol{v}\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{M}\|_2$, with probability at least $1-\delta$,

$$\begin{aligned}
s &= \frac{\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y} + \lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y} - \lambda_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y}}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1} \\
&\leq \frac{\frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big(1+\frac{C_2 n\lambda_1}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)}{1+\frac{n\lambda_1}{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})}} \\
&\leq \frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\cdot\Big(\frac{\tau+\|\boldsymbol{\lambda}\|_{-1}+C_2 n\lambda_1}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)\cdot\Big(\frac{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})}{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})+n\lambda_1}\Big) \\
&\leq \frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\cdot\Big(\frac{C_2C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})+C_2 n\lambda_1}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)\cdot\Big(\frac{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})}{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})+n\lambda_1}\Big) \\
&\leq \frac{C_4 n}{\tau+\|\boldsymbol{\lambda}\|_{-1}}.
\end{aligned}$$

For the lower bound of s, we need to show $\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y}$ is sufficiently small compared with $\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1$ and $\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y}$. We thus need the following Hanson-Wright inequality [140].

Lemma 18. *Let $\boldsymbol{z}$ be a random vector whose elements are IID zero-mean sub-Gaussian random variable with parameter at most 1. Then , there exists universal constant $c>0$ such that for any positive semi-definite matrix $\boldsymbol{M}$ and for every $t>0$, we have*

$$P\Big(|\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}]|>t\Big)\leq \exp\Big\{-c\min\Big\{\frac{t^2}{\|\boldsymbol{M}\|_F^2},\frac{t}{\|\boldsymbol{M}\|_2}\Big\}\Big\}.$$

Note $\|\boldsymbol{M}\|_F^2 \leq n\|\boldsymbol{M}\|_2^2$ and let $t=\frac{1}{C_0}n\|\boldsymbol{M}\|_2$ for sufficiently large constant C_0 to get with probability at least $1-2e^{-\frac{n}{c_1}}$,

$$|\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}^T\boldsymbol{M}\boldsymbol{z}]|\leq\frac{1}{C_0}n\|\boldsymbol{M}\|_2. \tag{A.50}$$

Then we use the similar trick as [119, D.3.1] and apply the parallelogram law to $\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y}$,

$$\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{y}=\frac{1}{4}((\boldsymbol{z}_1+\boldsymbol{y})^T\boldsymbol{U}_{-1,\tau}^{-1}(\boldsymbol{z}_1+\boldsymbol{y})-(\boldsymbol{z}_1-\boldsymbol{y})^T\boldsymbol{U}_{-1,\tau}^{-1}(\boldsymbol{z}_1-\boldsymbol{y})).$$

To use the Hanson-Wright inequality, we need to calculate the conditional expectation

$$\mathbb{E}[\boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{y} | \boldsymbol{U}_{-1,\tau}^{-1}] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{y}\boldsymbol{z}_1^T) | \boldsymbol{U}_{-1,\tau}^{-1}] = \mathrm{Tr}(\boldsymbol{U}_{-1,\tau}^{-1} \mathbb{E}[\boldsymbol{y}\boldsymbol{z}_1^T]),$$

where we use the fact that $\boldsymbol{y}$ and $\boldsymbol{z}_1$ are independent of $\boldsymbol{U}_{-1,\tau}^{-1}$. It is not hard to check that $\mathbb{E}[\boldsymbol{y}\boldsymbol{z}_1^T] = \boldsymbol{0}$, where $\boldsymbol{0}$ is the matrix with all elements 0. Now applying Lemma 18 to both $(\boldsymbol{z}_1 + \boldsymbol{y})^T \boldsymbol{U}_{-1,\tau}^{-1} (\boldsymbol{z}_1 + \boldsymbol{y})$ and $(\boldsymbol{z}_1 - \boldsymbol{y})^T \boldsymbol{U}_{-1,\tau}^{-1} (\boldsymbol{z}_1 - \boldsymbol{y})$ gives with probability at least $1 - 2e^{-\frac{n}{c_1}}$,

$$|\boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{y}| \leq \frac{2}{C_0} n \|\boldsymbol{U}_{-1,\tau}^{-1}\|_2.$$

Now for the numerator of s, using the bound of eigenvalues of $\boldsymbol{U}_{-1,\tau}$ in (A.49) and the fact that C_0 is sufficiently large gives with probability at least $1 - \delta$,

$$\lambda_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1 \boldsymbol{y}^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{y} - \lambda_1 \boldsymbol{y}^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{z}_1 \boldsymbol{z}_1^T \boldsymbol{U}_{-1,\tau}^{-1} \boldsymbol{y} \geq \frac{n^2 \lambda_1}{C_2 (\tau + \|\boldsymbol{\lambda}\|_{-1})^2},$$

for some large C_2. Therefore,

$$\begin{aligned}
s &\geq \frac{\frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_{-1})} \Big(1 + \frac{n\lambda_1}{C_2(\tau + \|\boldsymbol{\lambda}\|_{-1})}\Big)}{1 + \frac{C_3 n \lambda_1}{(\tau + \|\boldsymbol{\lambda}\|_{-1})}} \\
&\geq \frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_{-1})} \cdot \Big(\frac{C_2(\tau + \|\boldsymbol{\lambda}\|_{-1}) + n\lambda_1}{C_2(\tau + \|\boldsymbol{\lambda}\|_{-1})}\Big) \cdot \Big(\frac{(\tau + \|\boldsymbol{\lambda}\|_{-1})}{(\tau + \|\boldsymbol{\lambda}\|_{-1}) + C_3 n \lambda_1}\Big) \\
&\geq \frac{n}{C_1(\tau + \|\boldsymbol{\lambda}\|_{-1})} \cdot \Big(\frac{C_2(\tau + \|\boldsymbol{\lambda}\|_{-1}) + n\lambda_1}{C_2(\tau + \|\boldsymbol{\lambda}\|_{-1})}\Big) \cdot \Big(\frac{(\tau + \|\boldsymbol{\lambda}\|_{-1})}{(C_2 C_3(\tau + \|\boldsymbol{\lambda}\|_{-1}) + C_3 n \lambda_1}\Big) \\
&\geq \frac{n}{C_4(\tau + \|\boldsymbol{\lambda}\|_{-1})}.
\end{aligned}$$

The derivation of bounds for t_k is the same as the procedure above.

For f_k, with probability at least $1-\delta$,

$$\begin{aligned}
|f_k| &= |\frac{\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_k + \lambda_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_k\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1 - \lambda_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_k}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}| \\
&\leq \frac{\frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\Big(1+\frac{C_2 n\lambda_1}{\tau+\|\boldsymbol{\lambda}\|_{-1}}\Big)}{1+\frac{n\lambda_1}{C_3(\tau+\|\boldsymbol{\lambda}\|_{-1})}} \\
&\leq \frac{C_4 n}{\tau+\|\boldsymbol{\lambda}\|_{-1}}.
\end{aligned}$$

Similarly we can obtain upper bounds for $\|\boldsymbol{y}^T\boldsymbol{U}_\tau^{-1}\|_2$ and $\|\boldsymbol{z}_k^T\boldsymbol{U}_\tau^{-1}\|_2$.

For f_1,

$$\begin{aligned}
|f_1| &= |\frac{\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1 + \lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1 - \lambda_1\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}| \\
&= |\frac{\boldsymbol{y}^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}| \\
&\leq \frac{\frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}}{1+\frac{n\lambda_1}{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})}} \\
&\leq \frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\cdot\Big(\frac{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})}{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})+n\lambda_1}\Big) \\
&\leq \frac{C_3 n}{\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1}.
\end{aligned}$$

For g_1, we have

$$\begin{aligned}
|g_1| &= |\frac{\boldsymbol{z}_k^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1 + \lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{z}_k^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1 - \lambda_1\boldsymbol{z}_k^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}| \\
&= |\frac{\boldsymbol{z}_k^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}{1+\lambda_1\boldsymbol{z}_1^T\boldsymbol{U}_{-1,\tau}^{-1}\boldsymbol{z}_1}| \\
&\leq \frac{\frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}}{1+\frac{n\lambda_1}{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})}} \\
&\leq \frac{C_1 n}{(\tau+\|\boldsymbol{\lambda}\|_{-1})}\cdot\Big(\frac{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})}{C_2(\tau+\|\boldsymbol{\lambda}\|_{-1})+n\lambda_1}\Big) \\
&\leq \frac{C_3 n}{\tau+\|\boldsymbol{\lambda}\|_{-1}+n\lambda_1}.
\end{aligned}$$

This completes the proof.

A.8 Proofs for Section 2.7

The proofs follow similar conceptual steps to the noiseless case, but several technical adjustments are needed. This is because: on the on hand, the clean label vector $\boldsymbol{y}$ enters the features equation $\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\eta}^T + \boldsymbol{Q}$; on the other had, the estimator $\hat{\boldsymbol{\eta}}$ is generated according to the noisy label vector $\boldsymbol{y}_c$. We start from defining some additional primitive quadratic forms on $\boldsymbol{U}_0 = \boldsymbol{Q}\boldsymbol{Q}^T$:

$$\begin{aligned}
s_c &= \boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{y}, \\
h_c &= \boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{d}, \\
g_{c,i} &= \boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{e}_i, \quad i\in[n], \\
s_{cc} &= \boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{y}_c,
\end{aligned} \tag{A.51}$$

The subscript c here emphasizes that the corrupted noise vector enters these quantities (unlike the corresponding ones in Appendix A.1). The lemma below is our analogue to Lemma 8.

Lemma 19. *Recall* $D := s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2$, *then*

$$\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1} = \boldsymbol{y}_c^T\boldsymbol{U}_0^{-1} - \frac{1}{D}\boldsymbol{v}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_0^{-1},$$

where

$$\boldsymbol{v} = \Big[\|\boldsymbol{\eta}\|_2 s_c + \|\boldsymbol{\eta}\|_2(s_c h - s h_c), h_c h + h_c - s_c t - \|\boldsymbol{\eta}\|_2^2(s_c h - s h_c), s_c + s_c h - s h_c\Big].$$

Next, the lemma below gives upper/lower bounds for the newly defined quadratic forms in (A.51).

Lemma 20. *Assume* $\boldsymbol{\Sigma} = \boldsymbol{I}$ *and* $p > Cn\log n + n + 1$ *for a sufficiently large constant* C. *Fix* $\delta \in (0,1)$ *and suppose* n *is large enough such that* $n > c/\delta$ *for some* $c > 1$. *Then, there exist constants* C_i*'s* > 1 *such that with probability at least* $1-\delta$, *the following results hold:*

$$\begin{aligned}\frac{n}{C_1 p} \le& s_c \le \frac{C_1 n}{p},\\ -\frac{C_2 n\|\boldsymbol{\eta}\|_2}{p} \le& h_c \le \frac{C_2 n\|\boldsymbol{\eta}\|_2}{p},\\ \frac{n}{C_3 p} \le& s_{cc} \le \frac{C_3 n}{p}.\end{aligned}$$

Note that the bounds for the quadratic forms above are of the same order as those for the corresponding quadratic forms defined with $\boldsymbol{y}$, e.g., both s and s_c are at the order

of $\Theta(n/p)$. Now we are ready to prove the theorems.

A.8.1 Proof of Theorem 6

Similar to the proofs in Appendix A.2, we again start from the duality argument of [119] and so we need to find the conditions ensuring

$$y_{c,i}\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{e}_i > 0, \text{ for all } i \in [n], \tag{A.52}$$

where $y_{c,i}$ is the i-th element of $\boldsymbol{y}_c$. Lemma 19 and some algebra steps give:

$$\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{e}_i = g_{c,i} - \frac{D_2}{D} = \frac{A+B}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2},$$

where

$$A = g_{c,i} + 2g_{c,i}h - s_c f_i$$

$$B = \|\boldsymbol{\eta}\|_2^2(g_{c,i}s - g_i s_c) + g_i s_c t - g_{c,i}st + g_{c,i}h^2 - g_i h_c h - g_i h_c + s f_i h_c - s_c h f_i$$

$$D_2 =$$

$$\Big[\|\boldsymbol{\eta}\|_2 s_c + \|\boldsymbol{\eta}\|_2(s_c h - s h_c), h_c h + h_c - s_c t - \|\boldsymbol{\eta}\|_2^2(s_c h - s h_c), s_c + s_c h - s h_c\Big] \begin{bmatrix} \|\boldsymbol{\eta}\|_2 g_i \\ g_i \\ f_i \end{bmatrix}.$$

Let us start with an observation regarding the numerator $A+B$. We have already derived conditions making $y_{c,i}A > 0$ in Appendix A.2.2 (to be precise, Appendix A.2.2 considers $y_i(g_i + g_i h - s f_i)$, but the quadratic forms are of the same order, so the same results

apply). Specifically, when showing $y_{c,i}A > 0$, we first have

$$y_{c,i}g_{c,i} > 1/(Cp) > 0$$

with high probability (obtained by Lemma 12). Then, in Appendix A.2.2, we show that the rest of the terms in A, i.e., $|g_{c,i}h|, |s_c f_i|$, are sufficiently small compared to $1/(Cp)$. Note that when there is no label noise, i.e., $\gamma = 0$, then $g_{c,i} = g_i, s_c = s$, $h_c = h$ and $A + B = g_i + g_i h - s f_i$, which becomes the same as what we have in Appendix A.2.2.

Now, in order to derive conditions under which $y_{c,i}(A + B) > 0$, we first decompose

$$A + B = g_{c,i} + A_h - A_f + A_s,$$

where

$$\begin{aligned} A_h &= 2g_{c,i}h + g_{c,i}h^2 - g_i h_c h - g_i h_c \\ A_f &= s_c f_i - s f_i h_c + s_c h f_i \\ A_s &= g_{c,i}\|\boldsymbol{\eta}\|_2^2 s - g_i\|\boldsymbol{\eta}\|_2^2 s_c + g_i s_c t - g_{c,i} st. \end{aligned}$$

The idea is to that show that: (a) in A_h, the term $g_{c,i}h$ is dominant; (b) in A_f, $s_c f_i$ is the dominant term; (c) $|A_s|$ is sufficiently smaller than $1/(Cp)$. To achieve this, we need

$$p > C_0 \max\{n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\} \tag{A.53}$$

for a sufficiently large constant C_0. The reason is that in A_h and A_f, $|h|$ (and $|h_c|$) is upper bounded by $O(n\|\boldsymbol{\eta}\|_2/p)$ with high probability. In A_s, $s_c\|\boldsymbol{\eta}\|_2^2 \le O(n\|\boldsymbol{\eta}\|_2^2/p)$ and $st \le O(n^2\|\boldsymbol{\eta}\|_2^2/p^2)$ with high probability. Therefore, (A.53) ensures the terms mentioned above are sufficiently smaller than 1 as desired.

A.8.2 Proofs of Theorem 7 and Corollary 7.1

Again, similar to the proofs in Appendix A.3.2, we need to lower bound the ratio

$$\frac{(\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta})^2}{\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}_c}. \tag{A.54}$$

Here we will lower bound $\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta}$ and upper bound $\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}_c$. Lemma 19 and some algebra steps give:

$$\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y} + \boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Q}\boldsymbol{\eta} = \frac{s_c\|\boldsymbol{\eta}\|_2^2 - s_c t + h_c h + h_c}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}.$$

Similarly, the denominator $\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}_c$ is

$$\frac{s_{cc} + \|\boldsymbol{\eta}\|_2^2(s_{cc}s - s_c^2) + s_c^2 t - s_{cc}st + 2s_{cc}h + s_{cc}h^2 + sh_ch - 2s_{cc}h_ch - 2s_ch_c}{s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2}.$$

Combining the two expressions above gives that we need to lower bound:

$$\frac{\Big(s_c\|\boldsymbol{\eta}\|_2^2 - s_c t + h_c h + h_c\Big)^2}{D_s\Big(s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2\Big)}, \tag{A.55}$$

where

$$D_s = s_{cc} + \|\boldsymbol{\eta}\|_2^2(s_{cc}s - s_c^2) + s_c^2 t - s_{cc}st + 2s_{cc}h + s_{cc}h^2 + sh_ch - 2s_{cc}h_ch - 2s_ch_c.$$

Recall that in Appendix A.3.2, we have lower bounded $\frac{\Big(s(\|\boldsymbol{\eta}\|_2^2-t)+h^2+h\Big)^2}{s\Big(s(\|\boldsymbol{\eta}\|_2^2-t)+(h+1)^2\Big)}$ and since s_{cc}, s_c, s are of the same order and h_c, h are also of the same order, we actually have the same bound for $\frac{\Big(s_c\|\boldsymbol{\eta}\|_2^2-s_ct+h_ch+h_c\Big)^2}{s(\|\boldsymbol{\eta}\|_2^2-t)+(h+1)^2}$ in (A.55). The next step is to show that D_s is close to s_{cc}. This is true since due to the assumption $p > C\max\{n\sqrt{\log(2n)}\|\boldsymbol{\eta}\|_2, n\|\boldsymbol{\eta}\|_2^2\}$

for a large constant C, the bounds for terms such as $\|\boldsymbol{\eta}\|_2^2 s, st, h^2, h_c$ are sufficiently small compared to 1 (we also illustrate this under (A.53)). Therefore, in D_s, s_{cc} is the dominant term and we finally need to lower bound the term

$$\frac{\Big(s_c\|\boldsymbol{\eta}\|_2^2 - s_c t + h_c h + h_c\Big)^2}{s_{cc}\Big(s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2\Big)}.$$

This satisfies the same bound as $\frac{\Big(s(\|\boldsymbol{\eta}\|_2^2-t)+h^2+h\Big)^2}{s\Big(s(\|\boldsymbol{\eta}\|_2^2-t)+(h+1)^2\Big)}$ in Appendix A.3.2. Since $p > Cn\|\boldsymbol{\eta}\|_2^2$ falls into the low-SNR regime in Corollary 4.1, we can directly apply the results of low-SNR regime in Corollaries 4.1 and 5.2, which gives the desired.

A.8.3 Proofs of auxiliary lemmas

We first prove Lemma 19.

Proof of Lemma 19. The proof follows Appendix A.7.2 except for in the last steps, we have

$$\boldsymbol{y}_c{}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1} = \boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1} - \begin{bmatrix}\|\boldsymbol{\eta}\|_2 s_c & h_c & s_c\end{bmatrix}\boldsymbol{A}^{-1}\begin{bmatrix}\|\boldsymbol{\eta}\|_2\boldsymbol{y}^T\\ \boldsymbol{y}^T\\ \boldsymbol{d}^T\end{bmatrix}\boldsymbol{U}_\tau^{-1}, \tag{A.56}$$

where $\boldsymbol{A}^{-1}$ is

$$\frac{1}{D}\begin{bmatrix}(h+1)^2 - st & \|\boldsymbol{\eta}\|_2(st - h - h^2) & -\|\boldsymbol{\eta}\|_2 s\\ -\|\boldsymbol{\eta}\|_2 s & h+1+\|\boldsymbol{\eta}\|_2^2 s & -s\\ \|\boldsymbol{\eta}\|_2(st-h-h^2) & \|\boldsymbol{\eta}\|_2^2 h^2 - t(1+\|\boldsymbol{\eta}\|_2^2 s) & h+1+\|\boldsymbol{\eta}\|_2^2 s\end{bmatrix},$$

with $D = s(\|\boldsymbol{\eta}\|_2^2 - t) + (h+1)^2$. Then plugging the expression above in (A.56) completes

the proof. □

We now prove Lemma 20. We first start from a lemma bounding $\|\boldsymbol{y}_c + \boldsymbol{y}\|_2^2$ and $\|\boldsymbol{y}_c - \boldsymbol{y}\|_2^2$.

Lemma 21. *Assuming the probability γ of a label flipping is small enough such that $1 - \gamma \geq 1 - (1/C_0)$ for some large constant C_0, there exist large constants $C_1, C_2 > 1$ such that the event*

$$\mathcal{E}_y := \Big\{ \|\boldsymbol{y}_c + \boldsymbol{y}\|_2^2 \geq 4(1 - \frac{1}{C_1})n \quad \textit{and} \quad \|\boldsymbol{y}_c - \boldsymbol{y}\|_2^2 \leq \frac{4}{C_1} n \Big\}, \tag{A.57}$$

holds with probability at least $1 - 4e^{-\frac{n}{C_2}}$.

Proof. We first look at $(\tilde{y}_i + y_i)^2$, which evaluates to either 4 or 0. Since bounded, these variables are independent sub-Gaussians. The mean of $\|\boldsymbol{y}_c + \boldsymbol{y}\|_2^2$ is $4(1-\gamma)n$. Therefore, Hoeffding's bound [163, Ch. 2] gives

$$\mathbb{P}\Big(\big| \|\boldsymbol{y}_c + \boldsymbol{y}\|_2^2 - 4(1-\gamma)n \big| \geq t \Big) \leq 2 \exp\Big(- \frac{t^2}{Cn} \Big).$$

We complete the proof by setting $t = \frac{n}{C_3}$ for a large enough constant C_3. $(\tilde{y}_i - y_i)^2$ also evaluates to either 4 or 0 and the mean of $\|\boldsymbol{y}_c - \boldsymbol{y}\|_2^2$ is $4\gamma n$. Thus, we can repeat the previous derivation to obtain the advertised results. □

Now we are ready to prove Lemma 20.

Proof of Lemma 20. The bounds for h_c, s_{cc} and the upper bound for s_c follow exactly as in Lemma 9 since $\|\boldsymbol{y}_c\|_2^2 = n$ same as $\|\boldsymbol{y}\|_2^2 = n$. We now derive the lower bound for s_c. We will need the following standard lemma (here adapted from [119, Lemma 2]) to bound quadratic forms of a Wishart matrix.

Lemma 22. *Define* $p'(n) := (p - n + 1)$. *Let matrix* $\boldsymbol{M} \sim$ *Wishart*$(p, \mathbf{I}_n)$. *For any unit-frobenius norm vector* $\boldsymbol{v}$ *and any* $t > 0$, *we have*

$$\mathbb{P}\Big(\frac{1}{\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} > p'(n) + \sqrt{2tp'(n)} + 2t\Big) \leq e^{-t}$$
$$\mathbb{P}\Big(\frac{1}{\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} < p'(n) - \sqrt{2tp'(n)}\Big) \leq e^{-t},$$

provided that $p'(n) > 2\max\{t, 1\}$.

