université PARIS-SACLAY

NNT : 2019SACLC078

CentraleSupélec

Méthodes des matrices aléatoires pour l'apprentissage en grandes dimensions

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Mathématiques et Informatique

Thèse présentée et soutenue à Gif-sur-Yvette, le 16/10/2019, par

Xiaoyi MAI

Composition du Jury :

Marc Lelarge Directeur de recherche (INRIA), ENS Paris	Président
Paulo Goncalves Directeur de recherche (INRIA), ENS Lyon	Rapporteur
Jean-Philippe Vert Directeur de recherche (CNRS), Mines ParisTech	Rapporteur
Konstantin Avrachenkov Directeur de recherche (INRIA), Sophia Antipolis	Examinateur
Julien Perez Ingénieur de recherche, Naver Labs	Examinateur
Lenka Zdeborova Chargée de recherche (CNRS), CEA-Saclay	Examinatrice
Romain Couillet Professeur, CentraleSupélec	Directeur de thèse
Walid Hachem Directeur de recherche (CNRS), Université Paris-Est	Co-Directeur de thèse

Thèse de doctorat

Contents

Acknowledgements

I would like to start by thanking my supervisor Romain Couillet. Even though I was already impressed by him the day we first met, I did not fully realize at the time how lucky I was to encounter such a great supervisor. He guided my through every step of the way, without forgetting to give me sufficient freedom to develop on my own. From technical competences to presentation skills, from work ethic to career development, I have learned so much from him that I believe will benefit me my whole life. I also want to thank my co-advisor Walid Hachem. Despite his research field being less related to the topic of my PhD program, he supported me by showing a keen interest in my research and giving constructive feedback.

I want to take this opportunity to specially thank Laurent Le Brusquet for helping me embark on this adventure. In the last year of my master's studies, I began to consider seriously the option of pursuing a PhD. I then turned to him, who was in charge of my major at the time, for advice. He warmly received me and informed me about a career in research. More than that, he was the one who helped me get in contact with my PhD advisor, and was kind enough to ask every now and then about my progress and attend my PhD defense.

I must not forget to thank my fellow PhD colleagues, Lorenzo Dall'Amico, Cyprien Doz, Cosme Louart, Hafiz and Malik Tiomoko, Mohamed Seddik. Besides our vivid discussions about research work that I enjoyed tremendously, your companionship also meant a lot to me. Talking and laughing with you was the highlight of my daily life. Special thanks to Zhenyu Liao, a reliable collaborator and friend whom I was lucky enough to start this journey with and whom I could always talk to about various questions and doubts along the way.

I would like to thank GIPSA-lab for hosting me during my visit, where I had to chance to exchange with the CICS team members whose insightful opinions have contributed to my work. Also my sincere gratitude to the jury members of my defense for their supportive feedback encouraging me to pursue in the direction of my research.

In the end, I have to thank my parents for their love and support during my whole existence. There are obviously too many things that I should thank them for. As a exhaustive list would be too hard and unsuited for this occasion, I choose to focus on one specific matter. Romain told me once that one of my qualities was my natural tendency to think without presumptions. I think I owe that to my parents, for even when I was a little child, they were always so attentive to what I said, and never once tried to impose their opinions on me. Thank you, Mom and Dad, for making me feel free to have and express my own thinking.

Chapter 1

Introduction

1.1 Machine learning and the big data paradigm

Machine learning is a subfield of artificial intelligence focused on the automatic processing of data. Given a set of data samples and a learning task, the algorithms of machine learning extract information relevant to the task from the data set without explicit instructions. Naturally, the performance of machine learning algorithms is limited by the size of the input data set. The rapid growth of computational capacity has made it possible to collect and handle massive data sets with numerous features, resulting in many successful applications of machine learning methods, such as image classification, speech recognition and gene prediction. Even though superhuman performance has been achieved on certain tasks thanks to the power of big data, it is mostly done through supervised learning models and requires an immense amount of labeled samples. The costly labeling process and the limited access to data in many areas call for more efficient and flexible learning approaches. To improve on the current learning methods, it is needed to understand them on a profound level. However, the non-linear nature of the learning algorithms, which is at source of their empirical success, makes them also theoretically difficult to study. Indeed, most machine learning algorithms, even the most popular ones, have been motivated by intuitive reasoning and justified by heuristic arguments.

It has long been noticed that learning on large dimensional data presents some unique challenges, for which the term *curse of dimensionality* was used. Crucially, the intuitive arguments behind the proposition of many learning algorithms are only valid in small dimensions. An important phenomenon of the curse of dimensionality is the *concentration of distances*, which refers to the tendency of pairwise "distances" between data vectors to become indistinguishable in the limit of large dimensions. Since many learning techniques rely on the relation between geometric proximity and class "affinity" between data, their validity is in question under the distance concentration phenomenon. Consequently, many counterintuitive phenomena may occur, the explanation of which calls for a deeper understanding of high dimensional learning. Despite this strong need to unravel the learning process of large dimensional data, the theoretical research in this respect is rather underdeveloped in the literature. Most existing analyses of learning techniques assume in particular that the number n of data samples is infinitely large in comparison to their dimension p , i.e., $n/p \to \infty$, an assumption that is hardly adequate when the dimension is itself too large to be considered as negligible compared to the number of data

samples. The objective of this thesis is to analyze and improve learning methods in the modern regime of large and comparable n, p.

Since learning outcomes are random variables dependent of the input data, which only converge to deterministic values at $n \gg p$, analyzing learning algorithms with comparable n, p requires the non-trivial task of characterizing the randomness in learning results. Take linear discriminant analysis (LDA) , a simple and standard learning method, as an example. The LDA method approaches the learning problem by assuming that data instances $(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{R}^p$ the feature vectors and $y = \pm 1$ the class labels, follow a Gaussian mixture model with identical covariances, which is to say, for $y = (-1)^k$ with $k = \{1, 2\}$, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{C})$ (where we suppose that $\mathbf{C}$ has full rank). Under this assumption, the Bayes optimal solution is to assign an observation $\mathbf{x}$ to the class ± 1 by the sign of $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} - c$ for some threshold constant c, where $\boldsymbol{\beta} = \mathbf{C}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Since the statistical parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{C}$ are normally unknown in practice, they are estimated from a set of training data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ to get $\boldsymbol{\beta} = \hat{\mathbf{C}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, where $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$ is usually the maximal likelihood estimate (MLE) of $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{C})$, or some other estimates. While the performance of LDA is guaranteed to be optimal in the limit $n \gg p$ where $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}}) \to (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{C})$ for any consistent estimator $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$, the same cannot be said about the regime where n, p are commensurately large. Indeed, even with MLE, for which we have quite simple expressions of $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$ (also referred to as the sample means and the sample covariance):

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i, \quad k = \{1, 2\},$$

where we denote $i \in \mathcal{C}_k$ for $i \in \{1, \ldots, n\}$ such that $y_i = (-1)^k$; and

$$\hat{\mathbf{C}} = \frac{1}{n} \left[\sum_{i \in \mathcal{C}_1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^\mathsf{T} + \sum_{j \in \mathcal{C}_2} (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)^\mathsf{T} \right],$$

the statistical behavior of $\boldsymbol{\beta}$ is complicated to characterize at arbitrary n/p, mostly due to the existence of $\hat{\mathbf{C}}^{-1}$ in the expression of $\boldsymbol{\beta}$.

As explained earlier, the large dimensionality of modern data induces a need for theoretically advanced studies on the performance of learning algorithms away from the conventional asymptotic limit $n/p \to \infty$, at which the learned parameters, such as $\boldsymbol{\beta} = \hat{\mathbf{C}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$ in the above example of LDA, become deterministic constants. Meanwhile, it actually provides some technical advantages to place oneself under the high dimensional setting. Indeed, while the statistical behavior of $\hat{\mathbf{C}}^{-1}$, such as the distribution of its eigenvalues and the associated eigenvectors, is so far out of reach at finite n, p, it has come to the attention of some pioneer researchers that the statistical properties of random matrices like $\hat{\mathbf{C}}^{-1}$ are accessible for any finite n/p ratio (bounded away from zero) in the limit of large p. Indeed, by exploring the extra degrees of freedom provided by the large dimensionality of data, it was demonstrated first by Marchenko and Pastur in [1] that the density histogram[1] of the eigenvalues of the sample covariance matrix with data vectors $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ converges to a certain deterministic continuous distribution, now known as the Marchenko–Pastur distribution. The extension to the case

[1]This refers to the spectral measure defined by $\frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i}(t)$ where λ_i are the eigenvalues of the sample covariance matrix.

where the population covariance is allowed to be other than the identity matrix can be found in the works [2, 3] of Silverstein and Bai. Obviously, the knowledge of the spectral properties of the sample cavariance matrix $\hat{\mathbf{C}}$ is equivalent to knowing those of its inverse. As a matter of fact, many spectral investigations on random matrices like $\hat{\mathbf{C}}$ are conducted through technical manipulations involving their inverse. Based on results from random matrix theory (RMT), the performance of LDA was recently examined in [4], and its more elaborate variant QDA (quadratic discriminant analysis) in [5].

The solution of LDA is rather convenient for conducting theoretical analyses on account of its explicit form and the fact that its only non-linearity is due to the inverse of the sample covariance matrix $\hat{\mathbf{C}}^{-1}$, an extensively studied object in RMT. Most learning techniques, such as kernel methods, involve more complex non-linearities. Another complication in analyzing learning systems is that there may not exist an explicit expression of the system outcomes. Having no closed-form solution is actually common to a lot of widely used learning methods such as logistic regression, support vectors machines (SVMs), and neural networks, for which the solutions are stated as a point of minimization to some loss function. As RMT results concern usually the statistical properties of some specific explicit random matrix models, they are not adapted for characterizing implicit solutions to optimization problems which involve random matrices. Other approaches are thus needed for the study of learning algorithms with implicit solutions.

In this respect, the 'leave-one-out' perturbation technique has been proven effective by a series of contributions. The statistical behavior of robust regression with M-estimators, which in general does not assume the existence of a closed form solution, is captured in [6, 7], using such perturbation procedure. Following in the same line, the works of [8, 9] are focused on the logistic regression method for classification. The key idea of these studies is to establish statistical equations of learned parameters by capitalizing on the fact that the outcomes of algorithms remain practically unchanged after excluding one sample from the training data set, or one feature from the feature vector. This 'double leave-one-out' approach is applied in these works under the assumption that all data samples are centered with i.i.d. Gaussian features (i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$), which notably justifies the 'leave-one-feature-out' step as all features are statistically equivalent and independent. Contrarily to mixture models (like the one considered in LDA), there is no natural class separation in the cluster $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. In order to study classification problems under this setting, the authors of [8, 9] imposed the existence of a class separation signal inside the cluster $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. As such, the common classification scenarios with distinct class patterns (which are represented by different components in mixture models) are so far not covered by this kind of analyses.

Our contributions: The current big data paradigm lays grounds for the development of new mathematical tools to analyze learning algorithms in the modern regime of comparably large n, p. Unlike the existing analyses in this regime, the technical approaches developed in this thesis exploit both advanced tools of random matrix theory and leave-one-out arguments. As a result of combining the advantage of random matrix theory for handling structured data and the power of leave-one-out manipulation for tackling complex learning systems, we are able to conduct more involved analyses of machine learning algorithms under realistic mixture models. These analyses entail important consequences in applying learning methods, some of which have long been observed in practice without proper understanding, some are unknown to practitioners or even in contradiction to common beliefs. As a complete characterization of

the learning outcomes is available in our analyses, these problems can sometimes be directly addressed by simple measures of correction such as normalization (or rescaling) or improved parametrization. In some scenarios, large dimensional analyses can even spot fundamental flaws in the design of learning algorithms, and help inspire superior approaches, as was done in [10] and the follow-up work [11], as part of the contributions in this thesis. Remarkably, the theoretical results derived in this thesis closely predict the learning performance on both synthetic and real data sets, which suggests the adequacy of the mixture data models for describing the learning scenario in real applications. This observation is notably supported by the findings of [12] and [13] where the authors demonstrate that, under some conditions of concentration, a series of random objects concerning the sample covariance matrix converge in the regime of large n, p to the same limit irrespective of the actual data distribution.

1.2 Challenges for designing and understanding learning methods

1.2.1 Semi-supervised learning problem

Depending on whether the data fed into the learning model are *labelled* or *unlabelled*, the machine learning algorithms are broadly categorized as *supervised* or *unsupervised* respectively. The objective of unsupervised learning is to extract relevant structure or representation from a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, typically assumed to be drawn independently from a common distribution $\mathcal{X}$. Compared to unsupervised learning, the supervised learning approach has the advantage of being guided by the knowledge of desired outputs. Based a set of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where y_i is the target output (often called the label) of $\mathbf{x}_i$, supervised learning aims to construct a mapping from $\mathbf{x}$ to y. Again, the training samples $(\mathbf{x}_i, y_i)$ are considered as i.i.d. realizations from some underlying joint distribution $\mathcal{X} \times \mathcal{Y}$. The performance of supervised-learning algorithms is normally evaluated on the basis of their generalization capacity to unseen data points $\mathbf{x}$ outside the training samples.

Although the supervised approach has by now occupied a dominant place in real world applications thanks to its high level of accuracy, the cost of the labelling process, overly high in comparison to the collection of data, compels researchers to develop techniques using unlabelled data, as many popular learning tasks of these days, such as image classification, speech recognition and language translation, require enormous training data sets to achieve satisfying results.

The idea of semi-supervised learning comes from the expectation of improving the learning performance by combining labelled and unlabelled data, which is of significant practical value when the cost of supervised learning is too high and the performances of unsupervised approaches is too weak. Semi-supervised learning is a more accurate modelling of actual human learning process, and should surpass both supervised and unsupervised learning approaches as a result of utilizing all information, labelled and unlabelled. In spite of its natural idea, semi-supervised learning has not reached broad recognition, due to the difficult of designing methods that exploit properly labelled and unlabelled data at the same time. In fact, many standard semi-supervised learning techniques were found to exhibit worse performances than their one-sided counterparts [14, 15, 16], thereby hindering the interest for these methods.

Understandably, for semi-supervised learning to work, some assumptions should be met to ensure that both labelled and unlabelled information are beneficial to the learning task. Already in the supervised setting, since the requirement is to learn a mapping that generalizes well from a finite set of training samples to a potentially infinite set of unseen data points, *the smoothness assumption that if two data representations* $\mathbf{x}_1, \mathbf{x}_2$ *are close (with respect to a certain measure), then so are the corresponding outputs* y_1, y_2 should hold. Without the guidance of class labels, unsupervised learning can only aid in distinguishing different clusters of data points. For such information to be useful in classifying data vectors, *the cluster assumption that data points in the same cluster belong to the same class* should apply. A similar assumption is *the low density separation principle which states that data samples in different classes are separated by a low-density decision boundary.* It is easy to see the equivalence between the cluster assumption and the low-density separation assumption as the division of one cluster into different classes requires separation boundaries inside the cluster, which is a high-density region, and a decision boundary in a high density region would certainly go through a cluster, cutting it into different classes. Most learning algorithms can be seen as implementing one or several of these assumptions.

In a nutshell, there are two types of information in learning tasks: the global information contained in the underlying structure of data points $\mathbf{x}_i$, formulated by the cluster assumption (and the low-density separation principle); and the local information reflected by the connection between desired outputs y_i and feature vectors $\mathbf{x}_i$, corresponding to the smoothness assumption. With the purpose of learning from both global and local information, the meaningfulness of the semi-supervised approach relies on a combined version of all these assumptions [17]: *two data points that are close in a high-density region tend to be in the same class.*

Note in passing that the low-density separation assumption is incorporated sometimes in supervised learning methods, and so is the smoothness assumption in unsupervised approaches. For instance, the SVM algorithm [18], one of the most popular supervised method, tries to find a low-density region that separates the two classes of labelled binary data points and inside which there are few data samples. Another example is spectral clustering [19], which finds in an unsupervised manner a data representation by following the smoothness assumption (i.e., similarity in feature space implying closeness in representation space), then regroup data samples by some standard clustering technique based on this representation.

One of the challenges in semi-supervised learning is that, depending on whether or not (and how much) the learning process is guided by the presence of target outputs (i.e., the labels), implementing the learning assumptions can lead to different results that need to be reconciled. This thesis concerns specifically *graph-based methods*, which has been a highly active area of research in semi-supervised learning. The graph-based approach considers data points as nodes in a graph, connected by some edges weights w_{ij} that can be seen as a measure of similarity between any two data vectors $\mathbf{x}_i, \mathbf{x}_j$. In this context, the accordance with the smoothness assumption should be reflected by a small value of the penalty term

$$Q_s = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(y_i - y_j)^2 = \mathbf{y}^\mathsf{T}\mathbf{L}\mathbf{y}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the so-called Laplacian matrix with $\mathbf{W} = \{w_{ij}\}_{i,j=1}^n$ the weight matrix and $\mathbf{D}$ the diagonal degree matrix having $d_i = \sum_{j=1}^n w_{ij}$ as its diagonal elements, and $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}$ the label vectors containing the desired outputs y_i of data samples. In the absence of the target outputs y_i, the method of spectral clustering finds a substitution $\mathbf{f} = [f_1, \ldots, f_n]^\mathsf{T}$

of $\mathbf{y}$ as the one minimizing Q_s:

$$\min_{\mathbf{f}} \mathbf{f}^\top \mathbf{L}\mathbf{f}$$
$$\text{s.t. } \|\mathbf{f}\|^2 = n.$$

It is easy to see that $\mathbf{L}$ is a symmetric and positive definite matrix. Therefore, the solution of the above optimization is simply the eigenvector of $\mathbf{L}$ that is associated with the smallest eigenvalue 0. We obtain then $\mathbf{f} = \pm\mathbf{1}_n$, which is obviously useless for any learning tasks as it implies that all data samples are in the same class. That is why the algorithm of spectral clustering considers the second smallest eigenvector as its solution. It is also common practice to keep in addition a few other eigenvectors with smallest eigenvalues to construct a low-dimensional representation of the data samples upon which one can separate clusters using some standard low-dimensional clustering techniques, such as the k-means or expectation-maximization (EM) algorithms.

While the unsupervised method of spectral clustering is well understood [20, 21] and shown to achieve many empirical successes, the graph-based semi-supervised approach is less tractable. As said in the previous paragraph, it can be challenging to reconcile the label information and the global structure learned from the distribution of data features. Indeed, even though a smooth data representation $\mathbf{f}$ can be learned from the minimization of $\mathbf{f}^\top \mathbf{L}\mathbf{f}$, we still have to correlate it with the pre-known target outputs y_i of labelled points.

There are mainly two approaches to resolve this issue: the manifold-based methods and the regularization approach. The manifold-based methods consist in selecting a certain number m of the smallest eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of $\mathbf{L}$ expect the first one $\mathbf{v}_0$, then searching in the span of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ for an output vector $\mathbf{f}$ such that the f_i is closed to y_i for all labeled points $\mathbf{x}_i$. The complication about this approach is in the choice of the number m of the eigenvectors to take. On one hand, the more eigenvectors are selected, the more loyal is $\mathbf{f}$ to the label information. But choosing a greater m also means the inclusion of eigenvectors that correspond to less small eigenvalues, leading potentially to a less smooth $\mathbf{f}$. On the other hand, if only few of the smallest eigenvectors are taken, the label information may not be sufficiently exploited. Indeed, in the extreme scenario with $m = 1$, $\mathbf{f}$ is constrained to aligned with $\mathbf{v}_1$, leaving us with the same solution as spectral clustering. To ensure both the smoothness of $\mathbf{f}$ and its accordance with known labels y_i, the Laplacian regularization approach chooses to minimize $\mathbf{f}^\top \mathbf{L}\mathbf{f}$ while constraining $f_i = y_i$ (which is possible to relax by adding a penalty term to the optimization loss). As this optimization gives a unique solution, there is no direct way to eliminate the advantage of constant solutions $c\mathbf{1}_n$ in minimizing $\mathbf{f}^\top \mathbf{L}\mathbf{f}$. Although it is expected that the constraint $f_i = y_i$ would produce enough effect to pull $\mathbf{f}$ away from the direction of constant solutions, the power and limit of this effect remain an open question.

Due to these underlying complications induced by the step of bringing together the global and local information, it is vital to understand separately the impact of labeled and unlabeled data. Ideally, such understanding can be achieved through quantifying the learning performance as a function of the sizes of the labeled and unlabeled data sets. However, even for simple problem formulations, the solutions of which assume an explicit form, the analysis involves complicated-to-analyze mathematical objects (for instance the resolvent of kernel matrices), as is the case for the Laplacian regularization algorithms. As discussed in Section 1.1, the high dimensional analysis is particularly useful for the much needed understanding of these semi-supervised learning algorithms as it allows one to characterize the learning performance in the regime of comparably large n, p. Here in the semi-supervised setting, n refers to the sample

number of labelled or unlabelled set, we can thus quantify the effect of labelled and unlabelled data samples.

Moreover, since most semi-supervised algorithms are built upon low-dimensional reasoning, they may suffer the transition to the high dimensional setting. For example, in the ideal scenario of graph-based learning, data points $\mathbf{x}_i, \mathbf{x}_j$ in different classes are connected with extremely weak weights $w_{ij} \ll 1$. Under this assumption, all class-constant vectors $\mathbf{f}$ for which the points in the same class have the same value, such as the label vector $\mathbf{y}$, are almost as smooth as the constant solution $\mathbf{1}_n$, as the latter yields a zero value for the smoothness penalty term Q_s and the former to a value near zero (or exactly zero if $w_{ij} = 0$ for all $\mathbf{x}_i, \mathbf{x}_j$ in different classes). Therefore, the solution of Laplacian regularization should lean towards the label vector $\mathbf{y}$ as a result of imposing $f_i = y_i$. While this ideal scenario might be close to the actual situations in low dimensions, it is far from what happens for data of high dimensionality, due to the aforementioned distance concentration phenomenon. Since all large dimensional data vectors are basically at the same distance, they tend to be seen in the same class on accounts of the graph smoothness. It is then paramount to ask if the Laplacian regularization algorithm still works for high dimensional data. And if not, what are the key points in designing effective algorithms for large dimensional semi-supervised learning. These questions bring us respectively to the studies presented in Chapter 3 and Chapter 4. Other than identifying and correcting several important consequences of semi-supervised Laplacian regularization algorithms, the results of Chapter 3 point out that even though it is possible to achieve non-trivial classification performance with the Laplacian regularization approach, the method is inefficient in extracting information from large dimensional unlabelled data. In light of this critical remark, a new regularization approach is proposed in Chapter 4, allowing for an enhanced semi-supervised learning on high dimensional data.

1.2.2 Methods with implicit optimization

The previous section explained how the intricate nature of semi-supervised learning makes it difficult to study. It should be pointed out though that even the most basic algorithms in the simple setting of supervised learning can be hard to analyze, due to the lack of closed-form solution, such as the very popular methods of logistic regression and SVMs.

Indeed, recall that the objective of supervised learning is to learn a mapping $G(\mathbf{x}) = y$ from the feature space of $\mathbf{x}$ to the target output space of y, based on the knowledge of a set of training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. Usually in a particular supervised learning method, the mapping G is restricted to the space of a family of functions, controlled by a set of parameters $\mathcal{S}_p$. For instance, the function $G(\mathbf{x})$ can be constrained to take a linear form $G(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0$ with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$ being the parameters to be determined. Let $L(G(\mathbf{x}), y)$ be some pre-defined loss between the mapped output $G(\mathbf{x})$ with the target output y. The common goal of supervised learning methods is to find the best parametrization of $\mathcal{S}_p$ that minimizes the expected value for $L(G(\mathbf{x}), y)$ over the distribution $\mathcal{X} \times \mathcal{Y}$, i.e., by solving

$$\min_{\mathcal{S}_p} \mathbb{E}\{L(G(\mathbf{x}), y)\}.$$

Since we generally do not have access to the distribution $\mathcal{X} \times \mathcal{Y}$, which makes the computation of $\mathbb{E}\{L(G(\mathbf{x}), y)\}$ impossible, the average loss over the training set, called empirical risk, is used instead as an approximation of $\mathbb{E}\{L(G(\mathbf{x}), y)\}$. The actual optimization to solve in practice

becomes

$$\min_{S_p} \frac{1}{n} \sum_{i=1}^{n} L(G(\mathbf{x}_i), y_i).$$

These empirical risk minimization (ERM) algorithms are based on the convergence of the empirical risk to its expectation as $n/p \to \infty$.

A simple algorithm within this framework is the ridge regression method, which consists in training a linear classifier $G(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0$ by minimizing the average square loss $L(G(\mathbf{x}), y) = (G(\mathbf{x}) - y)^2$ over the training set, for which the solution has an explicit form[2]

$$\boldsymbol{\beta} = \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1} \mathbf{X}(\mathbf{y} - \beta_0 \mathbf{1}_n), \quad \beta_0 = \frac{\mathbf{1}_n^\mathsf{T}\mathbf{y} - \mathbf{1}_n^\mathsf{T}\mathbf{X}^\mathsf{T}\left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1}\mathbf{X}\mathbf{y}}{n - \mathbf{1}_n^\mathsf{T}\mathbf{X}^\mathsf{T}\left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1}\mathbf{X}\mathbf{1}_n}$$

where $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right]$. The performance analysis of this algorithm is within the reach of currently available tools in RMT, as the random matrices involved in the above expression of the solution have been (or similar in nature to) objects of interest in the literature.

The difficulty in understanding these seemingly elementary learning approaches is that having a closed-form solution as the ridge regression is actually a rare case which does not apply to some of the most popular algorithms. This is notably the case of the widely-used method of logistic regression, which is based on the maximal likelihood principle. Consider that the data distribution $\mathcal{X} \times \mathcal{Y}$ fits a logistic model having its conditional probability $\mathbb{P}(y|\mathbf{x})$ (where $y = \pm 1$) given by

$$\mathbb{P}(y|\mathbf{x}) = \frac{1}{1 + \mathbf{e}^{-y(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}^* + \beta_0^*)}}$$

for some unknown underlying parameters $\boldsymbol{\beta}^* \in \mathbb{R}^p$, $\beta_0^* \in \mathbb{R}$. It should be noted that the Gaussian mixture distribution $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ with identical covariance matrices $\mathbf{C}_1 = \mathbf{C}_2$ is a special case of the above logistic model. The logistic regression algorithm finds an estimate $(\boldsymbol{\beta}, \beta_0)$ of $(\boldsymbol{\beta}^*, \beta_0^*)$ that maximizes the joint conditional probability over the training set. This is equivalent to minimizing the sum of their negative log likelihood function:

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + e^{-y_i(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \beta_0)}),$$

which is retrieved by the ERM framework with the loss function $L(G(\mathbf{x}), y) = \ln(1 + e^{-y(\mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \beta_0)})$. The solution of the above optimization is given in practice as the outcome of some iterative procedures, making it hard to analyze using the standard tools of high dimensional statistics. Another example is the method of support vectors machines. Its idea of finding a hyperplane $\mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \beta_0 = 0$ separating two classes of data with a maximal distance between them is implemented by

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \beta_0)\} + \lambda \|\boldsymbol{\beta}\|^2,$$

[2]Here the solution is given under the condition $n > p$. Otherwise, the optimization problem is ill-posed with infinitely many solutions.

which is a regularized version of the ERM problem with the hinge loss $L(G(\mathbf{x}), y) = \max\{0, 1 - y_i \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta} + \beta_0)\}$ and an additional ridge regularization term $\lambda \|\boldsymbol{\beta}\|^2$.

One of the objectives of this thesis is to characterize the distribution of parameters $(\boldsymbol{\beta}, \beta_0)$ for these implicit methods in the regime of commensurately large n, p. Compared to the analyses in the limit of $n \gg p$ where the parameters $(\boldsymbol{\beta}, \beta_0)$ converge to deterministic constants, our high dimensional approach sheds light on the transitional regime where the performance of learning algorithms is sensible to the size of input data set, described by a finite n/p ratio, and where $(\boldsymbol{\beta}, \beta_0)$ remains random. Even though comprehending what happens in the setting of comparable n, p is crucial for finding the most efficient way to learn from a finite set of data samples, it remains a longstanding open question for which there exist much fewer results in the literature than for the well understood limiting case of $n \gg p$. Indeed, admitting no explicit solution is much more an issue when we are interested in an intermediate regime where the parameters stay random variables dependant of the input data, instead of converging to constant values. With the tools of high dimensional statistics, we are able to get a clear picture of the learning behavior of implicit methods such as SVMs (analyzed in Chapter 5) and logistic regression (study in Chapter 6) in this transitional regime, and answer a series of important questions about the choice of hyperparameters, the bias-variance trade-off, the optimality of learning performance, etc.

1.3 Outline and contributions

As mentioned earlier, this thesis aims to investigate involved learning methods like semi-supervised learning techniques and implicit algorithms under realistic mixture models of high dimensional data. The following chapters are organized as follows:

- On a technical level, the novelty of this thesis is the development of an approach combining the techniques of RMT and the leave-one-out procedure, adaptable to the analysis of a series of important learning problems as demonstrated by our main contributions. The basic tools of RMT and the concept of the leave-one-out manipulation are presented in Chapter 2, before the demonstration of how to combine them for more involved analyses through an illustrative example.

- Moving to the main contributions, the first part concerns semi-supervised learning on graphs, constituted of Chapter 3 and Chapter 4:

  - In Chapter 3, we present the high dimensional analysis of a family of graph-base semi-supervised learning algorithms, often referred to as the Laplacian regularization methods. Our analysis explains why most of these commonly used semi-supervised algorithms fail in high dimensions, except the one with the random walk Laplacian matrix (also known as the PageRank algorithm). The study also reveals several important consequences induced by the high dimensionality of data. Measures of correction and remarks providing practical guidance are given based on these findings. A very important conclusion from this analysis is that the performance of all the Laplacian regularization algorithms exhibits negligible growth as the size of unlabelled set increases. This suggests the existence of some fundamental flaw in the design of the Laplacian regularization approach, rendering it inadequate for performing effective

semi-supervised learning on high dimensional data. The results of this chapter are recollected from

> X. Mai, R. Couillet,"A Random Matrix Analysis and Improvement of Semi-Supervised Learning for Large Dimensional Data", Journal of Machine Learning Research, vol. 19, no. 79, pp. 1-27, 2018.
>
> X. Mai, R. Couillet,"The Counterintuitive Mechanism of Graph-based Semi-Supervised Learning in the Big Data Regime", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

– Following on the last remark from the analysis of Laplacian regularization, we proceed in Chapter 4 to design a superior regularization algorithm capable of learning effectively from both labelled and unlabelled data of high dimensionality, in the sense that the classification accuracy non-negligibly increases when one of the size ratios $n_{[l]}/p, n_{[u]}/p$ of labelled ($[l]$) and unlabelled sets ($[u]$) is larger. The proposed algorithm has an indisputable advantage over the Laplacian methods as the performance of the latter only depends on the size ratio $n_{[l]}/p$ of labelled set. Our new approach involves a key centering operation on the similarities. A thorough performance analysis is also conducted. The proposed method and its analysis presented in this chapter are based on the following contributions

> X. Mai, R. Couillet, "Revisiting and Improving Semi-Supervised Learning: A Large Dimensional Approach", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019.
>
> X. Mai, R. Couillet, "Consistent Semi-Supervised Graph Regularization for High Dimensional Data", submitted to Journal of Machine Learning Research, 2019.

- The second part focuses on the study of implicit algorithms, with Chapter 5 devoted to the method of SVM and Chapter 6 to logistic regression

  – The method of support vectors machines owes its name to the fact that the learned parameter $\boldsymbol{\beta}$ is determined by a subset of training samples, called the support vectors. In fact, since we have $\boldsymbol{\beta} = \sum_{i=1}^{n} c_i \mathbf{x}_i$ where $c_i \geq 0$, the training data vector $\mathbf{x}_i$ associated with a non-zero c_i is a support vector. We characterize in Chapter 5 the behavior of support vectors in high dimensions through the statistical distribution of c_i. Then we show how the statistical distribution of $\boldsymbol{\beta}$ is related to that of c_i, which allows us to derive a series of important conclusions about the impact of hyperparameter in the SVM method. This analysis is presented in the article

  > X. Mai, R. Couillet, "Statistical Behavior and Performance of Support Vector Machines for Large Dimensional Data", in preparation, 2019.

  – As explained in Section 1.2.2, logistic regression is one of the algorithms defined by the empirical risk minimization principle, with a negative log likelihood loss. Since logistic regression gives a maximum likelihood estimate of the parameters $\boldsymbol{\beta}, \beta_0$, an often used default option in practice which is commonly believed to be optimal when the assumption of data distribution is met, we propose to verify the optimality of logistic regression through a joint analysis of the empirical risk minimization algorithms with smooth loss function (as opposed to the non-smooth hinge loss function

in SVMs). Remarkably, our results prove that contrary to common belief, the maximum likelihood based logistic regression does not produce the best classification performance. We also devise strategies of improvement for these algorithms based on our theoretical findings before investigating the limitations of these strategies. The chapter gathers material from the following contributions

X. Mai, Z. Liao, R. Couillet, "A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019.

X. Mai, Z. Liao, "High Dimensional Classification via Empirical Risk Minimization: Statistical Analysis and Optimality", in preparation, 2019.

Chapter 2

Technical tools

As mentioned in the introduction of Chapter 1, the analyses presented in this dissertation are placed under the modern regime where the number n of data samples and their dimension p are large and comparable. As a consequence of comparable n, p, random objects in this regime exhibit different statistical properties than at the conventional asymptotic limit $n \gg p$. As an example, consider the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ with i.i.d. Gaussian vectors $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{C})$. Even though

$$\mathbb{E}\{\hat{\mathbf{C}}\} = \mathbf{C}$$

holds for all n, p as $\hat{\mathbf{C}}$ is an unbiased estimator of the population covariance $\mathbf{C}$, the same cannot be said about its resolvent $\left(\hat{\mathbf{C}} - z\mathbf{I}_p\right)^{-1}$ where $z \in \mathbb{C}$ is a value different from the eigenvalues of $\hat{\mathbf{C}}$. As a matter of fact, in the limit of large n, p, we have the following convergence

$$\left\| \mathbb{E}\left\{ \left(\hat{\mathbf{C}} - z\mathbf{I}_p\right)^{-1} \right\} - \left(\gamma(z)\mathbf{C} - z\mathbf{I}_p\right)^{-1} \right\| \to 0$$

for some $\gamma(z)$ dependent of the n/p ratio and different from 1 except at $n/p \to \infty$. The resolvent of $\hat{\mathbf{C}}$ and other (mainly polynomial) functionals of the data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ the are often involved in the solution of machine learning algorithms such as the famous LDA method (as explained in Chapter 1). As such, RMT can be used to study the performance of these algorithms by determining the statistical parameters of their solution. Another focus of RMT is the spectral properties of random matrices, e.g., the distribution of the eigenvalues. This kind of results are notably useful in the study of eigenvector-based learning methods like principal component analysis (PCA). In PCA, we find the directions upon which the data samples vary the most as being the eigenvectors of the sample covariance matrix $\hat{\mathbf{C}}$ associated with the largest eigenvalues. To understand the principle of PCA, imagine that there are two classes, represented by the two Gaussian distribution $\mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$, $k \in \{1, 2\}$. In the case of the null signal (i.e., $\boldsymbol{\mu} = \mathbf{0}_p$), it is a well-known result in RMT that the distribution of the eigenvalues of $\hat{\mathbf{C}}$ converges in the limit of large n, p to a certain distribution with a bounded support $\mathcal{S}$. In the presence of the class signal $\boldsymbol{\mu} \neq \mathbf{0}_p$, the sample covariance matrix $\hat{\mathbf{C}}$ falls under the spiked model where the matrix in question can be seen as the sum of a full-ranked non-informative random matrix plus a low-ranked perturbation of interest. There exists a threshold for the signal strength $\|\boldsymbol{\mu}\|$ under or above which we have respectively the absence or the existence of an isolated eigenvalue

outside the bounded support $\mathcal{S}$ with high probability, and only the isolated eigenvector (the one associated with the isolated eigenvalue) has a non-negligible alignment with the class signal $\boldsymbol{\mu}$. Thus PCA is asymptotically ineffective for extracting the class signal $\boldsymbol{\mu}$ below the threshold. This phase transition phenomenon is well-studied in the literature of RMT, where the theorems concerning the condition of the phase transition, the location of the isolated eigenvalue, the alignment of the isolated eigenvector with a certain deterministic direction were established.

To sum up, the tools of RMT are applicable to the study of learning methods with explicit solution involving random matrices (such as LDA) or those exploiting the spectral information of random matrices (like PCA). We will present some basic concepts of these tools in Section 2.1.

In contrast, if the solution of a learning method can be expressed neither explicitly nor in the form of spectral information, we may need to resort to other technical tools, such as the approximate message passing (AMP) approach [7], the replica method from statistical physics [22], and the leave-one-out procedure [6]. While the first two approaches consist in reapplying some mathematical models previously established in other contexts to the problem at hand, the leave-one-out procedure attacks the problem directly by adopting a basic idea of perturbation. We focus here on the leave-one-out procedure for its flexibility ad interpretability. This technique has been successfully employed in the study of robust regression with M-estimator [6, 23] and in the study of logistic regression for binary classification [8]. It is important to note that in these studies, the leave-one-out perturbation technique is conducted twice for the leave-one-observation-out step and the leave-one-feature (or predictor)-out step, hence referred to as the double leave-one-out procedure. A common point of these analyses [6, 23, 8] is that the solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ (the estimated regression vector) to the method under study is given implicitly in the form of an equation involving non-linear functions of $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i$, $i \in \{1, \ldots, n\}$, where $\mathbf{x}_i$ are the training samples. Due to the non-negligible dependence between $\hat{\boldsymbol{\beta}}$ and $\mathbf{x}_i$ in the regime of comparable n, p, there is no way to study the statistical distribution of $\hat{\boldsymbol{\beta}}$ through that of $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i$ without knowing this dependence.

Generally speaking, in the double leave-one-out procedure, the leave-one-observation-out step starts with the definition of the solution $\hat{\boldsymbol{\beta}}_{(i)}$ obtained by removing the i-th data sample from the training process. As $n \to \infty$, it is intuitively clear that the difference between $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}$ is negligible. With this argument and some standard concentration results, we can express $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i$ as a function of $\hat{\boldsymbol{\beta}}_{(i)}^\mathsf{T}\mathbf{x}_i$, with the help of an unknown constant κ_i. Once we replace $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i$ with $\hat{\boldsymbol{\beta}}_{(i)}^\mathsf{T}\mathbf{x}_i$ in the equation defining the solution of the learning algorithm, it is possible to study the distribution of $\hat{\boldsymbol{\beta}}_{(i)}$ (which is practically the same as $\hat{\boldsymbol{\beta}}$) from this equation, as $\hat{\boldsymbol{\beta}}_{(i)}$ is (by definition) independent of $\mathbf{x}_i$. The determination of the constant κ_i, and more importantly, the essential statistical properties of $\boldsymbol{\beta}$ was achieved in these previous studies through a leave-one-feature-out step. We will explain that in Section 2.2 through an example of M-estimation for regression.

Most of the arguments in the leave-one-out perturbation technique make sense when there are numerous independent copies of the random object concerned in the leave-one-out manipulation. Even though the leave-one-observation-out manipulation agrees with the common assumption in machine learning that data samples are i.i.d. realizations from a certain distribution, the double leave-one-out procedure implies also a leave-one-feature-out step, and thus relies on the statistical equivalence and independence among the data features, i.e., among the elements of

a data vector $\mathbf{x}_i$, which obviously does not apply to a lot of real-world applications. More than that, the statistical equivalence and independence among the entries of the underlying regression/classification signal $\boldsymbol{\beta}^*$ (the sought-for solution) should also be imposed. To work around these restrictive conditions, we propose instead to determine the unknown constants κ_i with advanced tools of RMT, in place of the restrictive leave-one-feature-out approach..

2.1 Basics of random matrix theory

RMT is motivated by the fact that matrix-formed random objects can not be seen as a mere collection of its random entries. As an example, even though we have the *joint point-wise convergence* for the sample covariance matrix $\hat{\mathbf{C}}_p$ (as defined in the previous paragraph) to the identity matrix in the limit of large n, p:

$$\max_{1\leq i,j\leq p}\left|\left\{\hat{\mathbf{C}}-\mathbf{I}_p\right\}_{ij}\right| = \max_{1\leq i,j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n}x_{ki}x_{kj}^*-\delta_{ij}\right| \xrightarrow{\text{a.s.}} 0,$$

the convergence in spectral norm of $\hat{\mathbf{C}}_p$ to $\mathbf{I}_p$ does not hold. One of the most important discoveries in the early development of RMT is that as n, p grow large with the ratio n/p converging to a positive value, the distribution of the eigenvalues of $\hat{\mathbf{C}}_p$ converges weakly and almost surely to some continuous deterministic distribution, which is spread far from 1, where the eigenvalues of $\mathbf{I}_p$ are concentrated. The limiting distribution of $\hat{\mathbf{C}}_p$ was mathematically formulated in [24], under the name of *Marcenko-Pastur law*.

This convergence result can be derived by the moment method (Section 30 of [25]) or by the Sieltjes transform method, the latter will be presented in this section. The basic idea of the Sieltjes transform method is that the limiting eigenvalue distribution of a random matrix, e.g., $\hat{\mathbf{C}}_p$, can be directly accessed from a limiting function $m_\mu(z)$ of $m_{\mu_p}(z) = \frac{1}{p}\operatorname{tr}\left(\hat{\mathbf{C}}_p - z\mathbf{I}_p\right)^{-1}$, which is the Sieltjes transform of $\hat{\mathbf{C}}_p$. The study on the spectral distribution of $\hat{\mathbf{C}}_p$ is thus reduced to finding the limit of $\frac{1}{p}\operatorname{tr}\left(\hat{\mathbf{C}}_p - z\mathbf{I}_p\right)^{-1}$. This brings us to another important remark (already mentioned in the beginning of this chapter) showing that $\hat{\mathbf{C}}_p$ behaves differently than $\mathbf{I}_p$ despite the joint point-wise convergence: $\frac{1}{p}\operatorname{tr}\left(\hat{\mathbf{C}}_p - z\mathbf{I}_p\right)^{-1}$ does not converge to $\frac{1}{p}\operatorname{tr}\left(\mathbf{I}_p - z\mathbf{I}_p\right)^{-1}$ in the limit of large n, p.

We will introduce in the following the results of the Marcenko-Pastur law, the technique of the Sieltjes transform used in its original derivation of [24], and the definition and the derivation of deterministic equivalents for the resolvent $\left(\hat{\mathbf{C}}_p - z\mathbf{I}_p\right)^{-1}$ of $\hat{\mathbf{C}}$, which gives directly the Sieltjes transform of $\hat{\mathbf{C}}_p$.