We use the parallelogram law to write

$$\boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{y} = \frac{1}{4}\Big((\boldsymbol{y}_c + \boldsymbol{y})^T\boldsymbol{U}_0^{-1}(\boldsymbol{y}_c + \boldsymbol{y}) - (\boldsymbol{y}_c - \boldsymbol{y})^T\boldsymbol{U}_0^{-1}(\boldsymbol{y}_c - \boldsymbol{y})\Big).$$

Let $t = \log n$ and recall that $p'(n) > Cn\log n$ for a sufficiently large constant C. To lower bound s_c, conditioned on event $\mathcal{E}_y$, we have with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned}
\boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{y} &\geq \frac{1}{4}\Big(\frac{4(1-1/C_1)n}{(p'(n) + \sqrt{2\log(n)p'(n)} + 2\log(n))} - \frac{(4/C_1)n}{(p'(n) - \sqrt{2\log(n)p'(n)})}\Big) \\
&\geq \frac{(1-1/C_1)n(p'(n) - \sqrt{2\log(n)p'(n)}) - (1/C_1)n(p'(n) + \sqrt{2\log(n)p'(n)} + 2\log(n))}{(p'(n) - \sqrt{2\log(n)p'(n)})(p'(n) + \sqrt{2\log(n)p'(n)})} \\
&\geq \frac{(1-1/C_3)np'(n)}{C_4p'(n)^2} \\
&\geq \frac{n}{C_5p},
\end{aligned}$$

where we replaced $\log(n)$ with $p'(n)/C$ using the fact that $p'(n) > Cn\log n$ for a sufficiently large constant C above. Let $\mathcal{E}$ be the desired event that $\boldsymbol{y}_c{}^T\boldsymbol{U}_0^{-1}\boldsymbol{y} \geq n/(Cp)$. We then complete the proof by adjusting the probability using $\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}(\mathcal{E}^c|\mathcal{E}_y) + \mathbb{P}(\mathcal{E}_y^c) \leq (1/n) + 4\exp(-n/c_1) \leq c_2/n$. □

A.9 On linear separability of GMM

The main result of this section Lemma 24 proves that GMM data are linearly separable with high-probability as long as $p > n + 2$. The arguments presented are pretty standard in the literature, but included here for completeness. Sharp separability thresholds for the GMM have been recently derived in [39].

We will first need the following technical lemma that lower bounds the minimum singular value of a non-zero mean isotropic Gaussian matrix. The result is a minor extension of the standard proof using Gordon's Gaussian min-max inequality for the case of a centered isotropic Gaussian matrix (e.g. see [162, Exercise 7.3.4]).

Lemma 23. *Let $Q \in \mathbb{R}^{p\times n}$ a matrix with IID standard normal entries and $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\eta} \in \mathbb{R}^p$ fixed vectors. Consider the matrix $\boldsymbol{A} = \boldsymbol{\eta}\mathbf{y}^T + \boldsymbol{Q}$. For every $t > 0$ it holds that*

$$\min_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{A}\boldsymbol{u}\|_2 \geq \sqrt{p-2} - \sqrt{n} - t \tag{A.58}$$

with probability at least $1 - 4e^{-t^2/8}$.

Proof. We now prove the lemma using Gordon's Gaussian comparison inequality [60]. Specifically, we apply a version that appears in [153]. We start by writing

$$\Phi(\boldsymbol{A}) := \min_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{A}\boldsymbol{u}\|_2 = \min_{\|\boldsymbol{u}\|_2=1} \max_{\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^T\boldsymbol{Q}\boldsymbol{u} + (\boldsymbol{w}^T\boldsymbol{\eta})(\boldsymbol{y}^T\boldsymbol{u})$$

Now, following [153, Thm. 3(i)] we focus on the following auxiliary problem where $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^d$ have iid standard normal entries:

$$\phi(\boldsymbol{g}, \boldsymbol{h}) := \min_{\|\boldsymbol{u}\|_2=1} \max_{\|\boldsymbol{w}\|_2=1} \boldsymbol{h}^T\boldsymbol{w} + \boldsymbol{g}^T\boldsymbol{u} + (\boldsymbol{w}^T\boldsymbol{\eta})(\boldsymbol{y}^T\boldsymbol{u}).$$

By decomposing $\boldsymbol{w} = \alpha\frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2} + \mathbf{P}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{w}$ for $\alpha := \frac{\boldsymbol{\eta}^T\boldsymbol{w}}{\|\boldsymbol{\eta}\|_2} \in [0,1]$ and $\mathbf{P}_{\boldsymbol{\eta}}^{\perp} = \mathbf{I}_d - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^T}{\|\boldsymbol{\eta}\|_2^2}$, we can see

that

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\|\boldsymbol{u}\|_2=1} \max_{\alpha\in[0,1]} \|\mathbf{P}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{h}\|_2\sqrt{1-\alpha^2} + \alpha\frac{\boldsymbol{\eta}^T\boldsymbol{h}}{\|\boldsymbol{\eta}\|_2} + \boldsymbol{g}^T\boldsymbol{u} + (\boldsymbol{y}^T\boldsymbol{u})\alpha\|\boldsymbol{\eta}\|_2 \tag{A.59}$$

$$\geq \min_{\|\boldsymbol{u}\|_2=1} \|\mathbf{P}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{h}^T\|_2 + \boldsymbol{g}^T\boldsymbol{u} \tag{A.60}$$

$$= \|\mathbf{P}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{h}\|_2 - \|\boldsymbol{g}\|_2 \tag{A.61}$$

But now from standard concentration arguments (e.g. see [124, Lemma B.2], for all $t > 0$ with probability at least $1 - 2e^{-t^2/2}$ it holds that $\|\mathbf{P}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{h}\|_2 - \|\boldsymbol{g}\|_2 \geq \sqrt{p-2} - \sqrt{n} - 2t$. We now invoke Gordon's inequality to complete the proof:

$$\Pr\left(\Phi(\boldsymbol{A}) \leq \sqrt{p-2} - \sqrt{n} - t\right) \leq 2\Pr\left(\phi(\boldsymbol{g}, \boldsymbol{h}) \leq \sqrt{p-2} - \sqrt{n} - t\right) \leq 4e^{-t^2/8}. \tag{A.62}$$

□

We are now ready to state and prove the main result of this section.

Lemma 24. *Let training data $\{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ be generated from the GMM in Equation (2.1). Assume $p > n + 2 + t$ for some $t > 0$. Then with probability at least $1 - 4e^{-t^2/8}$ the following statements hold:*

(i) The min-norm interpolator is feasible, i.e. there exists $\boldsymbol{\beta} \in \mathbb{R}^d$ such that for all $i \in [n] : y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

(ii) The training data are linearly separable, i.e. there exists $\boldsymbol{\beta} \in \mathbb{R}^d$ such that for all $i \in [n] : y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) \geq 1$.

Proof. To prove the first statement we need to show that the feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has full row-rank with high probability. Equivalently, we show that $\min_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{X}^T\boldsymbol{u}\|_2 > 0$ with high-probability. This is a direct application of Lemma 23 above for $\boldsymbol{A} = \boldsymbol{X}^T$.

Now we prove the second statement. From part (i), there exists $\boldsymbol{\beta}$ such that $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, i \in [n]$. Since $y_i \in \{\pm 1\}, i \in [n]$ it then holds that $y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) = 1, i \in [n]$. Thus, the same vector $\boldsymbol{\beta}$ from part (i) that interpolates the data is also a linear separator. $\square$

Appendix B

Appendix for Chapter 3

B.1 Lemmas used in the proof of Theorem 9

B.1.1 Auxiliary Lemmas

In this section, we state a series of auxiliary lemmas that we use to prove Lemma 7. The following result shows concentration of the norms of the label indicators $\boldsymbol{v}_c, c \in [k]$ under the equal-priors assumption (Assumption 2). Intuitively, in this balanced setting there are $\Theta(n/k)$ samples for each class; hence, $\Theta(n/k)$ non-zeros (in fact, 1's) in each label indicator vector $\boldsymbol{v}_c$.

Lemma 25. *Under the setting of Assumption 2, there exist large constants $C_1, C_2 > 0$ such that the event*

$$\mathcal{E}_v := \Big\{ \left(1 - \frac{1}{C_1}\right) \frac{n}{k} \leq \|\boldsymbol{v}_c\|_2^2 \leq \left(1 + \frac{1}{C_1}\right) \frac{n}{k} \ , \ \forall c \in [k] \Big\}, \tag{B.1}$$

holds with probability at least $1 - 2ke^{-\frac{n}{C_2 k^2}}$.

Next, we provide bounds on the "base case" 0-th order quadratic forms that involve

the Gram matrix $\boldsymbol{A}_0^{-1}$. We do this in three lemmas presented below. The first Lemma 26 follows by a direct application of [164, Lemma 4 and 5]. The only difference is that we keep track of throughout the proof is the scaling of $\mathcal{O}(1/k)$ arising from the multiclass case in the $\boldsymbol{v}_j$'s. For instance, the bound of the term $h_{mj}^{(0)} := \mathbf{v}_m^T \boldsymbol{A}_0^{-1} \mathbf{d}_j$ involves a term $\tilde{\rho}_{n,k} = \min\{1, \sqrt{\log(2n)/k}\}$ compared to the binary case. The other two Lemmas 27 and 28 are proved in Section B.1.3.

Lemma 26 (0-th order Quadratic forms, Part I)**.** *Under the event $\mathcal{E}_v$, there exist constants c_i's and C_i's > 1 such that the following bounds hold with probability at least $1 - c_1 k e^{-\frac{n}{c_2}}$.*

$$
\begin{aligned}
t_{jj}^{(0)} &\leq \frac{C_1 n \|\boldsymbol{\mu}\|_2^2}{p} \quad \textit{for all } j \in [k], \\
|h_{mj}^{(0)}| &\leq \tilde{\rho}_{n,k} \frac{C_2 n \|\boldsymbol{\mu}\|_2}{\sqrt{k} p} \quad \textit{for all } m, j \in [k], \\
|t_{mj}^{(0)}| &\leq \frac{C_3 n \|\boldsymbol{\mu}\|_2^2}{p} \quad \textit{for all } m \neq j \in [k], \\
\|\mathbf{d}_j\|_2^2 &\leq C_4 n \|\boldsymbol{\mu}\|_2^2 \quad \textit{for all } j \in [k], \\
\max_{i \in [n]} |f_{ji}^{(0)}| &\leq \frac{C_5 \sqrt{\log(2n)} \|\boldsymbol{\mu}\|_2}{p} \quad \textit{for all } j \in [k].
\end{aligned}
$$

To sharply characterize the forms $s_{ij}^{(0)}$ we need additional work, particularly for the cross-terms where $i \neq j$. We will make use of fundamental concentration inequalities on quadratic forms of inverse Wishart matrices. The following lemma controls these quadratic forms, and shows in particular that the $s_{ij}^{(0)}$ terms for $i \neq j$ are much smaller than the corresponding terms $s_{jj}^{(0)}$. This sharp control of the cross-terms is essential for several subsequent proof steps.

Lemma 27 (0-th order Quadratic forms, Part II)**.** *Working on the event $\mathcal{E}_v$ defined in Equation* (B.1)*, assume that $p > Cn\log(kn) + n - 1$ for large enough constant $C > 1$*

and large n. There exist constants C_i's > 1 such that with probability at least $1 - \frac{C_0}{n}$, the following bound holds:

$$\begin{aligned}\frac{C_1 - 1}{C_1} \cdot \frac{n}{kp} \leq & s_{jj}^{(0)} \leq \frac{C_1 + 1}{C_1} \cdot \frac{n}{kp}, \quad \textit{for } j \in [k], \\ -\frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp} \leq & s_{ij}^{(0)} \leq \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp}, \quad \textit{for } i \neq j \in [k].\end{aligned}$$

The proof of Lemma 27 for the cross terms with $i \neq j$ critically uses the in-built orthogonality of the label indicator vectors $\{\boldsymbol{v}_c\}_{c \in [k]}$. Finally, the following lemma controls the quadratic forms $g_{ji}^{(0)}$.

Lemma 28 (0-th order Quadratic forms, Part III)**.** *Working on the event $\mathcal{E}_v$ defined in Equation* (B.1)*, given $p > Ck^3 n \log(kn) + n - 1$ for a large constant C, there exist large enough constants C_1, C_2, such that with probability at least $1 - \frac{2}{kn}$, we have for every $i \in [n]$:*

$$\begin{aligned}\left(1 - \frac{1}{C_1}\right) \frac{1}{p} \leq & g_{(y_i)i}^{(0)} \leq \left(1 + \frac{1}{C_1}\right) \frac{1}{p}, \\ -\frac{1}{C_2} \cdot \frac{1}{k^2 p} \leq & g_{ji}^{(0)} \leq \frac{1}{C_2} \cdot \frac{1}{k^2 p}, \quad \textit{for } j \neq y_i.\end{aligned}$$

B.1.2 Proof of Lemma 7

In this section, we provide the full proof of Lemma 7. We begin with a proof outline.

Proof outline

As explained in Section 3.6.2, it suffices to consider the case where $j = k$, since when $j \neq k$ we can simply change the order of adding mean components, described in Equation (3.36), so that the j-th mean component is added last. For concreteness, we will also fix $i \in [n]$, $y_i = k$ and define as shorthand $m := k - 1$. These fixes are without

loss of generality. The reason why we fix $j = k$ and $m = k - 1$ is that when we do the proof, we want to add the $k-1$-th and k-th components last. This is for ease of reading and understanding.

For the case $j = k$, the leave-one-out quadratic forms in Lemma 7 are equal to the quadratic forms of order $k-1$, given by $s_{kk}^{(k-1)}, t_{kk}^{(k-1)}$, $h_{kk}^{(k-1)}$, $g_{ki}^{(k-1)}$ and $f_{ki}^{(k-1)}$. We will proceed recursively starting from the quadratic forms of order 1 building up all the way to the quadratic forms of order $k-1$. Specifically, starting from order 1, we will work on the event

$$\mathcal{E}_q := \{\text{all the inequalities in Lemmas 26, 27 and 28 hold}\}, \tag{B.2}$$

Further, we note that Lemma 28 shows that the bound for $g_{y_i i}^{(0)}$ is different from the bound for $g_{ji}^{(0)}$ when $j \neq y_i$. We will show the following set of upper and lower bounds:

$$\begin{aligned}
\left(\frac{C_{11}-1}{C_{11}}\right)\frac{n}{kp} \leq s_{kk}^{(1)} \leq \left(\frac{C_{11}+1}{C_{11}}\right)\frac{n}{kp},\\
-\left(\frac{C_{12}+1}{C_{12}}\right)\frac{\sqrt{n}}{kp} \leq s_{mk}^{(1)} \leq \left(\frac{C_{12}+1}{C_{12}}\right)\frac{\sqrt{n}}{kp},\\
t_{kk}^{(1)} \leq \frac{C_{13} n\|\boldsymbol{\mu}\|_2^2}{p},\\
|h_{mk}^{(1)}| \leq \tilde{\rho}_{n,k}\frac{C_{14} n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p},\\
|t_{mk}^{(1)}| \leq \frac{C_{15} n\|\boldsymbol{\mu}\|_2^2}{p},\\
\|\mathbf{d}_k\|_2^2 \leq C_{16} n\|\boldsymbol{\mu}\|_2^2,\\
|f_{ki}^{(1)}| \leq \frac{C_{17}\sqrt{n}\|\boldsymbol{\mu}\|_2}{p},\\
\left(1-\frac{1}{C_{18}}\right)\frac{1}{p} \leq g_{(y_i)i}^{(1)} \leq \left(1+\frac{1}{C_{18}}\right)\frac{1}{p}, \text{ and}\\
-\frac{1}{C_{19}k^2 p} \leq g_{mi}^{(1)} \leq \frac{1}{C_{19}k^2 p}
\end{aligned} \tag{B.3}$$

with probability at least $1 - \frac{c}{kn^2}$. Comparing the bounds on the terms of order 1 in Equation (B.3) with the terms in Lemmas 26, 27 and 28 of order 0, the key observation is that they are all at the same order. This allows us to repeat the same argument to now bound corresponding terms of order 2, and so on until order $k-1$. Note that for each $j \in [k]$, we have n terms of the form $g_{ji}^{(1)}$, corresponding to each value of $i \in [n]$. Thus, we will adjust the final probabilities by applying a union bound over the n training examples.

Proofs for 1-st order quadratic forms in Equation (B.3)

The proof makes repeated use of Lemmas 26, 27 and 28. In fact, we will throughout condition on the event $\mathcal{E}_q$, defined in Equation (B.2), which holds with probability at least $1 - \frac{c_1}{n} - c_2 e^{-\frac{n}{c_0 k^2}}$. Specifically, by Lemma 26 we have

$$h_{mj}^{(0)} \leq \tilde{\rho}_{n,k} \frac{C_1 \epsilon_n}{k^2 \sqrt{n}}, \qquad \max_{i \in [n]} |f_{mi}^{(0)}| \leq \frac{C_2 \epsilon_n}{k^{1.5} n}, \quad \text{and} \quad \frac{s_{mj}^{(0)}}{s_{kk}^{(0)}} \leq \frac{C}{\sqrt{n}} \text{ for } m, j \neq k, \tag{B.4}$$

where we recall from Equation (3.41) the notation $\epsilon_n := \frac{k^{1.5} n \sqrt{n} \|\boldsymbol{\mu}\|_2}{p}$. Also, recall that we choose $\epsilon_n \leq \tau$ for a sufficiently small constant τ.

In order to make use of Lemmas 26, 27 and 28, we need to relate the quantities of interest to corresponding quadratic forms involving $\boldsymbol{A}_0$. We do this recursively and make repeated use of the Woodbury identity. The recursions are proved in Appendix B.4.1. We now provide the proofs for the bounds on the terms in Equation (B.3) one-by-one.

Bounds on $s_{mk}^{(1)}$. By Equation (B.28) in Appendix B.4.1, we have

$$s_{mk}^{(1)} = s_{mk}^{(0)} - \frac{1}{\det_0} (\star)_s^{(0)}, \tag{B.5}$$

where we define

$$(\star)_s^{(0)} := (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1k}^{(0)}s_{1m}^{(0)} + s_{1m}^{(0)}h_{k1}^{(0)}h_{11}^{(0)} + s_{1k}^{(0)}h_{m1}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}h_{k1}^{(0)}h_{m1}^{(0)} + s_{1m}^{(0)}h_{k1}^{(0)} + s_{1k}^{(0)}h_{m1}^{(0)}$$
$$\det_0 := s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}) + (1 + h_{11}^{(0)})^2. \tag{B.6}$$

The essential idea is to show that $|\frac{(\star)_s^{(0)}}{\det_0}|$ is sufficiently small compared to $|s_{mk}^{(0)}|$. We first look at the first term given by $\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1k}^{(0)}s_{1m}^{(0)}\Big)/\det_0$. By Lemmas 26, 27 and the definition of $\det_0$, we have

$$\Big|\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1k}^{(0)}s_{1m}^{(0)}\Big)\Big| \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|s_{1k}^{(0)}s_{1m}^{(0)}|}{s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})} = \Big|\frac{s_{1k}^{(0)}s_{1m}^{(0)}}{s_{11}^{(0)}}\Big| \leq \frac{C_1}{\sqrt{n}} \cdot \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp},$$

where we use $\det_0 \geq s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})$ and $s_{mj}^{(0)}/s_{kk}^{(0)} \leq C/\sqrt{n}$ for all $m, j \neq k$. Now, we upper bound the other two dominant terms $|s_{1m}^{(0)}h_{k1}^{(0)}/\det_0|$ and $|s_{1k}^{(0)}h_{m1}^{(0)}/\det_0|$. Note that the same bound will apply to the remaining terms in Equation (B.6) because we trivially have $|h_{ij}^{(0)}| = \mathcal{O}(1)$ for all $(i,j) \in [k]$. Again, Lemmas 26 and 27 give us

$$\Big|\frac{s_{1m}^{(0)}h_{k1}^{(0)}}{\det_0}\Big| \leq \frac{|s_{1m}^{(0)}h_{k1}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{\tilde{\rho}_{n,k}C_3\epsilon_n}{\Big(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2 k^2\sqrt{n}} \cdot \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}.$$

The identical bound holds for $|s_{1k}^{(0)}h_{m1}^{(0)}|$. Noting that $|s_{mk}^{(0)}| \leq \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}$, we then have

$$\begin{aligned}
|s_{mk}^{(1)}| &\leq |s_{mk}^{(0)}| + \Big|\frac{(\star)_s^{(0)}}{\det_0}\Big| \\
&\leq \left(1 + \frac{C_6}{\sqrt{n}} + \frac{C_7\tilde{\rho}_{n,k}\epsilon_n}{\Big(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2 k^2\sqrt{n}}\right)\frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp} \\
&\leq (1+\alpha) \cdot \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp},
\end{aligned} \tag{B.7}$$

where in the last inequality, we use that $\epsilon \leq \tau$ for sufficiently small constant $\tau > 0$, and

defined

$$\alpha := \frac{C_6}{\sqrt{n}} + \frac{C_7\tau}{\left(1 - \frac{C_5\tau}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}.$$

Now, we pick τ to be sufficiently small and n to be sufficiently large such that $(1+\alpha)\frac{C_2+1}{C_2} \leq \frac{C_8+1}{C_8}$ for some constant $C_8 > 0$. Then, we conclude with the following upper bound:

$$|s_{mk}^{(1)}| \leq \frac{C_8+1}{C_8} \cdot \frac{\sqrt{n}}{kp}.$$

Bounds on $s_{kk}^{(1)}$. Equation (B.29) in Appendix B.4.1 gives us

$$s_{kk}^{(1)} = s_{kk}^{(0)} - \frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2 + 2s_{1k}^{(0)}h_{k1}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}{h_{k1}^{(0)}}^2 + 2s_{1k}^{(0)}h_{k1}^{(0)}\Big).$$

First, we lower bound $s_{kk}^{(1)}$ by upper bounding $\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2\Big)$. Lemmas 26 and 27 yield

$$\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2\Big) \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2}{s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}) + (1+h_{11}^{(0)})^2} \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2}{s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})} \leq \frac{C_1}{n} \cdot \frac{n}{kp}.$$

It suffices to upper bound the other dominant term $|s_{1k}^{(0)}h_{k1}^{(0)}|/\det_0$. For this term, we have

$$\Big|\frac{s_{1k}^{(0)}h_{k1}^{(0)}}{\det_0}\Big| \leq \frac{|s_{1k}^{(0)}h_{k1}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{C_2+1}{C_2} \cdot \frac{\sqrt{n}}{kp}.$$

Thus, we get

$$s_{kk}^{(1)} \geq \left(1 - \frac{C_1}{n} - \frac{C_5\tilde{\rho}_{n,k}n\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{C_6-1}{C_6} \cdot \frac{n}{kp} \geq (1-\alpha) \cdot \frac{C_6-1}{C_6} \cdot \frac{n}{kp}.$$

Next, we upper bound $s_{kk}^{(1)}$ by a similar argument, and get

$$\begin{aligned} s_{kk}^{(1)} &\leq |s_{kk}^{(0)}| + \frac{1}{\det_0}\Big|2s_{1k}^{(0)}h_{k1}^{(0)}h_{11}^{(0)} + s_{11}^{(0)}{h_{k1}^{(0)}}^2 + 2s_{1k}^{(0)}h_{k1}^{(0)}\Big| \\ &\leq \left(1 + \frac{C_7\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{C_8+1}{C_8}\cdot\frac{n}{kp} \leq (1+\alpha')\frac{C_8+1}{C_8}\cdot\frac{n}{kp}, \end{aligned}$$

where we used $\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}){s_{1k}^{(0)}}^2\Big) > 0$ in the first step. As above, we can tune ϵ and n such that $(1+\alpha')\frac{C_8+1}{C_8} \leq \frac{C_9+1}{C_9}$ and $(1-\alpha)\frac{C_6-1}{C_6} \geq \frac{C_9-1}{C_9}$ for sufficiently large constant $C_9 > 0$.