2.1.1 Limiting distribution of eigenvalues

Definition 2.1. *For a Hermitian matrix $\mathsf{H}_p \in \mathbb{C}^{p\times p}$, we define its empirical spectral measure μ^{H_p} for its eigenvalues as*

$$\mu^{\mathsf{H}_p}(A) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\lambda_i \in A}$$

for measurable $A \subset \mathbb{R}$, where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of H_p.

The Marcenko-Pastur law

Theorem 2.1.1. *Consider the sample covariance matrix $\hat{\mathbf{C}}_p = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}$ with $\mathbf{x}_i \in \mathbb{C}^p$ i.i.d. random vectors of independent entries with zero mean and unit variance. As $n, p \to \infty$ with $\frac{n}{p} \to c \in (0, \infty)$, the empirical spectral measure of $\hat{\mathbf{C}}_p$ converges almost surely to a deterministic measure μ_c with density f_c given by*

$$\mu_c(dx) = (1 - c^{-1})^+ 1_{\{0\}}(x) + \frac{1}{2\pi c x}\sqrt{(x-a)^+(b-x)^+}dx,$$

for all $x \in \mathbb{R}$, where $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$.

The Stieltjes transform

Definition 2.2. *Let μ be a probability measure. Then the Stieltjes transform $m_\mu(z)$, for $z \in \mathrm{Supp}(\mu)^c$,[1] the complex space complementary to the support of μ is defined as*

$$m_\mu(z) \triangleq \int_{-\infty}^{\infty} \frac{1}{\lambda - z} d\mu(\lambda).$$

The Stieltjes transform of the empirical spectral measure for the sample covariance matrix $\hat{\mathbf{C}}_p$ can be written as $m_{\mu^{\hat{\mathbf{C}}_p}}(z) = \frac{1}{p}\operatorname{tr}\left(\hat{\mathbf{C}} - z\mathbf{I}_p\right)^{-1}$, since

$$\begin{aligned} m_{\mu^{\hat{\mathbf{C}}_p}}(z) &= \int_{-\infty}^{\infty} \frac{1}{t-z} d\mu^{\hat{\mathbf{C}}_p}(t) \\ &= \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i - z} \\ &= \frac{1}{p}\operatorname{tr}\left(\hat{\mathbf{C}} - z\mathbf{I}_p\right)^{-1} \end{aligned}$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\hat{\mathbf{C}}_p$.

Importantly, the convergence of $m_{\mu^{\hat{\mathbf{C}}_p}}(z)$ to the Stieltjes transform $m_\mu(z)$ of a deterministic probability measure implies the convergence of $\mu^{\hat{\mathbf{C}}_p}$ to μ.

[1] The support of $\mathrm{Supp}(\mu)$ of a probability measure μ with density f is defined as the closure of the set $\{x \in \mathbb{R} | f(x) > 0\}$.

Theorem 2.1.2. *Let $\mu_1, \mu_2, \ldots$ be a series of probability measures with bounded support. If there exists a probability measure μ such that*

$$m_{\mu_p(z)} \to m_\mu(z)$$

for $z \in \mathcal{D}$ with $\mathcal{D}$ a subset of $\mathbb{C}^+ = \{z \in \mathbb{C} | \Im z > 0\}$ containing a limit point, then

$$\mu_p(A) \to \mu(A)$$

for $A \subset \mathbb{R}$.

The inverse mapping from the Stieltjes transform to the corresponding probability measure is given in the below theorem.

Theorem 2.1.3. *If a, b are continuity points of μ, i.e. $\mu(\{a\}) = \mu(\{b\}) = 0$, then*

$$\mu([a,b]) = \frac{1}{\pi} \lim_{y \to 0^+} \int_a^b \Im \left[m_\mu(x+iy) \right] dx.$$

And for all $x \in \mathbb{R}$,

$$\mu(\{x\}) = \lim_{y \to 0^+} y \Im \left[m_\mu(x+iy) \right].$$

With the above properties of the Stieltjes transform, we can study the convergence of the empirical spectral measure $\mu^{\hat{\mathbf{C}}_p}$ by working on the convergence of its Stieltjes transform $m_{\mu^{\hat{\mathbf{C}}_p}}(z)$, which can be achieved by finding a deterministic equivalent (as will be defined in the following subsection) of the resolvent $\left(\hat{\mathbf{C}} - z\mathbf{I}_p \right)^{-1}$.

2.1.2 Deterministic equivalents

Definition 2.3. *Consider a series of Hermitian random matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots$ with $\mathbf{A}_p \in \mathbb{C}^{p \times p}$. A deterministic equivalent of $\mathbf{A}_p$ is a series of deterministic matrices $\mathbf{B}_1, \mathbf{B}_2, \ldots$ with $\mathbf{B}_p \in \mathbb{C}^{p \times p}$, such that for a deterministic matrix $\mathbf{C}$ of bounded spectral norm and deterministic vectors $\mathbf{a}, \mathbf{b}$ of bounded norms, we have*

$$\frac{1}{p} \operatorname{tr} \mathbf{C}\mathbf{A}_p - \frac{1}{p} \operatorname{tr} \mathbf{C}\mathbf{A}_p \xrightarrow{\text{a.s.}} 0$$
$$\mathbf{a}^* \mathbf{A}_p \mathbf{b} - \mathbf{a}^* \mathbf{B}_p \mathbf{b} \xrightarrow{\text{a.s.}} 0$$

We will use the notation $\mathbf{A}_p \leftrightarrow \mathbf{B}_p$ to stand for the fact that $\mathbf{B}_p$ is a deterministic equivalent of $\mathbf{A}_p$.

As an example, here we are interested in the deterministic equivalent of the resolvent

$$\mathbf{Q}(z) = \left(\hat{\mathbf{C}} - z\mathbf{I}_p \right)^{-1} \tag{2.1}$$

for the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\mathsf{H}}$ with i.i.d. random vectors $\mathbf{x}_i = \mathbf{C}^{\frac{1}{2}} \mathbf{z}_i$ for $\mathbf{C} \in \mathbb{C}^{p \times p}$ some Hermitian deterministic matrix, $\mathbf{z}_i \in \mathbb{C}^{p \times 1}$ random vectors of i.i.d. entries with

zero mean, unit variance and finite fourth order moment. It was shown in [26] that $\mathbf{Q}(z)$ has a deterministic equivalent $\bar{\mathbf{Q}}(z)$, for which we have $\|\bar{\mathbf{Q}}(z) - \mathbb{E}\{\mathbf{Q}(z)\}\| = o(1)$. We refer the interested readers to Section 6.2 of [27] for more detailed results and a presentation of the *Bai and Silverstein method* that can be used to prove these results. Here we present a quick way to obtain the expression of $\bar{\mathbf{Q}}(z)$ through simple manipulations involving the lemmas below. The first lemma concerns a rank-1 perturbation of the resolvent $\mathbf{Q}$, obtained by the Sherman-Morrison formula.

Lemma 2.1. *Define* $\mathbf{Q}_{-i} = \left(\frac{1}{n}\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j^{\mathsf{H}} - z\mathbf{I}_p\right)^{-1}$, *we have*

$$\mathbf{Q} = \mathbf{Q}_{-i} - \frac{1}{n}\frac{\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}\mathbf{Q}_{-i}}{1+\frac{1}{n}\mathbf{x}_i^{\mathsf{H}}\mathbf{Q}_{-i}\mathbf{x}_i}$$

The particularity of $\mathbf{Q}_{-i}$ is that it is independent of $\mathbf{x}_i$. The second lemma states the convergence of $\mathbf{x}_i^{\mathsf{H}}\mathbf{Q}_{-i}\mathbf{x}_i$, which is a standard concentration result in RMT, known as the *trace lemma*.

Lemma 2.2. *For* $\mathbf{Q}_{-i}$ *defined in Lemma 2.1, we have*

$$\frac{1}{n}\mathbf{x}_i^{\mathsf{H}}\mathbf{Q}_{-i}\mathbf{x}_i - \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}_{-i} \to 0$$
$$\frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}_{-i} - \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}} \to 0$$

With the above lemmas, we get

$$\begin{aligned}
\mathbf{I}_p =& \mathbb{E}\left\{\mathbf{Q}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathsf{H}} - z\mathbf{I}_p\right)\right\}\\
=& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}\right\} - z\mathbb{E}\{\mathbf{Q}\}\\
=& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\frac{\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^{\mathsf{H}}}{1+\frac{1}{n}\mathbf{x}_i^{\mathsf{H}}\mathbf{Q}_{-i}\mathbf{x}_i}\right\} - z\mathbb{E}\{\mathbf{Q}\}\\
=& \mathbb{E}\{\mathbf{Q}\}\left(\frac{\mathbf{C}}{1+\frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}} - z\mathbf{I}_p\right) + o_{\|\cdot\|}(1)
\end{aligned} \tag{2.2}$$

Therefore,

$$\bar{\mathbf{Q}} = \left(\frac{\mathbf{C}}{1+\frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}} - z\mathbf{I}_p\right)^{-1}.$$

We summarize these results in the below theorem.

Theorem 2.1.4. *Let* $\mathbf{Q}(z)$ *be given by* (2.1) *for some* $z \in \mathbb{C}$ *such that* $\hat{\mathbf{C}} - z\mathbf{I}_p$ *is invertible, then*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = \left(\frac{\mathbf{C}}{1+\gamma(z)} - z\mathbf{I}_p\right)^{-1}$$

where $\gamma(z)$ satisfies the equation

$$\gamma(z) = \frac{1}{n} \operatorname{tr} \mathbf{C} \left(\frac{\mathbf{C}}{1+\gamma(z)} - z\mathbf{I}_p \right)^{-1} \tag{2.3}$$

for all such z.

Note that (2.3) can admit more than one solution of $\gamma(z)$. In fact, some other constraints should be imposed on $\gamma(z)$ (or a functional of $\gamma(z)$) to ensure the uniqueness of solution, again we refer to Section 6.2 of [27] for more information on this issue. In some cases, the uniqueness of $\gamma(z)$ is easy to see. For instances, when $n > p$ (i.e., the limit c of n/p ratio is greater than 1), we can take $z = 0$, Equation 2.3 then becomes

$$\gamma = c^{-1}(1+\gamma),$$

which assumes a unique solution $\gamma = \frac{c}{c-1}$.

The derivation of the theoretical results in Chapter 4 relies mainly on the RMT tools (rank-1 perturbation, concentration arguments like the trace lemma, etc), while the analyses of Chapters 4–6 resort to the leave-one-out procedure (which will be presented in the next section), in addition to the usage of a manipulation similar to (2.2) for finding the deterministic equivalents of some random matrices involved in these analyses.

2.2 Leave-one-out procedure for handling implicit solutions

2.2.1 General framework

Consider $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ the feature vectors of training samples and $y_1, \ldots, y_n$ the corresponding target outputs ($y_i = \pm$ for binary classification methods like logistic regression and SVMs, or $y_i \in \mathbb{R}$ in the case of regression). A large family of learning methods consists in finding a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and a residual term $\boldsymbol{\beta}_0$ that minimize the empirical loss

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i + \beta_0, y_i)$$

for some loss function ρ. It is often desired that $\boldsymbol{\beta}$ has a small norm, which is achieved by adding a ridge regularization term to the minimization problem:

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i + \beta_0, y_i) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|$$

where $\lambda \geq 0$ is a preset hyperparameter that controls the level of regularization (the unregularized solution, if well-defined, is simply retrieved at $\lambda = 0$).

In general, such learning methods do not admit an explicit solution. We consider thus the following framework of implicit solutions that is satisfied by several most important learning algorithms.

$$\lambda\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n} c_i \mathbf{x}_i \tag{2.4}$$

where c_i is given by (2.5) or (2.6) as detailed below, and for which we have $\sum_{i=1}^{n} c_i = 0$.

When $\rho(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i)$ is differentiable with respect to $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0$, we simply have

$$c_i = \psi(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) \tag{2.5}$$

where $\psi(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) = -\frac{\partial\rho(t,y_i)}{\partial t}|_{t=\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i+\beta_0}$.

Since $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i = \boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + c_i\|\mathbf{x}_i\|$ with $\boldsymbol{\eta}_{(-i)} = \frac{1}{n}\sum_{j\neq i} c_j\mathbf{x}_j$, we get $c_i = \psi(\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + c_i\|\mathbf{x}_i\|, y_i)$, which entails second expression of c_i:

$$c_i = g_{\|\mathbf{x}_i\|^2}(\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) \tag{2.6}$$

where

$$g_{\|\mathbf{x}_i\|^2}(\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) = \frac{\mathrm{prox}_{\|\mathbf{x}_i\|^2}(\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) - \boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \beta_0}{\|\mathbf{x}_i\|^2}$$

with the proximal mapping

$$\mathrm{prox}_t(a, b) = \mathrm{argmin}_{a'\in\mathbb{R}}\left(\rho(a', b) + \frac{(a-a')^2}{2t}\right). \tag{2.7}$$

The advantage of (2.6) is that it is well defined for non-differentiable ρ by introducing a proximal mapping.

This framework covers the methods studied in Chapters 4–6. For starters, in the logistic regression method and other algorithms of empirical loss minimization with smooth convex loss functions analyzed in Chapter 6, we have c_i directly given by (2.5).

In the SVM method investigated in Chapter 5, we recover the solution with c_i given by (2.6) for the hinge loss

$$\rho(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0, y_i) = \max\{0, 1 - y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0)\},$$

which is a non-smooth loss function.

For the kernel method discussed in Chapter 4, the framework (2.4) is satisfied not with $\mathbf{x}_i$, but with a known functional (mapping) of $\mathbf{x}_i$. We stick here to the notation of $\mathbf{x}_i$ for convenience. As a semi-supervised learning method involving labelled and unlabelled data, its solution is retrieved by letting $c_i = y_i$ if $\mathbf{x}_i$ is labelled with $y_i = \pm 1$ and $c_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$ (i.e., with $\beta_0 = 0$) if $\mathbf{x}_i$ is unlabelled, in which case c_i is the sought-for solution allowing for the classification of $\mathbf{x}_i$. Additionally, it is interesting to note that in the unsupervised method of PCA, the solution $\boldsymbol{\beta}$, which is an eigenvector of the matrix $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$, is also covered by this framework with λ the largest eigenvalue of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ and $c_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$.

The objective of this section is to explain the main arguments in the application of the leave-one-out procedure for studying the implicit solutions under the above framework, without going into the details of rigorous proofs.

2.2.2 Common step: leave-one-observation-out

Assumption 2.1. *The training data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are i.i.d. observations from a common distribution $\mathcal{D}$. For all $i \in \{1, \ldots, n\}$, $\mathbf{x}_i - \mathbb{E}\{\mathbf{x}_i\} = \mathbf{C}_i^{\frac{1}{2}}\mathbf{z}_i$ for some random vector $\mathbf{z}_i \in \mathbb{R}^p$ of i.i.d. entries with zero mean, unit variance and finite fourth order moment, $\mathbf{C}_i \in \mathbb{R}^{p\times p}$ some symmetric matrix with $\|\mathbf{C}_i\| = O(1)$ and $\|\mathbf{C}_i^{-1}\| = O(1)$, and $\|\mathbb{E}\{\mathbf{x}_i\}\| = O(1)$.*

The ratio $c = \frac{n}{p}$ is uniformly bounded in $(0, +\infty)$ for arbitrarily large p.

In the analyses of this thesis and other works falling under the common large dimensional setting described by Assumption 2.1, where the data vectors $\mathbf{x}_i \in \mathbb{R}^p$ have non-negligible, independent variations in all p directions, and with the number n of the data samples comparable to p, it is found that $\boldsymbol{\beta}$ is a random vector for which $\|\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\}\|$ remains non-negligible, while β_0 converges to a deterministic constant. Therefore, we can treat β_0 as a deterministic constant when studying the statistical behavior with the help of (2.4), then find the value of β_0 with the condition $\frac{1}{n}\sum_{i=1}^n c_i = 0$. For convenience, we let here $\beta_0 = 0$ so that $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0$ can be simplified as $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$. The general discussion on $\boldsymbol{\beta}$ with arbitrary β_0 follows exactly the same reasoning. Again, to improve readability we consider here c_i given by (2.5) (for differentiable ρ). The derivation for the more general form (2.6) can be done with a similar reasoning, leading to results of the same nature.

The main difficulty of studying $\boldsymbol{\beta}$ with (2.4) lies in the statistical characterization of the left-hand term in (2.4). Imagine that if $\boldsymbol{\beta}$ was independent of $\mathbf{x}_i$, then the statistical distribution of the left-hand term was known as a function of the statistical distributions of $(\mathbf{x}_i, y_i)$ and $\boldsymbol{\beta}$. Obviously, this is a false statement as $\boldsymbol{\beta}$ is implicitly dependent of $\mathbf{x}_i$. Thus a crucial task in these analyses is to characterize this implicit dependence between $\mathbf{x}_i$ and $\boldsymbol{\beta}$.

The key idea of the leave-one-observation-out step is to introduce a leave-one-observation-out version $\boldsymbol{\beta}_{(-i)}$ of the solution $\boldsymbol{\beta}$, obtained by removing a training sample $\mathbf{x}_i$ from the learning process. Precisely, $\boldsymbol{\beta}_{(-i)}$ is the solution given by the minimization below

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{j\neq i} \rho(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_j, y_j) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2.$$

Thus $\boldsymbol{\beta}_{(-i)}$ is by definition independent of $\mathbf{x}_i$. As there are numerous data samples, it makes sense that[2]

$$\boldsymbol{\beta}_{(-i)} \simeq \boldsymbol{\beta}.$$

While $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ is understandably quite different from $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$ due to the independence between $\boldsymbol{\beta}_{(-i)}$ and $\mathbf{x}_i$, for other data vectors $\mathbf{x}_j$ still included in the training of $\boldsymbol{\beta}_{(-i)}$, we should have

$$\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_j \simeq \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_j, \quad j \neq i.$$

To better understand this remark, it helps to point out that $\boldsymbol{\beta} - \boldsymbol{\beta}_{(-i)} \simeq \frac{1}{n}s_i x_i$ for some scalar s_i (as will be demonstrated in the following). Clearly, as $\mathbf{x}_i^\mathsf{T}\mathbf{x}_i \gg \mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ for $j \neq i$, $(\boldsymbol{\beta} - \boldsymbol{\beta}_{(i)})^\mathsf{T}\mathbf{x}_i$ is not negligible even when $(\boldsymbol{\beta} - \boldsymbol{\beta}_{(i)})^\mathsf{T}\mathbf{x}_j$ are so.

[2]The notation $\simeq$ is understood as follows. For two sequences $s_1(p), s_2(p)$ of scalars, $s_1 \simeq s_2$ if $|s_1 - s_2|/\min\{|s_1|, |s_2|\} \to 0$. As to the multidimensional objects, we write $\mathbf{v}_1 \simeq \mathbf{v}_2$ when $\mathbf{v}_1(p), \mathbf{v}_2(p) \in \mathbb{R}^p$ are two sequences of vectors with $\|\mathbf{v}_1 - \mathbf{v}_2\|/\min\{\|\mathbf{v}_1\|, \|\mathbf{v}_2\|\} \to 0$, and for $\mathbf{M}_1(p), \mathbf{M}_2(p) \in \mathbb{R}^{p\times p}$ two sequences of matrices, $\mathbf{M}_1 \simeq \mathbf{M}_2$ indicates $\mathrm{tr}(\mathbf{M}_1 - \mathbf{M}_2)^2/\min\{\mathrm{tr}(\mathbf{M}_1)^2, \mathrm{tr}(\mathbf{M}_2)^2\} \to 0$.

Since

$$\lambda \boldsymbol{\beta}_{(-i)} = \frac{1}{n}\sum_{j\neq i}^{n} \psi(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_j, y_j)\mathbf{x}_j,$$

subtracting the above equation from (2.4), we get

$$\begin{aligned}\lambda\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{(-i)}\right) &= \frac{1}{n}\sum_{j\neq i}^{n}\left[\psi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j, y_j) - \psi(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_j, y_j)\right]\mathbf{x}_j + \frac{1}{n}c_i\mathbf{x}_i \\ &= \frac{1}{n}\sum_{j\neq i}^{n}\frac{\partial\psi(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j, y_j)}{\partial\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j, y_)\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{(-i)}\right) + \frac{1}{n}c_i\mathbf{x}_i.\end{aligned}$$

It follows that

$$\boldsymbol{\beta}-\boldsymbol{\beta}_{(-i)} = \left(\lambda\mathbf{I}_p + \frac{1}{n}\sum_{j\neq i}^{n} d_j\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1}\frac{1}{n}c_i\mathbf{x}_i$$

where $d_j = \frac{\partial\psi(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_j, y_j)}{\partial\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_j} > 0$ (due to the convexity of ρ). Then,

$$\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i - \boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i = \frac{1}{n}c_i\mathbf{x}^T\left(\lambda\mathbf{I}_p + \frac{1}{n}\sum_{j\neq i}^{n} d_j\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1}\mathbf{x}_i.$$

Since $\mathbf{x}_i$ is independent of the matrix $\lambda\mathbf{I}_p + \frac{1}{n}\sum_{j\neq i}^{n} d_j\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}$, standard results from RMT such as the trace lemma (Lemma 2.2) suggest the following convergence (under some conditions on ρ and the data distribution)

$$\frac{1}{n}\mathbf{x}^T\left(\lambda\mathbf{I}_p + \frac{1}{n}\sum_{j\neq i}^{n} d_j\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1}\mathbf{x}_i \to \kappa_i.$$

for some deterministic constant $\kappa_i = O(1)$. This amounts to

$$\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i \simeq \boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i + \kappa_i c_i,$$

and consequently

$$c_i \simeq \psi(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i + \kappa_i c_i, y_i).$$

In the end, we get

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

where

$$g_{\kappa_i}(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i, y_i) = \frac{\mathrm{prox}_{\kappa_i}(\boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i, y_i) - \boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_i}{\kappa_i}$$

with the proximal mapping $\text{prox}_t(a, b)$ given by (2.7).