Bounds on $h_{mk}^{(1)}$. Equation (B.30) in Appendix B.4.1 gives us

$$h_{mk}^{(1)} = h_{mk}^{(0)} - \frac{1}{\det_0}(\star)_h^{(0)},$$

where we define

$$(\star)_h^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}h_{1k}^{(0)} + h_{m1}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} + h_{m1}^{(0)}h_{1k}^{(0)} + s_{1m}^{(0)}t_{k1}^{(0)} + s_{1m}^{(0)}t_{k1}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}t_{k1}^{(0)}h_{m1}^{(0)}.$$

We focus on the two dominant terms $((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}h_{1k}^{(0)})/\det_0$ and $s_{1m}^{(0)}t_{k1}^{(0)}/\det_0$. For the first dominant term $((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}h_{1k}^{(0)})/\det_0$, Lemmas 26 and 27 yield

$$\Big|\frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}h_{1k}^{(0)}\Big)\Big| \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|s_{1m}^{(0)}h_{1k}^{(0)}|}{s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})} \leq \Big|\frac{s_{1m}^{(0)}h_{1k}^{(0)}}{s_{11}^{(0)}}\Big| \leq \frac{C_1}{\sqrt{n}}|h_{1k}^{(0)}| \leq \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}.$$

For the second dominant term $s_{1m}^{(0)}t_{k1}^{(0)}/\det_0$, we have

$$\frac{1}{\det_0}s_{1m}^{(0)}t_{k1}^{(0)} \leq \frac{|s_{1m}^{(0)}t_{k1}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_3 n\sqrt{n}\|\boldsymbol{\mu}\|_2^2}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 kp^2} \leq \frac{C_5\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{1.5}\sqrt{n}}\cdot\frac{\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}},$$

where we use the fact $1/\sqrt{k} < \tilde{\rho}_{n,k}$ for $k > 1$. Thus, we get

$$|h_{mk}^{(1)}| \le |h_{mk}^{(0)}| + \left|\frac{1}{\det_0}(\star)_h^{(0)}\right| \le \left(1 + \frac{C_1}{\sqrt{n}} + \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{1.5}\sqrt{n}}\right)\frac{C_6\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}$$
$$\le (1+\alpha)\frac{C_7\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}},$$

and there exists constant C_8 such that $(1+\alpha)C_7 \le C_8$, which shows the desired upper bound.

Bounds on $t_{kk}^{(1)}$. Equation (B.32) in Appendix B.4.1 gives us

$$t_{kk}^{(1)} = t_{kk}^{(0)} - \frac{1}{\det_0}\left(\left(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}\right)h_{1k}^{(0)^2} + 2t_{1k}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}t_{1k}^{(0)^2} + 2t_{1k}^{(0)}h_{1k}^{(0)}\right).$$

We only need an upper bound on $t_{kk}^{(1)}$. The first dominant term $s_{11}^{(0)}t_{1k}^{(0)^2}/\det_0$ is upper bounded as follows:

$$\frac{s_{11}^{(0)}t_{1k}^{(0)^2}}{\det_0} \le \frac{s_{11}^{(0)}t_{1k}^{(0)^2}}{(1+h_{11}^{(0)})^2} \le \frac{C_6 n^3\|\boldsymbol{\mu}\|_2^4}{\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 kp^3} \le \frac{C_7\epsilon_n^2}{\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 pk^4 n}\cdot\frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Next, the second dominant term, $t_{1k}^{(0)}h_{1k}^{(0)}/\det_0$, is upper bounded as

$$\frac{t_{1k}^{(0)}h_{1k}^{(0)}}{\det_0} \le \frac{|t_{1k}^{(0)}h_{1k}^{(0)}|}{(1+h_{11}^{(0)})^2} \le \frac{C_8\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2}\cdot\frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Combining the results above gives us

$$t_{kk}^{(1)} \le t_{kk}^{(0)} + \frac{1}{\det_0}\left|2t_{1k}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} + s_{11}^{(0)}t_{1k}^{(0)^2} + 2t_{1k}^{(0)}h_{1k}^{(0)}\right|$$
$$\le \left(1 + \frac{C_9\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{n\|\boldsymbol{\mu}\|_2^2}{p} \le \frac{C_5 n\|\boldsymbol{\mu}\|_2^2}{p}.$$

This shows the desired upper bound.

Bounds on $t^{(1)}_{mk}$. Equation (B.31) in Appendix B.4.1 gives us

$$t^{(1)}_{mk} = t^{(0)}_{mk} - \frac{1}{\det_0}(\star)^{(0)}_t,$$

where we define

$$(\star)^{(0)}_t = (\|\boldsymbol{\mu}\|_2^2 - t^{(0)}_{11})h^{(0)}_{1m}h^{(0)}_{1k} + t^{(0)}_{m1}h^{(0)}_{1k}h^{(0)}_{11} + t^{(0)}_{k1}h^{(0)}_{1m}h^{(0)}_{11} + t^{(0)}_{1m}h^{(0)}_{1k} + t^{(0)}_{1k}h^{(0)}_{1m} - s^{(0)}_{11}t^{(0)}_{1m}t^{(0)}_{1k}.$$

Again, we only need an upper bound on $t^{(1)}_{mk}$. As in the previously derived bounds, we have

$$\frac{1}{\det_0}(\|\boldsymbol{\mu}\|_2^2 - t^{(0)}_{11})h^{(0)}_{1m}h^{(0)}_{1k} \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t^{(0)}_{11})|h^{(0)}_{1m}h^{(0)}_{1k}|}{s^{(0)}_{11}(\|\boldsymbol{\mu}\|_2^2 - t^{(0)}_{11})} \leq \frac{C_1\tilde{\rho}^2_{n,k}n^2\|\boldsymbol{\mu}\|_2^2}{kp^2} \cdot \frac{kp}{n} \leq \frac{C_1 n\|\boldsymbol{\mu}\|_2^2}{p}.$$

The other dominant term $t^{(0)}_{1m}h^{(0)}_{1m}/\det_0$ is upper bounded as:

$$\frac{t^{(0)}_{1m}h^{(0)}_{1m}}{\det_0} \leq \frac{|t^{(0)}_{1m}h^{(0)}_{1m}|}{(1+h^{(0)}_{11})^2} \leq \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}(\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p}.$$

Combining the results above yields

$$\begin{aligned}|t^{(1)}_{mk}| &\leq |t^{(0)}_{mk}| + \frac{1}{\det_0}\left|(\star)^{(0)}_t\right| \\ &\leq \left(C_1 + \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}}\right)\frac{n\|\boldsymbol{\mu}\|_2^2}{p} \leq \frac{C_4 n\|\boldsymbol{\mu}\|_2^2}{p}.\end{aligned}$$

Note that both $t^{(0)}_{kk}$ and $t^{(0)}_{mk}$ are much smaller than $\|\boldsymbol{\mu}\|_2^2$. The above upper bound shows that this continues to hold for $t^{(1)}_{kk}$ and $t^{(1)}_{mk}$ since $p \gg n$.

Bounds on $f^{(1)}_{ki}$. Consider $i \in [n]$ and fix $y_i = k$ without loss of generality. Equa-

tion (B.33) in Appendix B.4.1 gives us

$$f_{ki}^{(1)} = f_{ki}^{(0)} - \frac{1}{\det_0}(\star)_f^{(0)}, \tag{B.8}$$

where we define

$$(\star)_f^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)} + t_{1k}^{(0)}g_{1i}^{(0)} + t_{1k}^{(0)}h_{11}^{(0)}g_{1i}^{(0)} + h_{1k}^{(0)}f_{1i}^{(0)} + h_{1k}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}. \tag{B.9}$$

We only need an upper bound on $f_{ki}^{(1)}$. We consider the dominant terms $(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)}/\det_0$, $t_{1k}^{(0)}g_{1i}^{(0)}/\det_0$, $h_{1k}^{(0)}f_{1i}^{(0)}/\det_0$ and $s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}/\det_0$. Lemmas 26, 27 and 28 give us

$$\begin{aligned}
&\frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)}g_{1i}^{(0)}}{\det_0} \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|h_{1k}^{(0)}g_{1i}^{(0)}|}{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{11}^{(0)}} \\
&\leq \frac{C_1\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p} \cdot \frac{1}{C_2k^2p} \cdot \frac{kp}{n} \leq \frac{C_3}{k^{1.5}\sqrt{n}} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \\
&\frac{t_{1k}^{(0)}g_{1i}^{(0)}}{\det_0} \leq \frac{|t_{1k}^{(0)}g_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_4n\|\boldsymbol{\mu}\|_2^2}{\left(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon}{k^2\sqrt{n}}\right)^2 k^2p^2} \leq \frac{C_7\epsilon_n}{\left(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{3.5}n} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \\
&\frac{h_{1k}^{(0)}f_{1i}^{(0)}}{\det_0} \leq \frac{|h_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_6\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p}, \text{ and} \\
&\frac{s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}}{\det_0} \leq \frac{|s_{11}^{(0)}t_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_7\epsilon_n^2}{k^4n\left(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{\sqrt{n}\|\boldsymbol{\mu}\|_2}{p},
\end{aligned}$$

where, in the last two steps, we used the upper bound $C\sqrt{n}\|\boldsymbol{\mu}\|_2/p$ for $|f_{ji}^{(0)}|$ and previ-

ously derived bounds on $|h_{1k}^{(0)}|$ and $|s_{11}^{(0)}t_{1k}^{(0)}|$. Thus, we have

$$\begin{aligned}
|f_{ki}^{(1)}| &\leq |f_{ki}^{(0)}| + \Big|\frac{1}{\det_0}(\star)_f^{(0)}\Big| \\
&\leq \left(1 + \frac{C_3}{k^{1.5}\sqrt{n}} + \frac{C_8\epsilon_n}{\Big(1 - \frac{C_5\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2 k^2\sqrt{n}}\right)\frac{C_9\sqrt{n}\|\boldsymbol{\mu}\|_2}{p} \\
&\leq (1+\alpha)\frac{C_{10}\epsilon_n}{k^{1.5}n},
\end{aligned}$$

and we have $(1+\alpha)C_{10} \leq C_{11}$ for a large enough positive constant C_{11}. This shows the desired upper bound.

Bounds on $g_{ki}^{(1)}$ and $g_{mi}^{(1)}$. Equation (B.34) in Appendix B.4.1 gives

$$z_{ci}\mathbf{e}_i^T\boldsymbol{A}_1^{-1}\boldsymbol{u}_k = |z_{ci}|^2\Big(\mathbf{e}_i^T\boldsymbol{A}_0^{-1}\boldsymbol{v}_k - \frac{1}{\det_0}(\star)_{gk}^{(0)}\Big) = |z_{ci}|^2\Big(g_{ki}^{(0)} - \frac{1}{\det_0}(\star)_{gk}^{(0)}\Big), \tag{B.10}$$

where we define

$$(\star)_{gk}^{(0)} = (\|\boldsymbol{\mu}_1\|_2^2 - t_{11}^{(0)})s_{1k}^{(0)}g_{1i}^{(0)} + g_{1i}^{(0)}h_{11}^{(0)}h_{k1}^{(0)} + g_{1i}^{(0)}h_{k1}^{(0)} + s_{1k}^{(0)}f_{1i}^{(0)} + s_{1k}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}h_{k1}^{(0)}f_{1i}^{(0)}.$$

Lemmas 26, 27 and 28 give us

$$\begin{aligned}
&\frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|s_{1k}^{(0)}g_{1i}^{(0)}|}{\det_0} \leq \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})|s_{1k}^{(0)}g_{1i}^{(0)}|}{(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{11}^{(0)}} \leq \frac{C_1}{\sqrt{n}}\cdot\frac{1}{C_2k^2p}, \\
&\frac{|h_{k1}^{(0)}g_{1i}^{(0)}|}{\det_0} \leq \frac{|h_{k1}^{(0)}g_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\Big(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2 k^2\sqrt{n}}\cdot\frac{1}{C_2k^2p}, \text{ and} \\
&\frac{|s_{1k}^{(0)}f_{1i}^{(0)}|}{\det_0} \leq \frac{|s_{1k}^{(0)}f_{1i}^{(0)}|}{(1+h_{11}^{(0)})^2} \leq \frac{C_5\epsilon_n}{\Big(1 - \frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\Big)^2 \sqrt{k}\sqrt{n}}\cdot\frac{1}{C_2k^2p}.
\end{aligned}$$

We then have

$$
\begin{aligned}
g_{ki}^{(1)} &\geq g_{ki}^{(0)} - \frac{1}{\det_0}|(\star)_{gk}^{(0)}| \geq \left(1-\frac{1}{C}\right)\left(1-\frac{C_1}{k^2\sqrt{n}} - \frac{C_6\epsilon_n}{\left(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{2.5}\sqrt{n}}\right)\frac{1}{p} \\
&\geq \left(1-\frac{1}{C}\right)\frac{1-\alpha}{p} \\
g_{ki}^{(1)} &\leq g_{ki}^{(0)} + \frac{1}{\det_0}|(\star)_{gk}^{(0)}| \leq \left(1+\frac{1}{C}\right)\left(1+\frac{C_1}{k^2\sqrt{n}} + \frac{C_7\epsilon_n}{\left(1-\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^{2.5}\sqrt{n}}\right)\frac{1}{p} \\
&\leq \left(1+\frac{1}{C}\right)\frac{1+\alpha}{p},
\end{aligned}
$$

where for large enough n and positive constant C_9, we have $(1+\alpha)\frac{C+1}{C} \leq \frac{C_9+1}{C_9}$ and $(1-\alpha)\frac{C-1}{C} \geq \frac{C_9-1}{C_9}$. Similarly, for the case $m \neq k$, we have

$$
z_{ci}\mathbf{e}_i^T \boldsymbol{A}_1^{-1}\boldsymbol{u}_m = |z_{ci}|^2\left(\mathbf{e}_i^T\boldsymbol{A}_0^{-1}\boldsymbol{v}_m - \frac{1}{\det_0}(\star)_{gm}^{(0)}\right) = |z_{ci}|^2\Big(g_{mi}^{(0)} - \frac{1}{\det_0}(\star)_{gm}^{(0)}\Big), \tag{B.11}
$$

where we define

$$
(\star)_{gm}^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}g_{1i}^{(0)} + g_{1i}^{(0)}h_{11}^{(0)}h_{m1}^{(0)} + g_{1i}^{(0)}h_{m1}^{(0)} + s_{1m}^{(0)}f_{1i}^{(0)} + s_{1m}^{(0)}h_{11}^{(0)}f_{1i}^{(0)} - s_{11}^{(0)}h_{m1}^{(0)}f_{1i}^{(0)}.
$$

As a consequence of our equal energy and priors assumption (Assumption 2), we can directly use the bounds of the terms in $(\star)_{gk}^{(0)}$ to bound terms in $(\star)_{gm}^{(0)}$. We get

$$
|g_{mi}^{(1)}| \leq \frac{1}{C}\left(1+\frac{C_1}{\sqrt{n}} + \frac{C_8\epsilon_n}{(1-(\frac{C_4\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}))^2\sqrt{k}\sqrt{n}}\right)\frac{1}{k^2p} \leq \frac{1}{C}\cdot\frac{1+\alpha}{k^2p}.
$$

Finally, there exists a sufficiently large constant C_{10} such that $(1+\alpha)/C \leq 1/C_{10}$. This shows the desired bounds.

Completing the proof for k-th order quadratic forms

Notice from the above analysis that the 1-st order quadratic forms exhibit the same order-wise dependence on n, k and p as the 0-th order quadratic forms, e.g. both $s_{mk}^{(0)}$ and $s_{mk}^{(1)}$ are of order $\Theta(\frac{\sqrt{n}}{kp})$. Thus, the higher-order quadratic forms that arise by including more mean components will not change too much[1]. By Equation (3.36), we can see that we can bound the 2-nd order quadratic forms by bounding quadratic forms with order 1. We consider $s_{mk}^{(2)}$ as an example:

$$s_{mk}^{(2)} = s_{mk}^{(1)} - \frac{1}{\det_1}(\star)_s^{(1)},$$

where

$$(\star)_s^{(1)} := (\|\boldsymbol{\mu}\|_2^2 - t_{22}^{(1)})s_{2k}^{(1)}s_{2m}^{(1)} + s_{2m}^{(1)}h_{k2}^{(1)}h_{22}^{(1)} + s_{2k}^{(1)}h_{m2}^{(1)}h_{22}^{(1)} - s_{22}^{(1)}h_{k2}^{(1)}h_{m2}^{(1)} + s_{2m}^{(1)}h_{k2}^{(1)} + s_{2k}^{(1)}h_{m2}^{(1)},$$
$$\det_1 := s_{22}^{(1)}(\|\boldsymbol{\mu}\|_2^2 - t_{22}^{(1)}) + (1 + h_{22}^{(1)})^2.$$

We additionally show how $f_{ki}^{(2)}$ relates to the 1-st order quadratic forms:

$$f_{ki}^{(2)} = f_{ki}^{(1)} - \frac{1}{\det_1}(\star)_f^{(1)},$$

where we define

$$(\star)_f^{(1)} = (\|\boldsymbol{\mu}\|_2^2 - t_{22}^{(1)})h_{2k}^{(1)}g_{2i}^{(1)} + t_{2k}^{(1)}g_{2i}^{(1)} + t_{2k}^{(1)}h_{22}^{(1)}g_{2i}^{(1)} + h_{2k}^{(1)}f_{2i}^{(1)} + h_{2k}^{(1)}h_{22}^{(1)}f_{2i}^{(1)} - s_{22}^{(1)}t_{2k}^{(1)}f_{2i}^{(1)}.$$

Observe that the equations above are very similar to Equations (B.5) and (B.6) (for s),

[1]There are several low-level reasons for this. One critical reason is the aforementioned orthogonality of the label indicator vectors $\{\boldsymbol{v}_c\}_{c\in[k]}$, which ensures by Lemma 27 that the cross-terms $|s_{mk}^{(j)}|$ are always dominated by the larger terms $|s_{kk}^{(j)}|$. Another reason is that $h_{mk}^{(0)}$, which can be seen as the "noise" term in our analysis, is small and thus does not affect other terms.

and Equations (B.8) and (B.9) (for f), except that the quadratic forms are in terms of Gram matrix $\boldsymbol{A}_1$. We have shown that the quadratic forms with order 1 will not be drastically different different from the quadratic forms with order 0. Hence, we repeat the above procedures of bounding these quadratic forms $k-1$ times to obtain the desired bounds in Lemma 7. The only quantity that will change in each iteration is α, which nevertheless remains negligible[2].

Our analysis so far is conditioned on event $\mathcal{E}_q$. We define the *unconditional* event $\mathcal{E}_u := \{\text{all the inequalities in Lemma 7 hold}\}$. Then, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_u^c) &\leq \mathbb{P}(\mathcal{E}_u^c|\mathcal{E}_q) + \mathbb{P}(\mathcal{E}_q^c) \leq \mathbb{P}(\mathcal{E}_u^c|\mathcal{E}_q) + \mathbb{P}(\mathcal{E}_q^c|\mathcal{E}_v) + \mathbb{P}(\mathcal{E}_v^c) \\ &\leq \frac{c_1}{kn} + \frac{c_2}{n} + c_3 k(e^{-\frac{n}{c_4}} + e^{-\frac{n}{c_5 k^2}}) \\ &\leq \frac{c_6}{n} + c_7 k e^{-\frac{n}{c_5 k^2}},\end{aligned}$$

for constants c_i's > 1. This completes the proof.

B.1.3 Proofs of Auxiliary lemmas

We complete this section by proving the auxiliary Lemmas 25, 27 and 28, which were used in the proof of Lemma 7.

Proof of Lemma 25

Our goal is to upper and lower bound $\|\boldsymbol{v}_c\|_2^2$, for $c \in [k]$. Note that every entry of $\boldsymbol{v}_c$ is either 1 or 0, hence these entries are independent sub-Gaussian random variables with sub-Gaussian parameter 1 [163, Chapter 2]. Under the equal-prior Assumption 2, we

[2]To see this, recall that in the first iteration we had $\alpha_1 := \alpha = \frac{C_1}{\sqrt{n}} + \frac{C_2\tau}{(1-(C_5\tau/(k^2\sqrt{n})))^2 k^2\sqrt{n}}$ for the first-order terms. Thus, even if we repeat the procedure $k-1$ times, then we have $\alpha_k \leq Ck\alpha_1$, which remains small since we consider $n \gg k$.

have $\mathbb{E}[\|\boldsymbol{v}_c\|_2^2] = n/k$ when we assume equal priors. Thus, a straightforward application of Hoeffding's concentration inequality on bounded random variables [163, Chapter 2] gives us

$$\mathbb{P}\left(\left|\|\boldsymbol{v}_c\|_2^2 - \frac{n}{k}\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2n}\right).$$

We complete the proof by setting $t = \frac{n}{C_1 k}$ for a large enough constant C_1 and applying the union bound over all $c \in [k]$.

Proof of Lemma 27

We use the following lemma adapted from [120, Lemma 2] to bound quadratic forms of inverse Wishart matrices.

Lemma 29. *Define $p'(n) := (p - n + 1)$, and consider matrix $\boldsymbol{M} \sim$ Wishart$(p, \mathbf{I}_n)$. For any unit Euclidean norm vector $\boldsymbol{v}$ and any $t > 0$, we have*

$$\mathbb{P}\Big(\frac{1}{\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} > p'(n) + \sqrt{2tp'(n)} + 2t\Big) \leq e^{-t} \quad \textit{and} \quad \mathbb{P}\Big(\frac{1}{\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} < p'(n) - \sqrt{2tp'(n)}\Big) \leq e^{-t},$$

provided that $p'(n) > 2\max\{t, 1\}$.

We first upper and lower bound $s_{cc}^{(0)}$ for a fixed $c \in [k]$. Recall that we assume $p > Cn\log(kn) + n - 1$ for sufficiently large constant $C > 1$ and this can be obtained by assuming $p'(n) > Cn\log(kn)$. Let $t = 2\log(kn)$. Working on the event $\mathcal{E}_v$ defined in (B.1), Lemma 29 gives us

$$s_{cc}^{(0)} \leq \frac{\|\boldsymbol{v}_c\|_2^2}{p'(n) - \sqrt{4\log(kn)p'(n)}} \leq \frac{C_1+1}{C_1} \cdot \frac{n/k}{p'(n)\left(1 - \frac{2}{\sqrt{Cn}}\right)} \leq \frac{C_2+1}{C_2} \cdot \frac{n}{kp}$$

with probability at least $1 - \frac{2}{k^2n^2}$. Here, the last inequality comes from the fact that p is

sufficiently large compared to n and C is large enough. Similarly, for the lower bound, we have

$$s_{cc}^{(0)} \geq \frac{\|\boldsymbol{v}_c\|_2^2}{p'(n) + \sqrt{4\log(kn)p'(n)} + 2\log(kn)} \geq \frac{C_1 - 1}{C_1} \cdot \frac{n/k}{p'(n)\left(1 + \frac{4}{\sqrt{Cn}}\right)} \geq \frac{C_2 - 1}{C_2} \cdot \frac{n}{kp}$$

with probability $1 - \frac{2}{k^2n^2}$.

Now we upper and lower bound $s_{cj}^{(0)}$ for a fixed choice $j \neq c \in [k]$. We use the parallelogram law to get

$$\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j = \frac{1}{4}\Big((\boldsymbol{v}_c + \boldsymbol{v}_j)^T \boldsymbol{A}_0^{-1}(\boldsymbol{v}_c + \boldsymbol{v}_j) - (\boldsymbol{v}_c - \boldsymbol{v}_j)^T \boldsymbol{A}_0^{-1}(\boldsymbol{v}_c - \boldsymbol{v}_j)\Big).$$

Because of the orthogonality of the label indicator vectors ($\boldsymbol{v}_c^T \boldsymbol{v}_j = 0$ for any $j \neq c$), we have $\|\boldsymbol{v}_c + \boldsymbol{v}_j\|_2^2 = \|\boldsymbol{v}_c - \boldsymbol{v}_j\|_2^2$, which we denote by $\tilde{n}$ as shorthand. Then, we have

$$\begin{aligned}
\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\leq \frac{1}{4}\left(\frac{\tilde{n}}{p'(n) - \sqrt{4\log(kn)p'(n)}} - \frac{\tilde{n}}{p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn)}\right) \\
&\leq \frac{1}{4} \cdot \frac{2\tilde{n}\sqrt{4\log(kn)p'(n)} + 4\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})} \\
&\leq \frac{C_1 + 1}{2C_1 k} \cdot \frac{2n\sqrt{4\log(kn)p'(n)} + 4n\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})}
\end{aligned}$$

with probability at least $1 - \frac{2}{k^2n^2}$ Here, the last inequality follows because we have $\tilde{n} \leq \frac{2(C_1+1)}{C_1} \cdot \frac{n}{k}$ on $\mathcal{E}_v$. Because $p'(n) > Cn\log(kn)$, we have

$$\begin{aligned}
\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\leq \frac{C_1 + 1}{2C_1 k} \cdot \frac{2\sqrt{n}p'(n) \cdot \sqrt{4/C} + 4/C \cdot p'(n)}{\left(1 - \sqrt{4/(Cn)}\right) p'(n)^2} \\
&\leq \frac{C_1 + 1}{2C_1} \cdot \frac{\sqrt{n}}{k} \cdot \frac{2\sqrt{4/C} + \sqrt{4/(Cn)}}{p'(n)(1 - \sqrt{4/(Cn)})} \\
&\leq \frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp},
\end{aligned}$$

where in the last step we use the fact that $C > 1$ is large enough. To lower bound $s_{cj}^{(0)}$, we get

$$\begin{aligned}
\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\geq \frac{1}{4} \left(\frac{\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} - \frac{\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})} \right) \\
&\geq \frac{1}{4} \cdot \frac{-2\tilde{n}\sqrt{4\log(kn)p'(n)} - 4\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})} \\
&\geq -\frac{C_1 + 1}{2C_1 k} \cdot \frac{2n\sqrt{4\log(kn)p'(n)} + 4n\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})(p'(n) + \sqrt{4\log(kn)p'(n)})}
\end{aligned}$$

with probability at least $1 - \frac{2}{k^2 n^2}$. Then following similar steps to the upper bound of $\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j$ gives us

$$\begin{aligned}
\boldsymbol{v}_c^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\geq -\frac{C_1 + 1}{2C_1 k} \cdot \frac{2\sqrt{n}p'(n)\sqrt{4/C} + (4/C)p'(n)}{(p'(n) - \sqrt{4/(Cn)}p'(n))p'(n)} \\
&\geq -\frac{C_1 + 1}{2C_1} \cdot \frac{\sqrt{n}}{k} \cdot \frac{2\sqrt{4/C} + (4/C\sqrt{n})}{p'(n)(1 - \sqrt{4/(Cn)})} \\
&\geq -\frac{C_2 + 1}{C_2} \cdot \frac{\sqrt{n}}{kp}.
\end{aligned}$$

We finally apply the union bound on all pairs of $c, j \in [k]$ and complete the proof.

Proof of Lemma 28

We first lower and upper bound $g_{(y_i)i}^{(0)}$. Recall that we assumed $y_i = k$ without loss of generality. With a little abuse of notation, we define $\|\boldsymbol{v}_k\|_2^2 = \tilde{n}$ and $\boldsymbol{u} := \sqrt{\tilde{n}}\mathbf{e}_i$. We use the parallelogram law to get

$$\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_k = \frac{1}{4\sqrt{\tilde{n}}} \left((\boldsymbol{u} + \boldsymbol{v}_k)^T \boldsymbol{A}_0^{-1} (\boldsymbol{u} + \boldsymbol{v}_k) - (\boldsymbol{u} - \boldsymbol{v}_k)^T \boldsymbol{A}_0^{-1} (\boldsymbol{u} - \boldsymbol{v}_k) \right).$$

Note that $\|\boldsymbol{u}+\boldsymbol{v}_k\|_2^2 = 2(\tilde{n}+\sqrt{\tilde{n}})$ and $\|\boldsymbol{u}-\boldsymbol{v}_k\|_2^2 = 2(\tilde{n}-\sqrt{\tilde{n}})$. As before, we apply Lemma 29 with $t = 2\log(kn)$ to get with probability at least $1-\frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1}\boldsymbol{v}_k &\geq \frac{1}{4\sqrt{\tilde{n}}}\left(\frac{2(\tilde{n}+\sqrt{\tilde{n}})}{(p'(n)+\sqrt{4\log(kn)p'(n)}+4\log(kn))} - \frac{2(\tilde{n}-\sqrt{\tilde{n}})}{(p'(n)-\sqrt{4\log(kn)p'(n)})}\right)\\
&\geq \frac{1}{4\sqrt{\tilde{n}}}\cdot\frac{4\sqrt{\tilde{n}}p'(n)-4\tilde{n}\sqrt{4\log(kn)p'(n)}-8\tilde{n}\log(kn)}{(p'(n)+\sqrt{4\log(kn)p'(n)}+4\log(kn))p'(n)}\\
&\geq \frac{p'(n)-\sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)}-2\sqrt{\tilde{n}}\log(kn)}{(p'(n)+\sqrt{4\log(kn)p'(n)}+4\log(kn))p'(n)},\\
&\geq \frac{p'(n)-\sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)}-2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n)+\sqrt{4\log(kn)p'(n)}+4\log(kn))p'(n)}.
\end{aligned}$$

The last inequality works on event $\mathcal{E}_v$, by which we have $\tilde{n} \leq \frac{2(C_1+1)n}{C_1k}$. Then, $p'(n) > Ck^3n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1}\boldsymbol{v}_k &\geq \frac{p'(n)-\sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n)-\sqrt{(1+1/C_1)n/k}(2/Ck^3n)p'(n)}{(p'(n)+\sqrt{4\log(kn)p'(n)}+4\log(kn))p'(n)}\\
&\geq \frac{1-(1/(C_2\sqrt{k^4}))-(1/(C_3k^{3.5}\sqrt{n}))}{p'(n)(1+2\sqrt{4/(Ck^3n)})}\\
&\geq \frac{C_4-1}{C_4}\cdot\frac{1}{p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. To upper bound $g^{(0)}_{(y_i)i}$, we have with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_k &\le \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2(\tilde{n} + \sqrt{\tilde{n}})}{(p'(n) - \sqrt{4\log(kn)p'(n)})} - \frac{2(\tilde{n} - \sqrt{\tilde{n}})}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} \right) \\
&\le \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{4\sqrt{\tilde{n}}p'(n) + 4\tilde{n}\sqrt{4\log(kn)p'(n)} + 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{p'(n) + \sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\le \frac{p'(n) + \sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

Then $p'(n) > Ck^3 n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_k &\le \frac{p'(n) + \sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) + 2\sqrt{(1+1/C_1)n/k}(4/Ck^3n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\le \frac{1 + (1/(C_2\sqrt{k^4})) + (1/(C_3k^{3.5}\sqrt{n}))}{p'(n)(1 - 2\sqrt{4/(Ck^3n)})} \\
&\le \frac{C_4 + 1}{C_4} \cdot \frac{1}{p}.
\end{aligned}$$

We now upper and lower bound $g^{(0)}_{ji}$ for a fixed $j \neq y_i$. As before, we have

$$\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j = \frac{1}{4\sqrt{\tilde{n}}} \left((\boldsymbol{u} + \boldsymbol{v}_j)^T \boldsymbol{A}_0^{-1} (\boldsymbol{u} + \boldsymbol{v}_j) - (\boldsymbol{u} - \boldsymbol{v}_j)^T \boldsymbol{A}_0^{-1} (\boldsymbol{u} - \boldsymbol{v}_j) \right).$$

Since $\mathbf{e}_i^T \boldsymbol{v}_j = 0$, we now have $\|\boldsymbol{u} + \boldsymbol{v}_j\|_2^2 = \|\boldsymbol{u} - \boldsymbol{v}_j\|_2^2 = 2\tilde{n}$. We apply Lemma 29 with $t = 2\log(kn)$ to get, with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\leq \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})} - \frac{2\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n) + 4\log(kn)})} \right) \\
&\leq \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{4\tilde{n}\sqrt{4\log(kn)p'(n)} + 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\leq \frac{\sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\leq \frac{\sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

The last inequality works on event $\mathcal{E}_v$, by which we have $\tilde{n} \leq \frac{2(C_1+1)n}{C_1 k}$. Then, $p'(n) > Ck^3 n\log(kn)$ gives us

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\leq \frac{\sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) + \sqrt{(1+1/C_1)n/k}(2/Ck^3n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\leq \frac{(1/(C_2\sqrt{k^4})) + (1/(C_3 k^{3.5}\sqrt{n}))}{p'(n)(1 - \sqrt{4/(Ck^3n)})} \\
&\leq \frac{C_4 + 1}{C_4} \cdot \frac{1}{k^2 p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. To lower bound $g_{ij}^{(0)}$, we have with probability at least $1 - \frac{2}{k^2n^2}$,

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\geq \frac{1}{4\sqrt{\tilde{n}}} \left(\frac{2\tilde{n}}{(p'(n) + \sqrt{4\log(kn)p'(n)} + 4\log(kn))} - \frac{2\tilde{n}}{(p'(n) - \sqrt{4\log(kn)p'(n)})} \right) \\
&\geq \frac{1}{4\sqrt{\tilde{n}}} \cdot \frac{-4\tilde{n}\sqrt{4\log(kn)p'(n)} - 8\tilde{n}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\geq -\frac{\sqrt{\tilde{n}}\sqrt{4\log(kn)p'(n)} + 2\sqrt{\tilde{n}}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}, \\
&\geq -\frac{\sqrt{(1+1/C_1)n/k}\sqrt{4\log(kn)p'(n)} + 2\sqrt{(1+1/C_1)n/k}\log(kn)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)}.
\end{aligned}$$

Because $p'(n) > Ck^3 n\log(kn)$, we get

$$\begin{aligned}
\mathbf{e}_i^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_j &\geq -\frac{\sqrt{(1+1/C_1)n/k}\sqrt{4/(Ck^3n)}p'(n) + \sqrt{(1+1/C_1)n/k}(2/Ck^3n)p'(n)}{(p'(n) - \sqrt{4\log(kn)p'(n)})p'(n)} \\
&\geq -\frac{(1/(C_2\sqrt{k^4})) + (1/(C_3k^{3.5}\sqrt{n}))}{p'(n)(1 - \sqrt{4/(Ck^3n)})} \\
&\geq -\frac{C_4+1}{C_4} \cdot \frac{1}{k^2p},
\end{aligned}$$

where in the last step we use the fact that $C, C_2, C_3 > 1$ are large enough. We complete the proof by applying a union bounds over all k classes and n training examples.

B.2 Proof of Theorem 10

In this section, we provide the proof of Theorem 10, which was discussed in Section 3.3.2. After having derived the interpolation condition in Equation (3.12) for multiclass SVM, the proofs is in fact a rather simple extension of the arguments provided in [120, 72] to the multiclass case. This is unlike the GMM case that we considered in Section 3.6.2, which required substantial additional effort over and above the binary

case [164].

For this section, we define $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ as shorthand (we denoted the same quantity as $\boldsymbol{A}_k$ in Section 3.6.2). Recall that the eigendecomposition of the covariance matrix is given by $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$. By rotation invariance of the standard normal variable, we can write $\boldsymbol{A} = \boldsymbol{Q}^T\boldsymbol{\Lambda}\boldsymbol{Q}$, where the entries of $\boldsymbol{Q} \in \mathbb{R}^{p\times n}$ are IID $\mathcal{N}(0,1)$ random variables. Finally, recall that we denoted $\boldsymbol{\lambda} = \begin{bmatrix}\lambda_1 & \cdots & \lambda_p\end{bmatrix}$ and defined the effective dimensions $d_2 = \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and $d_\infty = \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$. Observe that Equation (3.12) in Theorem 8 is equivalent to the condition

$$z_{ci}\mathbf{e}_i^T\boldsymbol{A}^{-1}\boldsymbol{z}_c > 0, \quad \text{for all} \quad c \in [k] \quad \text{and} \quad i \in [n]. \tag{B.12}$$

We fix $c \in [k]$ and drop the subscript c, using $\overline{\boldsymbol{z}}$ to denote the vector $\boldsymbol{z}_c$. We first provide a deterministic equivalence to Equation (3.12) that resembles the condition provided in [72, Lemma 1]. Our proof is slightly modified compared to [72, Lemma 1] and relies on elementary use of block matrix inversion identity.

Lemma 30. *Let $\boldsymbol{Q} \in \mathbb{R}^{p\times n} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n]$. In our notation, Equation* (3.12) *holds for a fixed c if and only if:*

$$\frac{1}{z_i}\overline{\boldsymbol{z}}_{\setminus i}^T\Big(\boldsymbol{Q}_{\setminus i}^T\boldsymbol{\Lambda}\boldsymbol{Q}_{\setminus i}\Big)^{-1}\boldsymbol{Q}_{\setminus i}^T\boldsymbol{\Lambda}\boldsymbol{q}_i < 1, \quad \textit{for all} \quad i = 1, \cdots, n. \tag{B.13}$$

Above, $\overline{\boldsymbol{z}}_{\setminus i} \in \mathbb{R}^{(n-1)\times 1}$ is obtained by removing the i-th entry from vector $\overline{\boldsymbol{z}}$ and $\boldsymbol{Q}_{\setminus i} \in \mathbb{R}^{d\times(n-1)}$ is obtained by removing the i-th column from $\boldsymbol{Q}$.

Proof. By symmetry, it suffices to consider the case $i = 1$. We first write

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{q}_1^T\boldsymbol{\Lambda}\boldsymbol{q}_1 & \boldsymbol{q}_1^T\boldsymbol{\Lambda}\boldsymbol{Q}_{\setminus 1}\\ \boldsymbol{Q}_{\setminus 1}^T\boldsymbol{\Lambda}\boldsymbol{q}_1 & \boldsymbol{Q}_{\setminus 1}^T\boldsymbol{\Lambda}\boldsymbol{Q}_{\setminus 1}\end{bmatrix} \triangleq \begin{bmatrix}\alpha & \mathbf{b}^T\\ \mathbf{b} & \mathbf{D}\end{bmatrix}.$$

By Schur complement [17], we have

$$\boldsymbol{A} \succ \mathbf{0} \textit{ iff} \text{ either } \left\{\alpha > 0 \text{ and } \mathbf{D} - \frac{\mathbf{b}\mathbf{b}^T}{\alpha} \succ \mathbf{0}\right\} \text{ or } \left\{\mathbf{D} \succ \mathbf{0} \text{ and } \alpha - \mathbf{b}^T\mathbf{D}^{-1}\mathbf{b} > 0\right\}.$$

Since the entries of $\boldsymbol{Q}$ are drawn from a continuous distribution (IID standard Gaussian), both $\boldsymbol{A}$ and $\boldsymbol{D} = \boldsymbol{Q}_{\setminus 1}^T \boldsymbol{\Lambda} \boldsymbol{Q}_{\setminus 1}$ are positive definite almost surely. Therefore, $\alpha - \mathbf{b}^T\mathbf{D}^{-1}\mathbf{b} > 0$ almost surely.

Thus, by block matrix inversion identity [17], we have

$$\boldsymbol{A}^{-1} = \begin{bmatrix} (\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1} & -(\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1}\boldsymbol{b}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{b}(\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{b}(\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1}\boldsymbol{b}^T\boldsymbol{D}^{-1} \end{bmatrix}.$$

Therefore,$\mathbf{e}_1^T\boldsymbol{A}^{-1} = (\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1}\begin{bmatrix} 1 & -\boldsymbol{b}^T\boldsymbol{D}^{-1} \end{bmatrix}$. Hence we have

$$z_1\mathbf{e}_1^T\boldsymbol{A}^{-1}\overline{\boldsymbol{z}} = (\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1}(z_1^2 - \boldsymbol{b}^T\boldsymbol{D}^{-1}(z_1\overline{\boldsymbol{z}}_{\setminus 1})),$$

where we use the fact that $\overline{\boldsymbol{z}}_1 = z_1$. Since $\alpha - \mathbf{b}^T\mathbf{D}^{-1}\mathbf{b} > 0$ almost surely, we have

$$\begin{aligned} z_1\mathbf{e}_1^T\boldsymbol{A}^{-1}\overline{\boldsymbol{z}} > 0 &\iff (\alpha - \boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b})^{-1}(z_1^2 - \boldsymbol{b}^T\boldsymbol{D}^{-1}(z_1\overline{\boldsymbol{z}}_{\setminus 1})) > 0 \\ &\iff \frac{1}{z_1}\boldsymbol{b}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{z}}_{\setminus 1} < 1. \end{aligned}$$

Recall that $\boldsymbol{b}^T = \boldsymbol{q}_1^T\boldsymbol{\Lambda}\boldsymbol{Q}_{\setminus 1}$ and $\boldsymbol{D} = \boldsymbol{Q}_{\setminus 1}^T\boldsymbol{\Lambda}\boldsymbol{Q}_{\setminus 1}$. This completes the proof. □

Next, we define the following events:

1. For $i \in [n]$, $\mathcal{B}_i := \left\{\frac{1}{z_i}\overline{\boldsymbol{z}}_{\setminus i}^T\boldsymbol{A}_{\setminus i}^{-1}\boldsymbol{Q}_{\setminus i}^T\boldsymbol{\Lambda}\boldsymbol{q}_i \geq 1\right\}$.
2. For $i \in [n]$, given $t > 0$, $\mathcal{E}_i(t) := \left\{\|(\overline{\boldsymbol{z}}_{\setminus i}^T\boldsymbol{A}_{\setminus i}^{-1}\boldsymbol{Q}_{\setminus i}^T\boldsymbol{\Lambda})^T\|_2^2 \geq \frac{1}{t}\right\}$.

3. $\mathcal{B} := \cup_{i=1}^{n} \mathcal{B}_i$.

We know all the data points are support vectors i.e., Equation (B.12) holds, if none of the events $\mathcal{B}_i$ happens; hence, $\mathcal{B}$ is the undesired event. We want to upper bound the probability of event $\mathcal{B}$. As in the argument provided in [72], we have

$$\mathbb{P}(\mathcal{B}) \leq \sum_{i=1}^{n} \Big(\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) + \mathbb{P}(\mathcal{E}_i(t)) \Big). \tag{B.14}$$

The lemma below gives an upper bound on $\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c)$.