As such, the leave-one-observation-out step allows us to express c_i as a function of the product $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ of two independent random vectors $\boldsymbol{\beta}_{(-i)}$ and $\mathbf{x}_i$, parametrized by a certain constant κ_i (the value of which remains to be determined). As a result, (2.4) becomes

$$\lambda\boldsymbol{\beta} \simeq \frac{1}{n}\sum_{i=1}^{n} g_{\kappa_i}(\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i, y_i)\mathbf{x}_i. \tag{2.8}$$

2.2.3 Establishing systems of equations: double leave-one-out method with a second leave-one-feature-out step

After dealing with the statistical dependence inside $c_i = \psi(\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_j, y_j)$ by approximating it with $g_{\kappa_i}(\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i, y_i)$ in the above leave-one-observation-out step, we shall try to address the dependence between c_i and $\mathbf{x}_i$ for a further treatment of the left-hand term in (2.4). Other than that, recall that we still need to determine the value of κ_i.

In the double leave-one-out method with a second leave-one-feature-out step (employed in [6, 23, 8]), these two aforementioned objectives are achieved by mathematical manipulations relying on the condition that the entries of $\boldsymbol{\beta}$ have a symmetric role, which entails the same condition on the feature vectors $\mathbf{x}_i$ and the underlying classification/regression signal $\boldsymbol{\beta}^*$ that we hope to find. For instance, this can be achieved by assuming that $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ and $\boldsymbol{\beta}^* = \mathbf{0}_p$ (as in [23]) or $\boldsymbol{\beta}^*$ has almost i.i.d. entries (as in [8]).

We show here the example of the M-estimation regression method discussed in [6], with more rigorous results subsequently given by [23]. The regression problem of [6] is formulated as

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$$

for some convex loss function ρ and with continuous $y_i \in \mathbb{R}$ given by

$$y_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}^* + \epsilon_i$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a deterministic vector and $\epsilon_i \in \mathbb{R}$ a random scalar independent of $\mathbf{x}_i$. It suffices to take $\lambda = 0$ and

$$c_i = \psi(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i) = -\rho'(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$$

for the solution of the M-estimation problem to be expressed by (2.4). Here we do not consider the case of $n < p$, where the above repression problem is clearly not well defined.

Conforming to the assumptions required for the leave-one-feature-out manipulation in [6], we let $\boldsymbol{\beta}^* = \mathbf{0}_p$ and $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. Under these conditions, the entries of the solution $\boldsymbol{\beta}$ play symmetric roles. Recall also that κ_i is the limiting value of

$$\frac{1}{n}\mathbf{x}^T \left(\frac{1}{n}\sum_{j\neq i}^{n} \psi'(y_j - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_j)\mathbf{x}_j\mathbf{x}_j^\mathsf{T} \right)^{-1} \mathbf{x}_i.$$

As $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, the above term is approximated by

$$\frac{1}{n}\operatorname{tr}\left(\frac{1}{n}\sum_{j\neq i}^{n}\psi'(y_j - \boldsymbol{\beta}_{(-i)}^{\mathsf{T}}\mathbf{x}_j)\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1},$$

and it is easy to see that all κ_i have the same value κ. Without loss of generality, we focus on the first entry $\boldsymbol{\beta}(1)$ of $\boldsymbol{\beta}$. By considering separately $\boldsymbol{\beta}(1)$ and the vector of remaining entries $\boldsymbol{\beta}(\mathcal{S}_1) = \boldsymbol{\beta}(2,\ldots,p)$, (2.4) becomes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} c_i\mathbf{x}_i(1) &= 0\\ \frac{1}{n}\sum_{i=1}^{n} c_i\mathbf{x}_i(\mathcal{S}_1) &= \mathbf{0}_{p-1}.\end{aligned} \tag{2.9}$$

The main concept of the leave-one-feature-out step is to bring in the notion of a solution $\boldsymbol{\nu} \in \mathbb{R}^{p-1}$ obtained by removing one element (e.g., the first) from the feature vectors $\mathbf{x}_i$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1) = 0. \tag{2.10}$$

The interest of $\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})$, in contrast to c_i, is that $\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu}_1)$ is (by definition) independent of $\mathbf{x}_i(1)$. We hope then to relate c_i to $\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})$, in an attempt to characterize the dependence between c_i and $\mathbf{x}_i(1)$.

Intuitively, in the regime of large p, we have

$$\boldsymbol{\nu}^{\mathsf{T}}\mathbf{x}_i(\mathcal{S}_1) \simeq \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i,$$

implying that

$$c_i \simeq \psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu}).$$

Therefore, by subtracting (2.10) from (2.9), we get

$$\left[\frac{1}{n}\sum_{i=1}^{n} -\psi'(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)\mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\right](\boldsymbol{\beta}(\mathcal{S}_1) - \boldsymbol{\nu}) - \left[\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)\mathbf{x}_i(1)\right]\boldsymbol{\beta}(1)$$
$$\simeq 0.$$

Then,

$$\boldsymbol{\nu} - \boldsymbol{\beta}(\mathcal{S}_1) \simeq \left[\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)\mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)\mathbf{x}_i(1)\right]\boldsymbol{\beta}(1). \tag{2.11}$$

Similarly, the first line of (2.9) reads

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} c_i\mathbf{x}_i(1) &\simeq \frac{1}{n}\sum_{i=1}^{n}\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\mathbf{x}_i(1)\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - \mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})\left[\mathbf{x}_i(\mathcal{S}_1)^{\mathsf{T}}(\boldsymbol{\nu} - \boldsymbol{\beta}(\mathcal{S}_1)) - \boldsymbol{\beta}(1)\mathbf{x}_i(1)\right]\mathbf{x}_i(1)\\ &\simeq 0.\end{aligned}$$

Substituting (2.11) into the above equation, we get

$$\frac{1}{n}\sum_{i=1}^{n} c_i \mathbf{x}_i(1) \simeq \frac{1}{n}\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(1) - \xi_1\boldsymbol{\beta}(1) \simeq 0$$

where

$$\xi_1^r = \mathbf{u}_1^\mathsf{T}\left[\mathbf{D}_{(-1)} - \mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\left(\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\right)^{-1}\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\right]\mathbf{u}_1$$

with

$$\begin{aligned}
\mathbf{u}_1 &= \frac{1}{\sqrt{n}}[\mathbf{x}_1(1), \ldots, \mathbf{x}_n(1)]^\mathsf{T} \in \mathbb{R}^n \\
\mathbf{U}_{(-1)} &= \frac{1}{\sqrt{n}}[\mathbf{x}_1(\mathcal{S}_1), \ldots, \mathbf{x}_n(\mathcal{S}_1)]^\mathsf{T} \in \mathbb{R}^{n\times(p-1)} \\
\mathbf{D}_{(-1)} &= \operatorname{diag}\left[\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu}), \ldots, \psi'(y_n - \mathbf{x}_n(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\right] \in \mathbb{R}^{n\times n}.
\end{aligned}$$

As $\mathbf{U}_{(-1)}$ and $\mathbf{D}_{(-1)}$ are evidently independent of $\mathbf{u}_1$, experience from the trace lemma (Lemma 2.2) of RMT suggests that

$$\xi_1^r \simeq \frac{1}{n}\operatorname{tr}\left[\mathbf{D}_{(-1)} - \mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\left(\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\right)^{-1}\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\right] \to \xi$$

for some deterministic constant $\xi = O(1)$. Finally, we get

$$\frac{1}{n}\sum_{i=1}^{n} c_i \mathbf{x}_i(1) \simeq \frac{1}{n}\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(1) - \xi\boldsymbol{\beta}(1) \simeq 0.$$

Hence, the average of $c_i\mathbf{x}_i(1)$, which is the product of two dependent random variables c_i and $\mathbf{x}_i(1)$ can be asymptotically replaced by the average of $\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(1)$, where $\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})$ is independent of $\mathbf{x}_i(1)$, minus a rescaling of $\boldsymbol{\beta}(1)$. This remark is similar in spirit to the key result $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i \simeq \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \kappa c_i$ in the leave-one-observation-out step.

Since

$$\xi\boldsymbol{\beta}(d) \simeq \frac{1}{n}\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i(\mathcal{S}_d)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(d), \quad d \in \{1, \ldots, p\}, \tag{2.12}$$

we have by the central limit theorem that $\boldsymbol{\beta}$ follows asymptotically a normal distribution $\mathcal{N}(\mathbf{0}_p, \frac{\sigma^2}{p}\mathbf{I}_p)$ for some deterministic (positive) constant $\sigma = O(1)$ satisfying

$$\begin{aligned}
\sigma^2 \simeq \|\boldsymbol{\beta}\|^2 &\simeq \frac{1}{\xi}\mathbb{E}\{\psi(y_i - \mathbf{x}_i(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})^2\} \simeq \frac{1}{\xi}\mathbb{E}\{c_i^2\} \simeq \frac{1}{\xi}\mathbb{E}\{g_\kappa(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)^2\} \\
&\simeq \frac{1}{\xi}\mathbb{E}\{g_\kappa(y_i - \sigma z_i)^2\}
\end{aligned} \tag{2.13}$$

for some random variable $z_i \sim \mathcal{N}(0,1)$ independent of y_i.

Now it remains to study the unknown deterministic constants κ and ξ. Notice first that

$$\mathbf{D}_{(-1)} - \mathbf{D}_{(-1)}\mathbf{U}_{(-1)} \left(\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\right)^{-1} \mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)} = \mathbf{D}_{(-1)}^{\frac{1}{2}} \left[\mathbf{I}_n - \mathbf{P}_{(-1)}\right] \mathbf{D}_{(-1)}^{\frac{1}{2}}$$

where

$$\mathbf{P}_{(-1)} = \mathbf{D}_{(-1)}^{\frac{1}{2}}\mathbf{U}_{(-1)} \left(\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\right)^{-1} \mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}^{\frac{1}{2}}$$

is a projection matrix of rank $p-1$.

Using the rank-one perturbation result in Lemma 2.1 (an application of the Sherman-Morrison formula), we have the following development for the diagonal elements $\mathbf{P}_{(-1)}(i,i)$ of $\mathbf{P}_{(-1)}$:

$$\begin{aligned}\mathbf{P}_{(-1)}(i,i) &= \frac{1}{n}\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)^\mathsf{T} \left(\mathbf{U}_{(-1)}^\mathsf{T}\mathbf{D}_{(-1)}\mathbf{U}_{(-1)}\right)^{-1} \mathbf{x}_i(\mathcal{S}_1) \\ &= \frac{\frac{1}{n}\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)^\mathsf{T} \left(\sum_{j\neq i}\psi'(y_j - \mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_j(\mathcal{S}_1)\mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\right)^{-1} \mathbf{x}_i(\mathcal{S}_1)}{1 + \frac{1}{n}\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)^\mathsf{T} \left(\sum_{j\neq i}\psi'(y_j - \mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_j(\mathcal{S}_1)\mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\right)^{-1} \mathbf{x}_i(\mathcal{S}_1)}.\end{aligned}$$

With the concentration arguments already given in the previous discussion, we note that

$$\begin{aligned}&\frac{1}{n}\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_i(\mathcal{S}_1)^\mathsf{T} \left(\sum_{j\neq i}\psi'(y_j - \mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_j(\mathcal{S}_1)\mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\right)^{-1} \mathbf{x}_i(\mathcal{S}_1) \\ &\simeq \frac{1}{n}\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu}) \operatorname{tr}\left(\sum_{j\neq i}\psi'(y_j - \mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})\mathbf{x}_j(\mathcal{S}_1)\mathbf{x}_j(\mathcal{S}_1)^\mathsf{T}\right)^{-1} \simeq \kappa\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu}).\end{aligned}$$

Therefore,

$$\mathbf{P}_{(-1)}(i,i) \simeq \frac{\kappa\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})}{1 + \kappa\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})}.$$

Since $\mathbf{P}_{(-1)}$ is a projection matrix of rank $p-1$, we have the following equation of κ:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{P}_{(-1)}(i,i) \simeq \frac{1}{n}\sum_{i=1}^n \frac{\kappa\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})}{1 + \kappa\psi'(y_1 - \mathbf{x}_1(\mathcal{S}_1)^\mathsf{T}\boldsymbol{\nu})} \simeq \frac{p}{n}.$$

Recall that

$$\boldsymbol{\nu}^\mathsf{T}\mathbf{x}_i(\mathcal{S}_1) \simeq \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i \simeq \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \kappa c_i \simeq \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \kappa g_\kappa(\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i, y_i),$$

we get

$$\mathbb{E}\left\{\frac{1}{1 + \kappa\psi'(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i + \kappa g_\kappa(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i))}\right\} \simeq \frac{n-p}{n}. \tag{2.14}$$

Furthermore, as

$$\frac{1}{n}\operatorname{tr}\mathbf{D}_{(-1)}^{\frac{1}{2}}\left[\mathbf{I}_n-\mathbf{P}_{(-1)}\right]\mathbf{D}_{(-1)}^{\frac{1}{2}}\simeq\frac{1}{n}\sum_{i=1}^{n}\frac{\psi'(y_1-\mathbf{x}_1(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})}{1+\kappa\psi'(y_1-\mathbf{x}_1(\mathcal{S}_1)^{\mathsf{T}}\boldsymbol{\nu})}\simeq\frac{1}{\kappa}\frac{p}{n},$$

we get the following relation between κ and ξ:

$$\xi\simeq\frac{1}{\kappa}\frac{p}{n}. \tag{2.15}$$

Combining the two equations (2.15)–(2.14) of κ and ξ with the final result (2.13) given in the end of the leave-one-feature-out reasoning, we can determine the parameter σ in the limiting distribution $\mathcal{N}(\mathbf{0}_p,\frac{\sigma}{p}\mathbf{I}_p)$ of $\boldsymbol{\beta}$ by the following system of equations on σ and κ:

$$\begin{aligned}&\sigma^2=c\kappa\mathbb{E}\{g_\kappa(y_i-\sigma z_i)^2\}\\&\mathbb{E}\left\{\frac{1}{1+\kappa\psi'(y_i-\sigma z_i+\kappa g_\kappa(y_i-\sigma z_i))}\right\}=1-c^{-1}\end{aligned}$$

for some random variable $z_i\sim\mathcal{N}(0,1)$ independent of y_i.

2.2.4 Establishing systems of equations: our approach with advanced RMT tools

In this thesis, we consider more general settings than the assumptions required for the leave-one-feature-out manipulation explained in the previous section. The generalization is twofold: 1) there is no constraint on the correlations between features, i.e., the covariance matrix of the feature vectors $\mathbf{x}_i$ is of arbitrary form; 2) to better describe real-world classification problems, we allow the feature vectors $\mathbf{x}_i$ to have different statistical behaviors according to their classes by considering a Gaussian mixture model where

$$\mathbf{x}_i\sim\mathcal{N}(\boldsymbol{\mu}_k,\mathbf{C}_k),\quad\text{for }\mathbf{x}_i\in\mathcal{C}_k \tag{2.16}$$

with $\mathcal{C}_1,\ldots,\mathcal{C}_K$ standing for the K classes. Even though the learning methods studied in this dissertation are for classification, we stick to the example of M-estimation for regression discussed in the previous section for better readability.

As explained in the beginning of Section 2.2.3, to further study the statistical behavior of $\boldsymbol{\beta}$ after the leave-one-observation-out step (presented in Section 2.2.2), there are two remaining tasks: determining κ_i and treating the dependence between c_i and $\mathbf{x}_i$ in the sum of $c_i\mathbf{x}_i$.

We consider the Gaussian mixture data model (2.16), with n_k be the count of $\mathbf{x}_i\in\mathcal{C}_k$ for $k\in\{1,\ldots,K\}$. Also, we use the notations $\boldsymbol{\mu}_{(i)}=\boldsymbol{\mu}_k$ and $\mathbf{C}_{(i)}=\mathbf{C}_k$ for $\mathbf{x}_i\in\mathcal{C}_k$, $i\in\{1,\ldots,n\}$. Remark first that under the Gaussian mixture model (2.16), we have

$$\begin{aligned}\kappa_i&\simeq\frac{1}{n}\mathbf{x}^T\left(\frac{1}{n}\sum_{j\neq i}^{n}\psi'(y_j-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j)\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1}\mathbf{x}_i\simeq\frac{1}{n}\operatorname{tr}\mathbf{C}_k\left(\frac{1}{n}\sum_{j\neq i}^{n}\psi'(y_j-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_j)\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1}\\&\simeq\frac{1}{n}\operatorname{tr}\mathbf{C}_k\left(\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i-\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)^{-1}\to\kappa_{[k]}\end{aligned}$$

for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1, \ldots, K\}$. As κ_i can have K different values $\kappa_{[1]}, \ldots, \kappa_{[K]}$, the mathematical arguments of Section 2.2.3 are no longer enough to determine all the $\kappa_{[k]}$. We use here the RMT tool of deterministic equivalents discussed in Section 2.1.2. Denote $\bar{\mathbf{R}}$ the deterministic equivalent of

$$\mathbf{R} = \left(\frac{1}{n} \sum_{i=1}^{n} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right)^{-1},$$

and let

$$\mathbf{R}_{(-i)} = \left(\frac{1}{n} \sum_{j \neq i}^{n} \psi'(y_j - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_j) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} \right)^{-1}.$$

Using Lemmas 2.1–2.2, we have the following development similar to (2.2)

$$\begin{aligned}
\mathbf{I}_p =& \mathbb{E}\left\{ \mathbf{R} \frac{1}{n} \sum_{i=1}^{n} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \mathbf{R} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right\} \\
=& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \frac{\mathbf{R}_{-i} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\mathsf{T}}{1 + \frac{1}{n} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) \mathbf{x}_i^\mathsf{T} \mathbf{R}_{-i} \mathbf{x}_i} \right\} \simeq \mathbb{E}\{\mathbf{R}\} \left(\sum_{k=1}^{K} \frac{n_k}{n} \mathbb{E}\left\{ \frac{q_{[k]}}{1 + q_{[k]} \kappa_{[k]}} \right\} \mathbf{C}_k \right)
\end{aligned}$$

where $q_{[k]} \overset{\mathcal{L}}{=} \psi'(y_i - \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i)$ for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1, \ldots, K\}$. Hence,

$$\bar{\mathbf{R}} \simeq \left(\sum_{k=1}^{K} \frac{n_k}{n} \mathbb{E}\left\{ \frac{q_k}{1 + q_k \kappa_{[k]}} \right\} \mathbf{C}_k \right)^{-1}.$$

We get thus these K equations on $\kappa_{[1]}, \ldots, \kappa_{[K]}$:

$$\kappa_{[k]} \simeq \frac{1}{n} \operatorname{tr} \mathbf{C}_k \left(\sum_{k'=1}^{K} \frac{n_{k'}}{n} \mathbb{E}\left\{ \frac{q_{k'}}{1 + q_{k'} \kappa_{[k']}} \right\} \mathbf{C}'_k \right)^{-1}, \quad k \in \{1, \ldots, K\}.$$

We now turn our attention to the dependence between c_i and $\mathbf{x}_i$ in the product $c_i \mathbf{x}_i$. Since $c_i \simeq g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T} \mathbf{x}_i)$ (an approximation from the leave-one-observation-out step), we approach this problem by treating the dependence between $\boldsymbol{\beta}_{(-i)}^\mathsf{T} \mathbf{x}_i$ and $\mathbf{x}_i$. As a matter of fact, the properties of Gaussian vectors allow us to decompose $\mathbf{x}_i$ into an independent random vector of $\boldsymbol{\beta}_{(-i)}^\mathsf{T} \mathbf{x}_i$ and a dependent part, as stated in the following lemma.

Lemma 2.3. *Let $\mathbf{w} \in \mathbb{R}^p$ be a centred Gaussian vector with $\operatorname{Cov}\{\mathbf{w}\} = \mathbf{C}$ and $\mathbf{v} \in \mathbb{R}^p$ some deterministic vector. Then,*

$$\mathbf{w}' = \mathbf{w} - \frac{\mathbf{v}^\mathsf{T} \mathbf{w}}{\mathbf{v}^\mathsf{T} \mathbf{C} \mathbf{v}} \mathbf{C} \mathbf{v}$$

is independent of $\mathbf{v}^\mathsf{T} \mathbf{w}$.

Lemma 2.3 is proven as follows: since all elements of $\mathbf{w}'$ are jointly Gaussian with $\mathbf{v}^\mathsf{T}\mathbf{w}$ and

$$\mathbb{E}\{(\mathbf{v}^\mathsf{T}\mathbf{w})\mathbf{w}'\} = \mathbb{E}\{(\mathbf{v}^\mathsf{T}\mathbf{w})\mathbf{w}\} - \mathbb{E}\left\{\frac{(\mathbf{v}^\mathsf{T}\mathbf{w})^2}{\mathbf{v}^\mathsf{T}\mathbf{C}\mathbf{v}}\mathbf{C}\mathbf{v}\right\} = \mathbb{E}\{\mathbf{w}\mathbf{w}^\mathsf{T}\}\mathbf{v} - \frac{\mathbf{v}^\mathsf{T}\mathbb{E}\{\mathbf{w}\mathbf{w}^\mathsf{T}\}\mathbf{v}}{\mathbf{v}^\mathsf{T}\mathbf{C}\mathbf{v}}\mathbf{C}\mathbf{v} = \mathbf{0}_p,$$

the independence between the entries of $\mathbf{w}'$ and $\mathbf{v}^\mathsf{T}\mathbf{w}$ is thus proven by the property that two uncorrelated jointly Gaussian variables are also independent.