Lemma 31. *For any $t > 0$, $\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2ck^2}\right)$.*

Proof. On the event $\mathcal{E}_i(t)^c$, we have $\|(\overline{\boldsymbol{z}}_{\setminus i}^T \boldsymbol{A}_{\setminus i}^{-1} \boldsymbol{Q}_{\setminus i}^T \boldsymbol{\Lambda})^T\|_2^2 \leq \frac{1}{t}$. Since, by its definition, $|\frac{1}{z_i}| \leq k$, we have $\frac{1}{z_i} \overline{\boldsymbol{z}}_{\setminus i}^T \boldsymbol{A}_{\setminus i}^{-1} \boldsymbol{Q}_{\setminus i}^T \boldsymbol{\Lambda} \boldsymbol{q}_i$ is conditionally sub-Gaussian [163, Chapter 2] with parameter at most $ck^2 \|(\overline{\boldsymbol{z}}_{\setminus i}^T \boldsymbol{A}_{\setminus i}^{-1} \boldsymbol{Q}_{\setminus i}^T \boldsymbol{\Lambda})^T\|_2^2 \leq ck^2/t$. Then the sub-Gaussian tail bound gives

$$\mathbb{P}(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2ck^2}\right), \tag{B.15}$$

which completes the prof. □

Next we upper bound $\mathbb{P}(\mathcal{E}_i(t))$ with $t = d_\infty/(2n)$. Since $\|\boldsymbol{z}_{\setminus i}\|_2 \leq \|\boldsymbol{y}_{\setminus i}\|_2$, we can directly use [72, Lemma 4].

Lemma 32 (Lemma 4, [72])**.** $\mathbb{P}\left(\mathcal{E}_i\left(\frac{d_\infty}{2n}\right)\right) \leq 2 \cdot 9^{n-1} \cdot \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\}\right)$.

The results above are proved for fixed choices of $i \in [n]$ and $c \in [k]$. We combine Lemmas 31 and 32 with a union bound over all n training examples and k classes to

upper bound the probability of the undesirable event $\mathcal{B}$ over all k classes by:

$$kn9^{n-1} \cdot \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\}\right) \leq \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right)$$
$$\text{and } 2kn \cdot \exp\left(-\frac{d_\infty}{2ck^2 n}\right) \leq \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right).$$

Thus, the probability that every data point is a support vector is at least

$$1 - \exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right) - \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right).$$

To ensure that $\exp\left(-c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} + C_1 \log(kn) + C_2 n\right) + \exp\left(-\frac{c_2 d_\infty}{ck^2 n} + C_3 \log(kn)\right) \leq \frac{c_4}{n}$, we consider the conditions $c_1 \min\left\{\frac{d_2}{4c^2}, \frac{d_\infty}{c}\right\} - C_1 \log(kn) - C_2 n \geq \log(n)$ and $\frac{c_2 d_\infty}{ck^2 n} - C_3 \log(kn) \geq \log(n)$ to be satisfied. These are equivalent to the conditions provided in Equation (3.17). This completes the proof. Note that throughout the proof, we did not use any generative model assumptions on the labels given the covariates, so in fact our proof applies to scenarios beyond the MLM. □

B.3 Classification error proofs

In this section, we provide the proofs of classification error of the MNI under both GMM and MLM models. We begin with the proof for the GMM case (Theorem 11).

B.3.1 Proof of Theorem 11

Proof strategy and notations

The notation and main arguments of this proof follow closely the content of Section 3.6.2.

Our starting point here is the lemma below (adapted from [155, D.10]) that provides a simpler upper bound on the class-wise error $\mathbb{P}_{e|c}$.

Lemma 33. *Under GMM,* $\mathbb{P}_{e|c} \leq \sum_{j\neq c} Q\left(\frac{(\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c}{\|\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j\|_2}\right)$. *In particular, if* $(\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c > 0$, *then* $\mathbb{P}_{e|c} \leq \sum_{j\neq c} \exp\left(-\frac{((\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c)^2}{4(\widehat{\boldsymbol{w}}_c^T\widehat{\boldsymbol{w}}_c+\widehat{\boldsymbol{w}}_j^T\widehat{\boldsymbol{w}}_j)}\right)$.

Proof. [155, D.10] shows $\mathbb{P}_{e|c}$ is upper bounded by $\sum_{j\neq c} Q\left(\frac{(\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c}{\|\widehat{\boldsymbol{w}}_c-\widehat{\boldsymbol{w}}_j\|_2}\right)$. Then if $(\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c > 0$, the Chernoff bound [163, Ch. 2] gives

$$\mathbb{P}_{e|c} \leq \sum_{j\neq c} \exp\left(-\frac{((\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c)^2}{2\|\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j\|_2^2}\right) \leq \sum_{j\neq c} \exp\left(-\frac{((\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c)^2}{4(\widehat{\boldsymbol{w}}_c^T\widehat{\boldsymbol{w}}_c + \widehat{\boldsymbol{w}}_j^T\widehat{\boldsymbol{w}}_j)}\right),$$

where the last inequality uses the identity $\mathbf{a}^T\mathbf{b} \leq 2(\mathbf{a}^T\mathbf{a} + \mathbf{b}^T\mathbf{b})$. □

Thanks to Lemma 33, we can upper bound $P_{e|c}$ by lower bounding the terms

$$\frac{((\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T\boldsymbol{\mu}_c)^2}{(\widehat{\boldsymbol{w}}_c^T\widehat{\boldsymbol{w}}_c + \widehat{\boldsymbol{w}}_j^T\widehat{\boldsymbol{w}}_j)}, \quad \text{for all} \quad c \neq j \in [k]. \tag{B.16}$$

Our key observation is that this can be accomplished without the need to control the more intricate cross-correlation terms $\widehat{\boldsymbol{w}}_c^T\widehat{\boldsymbol{w}}_j$ for $c \neq j \in [k]$.

Without loss of generality, we assume onwards that $c = k$ and $j = k - 1$ (as in Section 3.6.2). Similar to Section 3.6.2, the quadratic forms introduced in Equation (3.37) play key role here, as well. For convenience, we recall the definitions of the *c-th order*

quadratic forms for $c, j, m \in [k]$ and $i \in [n]$:

$$
\begin{aligned}
s_{mj}^{(c)} &:= \boldsymbol{v}_m^T \boldsymbol{A}_c^{-1} \boldsymbol{v}_j, \\
t_{mj}^{(c)} &:= \mathbf{d}_m^T \boldsymbol{A}_c^{-1} \mathbf{d}_j, \\
h_{mj}^{(c)} &:= \mathbf{v}_m^T \boldsymbol{A}_c^{-1} \mathbf{d}_j, \\
g_{ji}^{(c)} &:= \mathbf{v}_j^T \boldsymbol{A}_c^{-1} \mathbf{e}_i, \\
f_{ji}^{(c)} &:= \mathbf{d}_j^T \boldsymbol{A}_c^{-1} \mathbf{e}_i.
\end{aligned}
$$

Further, recall that $\widehat{\boldsymbol{w}}_c = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c$ and $\boldsymbol{X} = \sum_{j=1}^k \boldsymbol{\mu}_j \boldsymbol{v}_j^T + \boldsymbol{Q}$. Also, from orthogonality of the class mean vectors (Assumption 3), we have $\boldsymbol{\mu}_c^T \boldsymbol{X} = \|\boldsymbol{\mu}\|_2^2 \boldsymbol{v}_c^T + \boldsymbol{d}_c^T$. Thus,

$$
\begin{aligned}
&\widehat{\boldsymbol{w}}_c^T \boldsymbol{\mu}_c - \widehat{\boldsymbol{w}}_j^T \boldsymbol{\mu}_c \\
=&\|\boldsymbol{\mu}\|_2^2 \boldsymbol{v}_c^T (\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c + \boldsymbol{v}_c^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c - \|\boldsymbol{\mu}\|_2^2 \boldsymbol{v}_c^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j - \boldsymbol{v}_j^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c. \quad \text{(B.17)}
\end{aligned}
$$

Additionally,

$$
\widehat{\boldsymbol{w}}_c^T \widehat{\boldsymbol{w}}_c = \boldsymbol{v}_c^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c, \quad \text{and} \quad \widehat{\boldsymbol{w}}_j^T \widehat{\boldsymbol{w}}_j = \boldsymbol{v}_j^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j.
$$

Using the leave-one-out trick in Section 3.6.2 and the matrix-inversion lemma, we show in Appendix B.3.1 that

$$
\text{(B.16)} = \frac{A_2}{B_2}, \tag{B.18}
$$

where

$$A_2 = \left(\frac{\|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)} t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)} - \|\boldsymbol{\mu}\|_2^2 s_{jc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)} h_{cc}^{(j)} + s_{jc}^{(j)} t_{cc}^{(j)}}{\det_j} \right)^2$$

$$B_2 = \left(\frac{s_{cc}^{(j)}}{\det_j} + \frac{s_{jj}^{(-j)}}{\det_{-j}} \right)$$

$$\det_j = (\|\boldsymbol{\mu}\|_2^2 - t_{cc}^{(j)}) s_{cc}^{(j)} + (h_{cc}^{(j)} + 1)^2.$$

Note that $\det_j = \det_{-c}$ when $c = k$ and $j = k - 1$.

Next, we will prove that

$$\text{(B.18)} \geq \|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p} \right) \|\boldsymbol{\mu}\|_2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\} \right)^2}{C_6 \left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n} \right)}. \tag{B.19}$$

Proof of Equation (B.19)

We will lower bound the numerator and upper bound the denominator of Equation (B.18). We will work on the high-probability event $\mathcal{E}_v$ defined in Equation (B.1) in Appendix B.1.1. For quadratic forms such as $s_{cc}^{(j)}, t_{cc}^{(j)}$ and $h_{cc}^{(j)}$, the Gram matrix $\boldsymbol{A}_j^{-1}$ does not "include" the c-th mean component because we have fixed $c = k, j = k - 1$. Thus, we can directly apply Lemma 7 to get

$$\begin{aligned}
\frac{C_1 - 1}{C_1} \cdot \frac{n}{kp} \leq & s_{cc}^{(j)} \leq \frac{C_1 + 1}{C_1} \cdot \frac{n}{kp}, \\
t_{cc}^{(j)} \leq & \frac{C_2 n \|\boldsymbol{\mu}\|_2^2}{p}, \\
-\tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{kp}} \leq & h_{cc}^{(j)} \leq \tilde{\rho}_{n,k} \frac{C_3 n \|\boldsymbol{\mu}\|_2}{\sqrt{kp}},
\end{aligned}$$

on the event $\mathcal{E}_v$. We need some additional work to bound $s_{jc}^{(j)} = \boldsymbol{v}_j \boldsymbol{A}_j^{-1} \boldsymbol{v}_c$ and $h_{jc}^{(j)} = \boldsymbol{v}_j \boldsymbol{A}_j^{-1} \boldsymbol{d}_c$, since the Gram matrix $\boldsymbol{A}_j^{-1}$ "includes" $\boldsymbol{v}_j$. The proof here follows the machin-

ery introduced in Appendix B.1.2 for proving Lemma 7. We provide the core argument and refer the reader therein for additional justifications. By Equation (B.28) in Appendix B.4.1 (with the index $j-1$ replacing the index 0), we first have

$$s_{jc}^{(j)} = s_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(\star)_s^{(j-1)},$$

where we define

$$(\star)_s^{(j-1)} = (\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)}s_{jc}^{(j-1)} + s_{jc}^{(j-1)}{h_{jj}^{(j-1)}}^2 + s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)},$$

and $\det_{j-1} = (\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + (h_{jj}^{(j-1)}+1)^2$. Further, we have

$$\begin{aligned}|s_{jc}^{(j)}| &= \left|\left(1 - \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + {h_{jj}^{(j-1)}}^2}{\det_{j-1}}\right)s_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)})\right| \\ &\leq \frac{1}{C}|s_{jc}^{(j-1)}| + \frac{1}{\det_{j-1}}|(s_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}h_{jc}^{(j-1)})|.\end{aligned}$$

We focus on the dominant term $|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|$. Using a similar argument to that provided in Appendix B.1.2, we get

$$\begin{aligned}\frac{|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|}{\det_{j-1}} \leq \frac{|s_{jj}^{(j-1)}h_{jc}^{(j-1)}|}{(1+h_{jj}^{(j-1)})^2} &\leq \frac{C_1}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n}{kp} \cdot \frac{\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}} \\ &\leq \frac{C_3\tilde{\rho}_{n,k}\epsilon_n}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2 k^2} \cdot \frac{\sqrt{n}}{kp}.\end{aligned}$$

Thus, we have

$$|s_{jc}^{(j-1)}| \leq \frac{C_4+1}{C_4} \cdot \frac{\sqrt{n}}{kp}.$$

Similarly, we bound the remaining term $h_{jc}^{(j)}$. Specifically, by Equation (B.30) in Section B.4.1, we have

$$h_{jc}^{(j)} = h_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(\star)_h^{(j-1)},$$

where we define

$$(\star)_h^{(j-1)} = (\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)}h_{jc}^{(j-1)} + h_{jc}^{(j-1)}h_{jj}^{(j-1)^2} + h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)}.$$

Furthermore,

$$\begin{aligned}|h_{jc}^{(j)}| &= \left|\left(1 - \frac{(\|\boldsymbol{\mu}\|_2^2 - t_{jj}^{(j-1)})s_{jj}^{(j-1)} + h_{jj}^{(j-1)^2}}{\det_{j-1}}\right)h_{jc}^{(j-1)} - \frac{1}{\det_{j-1}}(h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)})\right| \\ &\leq \frac{1}{C}|h_{jc}^{(j-1)}| + \frac{1}{\det_{j-1}}|(h_{jc}^{(j-1)}h_{jj}^{(j-1)} + s_{jj}^{(j-1)}t_{jc}^{(j-1)})|.\end{aligned}$$

We again consider the dominant term $|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|/\det_{j-1}$ and get

$$\begin{aligned}\frac{|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|}{\det_{j-1}} &\leq \frac{|s_{jj}^{(j-1)}t_{jc}^{(j-1)}|}{(1+h_{jj}^{(j-1)})^2} \leq \frac{C_1}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^2\sqrt{n}}\right)^2} \cdot \frac{n}{kp} \cdot \frac{n\|\boldsymbol{\mu}\|_2^2}{p} \\ &\leq \frac{C_3\epsilon_n}{\left(1 - \frac{C_2\tilde{\rho}_{n,k}\epsilon_n}{k^{1.5}\sqrt{n}}\right)^2 k^2\sqrt{n}} \cdot \frac{\tilde{\rho}_{n,k}n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.\end{aligned}$$

Thus, we find that

$$|h_{jc}^{(j-1)}| \leq \tilde{\rho}_{n,k}\frac{C_4 n\|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.$$

We are now ready to lower bound the RHS in Equation (B.18) by lower bounding its numerator and upper bounding its denominator.

First, for the numerator we have the following sequence of inequalities:

$$\begin{aligned}
&\|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)} t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)} - \|\boldsymbol{\mu}\|_2^2 s_{jc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)} h_{cc}^{(j)} + s_{jc}^{(j)} t_{cc}^{(j)} \\
\geq &\|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)} - \|\boldsymbol{\mu}\|_2^2 s_{jc}^{(j)} - s_{cc}^{(j)} t_{cc}^{(j)} + s_{jc}^{(j)} t_{cc}^{(j)} + h_{cc}^{(j)} - h_{jc}^{(j)} - h_{jc}^{(j)} h_{cc}^{(j)} \\
\geq &\frac{C_1 - 1}{C_1} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_2 + 1}{C_2} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 \sqrt{n}}{kp} - \frac{C_3 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_4 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 \sqrt{n}}{kp} - \frac{C_5 \tilde{\rho}_{n,k} n \|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.
\end{aligned}$$

Above, we use the fact that the terms $|h_{cc}^{(j)}|, |h_{jc}^{(j)}| \leq C\epsilon/(k^2\sqrt{n})$ are sufficiently small compared to 1. Consequently, the numerator is lower bounded by

$$\frac{D_3^2}{\det_j^2}, \tag{B.20}$$

where D_3 is

$$\frac{C_1 - 1}{C_1} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_2 + 1}{C_2} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 \sqrt{n}}{kp} - \frac{C_3 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 n}{kp} - \frac{C_4 n}{p} \cdot \frac{\|\boldsymbol{\mu}\|_2^2 \sqrt{n}}{kp} - \frac{C_5 \tilde{\rho}_{n,k} n \|\boldsymbol{\mu}\|_2}{\sqrt{k}p}.$$

Second, we upper bound the denominator. For this, note that under the assumption of equal energy and equal priors on class means (Assumption 2), there exist constants $C_1, C_2 > 0$ such that $C_1 \leq \det_j / \det_{-j} \leq C_2$. (In fact, a very similar statement was proved in Equation (3.42) and used in the proof of Theorem 9). Moreover, Lemma 7 shows that the terms $s_{cc}^{(j)}$ and $s_{jj}^{(-j)}$ are of the same order, so it suffices to upper bound $\frac{s_{cc}^{(j)}}{\det_j}$. Again applying Lemma 7, we have

$$\frac{s_{cc}^{(j)}}{\det_j} \leq \frac{C_6}{\det_j} \cdot \frac{n}{kp} \tag{B.21}$$

on the event $\mathcal{E}_v$. Then, combining Equations (B.20) and (B.21) gives us

$$\begin{aligned}(\text{B.18}) &\geq \frac{n}{C_0 kp} \cdot \frac{1}{\det_j} \left((1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}) \|\boldsymbol{\mu}\|_2^2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\} \|\boldsymbol{\mu}\|_2 \right)^2 \\ &\geq N_4 \\ &\geq \|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right) \|\boldsymbol{\mu}\|_2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\} \right)^2}{C_6 \left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n} \right)}, \end{aligned} \tag{B.22}$$

where

$$\begin{aligned} N_4 &= \frac{n}{C_0 kp} \cdot \frac{1}{\frac{C_4 \|\boldsymbol{\mu}\|_2^2 n}{kp} + 2 + \frac{C_5 n^2 \|\boldsymbol{\mu}\|_2^2}{kp^2}} \left(N_5\right)^2, \\ N_5 &= \left(1 - \frac{C_1}{\sqrt{n}} - \frac{C_2 n}{p}\right) \|\boldsymbol{\mu}\|_2^2 - C_3 \min\{\sqrt{k}, \sqrt{\log(2n)}\} \|\boldsymbol{\mu}\|_2, \end{aligned}$$

and the second inequality follows from the following upper bound on $\det_j$ on the event $\mathcal{E}_v$:

$$\begin{aligned} \det_j =& (\|\boldsymbol{\mu}\|_2^2 - t_{cc}^{(j)}) s_{cc}^{(j)} + (h_{cc}^{(j)} + 1)^2 \leq \|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)} + 2({h_{cc}^{(j)}}^2 + 1) \\ \leq& \frac{C_4 \|\boldsymbol{\mu}\|_2^2 n}{kp} + 2 + \frac{C_5 n^2 \|\boldsymbol{\mu}\|_2^2}{kp^2}. \end{aligned}$$

Completing the proof

Because of our assumption of equal energy on class means and equal priors, the analysis above can be applied to bound $\frac{((\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T \boldsymbol{\mu}_c)^2}{(\widehat{\boldsymbol{w}}_c^T \widehat{\boldsymbol{w}}_c + \widehat{\boldsymbol{w}}_j^T \widehat{\boldsymbol{w}}_j)}$, for every $j \neq c$ and $c \in [k]$. We define the *unconditional* event

$$\mathcal{E}_{u2} := \left\{ \frac{((\widehat{\boldsymbol{w}}_c - \widehat{\boldsymbol{w}}_j)^T \boldsymbol{\mu}_c)^2}{(\widehat{\boldsymbol{w}}_c^T \widehat{\boldsymbol{w}}_c + \widehat{\boldsymbol{w}}_j^T \widehat{\boldsymbol{w}}_j)} \text{ is lower bounded by (B.22) for every } j \neq c \right\}.$$

We have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{u2}^c) &\leq \mathbb{P}(\mathcal{E}_{u2}^c|\mathcal{E}_v) + \mathbb{P}(\mathcal{E}_v^c) \\ &\leq \frac{c_4}{n} + c_5 k(e^{-\frac{n}{c_6}} + e^{-\frac{n}{c_7 k^2}}) \leq \frac{c_4}{n} + c_8 k e^{-\frac{n}{c_7 k^2}}\end{aligned}$$

for constants c_i's > 1. Thus, the class-wise error $\mathbb{P}_{e|c}$ is upper bounded by

$$(k-1)\exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1-\frac{C_1}{\sqrt{n}}-\frac{C_2 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_3\min\{\sqrt{k},\sqrt{\log(2n)}\}\right)^2}{C_4\left(\|\boldsymbol{\mu}\|_2^2+\frac{kp}{n}\right)}\right)$$

with probability at least $1-\frac{c_4}{n}-c_8 k e^{-\frac{n}{c_7 k^2}}$. This completes the proof. □

Proof of Equation (B.18)

Here, using the results of Section B.4.1, we show how to obtain Equation (B.18) from Equation (B.16). First, by [164, Appendix C.2] (with $\boldsymbol{y}$ replaced by $\boldsymbol{v}_m$), we have

$$\boldsymbol{v}_m(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_m = \frac{s_{mm}^{(-m)}}{\det_{-m}}, \quad \text{for all} \quad m \in [k],$$

where $\det_{-m} = (\|\boldsymbol{\mu}\|_2^2 - t_{mm}^{(-m)})s_{mm}^{(-m)} + (h_{mm}^{(-m)}+1)^2$. Then [164, Equation (44)] gives

$$\|\boldsymbol{\mu}_c\|_2^2 \cdot \boldsymbol{v}_c(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c + \boldsymbol{v}_c(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c = \frac{\|\boldsymbol{\mu}_c\|_2^2 s_{cc}^{(j)} - s_{cc}^{(j)} t_{cc}^{(j)} + {h_{cc}^{(j)}}^2 + h_{cc}^{(j)}}{\det_j},$$

where $\det_j = (\|\boldsymbol{\mu}\|_2^2 - t_{cc}^{(j)})s_{cc}^{(j)} + (h_{cc}^{(j)}+1)^2$. Note that $\det_j = \det_{-c}$ when $c = k$ and $j = k-1$.

For $\boldsymbol{v}_c(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j$ and $\boldsymbol{v}_j(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c$, we can again express the k-th order quadratic

forms in terms of j-th order quadratic forms as follows:

$$\boldsymbol{v}_c(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j = \frac{s_{cj}^{(j)} + s_{cj}^{(j)}h_{cc}^{(j)} - s_{cc}^{(j)}h_{jc}^{(j)}}{\det_j},$$
$$\boldsymbol{v}_j(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c = \frac{\|\boldsymbol{\mu}\|_2^2 s_{cc}^{(j)}h_{jc}^{(j)} - \|\boldsymbol{\mu}\|_2^2 s_{cj}^{(j)}h_{cc}^{(j)} + h_{cc}^{(j)}h_{jc}^{(j)} + h_{jc}^{(j)} - s_{cj}^{(j)}t_{cc}^{(j)}}{\det_j}.$$

Thus, we have

$$\|\boldsymbol{\mu}\|_2^2\boldsymbol{v}_c(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j + \boldsymbol{v}_j(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c = \frac{\|\boldsymbol{\mu}_c\|_2^2 s_{jc}^{(j)} + h_{jc}^{(j)} + h_{jc}^{(j)}h_{cc}^{(j)} - s_{jc}^{(j)}t_{cc}^{(j)}}{\det_j}.$$

This completes the proof. □

Extensions to not-orthogonal means

While we made the orthogonality assumption on class means (Assumption 3) for simplicity, our error analysis can conceivably be extended to the more general unorthogonal setting. We provide a brief discussion of this extension here. To upper bound the class-wise error $\mathbb{P}_{e|c}$, recall that we need to lower bound the quantity in Equation (B.16). As with the orthogonal case, we consider $c = k, j = k-1$ without loss of generality. Recall that $\widehat{\boldsymbol{w}}_c = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c$ and $\boldsymbol{X} = \sum_{j=1}^k \boldsymbol{\mu}_j\boldsymbol{v}_j^T + \boldsymbol{Q}$. Thus

$$\widehat{\boldsymbol{w}}_c^T\boldsymbol{\mu}_c = \|\boldsymbol{\mu}\|_2^2\boldsymbol{v}_c^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c + \sum_{m\neq c}\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\boldsymbol{v}_m^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c + \boldsymbol{v}_c^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c \text{ and}$$
$$\widehat{\boldsymbol{w}}_j^T\boldsymbol{\mu}_c = N_6 + N_7,$$
$$N_6 = \|\boldsymbol{\mu}\|_2^2\boldsymbol{v}_j^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_c + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_c\boldsymbol{v}_j^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j$$
$$N_7 = \sum_{m\neq c,j}\boldsymbol{\mu}_m^T\boldsymbol{\mu}_c\boldsymbol{v}_m^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{v}_j + \boldsymbol{v}_j^T(\mathbf{X}^T\boldsymbol{X})^{-1}\boldsymbol{d}_c.$$

In Appendix B.3.1 we have already obtained the bounds for

$$\|\boldsymbol{\mu}\|_2^2 \boldsymbol{v}_c^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{v}_c + \boldsymbol{v}_c^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{d}_c - \|\boldsymbol{\mu}\|_2^2 \boldsymbol{v}_c^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{v}_j - \boldsymbol{v}_j^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{d}_c.$$

Moreover, under the equal energy and priors assumption, the $\boldsymbol{v}_j^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{v}_j$ terms have the same bound for every $j \in [k]$. Similarly, the $\boldsymbol{v}_j^T (\mathbf{X}^T \boldsymbol{X})^{-1} \boldsymbol{v}_m$ terms also have the same bound for all $j \neq m \in [k]$. An upper bound on classification error can then be derived in terms of the inner products between the mean vectors. We leave the detailed derivation to the reader. Expressions are naturally more complicated.

B.3.2 Proof of Corollary 11.1

We now prove the condition for benign overfitting provided in Corollary 11.1. Note that following Theorem 9, we assume that

$$p > C_1 k^3 n \log(kn) + n - 1 \quad \text{and} \quad p > C_2 k^{1.5} n^{1.5} \|\boldsymbol{\mu}\|_2. \tag{B.23}$$

We begin with the setting where $\|\boldsymbol{\mu}\|_2^2 > C\frac{kp}{n}$, for some $C > 1$. In this case, we get that Equation (B.22) is lower bounded by $\frac{1}{c}\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_5\sqrt{k}\right)^2$, and we have

$$\begin{aligned}\left(\left(1 - \frac{C_3}{\sqrt{n}} - \frac{C_4 n}{p}\right)\|\boldsymbol{\mu}\|_2 - C_5\sqrt{k}\right)^2 &> \|\boldsymbol{\mu}\|_2^2 - 2\|\boldsymbol{\mu}\|_2^2 \frac{C_3}{\sqrt{n}} - 2\|\boldsymbol{\mu}\|_2^2 \frac{C_4 n}{p} - 2C_5\sqrt{k}\|\boldsymbol{\mu}\|_2 \\ &> \left(1 - \frac{2C_3}{\sqrt{n}}\right)\frac{kp}{n} - 2\|\boldsymbol{\mu}\|_2^2 \frac{C_4 n}{p} - 2C_5\sqrt{k}\|\boldsymbol{\mu}\|_2.\end{aligned} \tag{B.24}$$

Then Equation (B.23) gives

$$\begin{aligned}(\text{B.24}) &> \left(1-\frac{2C_3}{\sqrt{n}}\right)\frac{kp}{n} - \left(\frac{p}{k^{1.5}n^{1.5}}\right)^2 \frac{C_6 n}{p} - \frac{C_7\sqrt{k}p}{k^{1.5}n^{1.5}} \\ &= \frac{kp}{n}\left(1-\frac{2C_3}{\sqrt{n}} - \frac{C_6}{k^4 n} - \frac{C_7}{k^2\sqrt{n}}\right), \end{aligned} \tag{B.25}$$

which goes to $+\infty$ as $\left(\frac{p}{n}\right) \to \infty$.

Next, we consider the case $\|\boldsymbol{\mu}\|_2^2 \leq \frac{kp}{n}$. Moreover, we assume that $\|\boldsymbol{\mu}\|_2^4 = C_2\left(\frac{p}{n}\right)^\alpha$, for $\alpha > 1$. Then, Equation (B.22) is lower bounded by $\frac{n}{ckp}\|\boldsymbol{\mu}\|_2^4\left(\left(1-\frac{C_3}{\sqrt{n}}-\frac{C_4 n}{p}\right) - \frac{C_5\sqrt{k}}{\|\boldsymbol{\mu}\|_2}\right)^2$, and we get

$$\begin{aligned} &\frac{n}{kp}\|\boldsymbol{\mu}\|_2^4\left(\left(1-\frac{C_3}{\sqrt{n}}-\frac{C_4 n}{p}\right) - \frac{C_5\sqrt{k}}{\|\boldsymbol{\mu}\|_2}\right)^2 \\ &> \left(1-\frac{2C_3}{\sqrt{n}}\right)\frac{n}{kp}\|\boldsymbol{\mu}\|_2^4 - \frac{C_6 n^2}{kp^2}\|\boldsymbol{\mu}\|_2^4 - \frac{C_7 n}{\sqrt{kp}}\|\boldsymbol{\mu}\|_2^3 \\ &\geq \left(1-\frac{2C_3}{\sqrt{n}}\right)\frac{1}{k}\left(\frac{p}{n}\right)^{\alpha-1} - \frac{C_6}{k}\left(\frac{p}{n}\right)^{\alpha-2} - \frac{C_7}{\sqrt{k}}\left(\frac{p}{n}\right)^{0.75\alpha-1}, \end{aligned} \tag{B.26}$$

where the last inequality uses Equations (B.23) and condition $\|\boldsymbol{\mu}\|_2^2 \leq \frac{kp}{n}$. Consequently, the RHS of Equation (B.26) will go to $+\infty$ as $\left(\frac{p}{n}\right) \to \infty$, provided that $\alpha > 1$. Overall, it suffices to have

$$\begin{aligned} &p > \max\left\{C_1 k^3 n\log(kn) + n - 1, C_2 k^{1.5} n^{1.5}\|\boldsymbol{\mu}\|_2, \frac{n\|\boldsymbol{\mu}\|_2^2}{k}\right\}, \\ &\text{and } \|\boldsymbol{\mu}\|_2^4 \geq C_8\left(\frac{p}{n}\right)^\alpha, \quad \text{for } \alpha \in (1,2]. \end{aligned}$$

All of these inequalities hold provided that $\|\boldsymbol{\mu}\|_2 = \Theta(p^\beta)$ for $\beta \in (1/4, 1/2]$ for finite k and n. This completes the proof. □

B.4 Recursive formulas for higher-order quadratic forms

We first show how quadratic forms involving the j-th order Gram matrix $\boldsymbol{A}_j^{-1}$ can be expressed using quadratic forms involving the $(j-1)$-th order Gram matrix $\boldsymbol{A}_{j-1}^{-1}$. For concreteness, we consider $j = 1$; identical expressions hold for any $j > 1$ with the only change being in the superscripts. Recall from Section 3.6.2 that we can write

$$\boldsymbol{A}_1 = \boldsymbol{A}_0 + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{Q}^T \boldsymbol{\mu}_1 & \boldsymbol{v}_1 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{\mu}_1^T \boldsymbol{Q} \end{bmatrix} = \boldsymbol{Q}^T \boldsymbol{Q} + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{d}_1 & \boldsymbol{v}_1 \end{bmatrix} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{d}_1^T \end{bmatrix}.$$

The first step is to derive an expression for $\boldsymbol{A}_1^{-1}$. By the Woodbury identity [70], we get

$$\boldsymbol{A}_1^{-1} = \boldsymbol{A}_0^{-1} - \boldsymbol{A}_0^{-1} N_8 \boldsymbol{A}_0^{-1}, \tag{B.27}$$

$$N_8 = \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{d}_1 & \boldsymbol{v}_1 \end{bmatrix} \left[\mathbf{I} + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{d}_1^T \end{bmatrix} \boldsymbol{A}_0^{-1} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{d}_1 & \boldsymbol{v}_1 \end{bmatrix} \right]^{-1} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{d}_1^T \end{bmatrix}$$

We first compute the inverse of the 3×3 matrix

$$\mathbf{B} := \left[\mathbf{I} + \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1^T \\ \boldsymbol{v}_1^T \\ \boldsymbol{d}_1^T \end{bmatrix} \boldsymbol{A}_0^{-1} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 \boldsymbol{v}_1 & \boldsymbol{d}_1 & \boldsymbol{v}_1 \end{bmatrix} \right].$$

Recalling our definitions of the terms $s_{mj}^{(c)}, h_{mj}^{(c)}$ and $t_{mj}^{(c)}$ in Equation (3.37) in Section 3.6.2, we have:

$$\mathbf{B} = \begin{bmatrix} 1 + \|\boldsymbol{\mu}\|_2^2 s_{11}^{(0)} & \|\boldsymbol{\mu}\|_2 h_{11}^{(0)} & \|\boldsymbol{\mu}\|_2 s_{11}^{(0)} \\ \|\boldsymbol{\mu}\|_2 s_{11}^{(0)} & 1 + h_{11}^{(0)} & s_{11}^{(0)} \\ \|\boldsymbol{\mu}\|_2 h_{11}^{(0)} & t_{11}^{(0)} & 1 + h_{11}^{(0)} \end{bmatrix}.$$

Recalling $\mathbf{B}^{-1} = \frac{1}{\det_0}\text{adj}(\mathbf{B})$, where $\det_0$ is the determinant of $\mathbf{B}$ and $\text{adj}(\mathbf{B})$ is the adjoint of $\mathbf{B}$, simple algebra gives us

$$\det_0 = s_{11}^{(0)}(\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}) + (h_{11}^{(0)} + 1)^2,$$

and $\text{adj}(\mathbf{B})$ is

$$\begin{bmatrix} (h_{11}^{(0)} + 1)^2 - s_{11}^{(0)} t_{11}^{(0)} & \|\boldsymbol{\mu}\|_2 (s_{11}^{(0)} t_{11}^{(0)} - h_{11}^{(0)} - {h_{11}^{(0)}}^2) & -\|\boldsymbol{\mu}\|_2 s_{11}^{(0)} \\ -\|\boldsymbol{\mu}\|_2 s_{11}^{(0)} & h_{11}^{(0)} + 1 + \|\boldsymbol{\mu}\|_2^2 s_{11}^{(0)} & -s_{11}^{(0)} \\ \|\boldsymbol{\mu}\|_2 (s_{11}^{(0)} t_{11}^{(0)} - h_{11}^{(0)} - {h_{11}^{(0)}}^2) & \|\boldsymbol{\mu}\|_2^2 {h_{11}^{(0)}}^2 - t_{11}^{(0)}(1 + \|\boldsymbol{\mu}\|_2^2 s_{11}^{(0)}) & h_{11}^{(0)} + 1 + \|\boldsymbol{\mu}\|_2^2 s_{11}^{(0)} \end{bmatrix}.$$

We will now use these expressions to derive expressions for the 1-order quadratic forms that are used in Appendix B.1.2.

B.4.1 Expressions for 1-st order quadratic forms

We now show how quadratic forms of order 1 can be expressed as a function of quadratic forms of order 0. All of the expressions are derived as a consequence of plugging in the expression for $\mathbf{B}^{-1}$ together with elementary matrix algebra.

First, we have

$$s_{mk}^{(1)} = \boldsymbol{v}_m^T \boldsymbol{A}_1^{-1} \boldsymbol{v}_k = \boldsymbol{v}_m^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_k - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{m1}^{(0)} & h_{m1}^{(0)} & s_{m1}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{k1}^{(0)} \\ s_{k1}^{(0)} \\ h_{k1}^{(0)} \end{bmatrix}$$
$$= s_{mk}^{(0)} - \frac{1}{\det_0} (\star)_s^{(0)}, \tag{B.28}$$

where we define

$$(\star)_s^{(0)}$$
$$:= (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}) s_{1k}^{(0)} s_{1m}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} h_{11}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} h_{k1}^{(0)} h_{m1}^{(0)} + s_{1m}^{(0)} h_{k1}^{(0)} + s_{1k}^{(0)} h_{m1}^{(0)}.$$

Thus, for the case $m = k$ we have

$$s_{kk}^{(1)} = \boldsymbol{v}_k^T \boldsymbol{A}_1^{-1} \boldsymbol{v}_k = \boldsymbol{v}_k^T \boldsymbol{A}_0^{-1} \boldsymbol{v}_k - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{k1}^{(0)} & h_{k1}^{(0)} & s_{k1}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{k1}^{(0)} \\ s_{k1}^{(0)} \\ h_{k1}^{(0)} \end{bmatrix}$$
$$= s_{kk}^{(0)} - \frac{1}{\det_0} \Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)}) {s_{1k}^{(0)}}^2 + 2 s_{1k}^{(0)} h_{k1}^{(0)} h_{11}^{(0)} - s_{11}^{(0)} {h_{k1}^{(0)}}^2 + 2 s_{1k}^{(0)} h_{k1}^{(0)} \Big). \tag{B.29}$$

Next, we have

$$h_{mk}^{(1)} = \boldsymbol{v}_m^T \boldsymbol{A}_1^{-1} \boldsymbol{d}_k = \boldsymbol{v}_m^T \boldsymbol{A}_0^{-1} \boldsymbol{d}_k - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{m1}^{(0)} & h_{m1}^{(0)} & s_{m1}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1k}^{(0)} \\ h_{1k}^{(0)} \\ t_{1k}^{(0)} \end{bmatrix}$$
$$= h_{mk}^{(0)} - \frac{1}{\det_0} (\star)_h^{(0)}, \tag{B.30}$$

where we define

$$(\star)_h^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1m}^{(0)}h_{1k}^{(0)} + h_{m1}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} + h_{m1}^{(0)}h_{1k}^{(0)} + s_{1m}^{(0)}t_{k1}^{(0)} + s_{1m}^{(0)}t_{k1}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}t_{k1}^{(0)}h_{m1}^{(0)}.$$

Next, we have

$$\begin{aligned} t_{km}^{(1)} = \boldsymbol{d}_k^T \boldsymbol{A}_1^{-1}\boldsymbol{d}_m &= \boldsymbol{d}_k^T \boldsymbol{A}_0^{-1}\boldsymbol{d}_m - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1m}^{(0)} \\ h_{1m}^{(0)} \\ t_{1m}^{(0)} \end{bmatrix} \\ &= t_{km}^{(0)} - \frac{1}{\det_0}(\star)_t^{(0)}, \end{aligned} \tag{B.31}$$

where we define

$$(\star)_t^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1m}^{(0)}h_{1k}^{(0)} + t_{m1}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} + t_{k1}^{(0)}h_{1m}^{(0)}h_{11}^{(0)} + t_{1m}^{(0)}h_{1k}^{(0)} + t_{1k}^{(0)}h_{1m}^{(0)} - s_{11}^{(0)}t_{1m}^{(0)}t_{1k}^{(0)}.$$

Thus, for the case $m = k$ we have

$$\begin{aligned} t_{kk}^{(1)} = \boldsymbol{d}_k^T \boldsymbol{A}_1^{-1}\boldsymbol{d}_k &= \boldsymbol{d}_k^T \boldsymbol{A}_0^{-1}\boldsymbol{d}_k - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1k}^{(0)} \\ h_{1k}^{(0)} \\ t_{1k}^{(0)} \end{bmatrix} \\ &= t_{kk}^{(0)} - \frac{1}{\det_0}\Big((\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)^2} + 2t_{1k}^{(0)}h_{1k}^{(0)}h_{11}^{(0)} - s_{11}^{(0)}t_{1k}^{(0)^2} + 2t_{1k}^{(0)}h_{1k}^{(0)}\Big). \end{aligned} \tag{B.32}$$

Next, we have

$$f_{ki}^{(1)} = \boldsymbol{d}_k^T \boldsymbol{A}_1^{-1}\mathbf{e}_i = \boldsymbol{d}_k^T \boldsymbol{A}_0^{-1}\mathbf{e}_i - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 h_{1k}^{(0)} & t_{1k}^{(0)} & h_{1k}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 g_{1i}^{(0)} \\ g_{1i}^{(0)} \\ f_{1i}^{(0)} \end{bmatrix}$$
$$= f_{ki}^{(0)} - \frac{1}{\det_0}(\star)_f^{(0)}, \tag{B.33}$$

where we define

$$(\star)_f^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})h_{1k}^{(0)} g_{1i}^{(0)} + t_{1k}^{(0)} g_{1i}^{(0)} + t_{1k}^{(0)} h_{11}^{(0)} g_{1i}^{(0)} + h_{1k}^{(0)} f_{1i}^{(0)} + h_{1k}^{(0)} h_{11}^{(0)} f_{1i}^{(0)} - s_{11}^{(0)} t_{1k}^{(0)} f_{1i}^{(0)}.$$

Finally, we have

$$g_{ji}^{(1)} = \boldsymbol{v}_j^T \boldsymbol{A}_1^{-1}\mathbf{e}_i = \boldsymbol{v}_j^T \boldsymbol{A}_0^{-1}\mathbf{e}_i - \begin{bmatrix} \|\boldsymbol{\mu}\|_2 s_{j1}^{(0)} & h_{j1}^{(0)} & s_{j1}^{(0)} \end{bmatrix} \frac{\mathrm{adj}(\mathbf{B})}{\det_0} \begin{bmatrix} \|\boldsymbol{\mu}\|_2 g_{1i}^{(0)} \\ g_{1i}^{(0)} \\ f_{1i}^{(0)} \end{bmatrix}$$
$$= g_{ji}^{(0)} - \frac{1}{\det_0}(\star)_{gj}^{(0)}, \tag{B.34}$$

where we define

$$(\star)_{gj}^{(0)} = (\|\boldsymbol{\mu}\|_2^2 - t_{11}^{(0)})s_{1j}^{(0)} g_{1i}^{(0)} + g_{1i}^{(0)} h_{11}^{(0)} h_{j1}^{(0)} + g_{1i}^{(0)} h_{j1}^{(0)} + s_{1j}^{(0)} f_{1i}^{(0)} + s_{1j}^{(0)} h_{11}^{(0)} f_{1i}^{(0)} - s_{11}^{(0)} h_{j1}^{(0)} f_{1i}^{(0)}.$$

□

B.5 One-vs-all SVM

In this section, we derive conditions under which the OvA solutions $\boldsymbol{w}_{\mathrm{OvA},c}$ interpolate, i.e, all data points are support vectors in Equation (3.8).

B.5.1 Gaussian mixture model

As in the case of the multiclass SVM, we assume equal priors on the class means and equal energy (Assumption 2).

Theorem 16. *Assume that the training set follows a multiclass GMM with noise covariance* $\boldsymbol{\Sigma} = \mathbf{I}_p$ *and Assumption 2 holds. Then, there exist constants* $c_1, c_2, c_3 > 1$ *and* $C_1, C_2 > 1$ *such that the solutions of the OvA-SVM and MNI are identical with probability at least* $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ *provided that*

$$p > C_1 kn \log(kn) + n - 1 \quad \text{and} \quad p > C_2 n^{1.5} \|\boldsymbol{\mu}\|_2. \tag{B.35}$$

We can compare Equation (B.35) with the corresponding condition for multiclass SVM in Theorem 9 (Equation (3.16)). Observe that the right-hand-side of Equation (B.35) above does not scale with k, while the right-hand-side of Equation (3.16) scales with k as k^3. Otherwise, the scalings with n and energy of class means $\|\boldsymbol{\mu}\|_2$ are identical. This discrepancy with respect to k-dependence arises because the multiclass SVM is equivalent to the OvA-SVM in Equation (3.31) with unequal margins $1/k$ and $(k-1)/k$ (as we showed in Theorem 8).

Proof sketch. Recall from Section 3.6.2 that we derived conditions under which the multiclass SVM interpolates the training data by studying the related symmetric OvA-type classifier defined in Equation (3.15). Thus, this proof is similar to the proof of Theorem 9 provided in Section 3.6.2. The only difference is that the margins for the OvA-SVM

are not $1/k$ and $(k-1)/k$, but 1 for all classes. Owing to the similarity between the arguments, we restrict ourselves to a proof sketch here.