Let us write

$$\mathbf{x}_i = \boldsymbol{\mu}_{(i)} + \mathbf{w}_i = \boldsymbol{\mu}_{(i)} + \frac{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{w}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)} + \mathbf{w}_i', \tag{2.17}$$

then we have, by Lemma 2.3, that $\mathbf{w}_i'$ is independent of $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{w}_i$ (conditioned on $\boldsymbol{\beta}_{(-i)}$), thus also independent of $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$. Substituting (2.17) into $\sum_{i=1}^n c_i\mathbf{x}_i = 0$ and using $c_i \simeq g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)$, we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\left[\boldsymbol{\mu}_{(i)} + \frac{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{w}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)} + \mathbf{w}_i'\right] \\
\simeq &\left[\frac{1}{n}\sum_{i=1}^n g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\frac{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{w}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\mathbf{C}_{(i)}\right]\boldsymbol{\beta} + \frac{1}{n}\sum_{i=1}^n g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)(\boldsymbol{\mu}_{(i)} + \mathbf{w}_i') \\
\simeq &-\left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]\boldsymbol{\beta} + \sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{c_{[k]}\}\boldsymbol{\mu}_k + \frac{1}{n}\sum_{i=1}^n g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\mathbf{w}_i' \simeq 0
\end{aligned}$$

where $r_{[k]} \overset{\mathcal{L}}{=} \frac{-g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{w}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}$ and $c_{[k]} \overset{\mathcal{L}}{=} c_i \simeq g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)$, for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1, \ldots, K\}$. Therefore,

$$\boldsymbol{\beta} \simeq \left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1}\left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{c_{[k]}\}\boldsymbol{\mu}_k + \frac{1}{n}\sum_{i=1}^n g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\mathbf{w}_i'\right]. \tag{2.18}$$

Hence, by treating the dependence between c_i and $\mathbf{x}_i$ in $\sum_{i=1}^n c_i\mathbf{x}_i$ with Lemma 2.3, we show that $\boldsymbol{\beta}$ can be approximately expressed as a deterministic vector plus a sum of $g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\mathbf{w}_i'$ with $g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)$ independent of $\mathbf{w}_i'$ (conditioned on $\boldsymbol{\beta}_{(-i)}$). This reminds us of a similar result (2.12) given in the leave-one-feature-out derivation of the previous section, for $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$.

It can be shown from (2.18) (as an extension of the central limit theorem) that $\boldsymbol{\beta}$ follows asymptotically a multivariate normal distribution. Notice also from (2.18) that

$$\begin{aligned}
\mathbb{E}\{\boldsymbol{\beta}\} &\simeq \left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1}\left(\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{c_{[k]}\}\boldsymbol{\mu}_k\right) \\
\mathrm{Cov}\{\boldsymbol{\beta}\} &\simeq \left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1}\left(\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\{g_{\kappa_i}(y_i - \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)^2\}\mathbb{E}\{\mathbf{w}_i'\mathbf{w}_i'^\mathsf{T}\}\right)\left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1} \\
&\simeq \left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1}\left(\frac{1}{n}\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{c_{[k]}^2\}\mathbf{C}_k\right)\left[\sum_{k=1}^K \frac{n_k}{n}\mathbb{E}\{r_{[k]}\}\mathbf{C}_k\right]^{-1}.
\end{aligned}$$

In view of the above discussion, the limiting distribution of $\boldsymbol{\beta}$ is given as

$$\left(\sum_{k=1}^{K}\frac{n_k}{n}\theta_k\mathbf{C}_k\right)^{-1}\mathcal{N}\left(\sum_{k=1}^{K}\frac{n_k}{n}\alpha_k\boldsymbol{\mu}_k,\sum_{k=1}^{K}\frac{n_k}{n}\frac{\gamma_k\mathbf{C}_k}{p}\right)$$

with the parameters $\theta_k,\alpha_k,\gamma_k$ determined by the following system of equations: for $k\in\{1,\ldots,K\}$,

$$\begin{aligned}
\theta_k &= \mathbb{E}\left\{\frac{-g_{\kappa_{[k]}}(y_i-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i)\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{w}_i}{\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{C}_k\tilde{\boldsymbol{\beta}}}\right\}\\
\alpha_k &= \mathbb{E}\{g_{\kappa_{[k]}}(y_i-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i)\}\\
\gamma_k &= \frac{p}{n}\mathbb{E}\{g_{\kappa_{[k]}}(y_i-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i)^2\}\\
\kappa_{[k]} &= \frac{1}{n}\operatorname{tr}\mathbf{C}_k\left(\sum_{k'=1}^{K}\frac{n_{k'}}{n}\mathbb{E}\left\{\frac{\psi'(y_j-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_j+\kappa_{[k']}g_{\kappa_{[k']}}(y_j-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_j))}{1+\psi'(y_j-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_j+\kappa_{[k']}g_{\kappa_{[k']}}(y_j-\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_j))\kappa_{[k']}}\right\}\mathbf{C}_k'\right)^{-1}
\end{aligned}$$

where $\mathbf{x}_i\in\mathcal{C}_k$, $\mathbf{x}_j\in\mathcal{C}_k'$ and

$$\tilde{\boldsymbol{\beta}}\sim\left(\sum_{k=1}^{K}\frac{n_k}{n}\theta_k\mathbf{C}_k\right)^{-1}\mathcal{N}\left(\sum_{k=1}^{K}\frac{n_k}{n}\alpha_k\boldsymbol{\mu}_k,\sum_{k=1}^{K}\frac{n_k}{n}\frac{\gamma_k\mathbf{C}_k}{p}\right)$$

some random vector independent of $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$.

As such, by combining the leave-one-out method with advanced RMT results, we can understand how the structured distribution of data samples is reflected by that of $\boldsymbol{\beta}$, providing thus more insightful remarks on the interaction between the statistical parameters of data features and the learning performance.

Part I

Semi-supervised learning on graphs

Chapter 3

Large dimensional behavior of semi-supervised Laplacian regularization algorithms

3.1 Introduction of graph-based semi-supervised learning

Graph-based methods are an important subset of semi-supervised learning. In these, one considers data instances $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ as vertices on a graph with edge weights w_{ij} encoding their similarity, which is usually defined through a kernel function h, as with radial kernels of the type $w_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ which we shall focus on in this chapter. The motivation follows from one's expectation that two instances with a strong edge weight tend to belong to the same class and thus vertices of a common class tend to aggregate. Standard methods for recovering the classes of the unlabelled data then consist in various random walk [28] or label propagation [29] algorithms on the graph which softly allocate "scores" for each node to belong to a particular class. These scores are then compared for each class in order to obtain a hard decision on the individual unlabelled node class. A popular, and widely recognized as highly performing, example is the PageRank approach [30].

Many of these algorithms also have the particularity of having a closed-form and quite interrelated expression for their stationary points. These stationary points are also often found to coincide with the solutions to optimization problems under constraints, independently established. This is notably the case of [31] where the pre-known labels are imposed on the labelled nodes or of [32] where a relaxation approach is used instead to allow for modifications of the value of labelled nodes – this ensuring that erroneously labelled data or poorly informative labelled data do not hinder the algorithm performance. These algorithms are also known as the semi-supervised Laplacian regularization methods. As is often the case in graph-related optimization, a proper choice of the matrix representative of the inter-data affinity is at the core of scientific research and debates and mainly defines the differences between any two schemes. In particular, [33] suggests the use of a standard Laplacian representative, where [34] advises for a normalized Laplacian approach. These individual choices correspondingly lead to different versions of the label propagation methods on the graph, as discussed in [30].

There also exists another branch of manifold based semi-supervised learning [35, 36, 37]. In

contrast to the methods discussed in this chapter, these approaches involve a step of manifold learning, which plays a decisive role in the success of the learning task. These methods have been theoretically investigated in [38, 39, 37, 40]. Another recent line of alternative works consider SSL from a graph signal processing perspective [41, 42, 43, 44], where the classification scores are viewed as smooth signals on the similarity graph and the learning task then consists in recovering a bandlimited (understood in the graph Fourier transform domain) graph signal from its known sample values.

3.2 Motivation and main findings

A likely key reason for the open-ended question of a most natural choice for the graph representative arises from these methods being essentially built upon intuitive reasoning arising from low dimensional data considerations rather than from mostly inaccessible theoretical results. Indeed, the non-linear expression of the affinity matrix $\mathbf{W}$ as well as the rather involved form assumed by the algorithm output (although explicit) hinder the possibility to statistically evaluate the algorithm performances for all finite n, p, even for simple data assumptions. The present analysis is placed instead under a large dimensional data assumption, thus appropriate to the present big-data paradigm, and proposes instead to derive, for the first time to the best of the authors' knowledge, theoretical results on the performance of the aforementioned algorithms in the large n, p limit for a certain class of statistically distributed data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Precisely due to the large data assumption, as we shall observe, most of the intuition leading up to the aforementioned algorithms collapse as $n, p \to \infty$ at a similar rate, and we shall prove that few algorithms remain consistent in this regime.

Specifically, recall that the idea behind graph-based semi-supervised learning is to exploit the similarity between data points and thus expect a clustering behavior of close-by data nodes. In the large data assumption (i.e., $p \gg 1$), this similarity-based approach suffers a curse of dimensionality. As the span of $\mathbb{R}^p$ grows exponentially with the data dimension p, when p is large, the data points $\mathbf{x}_i$ (if not too structured) are in general so sparsely distributed that their pairwise distances tend to be similar regardless of their belonging to the same class or not. The Gaussian mixture model that we define in Subsection 3.3.2 and will work on is a telling example of this phenomenon; as we show, in a regime where the classes ought to be separable (even by unsupervised methods as shown by [45]), the normalized distance $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{p}$ of two random different data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ generated from this model converges to a constant τ *irrespective of the class of* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ in the Gaussian mixture and, consequently, the similarity defined by $w_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ is asymptotically the same for all pairs of data instances. This phenomenon of high dimensional data is known under the name of *distance concentration*, studied in many works [46, 47, 48, 49, 50]. The distance concentration of large dimensional data should invalidate the notion of similarity, hence likely render graph-based methods ineffective. As a direct consequence, the predicted outputs for classification, here referred to as *scores*, are *flat* in the sense that they have the same asymptotic values, irrespective of the class. Nonetheless, we will show that provided that appropriate amendments to the classification algorithms are enforced, sensible classification on data sets generated from this model can still be achieved, thanks to the information contained in the small fluctuations around these flat asymptotic limit of scores. This flat limit is reminiscent of the work by [51] where the authors show that the scores indeed share the same limit, irrespective of the class, in the presence of infinitely

many unlabelled samples but for $p \geq 2$ fixed. Yet, despite the scores flatness, the authors experimentally observed non-trivial classification in binary tasks thanks to the small difference between scores; they however did not provide any theoretical support for such behavior, for their analysis failed to recover the small fluctuations.

In the same spirit as [30], we consider here a common framework for semi-supervised Laplacian regularization algorithms, with the help of a normalization parameter γ which allows ones to retrieve commonly used algorithms by choosing specific values of γ. The generalized optimization framework is presented in Section 3.3.

The main contribution of the present chapter is to provide a quantitative performance study of the generalized graph-based semi-supervised algorithm for large dimensional Gaussian-mixture data and radial kernels, technically following the random matrix approach developed by [45]. Our main findings are summarized as follows:

- Irrespective of the choice of the data affinity matrix, the classification outcome is strongly biased by *the number of labelled data from each class* and unlabelled data tend to be classified into the class with most labelled nodes; we propose a normalization update of the standard algorithms to correct this limitation.

- Once the aforementioned bias corrected, the choice of the affinity matrix (and thus of the parameter γ) strongly impacts the performances; most importantly, within our framework, both *standard Laplacian* ($\gamma = 0$ here) and *normalized Laplacian-based* ($\gamma = -\frac{1}{2}$) methods, although widely discussed in the literature, fail in the large dimensional data regime. Of the family of algorithms discussed above, only the *PageRank* approach ($\gamma = -1$) is shown to provide asymptotically acceptable results.

- The scores of belonging to each class attributed to individual nodes by the algorithms are shown to asymptotically follow a *Gaussian distribution* with mean and covariance depending on the statistical properties of classes, the ratio of labelled versus unlabelled data, and the value of the first derivatives of the kernel function at the limiting value τ of $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2$ (which we recall is irrespective of the genuine classes of $\mathbf{x}_i, \mathbf{x}_j$). This last finding notably allows one to *predict the asymptotic performances* of the semi-supervised learning algorithms.

- From the latter result, three main outcomes unfold:

  - when three classes or more are considered, there exist Gaussian mixture models for which classification is shown to be *impossible*;
  - despite PageRank's consistency, we further justify that the choice $\gamma = -1$ is not in general optimal. For the case of 2-class learning, we provide a method to approach the optimal value of γ; this method is demonstrated on real data sets to convey sometimes *dramatic improvements* in correct classification rates.
  - for a 2-class learning task, necessary and sufficient conditions for asymptotic consistency are: $h'(\tau) < 0$, $h''(\tau)h(\tau) > h'(\tau)^2$; in particular, Gaussian kernels, failing to meet the last condition, cannot deal with the large dimensional version of the "concentric spheres" task where the classes have the same means for the Gaussian mixture model.

In this chapter, theoretical results and related discussions are confirmed and illustrated with simulations on Gaussian-mixture data as well as the popular MNIST data [52], which serves as a comparison for our theoretical study on real world data sets. The consistent match of our theoretical findings on MNIST data, despite their departing from the very large dimensional and Gaussian-mixture assumption, suggests that our results have a certain robustness to these assumptions and can be applied to a larger range of data. We indeed believe that, while only the limiting behavior of Gaussian mixture inputs is characterized in this chapter (mostly for technical reasons), the analysis reveals certain properties inherent to graph-based SSL methods, which extend well beyond the Gaussian hypothesis.

3.3 Problem formulation

3.3.1 Optimization framework

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n data vectors belonging to K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$. The class association of the $n_{[l]}$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]}}$ is known (these vectors will be referred to as *labelled*), while the class of the remaining $n_{[u]}$ vectors $\mathbf{x}_{n_{[l]}+1}, \ldots, \mathbf{x}_n$ ($n_{[l]} + n_{[u]} = n$) is unknown (these are referred to as *unlabelled* vectors). Within both labelled and unlabelled subsets, the data are organized in such a way that the $n_{[l]1}$ first vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]1}}$ belong to class $\mathcal{C}_1$, $n_{[l]2}$ subsequent vectors to $\mathcal{C}_2$, and so on, and similarly for the $n_{[u]1}, n_{[u]2}, \ldots$ first vectors of the set $\mathbf{x}_{n_{[l]}+1}, \ldots, \mathbf{x}_n$. Note already that this ordering is for notational convenience and shall not impact the generality of our results.

The affinity relation between the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is measured from the weight matrix $\mathbf{W}$ defined by

$$\mathbf{W} \equiv \left\{ w_{ij} = h\left(\frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right) \right\}_{i,j=1}^{n}$$

for some non-negative function h. The matrix $\mathbf{W}$ may be seen as the adjacency matrix of the n-node graph indexed by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We further denote by $\mathbf{D}$ the diagonal matrix with $\{\mathbf{D}\}_{ii} \equiv d_i = \sum_{j=1}^{n} w_{ij}$ the degree of the node associated to $\mathbf{x}_i$.

We next define a score matrix $\mathbf{F} = \{f_{ik}\}_{\substack{i=1,\ldots,n \\ k=1,\ldots,K}}$ with f_{ik} representing the evaluated score for $\mathbf{x}_i$ to belong to $\mathcal{C}_k$. In particular, following the conventions typically used in graph-based semi-supervised learning [53], we shall affect a unit score $f_{ik} = 1$ if $\mathbf{x}_i$ is a labelled data of class $\mathcal{C}_k$ and a null score for all $f_{ik'}$ with $k' \neq k$. In order to attribute classes to the unlabelled data, scores are first affected by means of the resolution of an optimization framework. We propose here

$$\begin{aligned} \mathbf{F} &= \operatorname{argmin}_{\mathbf{F} \in \mathbb{R}^{n \times k}} \sum_{k=1}^{K} \sum_{i,j=1}^{n} w_{ij} \left\| d_i^{\gamma} f_{ik} - d_j^{\gamma} f_{jk} \right\|^2 \\ \text{s.t. } f_{ik} &= \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{C}_k \\ 0, & \text{otherwise} \end{cases}, \; 1 \leq i \leq n_{[l]}, \; 1 \leq k \leq K \end{aligned} \tag{3.1}$$

where $\gamma \in \mathbb{R}$ is a given parameter. The interest of this generic formulation is that it coincides with the standard Laplacian-based approach for $\gamma = 0$ and with the normalized Laplacian-based

approach for $\gamma = -\frac{1}{2}$, both discussed in Section 3.2. Note importantly that Equation (3.1) is naturally motivated by the observation that large values of w_{ij} (thus "close" $\mathbf{x}_i$ and $\mathbf{x}_j$) enforce close values for f_{ik} and f_{jk}, while small values for w_{ij} allow for more freedom in the choice of f_{ik} and f_{jk}.

By denoting

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{[l]} \\ \mathbf{F}_{[u]} \end{bmatrix}, \ \mathbf{W} = \begin{bmatrix} \mathbf{W}_{[ll]} & \mathbf{W}_{[lu]} \\ \mathbf{W}_{[ul]} & \mathbf{W}_{[uu]} \end{bmatrix}, \text{ and } \mathbf{D} = \begin{bmatrix} \mathbf{D}_{[l]} & 0 \\ 0 & \mathbf{D}_{[u]} \end{bmatrix}$$

with $\mathbf{F}_{[l]} \in \mathbb{R}^{n_{[l]}}$, $\mathbf{W}_{[ll]} \in \mathbb{R}^{n_{[l]} \times n_{[l]}}$, $\mathbf{D}_{[l]} \in \mathbb{R}^{n_{[l]} \times n_{[l]}}$, one easily finds (since the problem is a convex quadratic optimization with linear equality constraints) the solution to (3.1) is explicitly given by

$$\mathbf{F}_{[u]} = \left(\mathbf{I}_{n_u} - \mathbf{D}_{[u]}^{-1-\gamma} \mathbf{W}_{[uu]} \mathbf{D}_{[u]}^{\gamma} \right)^{-1} \mathbf{D}_{[u]}^{-1-\gamma} \mathbf{W}_{[ul]} \mathbf{D}_{[l]}^{\gamma} \mathbf{F}_{[l]}. \tag{3.2}$$

Once these scores are affected, a mere comparison between all scores $f_{i1}, \ldots, f_{iK}$ for unlabelled data $\mathbf{x}_i$ (i.e., for $i > n_{[l]}$) is performed to decide on its class, i.e., the allocated class index $\hat{\mathcal{C}}_{\mathbf{x}_i}$ for vector $\mathbf{x}_i$ is given by

$$\hat{\mathcal{C}}_{\mathbf{x}_i} = \mathcal{C}_{\hat{k}} \text{ for } \hat{k} = \text{argmax}_{1 \leq k \leq K} f_{ik}.$$

Note in passing that the formulation (3.2) implies in particular that

$$\mathbf{F}_{[u]} = \mathbf{D}_{[u]}^{-1-\gamma} \mathbf{W}_{[uu]} \mathbf{D}_{[u]}^{\gamma} \mathbf{F}_{[u]} + \mathbf{D}_{[u]}^{-1-\gamma} \mathbf{W}_{[ul]} \mathbf{D}_{[l]}^{\gamma} \mathbf{F}_{[l]} \tag{3.3}$$

$$\mathbf{F}_{[l]} = \{ \boldsymbol{\delta}_{\mathbf{x}_i \in \mathcal{C}_k} \}_{\substack{1 \leq i \leq n_{[l]} \\ 1 \leq k \leq K}} \tag{3.4}$$

and thus the matrix $\mathbf{F}$ is a stationary point for the algorithm constituted of the updating rules (3.3) and (3.4) (when replacing the equal signs by affectations). In particular, for $\gamma = -1$, the algorithm corresponds to the standard label propagation method found in the PageRank algorithm for semi-supervised learning as discussed in [30], with the major difference that $\mathbf{F}_{[l]}$ is systematically reset to its known value while in the study of [30], $\mathbf{F}_{[l]}$ is allowed to evolve (for reasons related to robustness to pre-labeling errors).

The technical objective of the chapter is to analyze the behavior of $\mathbf{F}_{[u]}$ in the large n, p regime for a Gaussian mixture model for the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. To this end, we shall first need to design appropriate growth rate conditions for the Gaussian mixture statistics as $p \to \infty$ (in order to avoid trivializing the classification problem as p grows large) before proceeding to the evaluation of the behavior of $\mathbf{W}$, $\mathbf{D}$, and thus $\mathbf{F}$.

3.3.2 Model and Assumptions

In the remainder of the chapter, we shall assume that the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are extracted from a Gaussian mixture model composed of K classes. Specifically, for $k \in \{1, \ldots, K\}$,

$$\mathbf{x}_i \in \mathcal{C}_k \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{C}_k).$$

Consistently with the previous section, for each k, there are n_k instances of vectors of class $\mathcal{C}_k$, among which $n_{[l]k}$ are labelled and $n_{[u]k}$ are unlabelled.

As pointed out above, in the regime where $n, p \to \infty$, special care must be taken to ensure that the classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$, the statistics of which evolve with p, remain at a "somewhat constant" distance from each other. This is to ensure that the classification problem does not become asymptotically infeasible nor trivially simple at arbitrarily large p. Based on the earlier work [45] where similar considerations were made, the behavior of the class means, covariances, and cardinalities will follow the prescription below:

Assumption 3.1 (Growth Rate)**.** *Data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. observations from a generative model such that, for $k \in \{1, \ldots, K\}$, $\mathbb{P}(\mathbf{x}_i \in \mathcal{C}_k) = \rho_k$, and*

$$\mathbf{x}_i \in \mathcal{C}_k \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{C}_k).$$

The ratios $c_0 = \frac{n}{p}$, $c_{[l]} = \frac{n_{[l]}}{p}$ and $c_{[u]} = \frac{n_{[u]}}{p}$ are uniformly bounded in $(0, +\infty)$ for arbitrarily large p. Besides,

1. *For $\boldsymbol{\mu}^\circ \triangleq \sum_{k=1}^K \frac{n_k}{n} \boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_k^\circ \triangleq \boldsymbol{\mu}_k - \boldsymbol{\mu}^\circ$, $\|\boldsymbol{\mu}_k^\circ\| = O(1)$.*
2. *For $\mathbf{C}^\circ \triangleq \sum_{k=1}^K \frac{n_k}{n} \mathbf{C}_k$ and $\mathbf{C}_k^\circ \triangleq \mathbf{C}_k - \mathbf{C}^\circ$, $\|\mathbf{C}_k\| = O(1)$, $\|\mathbf{C}_k^{-1}\| = O(1)$, $\operatorname{tr} \mathbf{C}_k^\circ = O(\sqrt{p})$ and $\operatorname{tr} (\mathbf{C}_k^\circ)^2 = O(\sqrt{p})$.*

It will also be convenient in the following to define

$$t_k \equiv \frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_k^\circ$$
$$T_{kk'} \equiv \frac{1}{p} \operatorname{tr} C_k C_{k'}$$

as well as the labelled-data centered notations

$$\tilde{\boldsymbol{\mu}}_k \equiv \boldsymbol{\mu}_k - \sum_{k'=1}^K \frac{n_{[l]k'}}{n_{[l]}} \boldsymbol{\mu}_{k'}$$
$$\tilde{\mathbf{C}}_k \equiv \mathbf{C}_k - \sum_{k'=1}^K \frac{n_{[l]k'}}{n_{[l]}} \mathbf{C}_{k'}$$
$$\tilde{t}_k \equiv \frac{1}{\sqrt{p}} \operatorname{tr} \tilde{\mathbf{C}}_k.$$

Here are some remarks to interpret the conditions imposed on the data means $\boldsymbol{\mu}_k$ and covariance matrices $\mathbf{C}_k$ in Assumption 3.1. Firstly, as the discussion is placed under a large dimensional context, we need to ensure that the data vectors do not lie in a low dimensional manifold; the fact that $\|\mathbf{C}_k\| = O(1)$ along with $\|\mathbf{C}_k^{-1}\| = O(1)$ guarantees non-negligible variations in p linearly independent directions. Other conditions controlling the differences between the class statistics $\|\boldsymbol{\mu}_k^\circ\| = O(1)$,$\operatorname{tr} \mathbf{C}_k^\circ = O(\sqrt{p})$, and $\operatorname{tr} (\mathbf{C}_k^\circ)^2 = O(\sqrt{p})$ are made to ensure *non-trivial* scenarios where the classification of unlabelled data does not become impossible or overly easy at extremely large values of p.

Note also that, unlike in the previous works [51, 40] where the number of labelled data $n_{[l]}$ and data dimension p are considered fixed and the number of unlabelled data $n_{[u]}$ is supposed to be infinite, we assume a regime where $n_{[l]}, n_{[u]}$ and p are simultaneously large. Letting p

large allows us to investigate SSL in the context of large dimensional data. Further imposing that $n_{[l]}, n_{[u]}$ grow at a controlled rate with respect to p (here at the same rate) allows for an *exact characterization* of the limiting SSL performances, as a function of the hyperparameters γ, f and data statistics $\boldsymbol{\mu}_k, \mathbf{C}_k$, in non-trivial classification scenarios (i.e., when classification is neither asymptotically perfect nor impossible), instead of solely retrieving consistency bounds as a function of growth rates in $p, n_{[l]}, n_{[u]}$. This in turn allows for possible means of precise parameter setting to reach optimal performances (which is not possible with results based on bounds). While it may be claimed that SSL in practice often handles scenarios where $n_{[u]} \gg n_{[l]}$, assuming that $n_{[u]}, n_{[l]}$ are of the same order but that $n_{[u]}$ is multiple times $n_{[l]}$ actually maintains the validity of our results so long that $n_{[l]}$ is not too small.

As a by-product of imposing the growth constraints on the data to ensure non-trivial classification, Assumption 3.1 induces the following proposition of distance concentration, easily justified by a simple concentration of measure argument.

Proposition 3.3.1. *Define* $\tau = \frac{2}{p}\mathrm{tr}\mathbf{C}^\circ$. *Under Assumption 5.1, we have that, for all* $i, j \in \{1, \ldots, n\}$,

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \tau + O(p^{-\frac{1}{2}}). \tag{3.5}$$

Equation (3.5) is the cornerstone of our analysis and states that all vector pairs $\mathbf{x}_i, \mathbf{x}_j$ are essentially at the same distance from one another as p gets large, *irrespective of their classes*. This striking result evidently is in sharp opposition to the very motivation for the optimization formulation (3.1) as discussed in the beginning of this chapter. It thus immediately entails that the solution (3.2) to (3.1) is bound to produce asymptotically inconsistent results. We shall see that this is indeed the case for all but a short range of values of γ.

This being said, Equation (3.5) has an advantageous side as it allows for a Taylor expansion of $w_{ij} = h(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ around $h(\tau)$, provided h is sufficiently smooth around τ, which is ensured by our subsequent assumption.

Assumption 3.2 (Kernel function)**.** *The function* $h : \mathbb{R}^+ \to \mathbb{R}^+$ *is three-times continuously differentiable in a neighborhood of* τ.

Note that Assumption 3.2 does not constrain h aside from its local behavior around τ. In particular, we shall not restrict ourselves to matrices $\mathbf{W}$ arising from nonnegative definite kernels as standard machine learning theory would advise [54].

The core technical part of the chapter now consists in expanding $\mathbf{W}$, and subsequently all terms intervening in (3.2), in a Taylor expansion of successive matrices of *non-vanishing operator norm*. Note indeed that the magnitude of the individual entries in the Taylor expansion of $\mathbf{W}$ needs not follow the magnitude of the operator norm of the resulting matrices;[1] rather, great care must be taken to only retain those matrices of non-vanishing operator norm. These technical details call for advanced random matrix considerations and are discussed in the appendix and in [45].

[1]For instance, $\|\mathbf{I}_n\| = 1$, $\|\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| = n$, and $\|\mathbf{X}\| = O(\sqrt{n})$ for $\mathbf{X} \in \mathbb{R}^{n\times n}$ with i.i.d $\mathcal{N}(0,1)$ entries, despite all three matrices having entries of similar magnitude.

3.4 Performance analysis on large dimensional data

In the course of this section, we provide in parallel a series of technical results under the proposed setting (notably under Assumption 3.1) along with simulation results both on a 2-class Gaussian mixture data model with $\boldsymbol{\mu}_1 = [4; 0_{p-1}]$, $\boldsymbol{\mu}_2 = [0; 4; 0_{p-2}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + \frac{3}{\sqrt{p}})$, as well as on real data sets, here images of eights and nines from the MNIST database [52], for $h(t) = \exp(-\frac{t}{2})$, i.e., the classical Gaussian (or heat) kernel. For reasons that shall become clear in the following discussion, these figures depict the (size n) vectors of centered scores

$$\left\{\mathbf{F}^{\circ}_{[u]}\right\}_{\cdot k} \equiv \left\{\mathbf{F}_{[u]}\right\}_{\cdot k} - \frac{1}{K}\sum_{k'=1}^{K}\left\{\mathbf{F}_{[u]}\right\}_{\cdot k'}$$

for $k \in \{1, 2\}$. Obviously, the decision rule on $\mathbf{F}^{\circ}_{[u]}$ is the same as that on $\mathbf{F}_{[u]}$.

Our first hinging result concerns the behavior of the score matrix $\mathbf{F}$ in the large n, p regime, as per Assumption 3.1, and reads as follows.

Proposition 3.4.1. *Let Assumptions 3.1–3.2 hold. Then, for $i > n_{[l]}$ (i.e., for $\mathbf{x}_i$ an unlabelled vector),*

$$f_{ik} = \frac{n_{[l]k}}{n}\left[1 + \underbrace{(1+\gamma)\frac{h'(\tau)}{h(\tau)}\frac{t_k}{\sqrt{p}} + z_i}_{O(n^{-\frac{1}{2}})} + O(n^{-1})\right] \tag{3.6}$$

where $z_i = O(n^{-\frac{1}{2}})$ is a certain random variable, function of $\mathbf{x}_i$, but independent of k.

Proposition 3.4.1 provides a clear overview of the outcome of the semi-supervised learning algorithm. First note that $f_{ik} = c_{[l]k} + O(n^{-\frac{1}{2}})$. Therefore, irrespective of $\mathbf{x}_i$, f_{ik} is strongly biased towards $c_{[l]k}$. If the values $n_{[l]1}, \ldots, n_{[l]k}$ differ by $O(n)$, this induces a systematic asymptotic allocation of every $\mathbf{x}_i$ to the class having largest $c_{[l]k}$ value. Figure 3.1 illustrates this phenomenon, observed both on synthetic and real data sets, here for $n_{[l]1} = 3n_{[l]2}$.

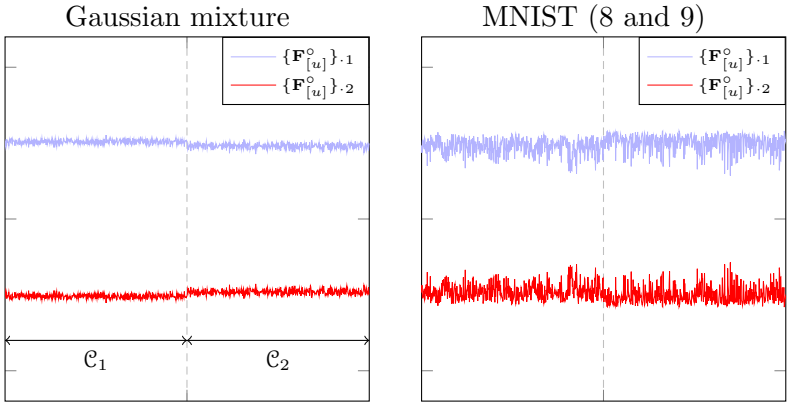

Figure 3.1: $\{\mathbf{F}^{\circ}_{[u]}\}_{\cdot 1}$ and $\{\mathbf{F}^{\circ}_{[u]}\}_{\cdot 2}$ for 2-class data, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = 3n_{[l]2}$, $\gamma = -1$, Gaussian kernel.

Pursuing the analysis of Proposition 3.4.1 by now assuming that $n_{[l]1} = \ldots = n_{[l]K}$, the comparison between $f_{i1}, \ldots, f_{iK}$ next revolves around the term of order $O(n^{-\frac{1}{2}})$. Since z_i only

depends on $\mathbf{x}_i$ and not on k, it induces a constant offset to the vector $\{\mathbf{F}\}_{i\cdot}$, thereby not intervening in the class allocation. On the opposite, the term t_k is independent of $\mathbf{x}_i$ but may vary with k, hereby possibly intervening in the class allocation, again an undesired effect. Figure 3.2 depicts the effect of various choices of γ for equal values of $n_{[l]k}$. This deleterious outcome can be avoided either by letting $h'(\tau) = O(n^{-\frac{1}{2}})$ or $\gamma = -1 + O(n^{-\frac{1}{2}})$. But, as discussed in the study of [45] and later in the chapter, the choice of h such that $h'(\tau) \simeq 0$, if sometimes of interest, is generally inappropriate.

The discussion above thus induces two important consequences to adapt the semi-supervised learning algorithm to large data.

1. The final comparison step *must* be made upon the normalized scores

$$\hat{f}_{ik} \equiv \frac{n}{n_{[l]k}} f_{ik} \tag{3.7}$$

   rather than upon the scores f_{ik} directly.

2. The parameter γ *must* be chosen in such a way that

$$\gamma = -1 + O(p^{-\frac{1}{2}}).$$

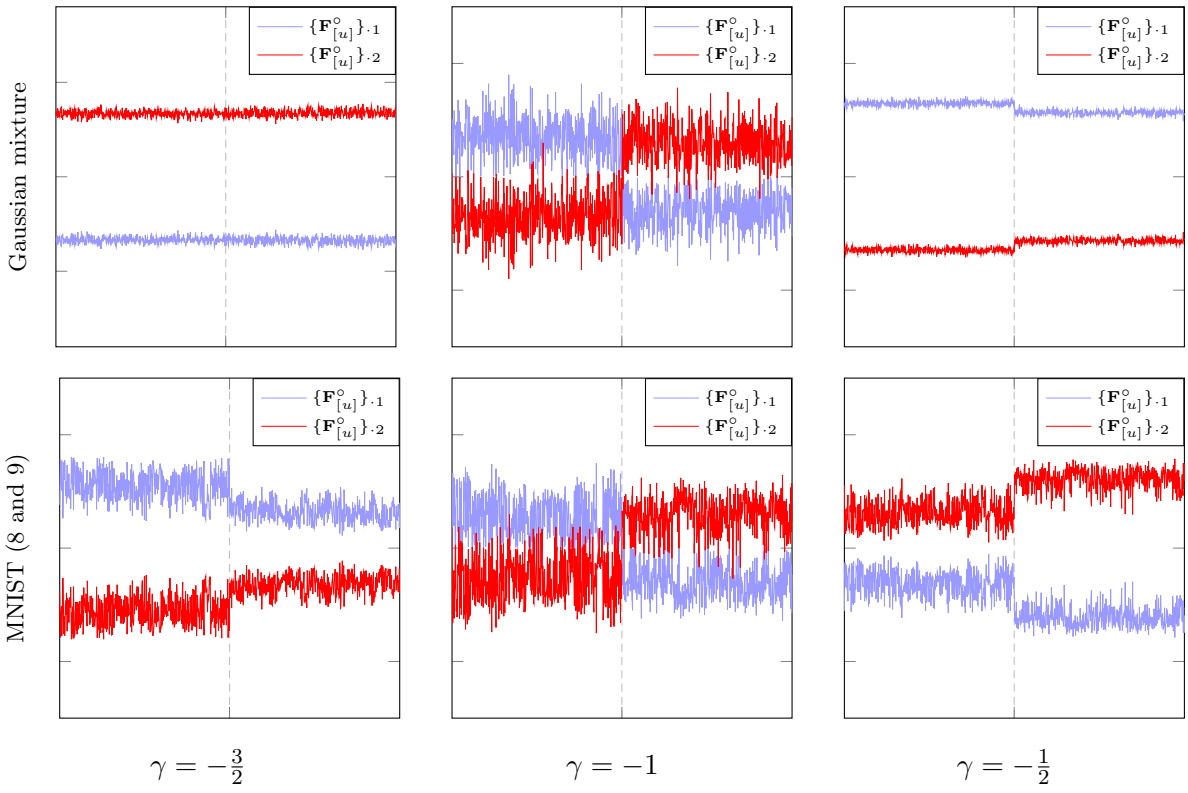

Figure 3.2: $\{\mathbf{F}^\circ_{[u]}\}_{\cdot 1}$, $\{\mathbf{F}^\circ_{[u]}\}_{\cdot 2}$ for 2-class data, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = n_{[l]2}$, Gaussian kernel.

Under these two amendments of the algorithm, according to Proposition 3.4.1, the performance of the semi-supervised learning algorithm now relies upon terms of magnitude $O(n^{-1})$, which are so far left undefined. A thorough analysis of these terms allows for a complete understanding of the asymptotic behavior of the normalized scores $\hat{f}_{ik}$, as presented in our next result.

Theorem 3.4.1. *For $\mathbf{x}_i \in \mathcal{C}_b$ an unlabelled vector (i.e., $i > n_{[l]}$), let $\hat{\mathbf{F}} = \{\hat{f}_{ik}\}_{\substack{i=1,\ldots,n \\ k=1,\ldots,K}}$ with $\hat{f}_{ik}$ given by (3.7), and $\gamma = -1 + \frac{\beta}{\sqrt{p}}$ for $\beta = O(1)$. Then, under Assumptions 3.1–3.2,*

$$p\{\hat{\mathbf{F}}\}_{i\cdot} = p(1+z_i)\mathbf{1}_K + \mathbf{g}_i + o_P(1)$$
$$\mathbf{g}_i \sim \mathcal{N}(\mathbf{m}_b, \mathbf{\Sigma}_b)$$

where z_i is as in Proposition 3.4.1 and

$$\{\mathbf{m}_b\}_a = -\frac{2h'(\tau)}{h(\tau)}\tilde{\boldsymbol{\mu}}_a^\mathsf{T}\tilde{\boldsymbol{\mu}}_b + \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau)^2}\right)\tilde{t}_a\tilde{t}_b + \frac{n\beta}{n_{[l]}}\frac{h'(\tau)}{h(\tau)}t_a \tag{3.8}$$

$$\{\mathbf{\Sigma}_b\}_{a_1a_2} = 2\left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau)^2}\right)^2 T_{bb}t_{a_1}t_{a_2} + 4\frac{h'(\tau)^2}{h(\tau)^2}\left[\boldsymbol{\mu}_{a_1}^{\circ\mathsf{T}}\mathbf{C}_b\boldsymbol{\mu}_{a_2}^{\circ} + \boldsymbol{\delta}_{a_1a_2}\frac{T_{b,a_1}}{\rho_{a_1}c_{[l]}}\right]. \tag{3.9}$$

Besides, there exists $\mathcal{A} \subset \sigma(\{\{x_1,\ldots,x_{n_{[l]}}\}, p = 1,2,\ldots\})$ (the σ-field induced by the labelled variables) with $P(\mathcal{A}) = 1$ over which (3.8)–(3.9) also hold conditionally to $\{\{x_1,\ldots,x_{n_{[l]}}\}, p = 1,2,\ldots\}$.

Note that the statistics of $\mathbf{g}_i$ are independent of the realization of $\mathbf{x}_1,\ldots,\mathbf{x}_{[l]}$ when $\gamma = -1 + O(\frac{1}{\sqrt{p}})$. This in fact no longer holds when α is outside this regime, as pointed out by Theorem A.1.1 in the appendix which provides the asymptotic behavior of $\{\hat{\mathbf{F}}\}_{i\cdot}$ for all values of γ (and thus generalizes Theorem 3.4.1).

Since the ordering of the entries of $\{\hat{\mathbf{F}}\}_{i\cdot}$ is the same as that of $\{\hat{\mathbf{F}}\}_{i\cdot} - (1+z_i)\mathbf{1}_K$, Theorem 3.4.1 amounts to saying that the probability of correctly classifying unlabeled vectors $\mathbf{x}_i$ genuinely belonging to class $\mathcal{C}_b$ is asymptotically given by the probability of $\{\mathbf{g}_i\}_b$ being the maximal element of $\mathbf{g}_i$. This is formulated in the following corollary.

Corollary 3.1. *Let Assumptions 3.1–3.2 hold. Then, under the notations of Theorem 3.4.1,*

$$\mathbb{P}\left(\mathbf{x}_i \to \mathcal{C}_b | \mathbf{x}_i \in \mathcal{C}_b\right) - \mathbb{P}\left(\{\mathbf{g}_i\}_b > \max_{a\neq b}\left(\{\mathbf{g}_i\}_a\right) | \mathbf{x}_i \in \mathcal{C}_b\right) \to 0.$$

In particular, for $K = 2$, and $a \neq b \in \{1,2\}$,

$$\mathbb{P}\left(\{\mathbf{g}_i\}_b > \max_{a\neq b}\{\mathbf{g}_i\}_a | \mathbf{x}_i \in \mathcal{C}_b\right) = \Phi(\theta_b^a), \quad \text{with } \theta_b^a \equiv \frac{\{\mathbf{m}_b\}_b - \{\mathbf{m}_b\}_a}{\sqrt{\{\mathbf{\Sigma}_b\}_{bb} + \{\mathbf{\Sigma}_b\}_{aa} - 2\{\mathbf{\Sigma}_b\}_{ab}}}$$

where $\Phi(u) = \frac{1}{2\pi}\int_{-\infty}^{u} e^{-\frac{t^2}{2}}dt$ is the Gaussian distribution function.

Corollary 3.1 allows us to approach the empirical classification accuracy as approximating it with the probability of correct classification given in the corollary. Figure 3.3 displays a comparison between simulated accuracy from various pairs of digits from the MNIST data

against our theoretical results; to apply our results, a 2-class Gaussian mixture model is assumed with means and covariances equal to the empirical means and covariances of the individual digits, evaluated from the full 60 000-image MNIST database. It is quite interesting to observe that, despite the obvious inadequacy of a Gaussian mixture model for this image database, the theoretical predictions are in strong agreement with the practical performances.

Remarkably, by substituting Equations 3.8–3.9 into the expression of the asymptotic performance given in Corollary 3.1, we conclude that with an optimally chosen β, the high dimensional probability of correct classification increases with larger $c_{[l]} = n_{[l]}/n$, while exhibiting a negligible growth with respect to $c_{[u]}$. This suggests an inefficient learning from the information of unlabelled data. We refer to Section 3.6 and Chapter 4 for further discussion on this issue.