Following Section 3.6.2 and Equation (3.43), we consider $y_i = k$. We will derive conditions under which the condition

$$\left((1+h_{kk}^{(-k)})g_{ki}^{(-k)} - s_{kk}^{(-k)} f_{ki}^{(-k)}\right) + C\sum_{j\neq k}\left((1+h_{jj}^{(-j)})g_{ji}^{(-j)} - s_{jj}^{(-j)} f_{ji}^{(-j)}\right) > 0, \tag{B.36}$$

holds with high probability for some $C > 1$. We define

$$\epsilon := \frac{n^{1.5}\|\boldsymbol{\mu}\|_2}{p} \leq \tau,$$

where τ is chosen to be a sufficiently small constant. Applying the same trick as in Lemma 7 (with the newly defined parameters ϵ and τ) gives us with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$:

$$\begin{aligned}
&\text{(B.36)} \\
&\geq \left(\left(1 - \frac{C_1\epsilon}{\sqrt{k}\sqrt{n}}\right)\left(1 - \frac{1}{C_2}\right)\frac{1}{p} - \frac{C_3\epsilon}{n}\cdot\frac{n}{kp}\right) - \frac{k}{C_4}\left(\left(1 + \frac{C_5\epsilon}{\sqrt{k}\sqrt{n}}\right)\frac{1}{kp} - \frac{C_6\epsilon}{n}\cdot\frac{n}{kp}\right) \\
&\geq \left(1 - \frac{1}{C_9} - \frac{C_{10}\epsilon}{\sqrt{k}\sqrt{n}} - \frac{C_{11}\epsilon}{k} - C_{12}\epsilon\right)\frac{1}{p} \\
&\geq \frac{1}{p}\left(1 - \frac{1}{C_9} - C_0\tau\right), \qquad \text{(B.37)}
\end{aligned}$$

for some constants C_i's > 1. We used the fact that $|g_{ji}^{(0)}| \leq (1/C)(1/(kp))$ for $j \neq y_i$ with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ provided that $p > C_1 kn\log(kn) + n - 1$, which is the first sufficient condition in the theorem statement. □

B.5.2 Multinomial logistic model

Recall that we defined the data covariance matrix $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and its spectrum $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \cdots & \lambda_p \end{bmatrix}$. We also defined the effective dimensions $d_2 := \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}$ and $d_\infty := \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$.

The following result provides sufficient conditions under which the OvA SVM and MNI classifier have the same solution with high probability under the MLM.

Theorem 17. *Assume that the training set follows a multiclass MLM. There exist constants c and $C_1, C_2 > 1$ such that, if the following conditions hold:*

$$d_\infty > C_1 n \log(kn) \quad \textit{and} \quad d_2 > C_2(\log(kn) + n), \tag{B.38}$$

the solutions of the OvA-SVM and MNI are identical with probability at least $(1 - \frac{c}{n})$. In the special case of isotropic covariance, the same result holds provided that

$$p > 10n \log(\sqrt{k}n) + n - 1, \tag{B.39}$$

Comparing this result to the corresponding results in Theorems 10, we observe that k now only appears in the log function (as a result of k union bounds). Thus, the unequal $1/k$ and $(k-1)/k$ margins that appear in the multiclass-SVM make interpolation harder than with the OvA-SVM, just as in the GMM case.

Proof sketch. For the OvA SVM classifier, we need to solve k binary max-margin classification problems, hence the proof follows directly from [120, Theorem 1] and [72, Theorem 1] by applying k union bounds. We omit the details for brevity. □

One-vs-one SVM

In this section, we first derive conditions under which the OvO solutions interpolate, i.e, all data points are support vectors. We then provide an upper bound on the classification error of the OvO solution.

In OvO classification, we solve $k(k-1)/2$ binary classification problems, e.g., for classes pair (c, j), we solve

$$\boldsymbol{w}_{\mathrm{OvO},(c,j)} := \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \quad \text{sub. to} \;\; \boldsymbol{w}^T\boldsymbol{x}_i \geq 1, \; \text{if } \boldsymbol{y}_i = c; \;\; \boldsymbol{w}^T\boldsymbol{x}_i \leq -1 \text{ if } \boldsymbol{y}_i = j, \; \forall i \in [n]. \tag{B.40}$$

Then we apply these $k(k-1)/2$ classifiers to a fresh sample and the class that got the highest $+1$ voting gets predicted.

We now present conditions under which every data point becomes a support vector over these $k(k-1)/2$ problems. We again assume equal priors on the class means and equal energy (Assumption 2).

Theorem 18. *Assume that the training set follows a multiclass GMM with noise covariance $\boldsymbol{\Sigma} = \mathbf{I}_p$ and Assumption 2 holds. Then, there exist constants $c_1, c_2, c_3 > 1$ and $C_1, C_2 > 1$ such that the solutions of the OvA-SVM and MNI are identical with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$ provided that*

$$p > C_1 n \log(kn) + (2n/k) - 1 \quad \textit{and} \quad p > C_2 n^{1.5} \|\boldsymbol{\mu}\|_2. \tag{B.41}$$

Proof sketch. Note that the margins of OvO SVM are 1 and -1, hence the proof is similar to the proof of Theorem 16. Recall that in OvO SVM, we solve $k(k-1)/2$ binary problems and each problems has sample size $2n/k$ with high probability. Therefore, compared to OvA SVM which solves k problems each with sample size n, OvO SVM

needs less overparameterization to achieve interpolation. Thus the first condition in Equation (B.35) reduces to $p > C_1 n \log(kn) + (2n/k) - 1$. □

We now derive the classification risk for OvO SVM classifiers. Recall that OvO classification solves $k(k-1)/2$ binary subproblems. Specifically, for each pair of classes, say $(i, j) \in [k] \times [k]$, we train a classifier $\boldsymbol{w}_{ij} \in \mathbb{R}^p$ and the corresponding decision rule for a fresh sample $\boldsymbol{x} \in \mathbb{R}^p$ is $\hat{y}_{ij} = \text{sign}(\boldsymbol{x}^T \hat{\boldsymbol{w}}_{ij})$. Overall, each class $i \in [k]$ gets a voting score $s_i = \sum_{j \neq i} \mathbf{1}_{\hat{y}_{ij}=+1}$. Thus, the final decision is given by majority rule that *decides the class with the highest score*, i.e., $\arg\max_{i \in [k]} s_i$. Having described the classification process, the total classification error $\mathbb{P}_e$ for balanced classes is given by the conditional error $\mathbb{P}_{e|c}$ given the fresh sample belongs to class c. Without loss of generality, we assume $c = 1$. Formally, $\mathbb{P}_e = \mathbb{P}_{e|1} = \mathbb{P}_{e|1}(s_1 < s_2 \text{ or } s_1 < s_3 \text{ or} \cdots \text{or } s_1 < s_k)$. Under the equal prior and energy assumption, by symmetry and union bound, the conditional classification risk given that true class is 1 can be upper bounded as:

$$\begin{aligned}
&\mathbb{P}_{e|1}(s_1 < s_2 \text{ or } s_1 < s_3 \text{ or} \cdots \text{or } s_1 < s_k) \\
&\leq \mathbb{P}_{e|1}(s_1 < k-1) = \mathbb{P}_{e|1}(\exists j \ s.t. \ \hat{y}_{1j} \neq 1) \leq (k-1)\mathbb{P}_{e|1}(\hat{y}_{12} \neq 1).
\end{aligned}$$

Therefore, it suffices to bound $\mathbb{P}_{e|1}(y_{12} \neq 1)$. We can directly apply Theorem 11 with changing k to 2 and n to $2n/k$.

Theorem 19. *Let Assumption 3 and the condition in Equation* (B.41) *hold. Further assume constants* $C_1, C_2, C_3 > 1$ *such that* $\left(1 - C_1\sqrt{\frac{k}{n}} - \frac{C_2 n}{kp}\right)\|\boldsymbol{\mu}\|_2 > C_3$. *Then, there exist additional constants* c_1, c_2, c_3 *and* $C_4 > 1$ *such that the OvO SVM solutions satisfies:*

$$\mathbb{P}_{e|c} \leq (k-1)\exp\left(-\|\boldsymbol{\mu}\|_2^2 \frac{\left(\left(1 - C_1\sqrt{\frac{k}{n}} - \frac{C_2 n}{kp}\right)\|\boldsymbol{\mu}\|_2 - C_3\right)^2}{C_4\left(\|\boldsymbol{\mu}\|_2^2 + \frac{kp}{n}\right)}\right) \tag{B.42}$$

with probability at least $1 - \frac{c_1}{n} - c_2 k e^{-\frac{n}{c_3 k^2}}$*, for every* $c \in [k]$*. Moreover, the same bound holds for the total classification error* $\mathbb{P}_e$*.*

Appendix C

Appendix for Chapter 4

C.1 Proofs

C.1.1 Proof outline and auxiliary lemmas

To prove Theorems 12 and 13, we first write $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$ as below:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} &= (\hat{\boldsymbol{\Sigma}}_{obs} - \hat{\lambda}_1 \hat{\boldsymbol{\theta}}_1 \hat{\boldsymbol{\theta}}_1^T) - (\boldsymbol{\Sigma}_{obs} - \lambda_1 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T + \lambda_k \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T) \\ &= (\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}) + (\lambda_1 \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T - \hat{\lambda}_1 \hat{\boldsymbol{\theta}}_1 \hat{\boldsymbol{\theta}}_1^T) - \lambda_k \boldsymbol{\theta}_1 \boldsymbol{\theta}_1^T, \end{aligned} \tag{C.1}$$

where λ_k is the k-th eigenvalue of $\boldsymbol{\Sigma}_{obs}$, and $\boldsymbol{\theta}_k$ is the k-th eigenvector of $\boldsymbol{\Sigma}_{obs}$. To bound $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$, we need to bound the norms of $\boldsymbol{\Sigma}_{obs} - \hat{\boldsymbol{\Sigma}}_{obs}$, $\lambda_1 - \hat{\lambda}_1$ and $\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1$, and the lemmas below show these bounds.

Lemma 34. *Assuming that $\boldsymbol{\Sigma}_{obs}^{-1}$ satisfies the maximum and minimum eigenvalue condition in (4.14), then*

$$P(\|\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}\|_\infty \geq t) \leq p^2 C_1 p^{-3}, \quad t = 3C_2 \sqrt{\frac{\log p}{n}},$$

where C_1 and C_2 depends on the eigenvalue bound M in (4.14).

Proof. The proof follows [19, Lemma A.3]. □

The next two lemmas provide bounds for $|\lambda_1 - \hat{\lambda}_1|$.

Lemma 35. *Under the assumptions of Theorem 12, we have*

$$|\lambda_1 - \hat{\lambda}_1| \leq C_1 \lambda_1 \sqrt{\frac{p}{n}},$$

with probability at least $1 - 2e^{-p/C_2}$ for some constants C_i's > 1.

Proof. By Weyl's lemma [70]

$$\max_{j=1,\ldots,p} |\lambda_j(\boldsymbol{\Sigma}_{obs}) - \lambda_j(\hat{\boldsymbol{\Sigma}}_{obs})| \leq \|\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}\|_2.$$

The bound on $\|\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}\|_2$ is then obtained from [163, Theorem 6.5]. □

Following [87, Theorem 5], a tighter bound can be obtained using the effective rank $r(\boldsymbol{\Sigma}_{obs})$ defined in (4.15).

Lemma 36. *Under the assumptions of Theorem 12, we have*

$$|\lambda_1 - \hat{\lambda}_1| \leq \|\hat{\boldsymbol{\Sigma}}_{obs} - \boldsymbol{\Sigma}_{obs}\|_2 \leq C_1 \lambda_1 (\sqrt{\frac{r(\boldsymbol{\Sigma})}{n}} \sqrt{\frac{p}{n}} \vee \frac{p}{n}),$$

with probability at least $1 - e^{-p/C_2}$ for some constants C_i's > 1.

Then the two lemmas below provide bounds for $\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_\infty$.

Lemma 37 (adapted from Wainwright (2019, Corollary 8.7))**.** *Under the assumptions of Theorem 12, suppose $n \geq p$ and $\|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{p}{n}} \leq \frac{1}{128}$, then*

$$\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_2 \leq C_1 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{p}{n}},$$

with probability at least $1 - C_2 e^{-p/C_3}$ *for some* C_i*'s* > 1*, where* $\nu = \lambda_1(\mathbf{\Sigma}_{obs}) - \lambda_2(\mathbf{\Sigma}_{obs})$.

The lemma below shows a tighter bound with a large eigengap $\nu = \lambda_1(\mathbf{\Sigma}_{obs}) - \lambda_2(\mathbf{\Sigma}_{obs})$ following [47, Section 3.1].

Lemma 38. *Under the assumptions of Theorem 13, suppose* $\sqrt{p}\nu \geq C_1(p\lambda_1(\mathbf{\Sigma}) \vee \sigma^2)$*, then*

$$\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_\infty \leq C_2\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - C_3/p$ *for some* C_i*'s* > 1*.*

Before moving to the proofs of Theorem 12 and 13, we first show an upper bound for $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty$ using (C.1).

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \|\mathbf{\Sigma}_{obs} - \hat{\mathbf{\Sigma}}_{obs}\|_\infty + \|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty + \|\lambda_k\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\|_\infty. \tag{C.2}$$

The term $\|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty$ can be expressed as:

$$\begin{aligned}
&\|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty \\
=&\|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \lambda_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T + \lambda_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T - \hat{\lambda}_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T + \hat{\lambda}_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty \\
\leq&\|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \lambda_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty + \|\lambda_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T - \hat{\lambda}_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty + \|\hat{\lambda}_1\boldsymbol{\theta}_1\hat{\boldsymbol{\theta}}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty \\
\leq&|\lambda_1|\|\boldsymbol{\theta}_1\|_\infty\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_\infty + |\lambda_1 - \hat{\lambda}_1|\|\boldsymbol{\theta}_1\|_\infty\|\hat{\boldsymbol{\theta}}_1\|_\infty + |\hat{\lambda}_1|\|\hat{\boldsymbol{\theta}}_1\|_\infty\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_\infty.
\end{aligned} \tag{C.3}$$

We can then use the bounds for $|\lambda_1 - \hat{\lambda}_1|$ and $\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_\infty$ in previous lemmas to bound $\|\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T - \hat{\lambda}_1\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_1^T\|_\infty$.

C.1.2 Proof of Theorems 12 and 13

Now we are ready to prove Theorem 12. We first plug in the bounds in Lemmas 34, 35 and 37 to (C.3). Since $\boldsymbol{\theta}_1$ is the eigenvector of a matrix, $\|\boldsymbol{\theta}_1\|_\infty \leq 1$. Combining these two results completes the proof.

To prove Theorem 13, we need to plug in the bounds in Lemmas 34, 36 and 38 to (C.3). Then we use the fact that $\|\boldsymbol{\theta}_1\|_\infty = O(1/\sqrt{p})$ to complete the proof.

C.1.3 Proof of Theorem 14

The proof follows the proof of [24, Theorem 6]. First we know,

$$\|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I}\|_\infty = \|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Omega}\|_\infty \leq \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \|\boldsymbol{\Omega}\|_{L_1}.$$

Then we have,

$$\begin{aligned}\|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1)\|_\infty &= \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I} + \boldsymbol{I} - \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Omega}}_1\|_\infty \\ &\leq \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I}\|_\infty + \|\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Omega}}_1\|_\infty + \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \|\boldsymbol{\Omega}\|_{L_1} + \lambda_n.\end{aligned}$$

We know,

$$\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1\|_\infty = \|\boldsymbol{\Omega}\boldsymbol{\Sigma}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1)\|_\infty \leq \|\boldsymbol{\Sigma}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1)\|_\infty \|\boldsymbol{\Omega}\|_{L_1}.$$

To bound the terms above, we need,

$$\|\boldsymbol{\Sigma}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1)\|_\infty \leq \|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}_1)\|_\infty + \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \|\boldsymbol{\Omega}\|_{L_1}.$$

We know $\|\boldsymbol{\Omega}\|_{L_1} \leq M_0$ from (4.17) and combining the relations above with the result of Theorem 12 or Theorem 13, with the choice of λ_n specified in Theorem 14, we can obtain the bound for $\hat{\boldsymbol{\Omega}}_1$. The bound of the same order can be obtained for $\hat{\boldsymbol{\Omega}}$, the symmetric version of $\hat{\boldsymbol{\Omega}}_1$.

C.2 Generalization of section 4.3

The analysis in section 4.3 assumes that the low-rank confounder is independent of $\boldsymbol{X}$ and the eigenvector of the covariance of the low-rank confounding is one of the eigenvectors of $\boldsymbol{\Sigma}$, the covariance of $\boldsymbol{X}$. Those two assumptions can be extended to the more general setups. In equation (4.11), when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are not independent, the convariance matrix for $\boldsymbol{X}_{obs}$ becomes

$$\boldsymbol{\Sigma}_{obs} = \boldsymbol{\Sigma} + \sigma Cov(\boldsymbol{X}, \boldsymbol{Z})\boldsymbol{v}^T + \sigma \boldsymbol{v} Cov(\boldsymbol{X}, \boldsymbol{Z})^T + \sigma^2 \boldsymbol{v}\boldsymbol{v}^T,$$

where $Cov(\boldsymbol{X}, \boldsymbol{Z})$ is a p-dimensional column vector. We can see that $\sigma Cov(\boldsymbol{X}, \boldsymbol{Z})\boldsymbol{v}^T + \sigma \boldsymbol{v} Cov(\boldsymbol{X}, \boldsymbol{Z})^T + \sigma^2 \boldsymbol{v}\boldsymbol{v}^T$ has rank at most 3, hence $\boldsymbol{\Sigma}_{obs}$ can still be expressed as the sum of $\boldsymbol{\Sigma}$ and a low-rank matrix. Here, to ensure that the confounding can be identified in PCA-based approach, we assume that both σ and σ^2 are large compared to the eigenvalues of $\boldsymbol{\Sigma}$. Then, our analysis in section 4.3 can still be applied here, but the eigenvectors of the low-rank matrix are not necessarily the eigenvectors of $\boldsymbol{\Sigma}$.

C.3 Eigenvalues of sparse graphs

Figure C.1 shows the first 25 eigenvalues of sparse graphs. We use `huge` package [171] to generate the sparse graphs with three different structures: scale-free, random

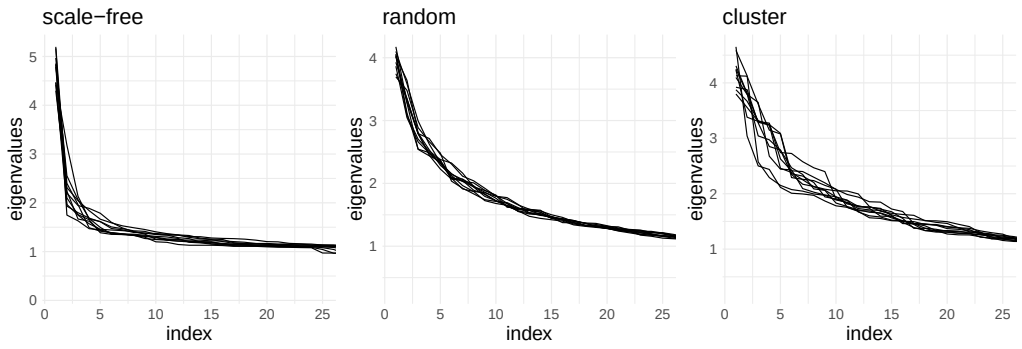


Figure C.1: The distribution of the top 25 eigenvalues created by scale-free, random and cluster graphs using `huge` package with $p = 100$ and $n = 10000$.

and clustered. We set $p = 100, n = 10000$ and assume default for all other parameters (see [171] for more detail). We generate 10 realizations for each graph structure and show the distribution of the first 25 eigenvalues of $\mathbf{\Sigma}$ in Figure C.1. We notice that the top eigenvalues are typically larger than the rest, especially for the scale-free graphs.

C.4 Gene co-expression networks data

Now we briefly introduce the data and the pre-processing procedure of gene co-expression networks in section 4.5.1. More detail can be found from [126]. We use the RNA-Seq data from Genotype-Tissue Expression (GTEx) project v6p release [1]. We consider three diverse tissues with sample sizes between 300 to 400 each: blood, lung and tibial nerve. We first filter the non-overlapping protein genes and perform a log transformation with base 2 to scale the data following [126, Appendix 2.4]. Since the underlying true network structure is unknown, we obtain the interaction information from some canonical pathway databases including KEGG, Biocarta, Reactome and Pathway Interaction Database. To make better use of those information, we pick 1000 high-variance genes which are included in all these databases, thus $p = 1000$ in this example.

[1] https://www.gtexportal.org/home/

C.5 Joint estimation of multiple graphs with latent confounders

Now we propose methods to estimate multiple graphs jointly when latent confouding exists. LVGGM is used to estimate a single graph with latent variables, assuming that all observations are drawn independently from the same distribution. In practice, we might need to estimate multiple related Gaussian graphical models with latent confounding. The graphs won't be estimated correctly if latent confounding is ignored. In this case, we can apply LVGGM jointly. In our model, we assume that all the classes share the same latent variable structure, this can be easily generalized to the case with different latent confounding structures. We will introduce two forms of penalties, corresponding to different graph structures that we expect. [123] study similar problems. Their goal is to estimate differential networks rather than graphs.

Suppose we are given K data sets $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(K)}$, with $K \geq 2$. $\boldsymbol{X}^{(k)}$ is an $n_k \times (p+h)$ matrix consisting of n_k observations with dimension $(p+h)$, which is common to all K data sets. We further assume $n_k = n$, for all k, and it is not hard to generalize current analysis to different sample sizes cases. Furthermore, we assume that $\sum_{k=1}^{K} n_k$ observations are independent, and that the observations within each data set are identically distributed: $\boldsymbol{x}_1^{(k)}, ..., \boldsymbol{x}_{n_k}^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}_{(O,H)}^{(k)})$. Without loss of generality, we assume that the features in each data set are centred such that $\boldsymbol{\mu}^{(k)} = \mathbf{0}$.