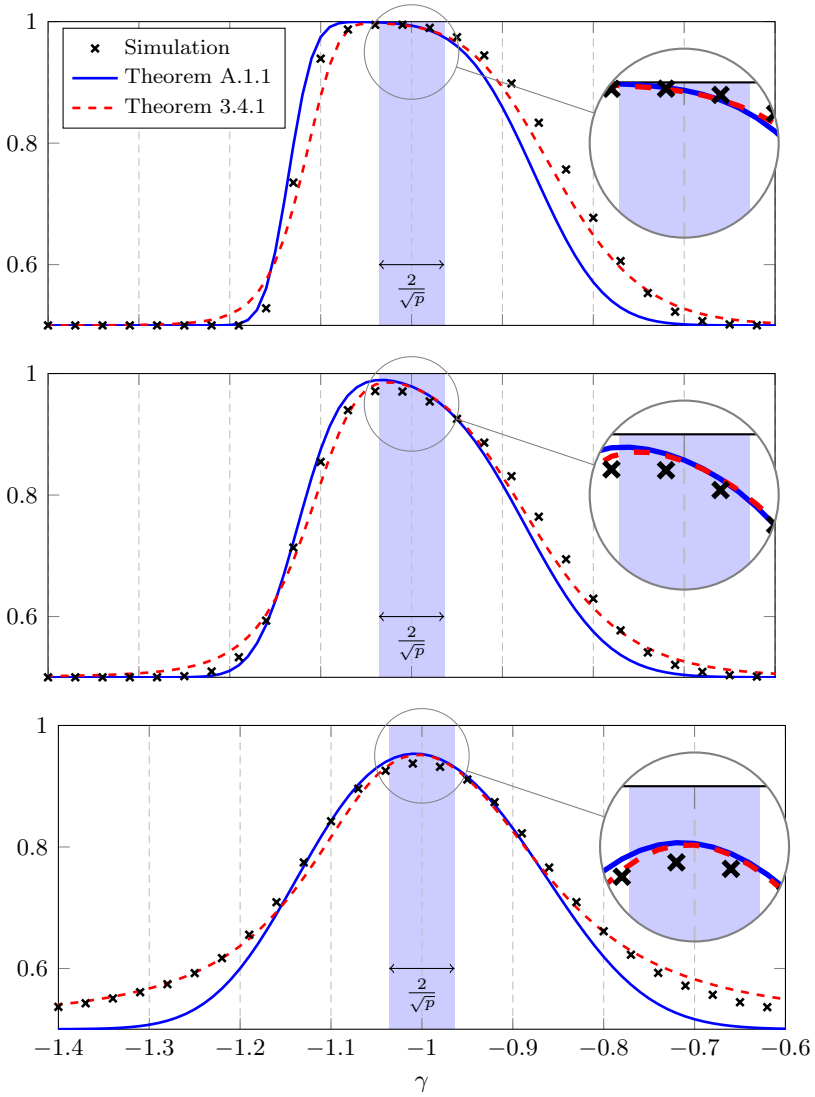

Figure 3.3: Theoretical and empirical accuracy as a function of γ for 2-class MNIST data (top: digits (0,1), middle: digits (1,7), bottom: digits (8,9)), $n = 1024$, $p = 784$, $n_{[l]}/n = 1/16$, $n_{[u]1} = n_{[u]2}$, Gaussian kernel. Averaged over 50 iterations.

3.5 Consequences

3.5.1 Semi-Supervised Learning beyond Two Classes

An immediate consequence of Corollary 3.1 is that, for $K > 2$, there exists a Gaussian mixture model for which the semi-supervised learning algorithms under study necessarily fail to classify at least one class. To see this, we consider $K = 3$ and let $\boldsymbol{\mu}_3 = 3\boldsymbol{\mu}_2 = 6\boldsymbol{\mu}_1$, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}_3$, $n_1 = n_2 = n_3$, $n_{[l]1} = n_{[l]2} = n_{[l]3}$. First, it follows from Corollary 3.1 that,

$$\mathbb{P}\left(\mathbf{x}_i \to \mathcal{C}_2 | \mathbf{x}_i \in \mathcal{C}_2\right) \leq \mathbb{P}\left(\{\mathbf{g}_i\}_2 > \{\mathbf{g}_i\}_1 | \mathbf{x}_i \in \mathcal{C}_2\right) + o(1) = \Phi(\theta_2^1) + o(1)$$
$$\mathbb{P}\left(\mathbf{x}_i \to \mathcal{C}_3 | \mathbf{x}_i \in \mathcal{C}_3\right) \leq \mathbb{P}\left(\{\mathbf{g}_i\}_3 > \{\mathbf{g}_i\}_1 | \mathbf{x}_i \in \mathcal{C}_3\right) + o(1) = \Phi(\theta_3^1) + o(1)$$

Then, under Assumptions 3.1–3.2 and the notations of Corollary 3.1,

$$\theta_2^1 = \operatorname{sgn}(h'(\tau)) \frac{\boldsymbol{\mu}_1^2}{\sqrt{\{\boldsymbol{\Sigma}_2\}_{22} + \{\boldsymbol{\Sigma}_2\}_{11} - 2\{\boldsymbol{\Sigma}_2\}_{12}}}$$
$$\theta_3^1 = -\operatorname{sgn}(h'(\tau)) \frac{15\boldsymbol{\mu}_1^2}{\sqrt{\{\boldsymbol{\Sigma}_3\}_{33} + \{\boldsymbol{\Sigma}_3\}_{11} - 2\{\boldsymbol{\Sigma}_3\}_{13}}}$$

so that $h'(\tau) < 0 \Rightarrow \theta_2^1 < 0$, $h'(\tau) > 0 \Rightarrow \theta_3^1 < 0$, while $h'(\tau) = 0 \Rightarrow \theta_2^1 = \theta_3^1 = 0$. As such, the correct classification rate of elements of $\mathcal{C}_2$ and $\mathcal{C}_3$ cannot be simultaneously greater than $\frac{1}{2}$, leading to necessarily inconsistent classifications.

It is nonetheless easy to check that this kind of inconsistency cannot occur if $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ are mutually orthogonal (which is often bound to occur with large dimensional data). Indeed, note that all first three terms at the right-hand side of (3.8) can be viewed as products of some centered vectors $\tilde{\mathbf{v}}_k = \mathbf{v}_k - \sum_{k'=1}^K r_{k'} \mathbf{v}_{k'}$ where $\sum_{k'=1}^K r_{k'} = 1$. Inconsistency occurs to class k if there exist $a, b \neq k$ such that $\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_b > \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_k > \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_a$. To better understand the cause of this inconsistency, let us consider two extreme scenarios: (i) the $\mathbf{v}_k$ differ by 'intensity', i.e., $\mathbf{v}_k = e_k \mathbf{v}$ for $k \in \{1, \ldots, K\}$, or (ii) the $\mathbf{v}_k$ differ by 'direction', i.e, $\mathbf{v}_k = \mathbf{v} + \mathbf{u}_k$ with orthogonal $\mathbf{u}_k$'s. In scenario (i), let $e_{\min} = \operatorname{argmin}_{k \in \{1,\ldots,K\}} e_k$ and $s_{\max} = \operatorname{argmax}_{k \in \{1,\ldots,K\}} e_k$; then, for $k \neq \{e_{\min}, e_{\max}\}$, $\min\{\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{e_{\min}}, \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{e_{\max}}\} < \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_k < \max\{\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{e_{\min}}, \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{e_{\max}}\}$ and inconsistency is thus observed for classes $k \neq \{e_{\min}, e_{\max}\}$. Contrarily, in scenario (ii), for all $k \neq k' \in \{1, \ldots, K\}$, $\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_k \geq \tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{k'}$ since $\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_k \geq 0$ and $\tilde{\mathbf{v}}_k^T \tilde{\mathbf{v}}_{k'} \leq 0$. As such, inconsistency is less likely to occur if the $\mathbf{v}_k$'s have very different directions.

3.5.2 Choice of h and Suboptimality of the Heat Kernel

As a consequence of the previous section, we shall from here on concentrate on the semi-supervised classification of $K = 2$ classes. In this case, it is easily seen that,

$$(K = 2) \quad \forall a \neq b \in \{1, 2\}, \ \|\tilde{\boldsymbol{\mu}}_b\|^2 \geq \tilde{\boldsymbol{\mu}}_b^{\mathsf{T}} \tilde{\boldsymbol{\mu}}_a, \quad \tilde{t}_b^2 \geq \tilde{t}_a \tilde{t}_b$$

with equalities respectively for $\boldsymbol{\mu}_a = \boldsymbol{\mu}_b$ and $t_a = t_b$. This result, along with Corollary 3.1, implies the necessity of the conditions

$$h'(\tau) < 0, \quad h''(\tau) h(\tau) > h'(\tau)^2$$

to fully discriminate Gaussian mixtures. As such, from Corollary 3.1, by letting $\gamma = -1$, semi-supervised classification of $K = 2$ classes is always consistent under these conditions.

A quite surprising outcome of the necessary conditions on the derivatives of h is that the widely used Gaussian (or heat) kernel $h(t) = \exp(-\frac{t}{2\sigma^2})$, while fulfilling the condition $h'(t) < 0$ for all t (and thus $h'(\tau) < 0$), only satisfies $h''(t)f(t) = h'(t)^2$. This indicates that discrimination over $t_1, \ldots, t_K$, under the conditions of Assumption 3.1, is asymptotically *not* possible with a Gaussian kernel. This remark is illustrated in Figure 3.4 for a discriminative task between two centered isotropic Gaussian classes only differing by the trace of their covariance matrices. There, irrespective of the choice of the bandwidth σ, the Gaussian kernel leads to a constant $1/2$ accuracy, where a mere second order polynomial kernel selected upon its derivatives at τ demonstrates good performances. Since p-dimensional isotropic Gaussian vectors tend to concentrate "close to" the surface of a sphere, this thus suggests that Gaussian kernels are not inappropriate to solve the large dimensional generalization of the "concentric spheres" task (for which they are very efficient in small dimensions). In passing, the right-hand side of Figure 3.4 confirms the need for $h''(\tau)h(\tau) - h'(\tau)^2$ to be positive (there $|h'(\tau)| < 1$) as an accuracy lower than $1/2$ is obtained for $h''(\tau)h(\tau) - h'(\tau)^2 < 0$.

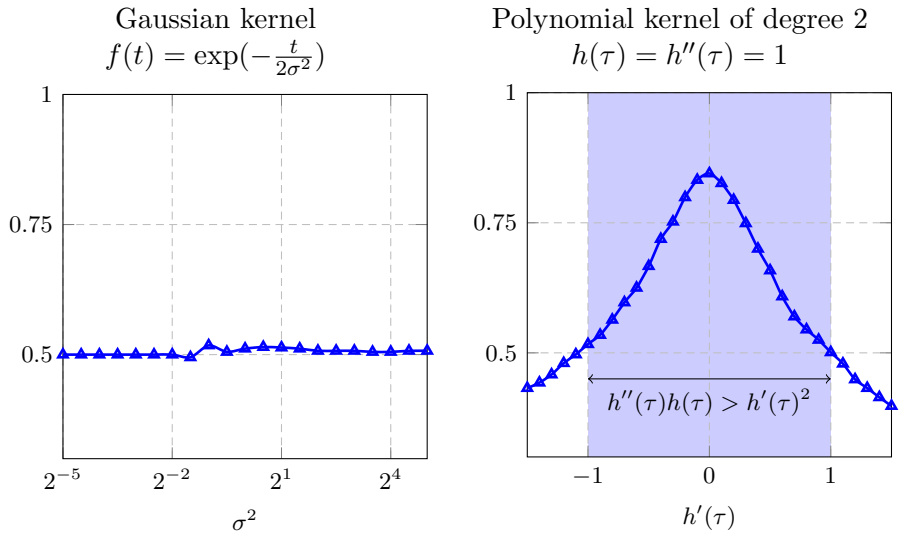

Figure 3.4: Empirical accuracy for 2-class Gaussian data with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathbf{C}_1 = I_p$ and $\mathbf{C}_2 = (1 + \frac{3}{\sqrt{p}})I_p$, $n = 1024$, $p = 784$, $n_l/n = 1/16$, $n_{[u]1} = n_{[u]2}$, $n_{[l]1} = n_{[l]2}$, $\gamma = -1$.

3.6 Summary and remarks

This chapter is part of a series of works evaluating the performance of kernel-based machine learning methods in the large dimensional data regime [45, 55, 56]. Relying on the derivations of [45] that provide a Taylor expansion of radial kernel matrices around the limiting common value τ of $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2$ for $i \neq j$ and at large p, we observed that the choice of the kernel function h merely affects the classification performances through the successive derivatives of h at τ. Of importance is the finding that, under a heat kernel assumption $h(t) = \exp(-\frac{t}{2\sigma^2})$, the studied

semi-supervised learning method fails to classify Gaussian mixtures of the type $\mathcal{N}(0, \mathbf{C}_k)$, which unsupervised learning or LS-SVMs are able to do [45, 55]. This paradox may deserve a more structural way of considering together methods on the spectrum from unsupervised to supervised learning.

The very fact that the kernel matrix $\mathbf{W}$ is essentially equivalent to the matrix $h(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ (the $n \times n$ matrix filled with $h(\tau)$ values), strongly disrupting the expected natural behavior of kernels. It is also quite instructive to note that, from the proof of our main results, the terms remaining after the expansion of $\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^{\gamma}$ appearing in (3.2) almost all vanish, strongly suggesting that similar results would be obtained if the inverse matrix in (3.2) were discarded altogether. This implies that the intra-unlabelled data kernel $\mathbf{W}_{[uu]}$ is of virtually no asymptotic use, leading to a vanishing classification rate at $c_{[l]} \to 0$. This consequence entails that even the (unsupervised) clustering performance obtained by [45] is not achieved, despite the presence of possibly numerous unlabelled data. A promising avenue of investigation would consist in finding a proper approach to ensure that $\mathbf{W}_{[uu]}$ is effectively used in the algorithm, which is the objective of the next chapter

Chapter 4

Improved semi-supervised learning with centering regularization

4.1 Motivation: inconsistency of existing algorithms on high dimensional data

In the analysis of semi-supervised Laplacian regularization algorithms presented in the above chapter, we proved that, regardless of the choice of Laplacian matrix, all these algorithms fail to learn effectively from large dimensional unlabelled data. This causes them to be surpassed by the unsupervised method of spectral clustering, hence the inconsistency issue raise in Section of the previous chapter. The objective of the present chapter is to search for a new graph regularization approach allowing for a consistent semi-supervised learning on very high dimensional data. To focus more on the question of consistency, we choose here to concentrate on the binary classification problem to avoid irrelevant details at this stage. Correspondingly, rather than a score matrix $\mathbf{F}$, we define here a vector $\mathbf{f} = [f_1, \ldots, f_n]^\mathsf{T} \in \mathbb{R}^n$ of scores with f_i being the score of $\mathbf{x}_i$. Let $\mathbf{f} = [\mathbf{f}_{[l]}; \mathbf{f}_{[u]}]$ where $\mathbf{f}_{[l]}$ stands for the score vector of labelled set and $\mathbf{f}_{[u]}$ for that of unlabelled set. To eliminate the bias discussed in the previous chapter, we shall use in the remainder of this chapter a class-balanced $\mathbf{f}_{[l]}$ defined as

$$\mathbf{f}_{[l]} = \left(\mathbf{I}_{n_{[l]}} - \frac{1}{n_{[l]}}\mathbf{1}_{n_{[l]}}\mathbf{1}_{n_{[l]}}^\mathsf{T}\right)\mathbf{y}_{[l]} \tag{4.1}$$

where $\mathbf{y}_{[l]} = [y_1, \ldots, y_{n_{[l]}}]^\mathsf{T}$ is the label vector with $y_i = (-1)^k$ if $\mathbf{x}_i \in \mathcal{C}_k$ for $i = \{1, \ldots, n_{[l]}\}$, $k = \{1, 2\}$. Then, from the optimization framework presented in (3.1), we have the solution of Laplacian regularization given by

$$\mathbf{f}_{[u]} = \mathbf{L}_{[uu]}^{(\gamma)-1}\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^{\gamma}\mathbf{f}_{[l]} \tag{4.2}$$

where $\mathbf{L}_{[uu]}^{(\gamma)}$ is the unlabelled subset of the generalized Laplacian matrix

$$\mathbf{L}^{(\gamma)} = \mathbf{I}_n - \mathbf{D}^{-1-\gamma}\mathbf{W}\mathbf{D}^{\gamma}.$$

Namely,

$$\mathbf{L}_{[uu]}^{(\gamma)} = \mathbf{I}_{n_{[u]}} - \mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^{\gamma}.$$

To gain more perspective on the cause of the inefficient learning from unlabelled data, we start with a discussion linking the issue to the data high dimensionality.

From a graph-signal processing perspective [57], since $L^{(\gamma)}_{[uu]}$ is the Laplacian matrix on the subgraph of unlabelled data, and a smooth signal $\mathbf{s}_{[u]}$ on the unlabelled data subgraph typically induces large values for the inverse smoothness penalty $\mathbf{s}^{\mathsf{T}}_{[u]}\mathbf{L}^{(\gamma)-1}_{[uu]}\mathbf{s}_{[u]}$, we may consider the operator $\mathcal{P}_u(\mathbf{s}_{[u]}) = \mathbf{L}^{(\gamma)-1}_{[uu]}\mathbf{s}_{[u]}$ as a "smoothness filter" strengthening smooth signals on the unlabelled data subgraph. The unlabelled scores $\mathbf{f}_{[u]}$ can be therefore seen as obtained by a two-step procedure:

1. propagating the predetermined labelled scores $\mathbf{f}_{[l]}$ through the graph with the γ-normalized weight matrix $\mathbf{D}^{-1-\gamma}_{[u]}\mathbf{W}_{[ul]}\mathbf{D}^{\gamma}_{[l]}$ through the label propagation operator

$$\mathcal{P}_l(\mathbf{f}_{[l]}) = \mathbf{D}^{-1-\gamma}_{[u]}\mathbf{W}_{[ul]}\mathbf{D}^{\gamma}_{[l]}\mathbf{f}_{[l]};$$

2. passing the received scores at unlabelled points from the above label propagation step through the smoothness filter $\mathcal{P}_u(\mathbf{s}_{[u]}) = \mathbf{L}^{(\gamma)-1}_{[uu]}\mathbf{s}_{[u]}$ to finally get $\mathbf{f}_{[u]} = \mathcal{P}_u\left(\mathcal{P}_l(\mathbf{f}_{[l]})\right)$.

It is easy to see that the first step is essentially a supervised learning process, whereas the second one makes it possible to capitalize on the global information contained in unlabelled data. However, as a consequence of the distance concentration phenomenon stated in Proposition 3.3.1, the similarities (weights) w_{ij} between high dimensional data vectors are dominated by the constant value $h(\tau)$ plus some small fluctuations:

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \tau + O(p^{-\frac{1}{2}}),$$

which results in the collapse of the smoothness filter:

$$\mathcal{P}_u(\mathbf{s}_{[u]}) = \mathbf{L}^{(\gamma)-1}_{[uu]}\mathbf{s}_{[u]} \simeq \left(\mathbf{I}_{n_{[u]}} - \frac{1}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}^{\mathsf{T}}_{n_{[u]}}\right)^{-1}\mathbf{s}_{[u]} = \mathbf{s}_{[u]} + \frac{1}{n_{[l]}}(\mathbf{1}^{\mathsf{T}}_{n_{[u]}}\mathbf{s}_{[u]})\mathbf{1}_{n_{[u]}},$$

meaning that at large values of p, only the constant signal direction $\mathbf{1}_{n_{[u]}}$ is amplified by the smoothness filter $\mathcal{P}_u$.

To understand this behavior, we recall that constant signals with the same value at all points are always considered to be the smoothest signal on the graph. This comes from the fact that all weights w_{ij} have non-negative values, so the smoothness penalty term $\Omega(\mathbf{s}) = \sum_{i,j} w_{ij}(s_i - s_j)^2$ is minimized at the value of zero if all elements of the signal $\mathbf{s}$ have the same value. Notice also that in perfect situations where the data points in different class subgraphs are connected with zero weights w_{ij}, class indicators (i.e., signals with constant values within class subgraphs which are different for each class) are just as smooth as constant signals for they also minimize the smoothness penalty term to zero. Even though such scenarios almost never happen in real life, it is hoped that the inter-class similarities are sufficiently weak so that the smoothness filter $\mathcal{P}_u$ is still effective. What is problematic for high dimensional learning is that, when the similarities w_{ij} tend to be indistinguishable due to the distance concentration issue of high dimensional data vectors, constant signals have overwhelming advantages to the point that they become the only direction privileged by the smoothness filter $\mathcal{P}_u$, with almost no discrimination between all

other directions. Consequently, there is nearly no utilization of the global information in high dimensional unlabelled data through Laplacian regularizations.

In this chapter, a novel semi-supervised graph regularization algorithm is proposed to address the aforementioned inconsistency problem of the traditional Laplacian approach with respect to unlabelled data. The proposed improvement is simple to implement and effective. To support the proposition of this new approach, we present a rigorous theoretical analysis placed under the large dimensional random matrix setting of large and numerous data (similar to the previous work [10] or to [58] in the context of spectral clustering). The proposed method is shown by the analysis to induce a consistent semi-supervised learning from high dimensional data, with labelled and unlabelled data learning efficiency lowered bounded respectively by Laplacian regularization and spectral clustering. As a matter of fact, the proposed method, featuring a tuning hyperparameter, consistently relates semi-supervised learning to both unsupervised and supervised learning in showing that, at the two extremes in the selection of the hyperparameter, the performance of unsupervised spectral clustering and that of Laplacian regularization, which is essentially a supervised learning method in high dimensions, are exactly recovered. With the hyperparameter optimally set somewhere between these two extremes, the algorithm fulfills precisely the semi-supervised learning goal of surpassing one-sided learning schemes by properly combining them, resulting in a significant advantage over the traditional Laplacian regularization. Beyond theoretical conclusions, the superiority of the new regularization method is also illustrated by simulations on various real data sets.