The following discussion is for all the K classes. Suppose that each $\boldsymbol{X}^{(k)}$ can be divided into the observed part and hidden part. $\boldsymbol{X}_O^{(k)} \in \mathbb{R}^p$ and $\boldsymbol{X}_H^{(k)} \in \mathbb{R}^h$ are subvectors of $\boldsymbol{X}^{(k)}$. Assume that we can only observe $\boldsymbol{X}_O^{(k)}$, $\boldsymbol{\Sigma}_{(O,H)}^{(k)}$ is the covariance matrix of $\boldsymbol{X}^{(k)}$, $\boldsymbol{\Sigma}_O^{(k)}$ is the marginal covariance matrix of $\boldsymbol{X}_O^{(k)}$ and $\boldsymbol{\Sigma}_H^{(k)}$ is the marginal covariance matrix of $\boldsymbol{X}_H^{(k)}$, the complete data covariance matrix $\boldsymbol{\Sigma}_{(O,H)}^{(k)}$ is

$$\begin{pmatrix} \boldsymbol{\Sigma}_O^{(k)} & \boldsymbol{\Sigma}_{O,H}^{(k)} \\ \boldsymbol{\Sigma}_{H,O}^{(k)} & \boldsymbol{\Sigma}_H^{(k)} \end{pmatrix}.$$

$\tilde{\boldsymbol{\Omega}}_O^{(k)} = (\boldsymbol{\Sigma}_O^{(k)})^{-1}$ is the marginal concentration matrix of $\boldsymbol{X}_O^{(k)}$. If we only observe $\boldsymbol{X}_O^{(k)}$, then we only have access to $\boldsymbol{\Sigma}_O^{(k)}$ (or $\tilde{\boldsymbol{\Omega}}_O^{(k)}$). $\boldsymbol{\Omega}_{(O,H)}^{(k)} = (\boldsymbol{\Sigma}_{(O,H)}^{(k)})^{-1}$ is the complete data concentration matrix:

$$\begin{pmatrix} \boldsymbol{\Omega}_O^{(k)} & \boldsymbol{\Omega}_{O,H}^{(k)} \\ \boldsymbol{\Omega}_{H,O}^{(k)} & \boldsymbol{\Omega}_H^{(k)} \end{pmatrix}.$$

By Schur complement, we have this decomposition

$$\tilde{\boldsymbol{\Omega}}_O^{(k)} = (\boldsymbol{\Sigma}_O^{(k)})^{-1} = \boldsymbol{\Omega}_O^{(k)} - \boldsymbol{\Omega}_{O,H}^{(k)}(\boldsymbol{\Omega}_H^{(k)})^{-1}\boldsymbol{\Omega}_{H,O}^{(k)} = \boldsymbol{S}^{(k)} - L.$$

$\boldsymbol{\Omega}_O^{(k)}$ is the concentration matrix of the conditional variables of the observed variables given latent variables. We assume that $\boldsymbol{\Omega}_O^{(k)}$ is sparse. $\boldsymbol{\Omega}_{O,H}^{(k)}(\boldsymbol{\Omega}_H^{(k)})^{-1}\boldsymbol{\Omega}_{H,O}^{(k)}$ is a summary of the effect of marginalization over the latent variables $\boldsymbol{X}_H^{(k)}$. This matrix has low rank if the number of latent variables is small relative to the number of observed variables. Here we assume the latent structures are the same across all K classes, so we use L to denote the low-rank component for all K classes.We let $\hat{\boldsymbol{\Sigma}}_O^{(k)} = (1/n_k)(\boldsymbol{X}^{(k)})^T\boldsymbol{X}^{(k)}$, the empirical covariance matrix for $\boldsymbol{X}^{(k)}$. We need to solve the following penalized log-likelihood problem:

$$\begin{aligned} \min \sum_{k=1}^{K} &[\mathrm{Tr}(\boldsymbol{S}^{(k)} - L)^{(k)}\hat{\boldsymbol{\Sigma}}_O^{(k)}) - \log\det(\boldsymbol{S}^{(k)} - L)] \\ &+ \lambda_1 \sum_k \sum_{i \neq j} |\boldsymbol{S}_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\boldsymbol{S}_{ij}^{(k)} - \boldsymbol{S}_{ij}^{(k')}| + \gamma K \mathrm{Tr}(L) \qquad \text{(C.4)} \\ \text{subject to } &\boldsymbol{S}^{(k)} \succ \boldsymbol{0}, L \succeq \boldsymbol{0}, \quad \text{for all } k. \end{aligned}$$

We use the proximal gradient-based alternating direction method (PGADM) in [106]

to solve (C.4). We first write the problem in this form

$$\begin{aligned}
&\min \sum_{k=1}^{K} [\mathrm{Tr}(\boldsymbol{R}^{(k)}\hat{\boldsymbol{\Sigma}}_O^{(k)}) - \log\det(\boldsymbol{R}^{(k)})] \\
&\quad + \lambda_1 \sum_k \sum_{i\neq j} |\boldsymbol{S}_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\boldsymbol{S}_{ij}^{(k)} - \boldsymbol{S}_{ij}^{(k')}| + \gamma K \mathrm{Tr}(L) \\
&\text{subject to } \ \boldsymbol{R}^{(k)} - \boldsymbol{S}^{(k)} + L = \boldsymbol{0},\ \boldsymbol{S}^{(k)} \succ \boldsymbol{0} \ for\ k = 1, ..., K \ and\ L \succeq \boldsymbol{0}.
\end{aligned}$$

Then we group two sets of variables $\{\boldsymbol{S}\}$ and L as one set of variables and solve:

$$\begin{aligned}
&\min\ f(\{\boldsymbol{R}\}) + \varphi(\{\boldsymbol{W}\}) \\
&\text{subject to } \boldsymbol{R}^{(k)} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}^{(k)} = 0,\ for\ all\ k,
\end{aligned}$$

where $\{\boldsymbol{S}\} = \{\boldsymbol{S}^{(1)}, ..., \boldsymbol{S}^{(K)}\}$, $\{\boldsymbol{R}\} = \{\boldsymbol{R}^{(1)}, ..., \boldsymbol{R}^{(K)}\}$, $\{\boldsymbol{W}\} = \{\boldsymbol{W}^{(1)}, ..., \boldsymbol{W}^{(K)}\}$, and $\boldsymbol{W}^{(k)} = [\boldsymbol{S}^{(k)}; L]$ and $f(\{\boldsymbol{R}\}) = \sum_{k=1}^{K} f(\{\boldsymbol{R}^{(k)}\}) = \sum_{k=1}^{K} [\mathrm{Tr}(\boldsymbol{R}^{(k)}\hat{\boldsymbol{\Sigma}}_O^{(k)}) - \log\det(\boldsymbol{R}^{(k)})]$ and $\varphi(\{\boldsymbol{W}\}) = \lambda_1 \sum_k \sum_{i\neq j} |\boldsymbol{S}_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\boldsymbol{S}_{ij}^{(k)} - \boldsymbol{S}_{ij}^{(k')}| + \gamma K \mathrm{Tr}(L) + \boldsymbol{I}\{L \succ \boldsymbol{0}\}$. The problem is decomposable for each class, so the ADMM procedure for the kth class becomes:

$$\begin{aligned}
&\boldsymbol{R}_{t+1}^{(k)} = \mathrm{argmin}_{\boldsymbol{R}^{(k)}} f(\boldsymbol{R}^{(k)}) - \langle \boldsymbol{\Lambda}^{(k)}, \boldsymbol{R}^{(k)} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}_t^{(k)} \rangle + \frac{1}{2\mu} \|\boldsymbol{R}^{(k)} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}_t^{(k)}\|_F^2, \\
&\boldsymbol{W}_{t+1}^{(k)} = \mathrm{argmin}_{\boldsymbol{W}^{(k)}} \varphi(\boldsymbol{W}^{(k)}) - \langle \boldsymbol{\Lambda}^{k}, \boldsymbol{R}_{t+1}^{(k)} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}^{(k)} \rangle + \frac{1}{2\mu} \|\boldsymbol{R}_{t+1}^{k} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}^{(k)}\|_F^2, \\
&\boldsymbol{\Lambda}_{t+1}^{(k)} = \boldsymbol{\Lambda}^{(k)} - (\boldsymbol{R}_{t+1}^{(k)} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}_{t+1}^{(k)})/\mu.
\end{aligned} \tag{C.5}$$

We need to solve the following four problems until convergence. More details can be found in [106, 123]. In the $t+i$ th iteration, we update $\{\boldsymbol{R}\}$ first

$$\boldsymbol{R}_{t+1}^{(k)} = \mathrm{argmin}_{\boldsymbol{R}} f(\boldsymbol{R}) - \langle \boldsymbol{\Lambda}_t^{(k)}, \boldsymbol{R} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}_t^{(k)} \rangle + \frac{1}{2\mu} \|\boldsymbol{R} - [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{W}_t^{(k)}\|_F^2,\ \forall k. \tag{C.6}$$

Then, when updating $\{\boldsymbol{S}\}$, we need to solve

$$\min_{\boldsymbol{S}_{ij}^{(1)},\ldots,\boldsymbol{S}_{ij}^{(K)}} \frac{1}{2\mu\tau}\sum_{k=1}^{K}(\boldsymbol{S}_{ij}^{(k)}-\boldsymbol{B}_{ij}^{(k)})^2+\lambda_1\boldsymbol{I}_{\{i\neq j\}}\sum_{k=1}^{K}|\boldsymbol{S}_{ij}^{(k)}|+\lambda_2\sum_{k\leq k'}|\boldsymbol{S}_{ij}^{(k)}-\boldsymbol{S}_{ij}^{(k')}|, \tag{C.7}$$

where $\boldsymbol{B}^{(k)}=\boldsymbol{S}_{(t)}^{(k)}+\boldsymbol{G}_{(t)}^{(k)}$ and $\boldsymbol{G}_{(t)}^{(k)}=\boldsymbol{R}_{t+1}^{(k)}-\boldsymbol{S}_{t}^{(k)}+L-\mu\boldsymbol{\Lambda}_{t}^{(k)}$. We next update L and $\boldsymbol{\Lambda}$

$$L_{t+1}=\text{argmin}_L K\gamma\text{Tr}(L)+\frac{1}{2\mu\tau}\sum_{k=1}^{K}\|L-(L_t-\tau\boldsymbol{G}_t^{(k)})\|_F^2 \tag{C.8}$$

$$\boldsymbol{\Lambda}_{t+1}^{(k)}=\boldsymbol{\Lambda}_{t}^{(k)}-\frac{1}{\mu}(\boldsymbol{R}_{t+1}^{(k)}-[\boldsymbol{I},-\boldsymbol{I}]\boldsymbol{W}_{t+1}^{(k)}). \tag{C.9}$$

(C.6) and (C.8) have analytic solutions, (C.7) can be solved using fused lasso algorithms [37, 67, 68].

Bibliography

[1] A. Agarwal, S. Negahban, M. J. Wainwright, et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, Sept. 2001. ISSN 1532-4435. doi: 10.1162/15324430152733133. URL `https://doi.org/10.1162/15324430152733133`.

[4] N. Ardeshir, C. Sanford, and D. J. Hsu. Support vector machines and linear regression coincide with very high-dimensional features. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] B. Aubin, F. Krzakala, Y. Lu, and L. Zdeborová. Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12199–12210. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/8f4576ad85410442a74ee3a7683757b3-Paper.pdf`.

[6] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International Conference on Learning Representations*, 2019.

[7] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80: 169–189, 2013.

[8] D. J. Bartholomew, M. Knott, and I. Moustaki. *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, 2011.

[9] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, Mar. 2003. ISSN 1532-4435.

[10] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL `https://www.pnas.org/content/117/48/30063`.

[11] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[12] M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

[13] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 540–548, 2018.

[14] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/content/116/32/15849`.

[15] M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[16] K. P. Bennett and O. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 3(1-3):27–39, 1994. doi: 10.1080/10556789408805554. URL `https://doi.org/10.1080/10556789408805554`.

[17] D. S. Bernstein. *Matrix mathematics: theory, facts, and formulas*. Princeton university press, 2009.

[18] P. J. Bickel, E. Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[19] P. J. Bickel, E. Levina, et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[20] A. S. Bosman, A. Engelbrecht, and M. Helbig. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions. *Neurocomputing*, 400:113–136, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.02.113. URL `https://www.sciencedirect.com/science/article/pii/S0925231220303593`.

[21] E. J. Bredensteiner and K. P. Bennett. Multicategory classification by support vector machines. In *Computational Optimization*, pages 53–79. Springer, 1999.

[22] A. Buhot and M. B. Gordon. Robust learning and generalization with support vector machines. *Journal of Physics A: Mathematical and General*, 34(21):4377–4388, May 2001. doi: 10.1088/0305-4470/34/21/301. URL `https://doi.org/10.1088/0305-4470/34/21/301`.

[23] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.

[24] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494):594–607, 2011.

[25] T. T. Cai, Z. Ren, H. H. Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[26] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[27] Y. Cao, Q. Gu, and M. Belkin. Risk bounds for over-parameterized maximum margin classification on sub-Gaussian mixtures. *arXiv preprint arXiv:2104.13628*, 2021.

[28] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[29] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, pages 1935–1967, 2012.

[30] X. Chang, Y. Li, S. Oymak, and C. Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

[31] N. S. Chatterji and P. M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

[32] N. S. Chatterji, P. M. Long, and P. L. Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *arXiv preprint arXiv:2012.02409*, 2020.

[33] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12(May):1771–1812, 2011.

[34] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844.

[35] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang. Structured prediction theory based on factor graph complexity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/535ab76633d94208236a2e829ea6d888-Paper.pdf`.

[36] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, Mar. 2002. ISSN 1532-4435.

[37] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[38] A. Demirkaya, J. Chen, and S. Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5, 2020. doi: 10.1109/CISS48834.2020.1570627167.

[39] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, April 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab002. URL `https://doi.org/10.1093/imaiai/iaab002`.

[40] O. Dhifallah and Y. M. Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

[41] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Physical Review Letters*, 82:2975–2978, Apr 1999. doi: 10.1103/PhysRevLett.82.2975. URL `https://link.aps.org/doi/10.1103/PhysRevLett.82.2975`.

[42] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2(1):263–286, Jan. 1995. ISSN 1076-9757.

[43] R. P. Duin. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 1–7. IEEE, 2000.

[44] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[45] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[46] E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004.

[47] J. Fan, W. Wang, and Y. Zhong. An l_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

[48] C. Fang, H. He, Q. Long, and W. J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[49] C. Fang, H. He, Q. Long, and W. J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[50] M. D. Fox and M. E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, 8(9): 700–711, 2007.

[51] S. Freytag, J. Gagnon-Bartsch, T. P. Speed, and M. Bahlo. Systematic noise degrades gene co-expression signals but can be corrected. *BMC bioinformatics*, 16 (1):309, 2015.

[52] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[53] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, Mar. 2002. ISSN 1532-4435. doi: 10.1162/153244302320884605. URL `https://doi.org/10.1162/153244302320884605`.

[54] J. A. Gagnon-Bartsch, L. Jacob, and T. P. Speed. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112, 2013.

[55] K. Gajowniczek, L. J. Chmielewski, A. Orlowski, and T. Zabkowski. Generalized entropy cost function in neural networks. In A. Lintas, S. Rovetta, P. F. Verschure, and A. E. Villa, editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, pages 128–136, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68612-7.

[56] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, February 2020. doi: 10.1088/1742-5468/ab633c. URL `https://doi.org/10.1088/1742-5468/ab633c`.

[57] S. Geng, M. Kolar, and O. Koyejo. Joint nonparametric precision matrix estimation with confounding. *arXiv preprint arXiv:1810.07147*, 2018.

[58] P. Germain, A. Lacoste, F. Laviolette, M. Marchand, and S. Shanian. A pac-bayes sample-compression approach to kernel methods. In *ICML*, 2011.

[59] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[60] Y. Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[61] T. Graepel, R. Herbrich, and J. Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[62] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[63] X. Han, V. Papyan, and D. L. Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[64] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[65] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[66] L. X. Hayden, R. Chachra, A. A. Alemi, P. H. Ginsparg, and J. P. Sethna. Canonical sectors and evolution of firms in the us stock markets. *arXiv preprint arXiv:1503.06205*, 2015.

[67] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, 2011.

[68] H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[69] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[70] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, USA, 2nd edition, 2012. ISBN 0521548233.

[71] L. Hou, C.-P. Yu, and D. Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

[72] D. Hsu, V. Muthukumar, and J. Xu. On the proliferation of support vectors in high dimensions. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 91–99. PMLR, 13–15 Apr 2021. URL `http://proceedings.mlr.press/v130/hsu21a.html`.

[73] H. Huang. Asymptotic behavior of support vector machine for spiked population model. *Journal of Machine Learning Research*, 18(45):1–21, 2017.

[74] L. Hui and M. Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[75] A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos. Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.11.099. URL `https://www.sciencedirect.com/science/article/pii/S0925231219301614`.

[76] L. Jacob, J. A. Gagnon-Bartsch, and T. P. Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, 2016.

[77] Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798, Phoenix, USA, 25–28 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v99/ji19a.html`.

[78] J. Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009.

[79] I. Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[80] A. Kammoun and M.-S. Alouini. On the precise error analysis of support vector machines. *IEEE Open Journal of Signal Processing*, 2:99–118, 2021. doi: 10.1109/OJSP.2021.3051849.

[81] A. Kammoun and M.-S. Alouini. On the precise error analysis of support vector machines. *IEEE Open Journal of Signal Processing*, 2:99–118, 2021.

[82] K. Khare, S.-Y. Oh, and B. Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4): 803–825, 2015.

[83] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[84] G. Kini and C. Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572*, 2020.

[85] G. R. Kini and C. Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4020–4024, 2021. doi: 10.1109/ICASSP39728.2021.9414099.

[86] D. Kobak, J. Lomond, and B. Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020. URL `http://jmlr.org/papers/v21/19-844.html`.

[87] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

[88] V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1 – 50, 2002. doi: 10.1214/aos/1015362183. URL `https://doi.org/10.1214/aos/1015362183`.

[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[90] H. Kumar and P. S. Sastry. Robust loss functions for learning multi-class classifiers. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 687–692, 2018. doi: 10.1109/SMC.2018.00125.

[91] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.

[92] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[93] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. *Journal of the American Statistical Association*, 99(465):67–81, 2004. doi: 10.1198/016214504000000098. URL https://doi.org/10.1198/016214504000000098.

[94] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9), 2007.

[95] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[96] Y. Lei, U. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/3a029f04d76d32e79367c4b3255dda4d-Paper.pdf.

[97] Y. Lei, U. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021, 2019. doi: 10.1109/TIT.2019.2893916.

[98] T. Liang and B. Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

[99] T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

[100] T. Liang, A. Rakhlin, and X. Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv preprint arXiv:1908.10292*, 2019.

[101] Z. Liao, R. Couillet, and M. W. Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *arXiv preprint arXiv:2006.05013*, 2020.

[102] S. Lim and S. Jahng. Determining the number of factors using parallel analysis and its recent variants. *Psychological methods*, 24(4):452, 2019.

[103] P. Lolas. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv preprint arXiv:2003.13723*, 2020.

[104] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

[105] J. Lu and S. Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 2022.

[106] S. Ma, L. Xue, and H. Zou. Alternating direction methods for latent variable gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

[107] X. Mai, Z. Liao, and R. Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361, 2019. doi: 10.1109/ICASSP.2019.8683376.

[108] D. Malzahn and M. Opper. A statistical physics approach for the analysis of machine learning algorithms on real data. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11001–P11001, nov 2005. doi: 10.1088/1742-5468/2005/11/p11001. URL `https://doi.org/10.1088/1742-5468/2005/11/p11001`.

[109] A. Maurer. A vector-contraction inequality for rademacher complexities. In R. Ortner, H. U. Simon, and S. Zilles, editors, *Algorithmic Learning Theory*, pages 3–17, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

[110] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[111] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[112] N. Meinshausen, P. Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[113] Z. Meng, B. Eriksson, and A. Hero. Learning latent variable gaussian graphical models. In *International Conference on Machine Learning*, pages 1269–1277, 2014.

[114] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová. The role of regularization in classification of high-dimensional noisy gaussian mixture. *arXiv preprint arXiv:2002.11544*, 2020.

[115] D. G. Mixon, H. Parshall, and J. Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[116] A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[117] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

[118] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[119] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

[120] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

[121] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[122] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11): L581, 1990.

[123] L. Ou-Yang, X.-F. Zhang, X.-M. Zhao, D. D. Wang, F. L. Wang, B. Lei, and H. Yan. Joint learning of multiple differential networks with latent variables. *IEEE transactions on cybernetics*, (99):1–13, 2018.

[124] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

[125] V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2015509117. URL `https://www.pnas.org/content/117/40/24652`.

[126] P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome biology*, 20(1):94, 2019.

[127] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104 (486):735–746, 2009.

[128] B. Á. Pires and C. Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

[129] B. A. Pires, M. Ghavamzadeh, and C. Szepesvári. Cost-sensitive multiclass classification risk bounds. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1391–III–1399. JMLR.org, 2013.

[130] T. Poggio and Q. Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[131] T. Poggio and Q. Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[132] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[133] T. Price, C.-Y. Wee, W. Gao, and D. Shen. Multiple-network classification of childhood autism using functional connectivity dynamics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 177–184. Springer, 2014.

[134] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[135] Z. Ren and H. H. Zhou. Discussion: Latent variable graphical model selection via convex optimization1. *The Annals of Statistics*, 40(4):1989–1996, 2012.

[136] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[137] R. M. Rifkin. *Everything old is new again: a fresh look at historical approaches in machine learning*. PhD thesis, MaSSachuSettS InStitute of Technology, 2002.

[138] S. Rosset, J. Zhu, and T. Hastie. Margin maximizing loss functions. In *NIPS*, pages 1237–1244, 2003.

[139] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[140] M. Rudelson, R. Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[141] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[142] F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ab49ef78e2877bfd2c2bfa738e459bf0-Paper.pdf.

[143] F. Salehi, E. Abbasi, and B. Hassibi. The performance analysis of generalized margin maximizers on separable data. In *International Conference on Machine Learning*, pages 8417–8426. PMLR, 2020.

[144] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[145] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[146] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[147] O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate gaussian models with iid observation noise. In *Advances in neural information processing systems*, pages 630–638, 2011.

[148] M. Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[149] P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[150] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3739–3749. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/taheri20a.html.

[151] H. Taheri, R. Pedarsani, and C. Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2773–2781. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/taheri21a.html.

[152] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

[153] C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. *Proceedings of Machine Learning Research*, 40:1683–1709, 2015.

[154] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64 (8):5592–5628, 2018.

[155] C. Thrampoulidis, S. Oymak, and M. Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8907–8920. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/6547884cea64550284728eb26b0947ef-Paper.pdf`.

[156] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[157] A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[158] F. Vallet, J.-G. Cailton, and P. Refregier. Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *EPL (Europhysics Letters)*, 9(4):315, 1989.

[159] K. R. Van Dijk, M. R. Sabuncu, and R. L. Buckner. The influence of head motion on intrinsic functional connectivity mri. *Neuroimage*, 59(1):431–438, 2012.

[160] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[161] A. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

[162] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[163] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[164] K. Wang and C. Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

[165] K. Wang, V. Muthukumar, and C. Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021.

[166] J. Weston and C. Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[167] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

[168] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[169] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

[170] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[171] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13:1059–1062, 2012.

[172] Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

[173] D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *arXiv preprint arXiv:2103.12692*, 2021.