4.2 Semi-supervised graph regularization with centered similarity matrix

Based on the discussion of the previous section, we shall try to eliminate the dominant advantages of constant signals in terms of graph smoothness, in an attempt to render the smoothness filter $\mathcal{P}_u$ effective in extracting class-structured signals from other non-informative directions. As constant signals always have a smoothness penalty of zero, a very easy way to break their optimal smoothness is to introduce negative weights in the graph so that the values of the smoothness regularizer can go below zero. More specifically, when the intra-class similarities are positive on average and the inter-class similarities are negative on average, class-structured signals are bound to have a lower smoothness penalty than constant signals. However, the implementation of such an idea using both positive and negative similarities is hindered by the fact that the positivity of the data points degrees $d_i = \sum_{j=1}^n w_{ij}$ is no longer ensured, and having negative degrees can lead to severely unstable results. Take for instance the label propagation step $\mathcal{P}_l(\mathbf{f}_{[l]}) = \mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^{\gamma}\mathbf{f}_{[l]}$, at an unlabelled point $\mathbf{x}_i$, the sum of the received scores after that step equals to $d_i^{-1-\gamma}\sum_{j=1}^{n_{[l]}}(w_{ij}d_j^{\gamma})f_j$, the sign of which obviously changes dramatically with the signs of the degree of that point and those of labelled data, thus leading to extremely unstable classification results.

To cope with this problem, we propose here the usage of centered similarities $\hat{w}_{ij}$, for which the positive and negative weights are balanced out at any data point, i.e., for all $i \in \{1, \ldots, n\}$, $d_i = \sum_{j=1}^n w_{ij} = 0$. Given any similarity matrix $\mathbf{W}$, its centered version $\hat{\mathbf{W}} = \{\hat{w}_{ij}\}_{i,j=1}^n$ is

easily obtained by applying a projection matrix $\mathbf{P} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\right)$ on both sides:

$$\hat{\mathbf{W}} = \mathbf{PWP}.$$

As a first advantage, the centering approach allows us to remove the degree matrix altogether (for the degrees are exactly zero now) from the updated smoothness penalty

$$\hat{Q}(s) = \sum_{i,j=1}^{n} \hat{w}_{ij}(s_i - s_j)^2 = -\mathbf{s}^\mathsf{T}\hat{\mathbf{W}}\mathbf{s}, \tag{4.3}$$

securing thus a stable behavior of graph regularization with both positive and negative weights.

Another merit of using centered similarities is that the distance between the intra-class similarities and inter-class similarities in the previous graph is preserved, in the sense that the average of inter-class similarities minus the average of inter-class similarities stays unchanged after centering. Since the total sum of centered similarities $\hat{w}_{ij}$ amounts to zero, the average of intra-class similarities is always positive while that of inter-class similarities negative as long as the former are greater on average than the latter, which remains a necessary condition for a functional semi-supervised graph regularization. Furthermore, in the common situations where the similarity matrices $\mathbf{W}$ are constructed through a kernel function, e.g., through the popular radial basis function (RBF) kernel $w_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}$, there exists (by definition of kernel functions) a mapping $x \mapsto \phi(x)$ such that

$$w_{ij} = \phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j).$$

Since

$$\hat{w}_{ij} = \left(\phi(\mathbf{x}_i) - \frac{1}{n}\sum_{t=1}^{n}\phi(x_t)\right)^\mathsf{T}\left(\phi(\mathbf{x}_j) - \frac{1}{n}\sum_{t=1}^{n}\phi(x_t)\right),$$

the centering operation is equivalent to translating the feature vectors $\phi(\mathbf{x}_i)$ by moving their center to the origin, meaning that the relative positions between feature vectors remain intact after the centering step.

This being said, a problematic consequence of regularization procedures employing positive and negative weights is that the convexity of the optimization problem is no longer ensured. In fact, the optimization may have an infinite solution. To deal with this issue, we add a constraint on the norm of the solution. Letting $\mathbf{f}_{[l]}$ be given by (4.1), the new optimization problem may now be posed as follows:

$$\begin{aligned} &\min_{\mathbf{f}_{[u]} \in \mathbb{R}^{n_{[u]}}} -\mathbf{f}^\mathsf{T}\hat{\mathbf{W}}\mathbf{f} \\ &\text{s.t.} \|\mathbf{f}_{[u]}\|^2 = n_{[u]}e^2. \end{aligned} \tag{4.4}$$

Naturally, the optimization can be solved by introducing a Laplacian multiplier α to the norm constraint $\|\mathbf{f}_{[u]}\|^2 = n_{[u]}e^2$ and the solution is given by

$$\mathbf{f}_{[u]} = \left(\alpha\mathbf{I}_{n_{[u]}} - \hat{\mathbf{W}}_{[uu]}\right)^{-1}\hat{\mathbf{W}}_{[ul]}\mathbf{f}_{[l]} \tag{4.5}$$

Algorithm 1 Semi-Supervised Graph Regularization with Centered Similarities

1: **Input:** $n_{[l]}$ pairs of labelled points and labels $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_{[l]}}, y_{n_{[l]}})\}$ with $y_i \in \{-1, 1\}$ the class label of $\mathbf{x}_i$, and $n_{[u]}$ unlabelled data $\{\mathbf{x}_{n_{[l]}+1}, \ldots, \mathbf{x}_n\}$.
2: **Output:** Classification of unlabelled data $\{\mathbf{x}_{n_{[l]}+1}, \ldots, \mathbf{x}_n\}$.
3: Compute the similarity matrix $\mathbf{W}$.
4: Compute the centered similarity matrix $\hat{\mathbf{W}} = \mathbf{PWP}$ with $\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, and define $\hat{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{W}}_{[ll]} & \hat{\mathbf{W}}_{[lu]} \\ \hat{\mathbf{W}}_{[ul]} & \hat{\mathbf{W}}_{[uu]} \end{bmatrix}$.
5: Set $\mathbf{f}_{[l]} = \left(\mathbf{I}_{n_{[l]}} - \frac{1}{n_{[l]}}\mathbf{1}_{n_{[l]}}\mathbf{1}_{n_{[l]}}^\mathsf{T}\right)\mathbf{y}_{[l]}$ with $\mathbf{y}_{[l]}$ the vector containing labelled y_i.
6: Compute the class scores of unlabelled data $\mathbf{f}_{[u]} = \left(\alpha\mathbf{I}_{n_{[u]}} - \hat{\mathbf{W}}_{[uu]}\right)^{-1}\hat{\mathbf{W}}_{[ul]}\mathbf{f}_{[l]}$ for some $\alpha > \|\hat{\mathbf{W}}_{[uu]}\|$.
7: Classify unlabelled data $\{\mathbf{x}_{n_{[l]}+1}, \ldots, \mathbf{x}_n\}$ by the signs of $\mathbf{f}_{[u]}$.

where α is determined by $\alpha > \|\hat{\mathbf{W}}_{[uu]}\|$ and $\|\mathbf{f}_{[u]}\|^2 = n_{[u]}e^2$. In practice, α can be used directly as a parameter for more convenient implementation. We summarize the method in Algorithm 1.

The proposed algorithm induces almost no extra cost to the classical Laplacian approach, except for the addition of the parameter α controlling the norm of $\mathbf{f}_{[u]}$. As will be demonstrated in the next section on performance analysis, the existence of this parameter, aside from making the regularization with centered similarities a well-posed problem, actually allows one to adjust the combination of labelled and unlabelled information in search for an optimal semi-supervised learning performance.

4.3 Performance Analysis

With the proposition of the centered similarities regularization intuitively justified in Subsection 4.2, the main purpose of this section is to provide mathematical support for its effective high dimensional learning capabilities from not only labelled data but also from unlabelled data, allowing for a theoretically guaranteed performance gain over the classical Laplacian approach (through an enhanced utilization of unlabelled data). The theoretical results also point out that the learning performance of the proposed method has an unlabelled data learning efficiency that is at least as good as spectral clustering, as opposed to Laplacian regularization.

We first enunciate the central theorem providing the statistical characterization of unlabelled data scores $\mathbf{f}_{[u]}$ obtained by the proposed updated algorithm. As the new algorithm will be shown to draw both on labelled and unlabelled data information, the complex interactions between these two types of data generate more intricate outcomes than in [10]. To facilitate the interpretation of the theoretical results without cumbersome notations, we restrict the theorem to the homoscedastic case as considered in linear discrimination analysis (i.e., the class covariances are taken equal, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$), without affecting the key messages of the conclusions given subsequently. We refer the interested reader to the appendix for an extended version of the theorem along with its proof.

Theorem 4.3.1. *Let Assumption 5.1 hold with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, h be three-times continuously differentiable in a neighborhood of τ, and $\mathbf{f}_{[u]}$ be the solution of (4.4) with fixed norm $n_{[u]}e^2$.*

Then, for $n_{[l]}+1 \leq i \leq n$ (i.e., $\mathbf{x}_i$ unlabelled) and $\mathbf{x}_i \in \mathcal{C}_k$,

$$f_i = g_i + o_P(1)$$

where

$$g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m, \sigma^2\right)$$

for some $m, \sigma^2 > 0$. More precisely, defining

$$\theta = \frac{c_{[u]}m}{2c_{[l]}} \tag{4.6}$$

and letting $s : (0, \|\mathbf{C} + \rho_1\rho_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\|) \to (0, +\infty)$ be the injective function given by

$$s(\xi) = \xi\rho_1\rho_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\left\{I_p - \xi\left[C + \rho_1\rho_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\right]\right\}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{4.7}$$

the values of m and σ^2 are determined by $\rho_1\rho_2 m^2 + \sigma^2 = e^2$ and

$$\frac{\sigma^2}{m^2} = \left[1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{(\rho_1\rho_2)^2 c_{[u]}}\right]^{-1}\left[\omega(\theta) + \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{\rho_1\rho_2 c_{[u]}} + \left(\frac{1}{1+\theta}\right)^2 \frac{q(\theta)}{\rho_1\rho_2 c_{[l]}}\right] \tag{4.8}$$

where

$$q(\theta) = \frac{\mathrm{tr}\left[(\mathbf{I}_p - s^{-1}(\theta)\mathbf{C})^{-1}C\right]^2}{p\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}(\mathbf{I}_p - s^{-1}(\theta)\mathbf{C})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2}$$
$$\omega(\theta) = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}(I_p - s^{-1}(\theta)\mathbf{C})^{-1}\mathbf{C}(\mathbf{I}_p - s^{-1}(\theta)\mathbf{C})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}(\mathbf{I}_p - s^{-1}(\theta)\mathbf{C})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^2}.$$

In the special cases where $\mathbf{C}_1 = \mathbf{C}_2 = \lambda^2\mathbf{I}_p$, the above theorem admits a much simpler form.

Corollary 4.1. *Under the conditions and notations of Theorem 4.3.1, let $\mathbf{C}_1 = \mathbf{C}_2 = \lambda^2\mathbf{I}_p$. Then the values of m, σ^2 are given by $\rho_1\rho_2 m^2 + \sigma^2 = e^2$ and*

$$\frac{\sigma^2}{m^2} = \left[1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{\lambda^4}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4(\rho_1\rho_2)^2 c_{[u]}}\right]^{-1}\left[\frac{\lambda^2}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2} + \left(\frac{\theta}{1+\theta}\right)^2 \frac{\lambda^4}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4\rho_1\rho_2 c_{[u]}}\right.$$
$$\left. + \left(\frac{1}{1+\theta}\right)^2 \frac{\lambda^4}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4\rho_1\rho_2 c_{[l]}}\right]. \tag{4.9}$$

Like the centered similarities regularization, the random walk normalized Laplacian algorithm, which is the only one ensuring non-trivial classification results among existing Laplacian algorithms for high dimensional data (as we recall from Section4.3 of Chapter3), also gives $g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m', \sigma'^2\right)$ for some other $m', \sigma'^2 > 0$ under the homoscedasticity assumption of $\mathbf{C}_1 = \mathbf{C}_2$. We shall use the variance over square mean ratio $r = \sigma^2/m^2$ as the inverse performance measure (i.e., lower r indicates better classification results for high dimensional data) in the following discussion. Denote by r_{lap} the ratio of the random walk normalized Laplacian algorithm, which is obtained from Theorem 3.4.1 in Chapter3 as

$$r_{\mathrm{lap}} = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\mathbf{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4} + \frac{\mathrm{tr}\mathbf{C}^2}{p\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4\rho_1\rho_2 c_{[l]}} \tag{4.10}$$

and by r_{ctr} the ratio for the centered similarities method, the expression of which has a rather complicated form given by (4.8).

Note importantly that the quantity $\theta(e)$ in fact reflects *the ratio between the labelled data scores* $\mathbf{f}_{[l]}$ *and the unlabelled data scores* $\mathbf{f}_{[u]}$ as

$$\theta = \frac{c_{[u]}m}{2c_{[l]}} \simeq \sqrt{\frac{\|\mathbb{E}\{\mathbf{f}_{[u]}\}\|^2}{\|\mathbf{f}_{[l]}\|^2}}.$$

We observe notably that, when $\|\mathbf{f}_{[l]}\|^2 \gg \|\mathbb{E}\{\mathbf{f}_{[u]}\}\|^2$, θ goes to zero, at which value the unlabelled data over dimension ratio $c_{[u]} = n_{[u]}/p$ disappears from the expression of r_{ctr}, suggesting the performance relies solely on the labelled data. Inversely, if θ goes to infinity, then it is the labelled data ratio $c_{[l]} = n_{[l]}/p$ that will be left out from (4.8), and the learning is only guided by the unlabelled data. In other words, the quantity θ can be seen as *a variable tuning the impacts of the two types of data* on the learning process, which is modified by changing the parameter e in the equality constraint $\|\mathbf{f}_{[u]}\| = n_{[u]}e^2$ of the optimization problem (4.4).

As stated in Subsection 4.2, the proposed method can be more conveniently implemented by Algorithm 1, with $\mathbf{f}_{[u]}$ computed by $\mathbf{f}_{[u]} = \left(\alpha\mathbf{I}_{n_{[u]}} - \hat{\mathbf{W}}_{[uu]}\right)^{-1}\hat{\mathbf{W}}_{[ul]}\mathbf{f}_{[l]}$ for some $\alpha > \|\hat{\mathbf{W}}_{[uu]}\|$. Obviously, the norm of $\mathbf{f}_{[u]}$ is controlled by the hyperparameter α with large α implying small $\|\mathbf{f}_{[u]}\|^2$, and consequently small θ, indicating that the labelled data information is emphasized at great values of α. By the same reasoning, the unlabelled data information becomes more influential as α gets close to its minimal limit $\alpha_{\text{inf}} = \|\hat{\mathbf{W}}_{[uu]}\|$. Actually, taking $\alpha \in \left(\|\hat{\mathbf{W}}_{[uu]}\|, +\infty\right)$ infinitely near the two extremes of its admissible range allows to retrieve respectively the performances of Laplacian regularization and spectral clustering, as will be demonstrated in the following.

Firstly, following the argument in Subsection 4.2 that using centered similarities should cause no loss of information as the difference between the intra-class and inter-class similarities is preserved, we indeed find, by comparing (4.8) and (4.10), that

$$\lim_{\theta\to 0} r_{\text{ctr}} = r_{\text{lap}},$$

meaning that the performance of the classical Laplacian regularization can be perfectly retrieved with the centered similarities approach by letting its learning process be completely guided with labelled data. In practice, this is achieved by letting $\alpha \to +\infty$, at which $\|\mathbb{E}\{\mathbf{f}_{[u]}\}\|^2 < \|\mathbf{f}_{[u]}\|^2 \to 0$, leading to $\theta \to 0$. We thus remark that, with an appropriately set α, the performance of the proposed method is *lowered bounded by that of Laplacian regularization.*

After ensuring the superiority of the new regularization method over the original approach, we now proceed to provide further guarantee on its unlabelled data learning efficiency by comparing it to spectral clustering, the standard unsupervised graph learning technique.

Recall that the regular graph smoothness penalty term $Q(\mathbf{s})$ of a signal $\mathbf{s}$ can be written as $Q(\mathbf{s}) = \mathbf{s}^\mathsf{T}\mathbf{L}^{(\gamma)}\mathbf{s}$. In an unsupervised spectral learning manner, we therefore seek the unit-norm vector that minimizes the smoothness penalty, which is the eigenvector of $\mathbf{L}$ associated with the smallest eigenvalue. However, as $Q(\mathbf{s})$ reaches its minimum at the trivial solution vector $\mathbf{s} = \mathbf{D}^{-\gamma}\mathbf{1}_n$, the sought-for solution is provided by the eigenvector associated with the second

smallest eigenvalue. Instead, by (4.3), the updated smoothness penalty term with centered similarities, that is $\hat{Q}(\mathbf{s}) = \mathbf{s}^{\mathsf{T}}\hat{\mathbf{W}}\mathbf{s}$, does not achieves its minimum for "flat" signals, and thus the eigenvector associated with the smallest eigenvalue is here a valid solution. Among the two common choices of Laplaican matrices in spectral clustering, the unnormalized Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$ has long been known to behave unstably [20], as opposed to the symmetric normalized Laplacian $\mathbf{L}_s = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$, so fair comparison should be made versus $\mathbf{L}_s$ rather than $\mathbf{L}$.

Let us define $d_{\text{inter}}(\mathbf{v})$ as the inter-cluster distance operator that takes as input a real-valued vector $\mathbf{v}$ of dimension n, then returns the distance between the centroids of the clusters formed by the set of points $\{\mathbf{v}_i | 1 \leq i \leq n, \mathbf{x}_i \in \mathcal{C}_k\}$, for $k \in \{1, 2\}$; and $d_{\text{intra}}(\mathbf{v})$ be the intra-cluster distance operator that returns the standard deviation within clusters. As the purpose of clustering analysis is to produce clusters conforming to the intrinsic classes of data points, with low variance within a cluster and large distance between clusters, the following proposition (see the proof in the appendix) shows that the performance of the classical normalized spectral clustering is practically the same as the one with centered similarities for high dimensional data. In other terms, the high dimensional performance of Laplacian spectral clustering on data samples of size $n_{[u]}$ is retrieved from the limiting results in Theorem 4.3.1 at $\theta \to +\infty$ (when spectral clustering leads to non-trivial partitioning). This remark is subsequently validated on simulations in Figure 4.2, where the empirical performance of Laplacian spectral clustering is found to closely match the theoretical performance of the centered similarity approach when letting the learning process guided completely by unlabelled data.

Proposition 4.3.1. *Under the conditions of Theorem 4.3.1, let* $\mathbf{v}_{\text{lap}}$ *be the eigenvector of* $\mathbf{L}_s$ *associated with the second smallest eigenvalue, and* $\mathbf{v}_{\text{ctr}}$ *the eigenvector of* $\hat{\mathbf{W}}$ *associated with the largest eigenvalue. Then,*

$$\frac{d_{\text{inter}}(\mathbf{v}_{\text{lap}})}{d_{\text{intra}}(\mathbf{v}_{\text{lap}})} = \frac{d_{\text{inter}}(\mathbf{v}_{\text{ctr}})}{d_{\text{intra}}(\mathbf{v}_{\text{ctr}})} + o_P(1)$$

for non-trivial clustering with $d_{\text{inter}}(\mathbf{v}_{\text{lap}})/d_{\text{intra}}(\mathbf{v}_{\text{lap}}), d_{\text{inter}}(\mathbf{v}_{\text{ctr}})/d_{\text{intra}}(\mathbf{v}_{\text{ctr}}) = O(1)$.

As explained before, the solution $\mathbf{f}_{[u]}$ of the centered similarities regularization can be expressed as $\mathbf{f}_{[u]} = \left(\alpha\mathbf{I}_{n_{[u]}} - \hat{\mathbf{W}}_{[uu]}\right)^{-1}\hat{\mathbf{W}}_{[ul]}\mathbf{f}_{[l]}$ for some $\alpha > \|\hat{\mathbf{W}}_{[uu]}\|$. Clearly, as $\alpha \downarrow \|\hat{\mathbf{W}}_{[uu]}\|$, $\mathbf{f}_{[u]}$ tends to align to the eigenvector of $\hat{\mathbf{W}}_{[uu]}$ associated with the largest eigenvalue, and we thus retrieve *the performance of spectral clustering on the unlabelled data subgraph.*

It is worth pointing out that, according to the results of [58], it may occur that the solution $\mathbf{v}_{[u]}$ obtained by spectral clustering be pure noise, i.e., $\mathbb{E}\{\mathbf{v}_{[u]}\} \simeq 0_{n_{[u]}}$ for all large n, p. For example, with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}$, we have $\mathbb{E}\{\mathbf{v}_{[u]}\} \simeq 0_{n_{[u]}}$ unless

$$c_{[u]} > \frac{1}{(\rho_1\rho_2)^2\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4},$$

suggesting that there exists a threshold for $c_{[u]}$ under which spectral clustering performs equally as random guess. This behavior of spectral clustering relates to an important phase transition phenomenon on spiked random matrix models discussed in [58] (see, e.g., [59, 60]). The phase transition phenomenon imples that the proposed semi-supervised learning scheme cannot produce reasonable classification results (i.e., bounded values of r_{ctr}) by solely relying on unlabelled

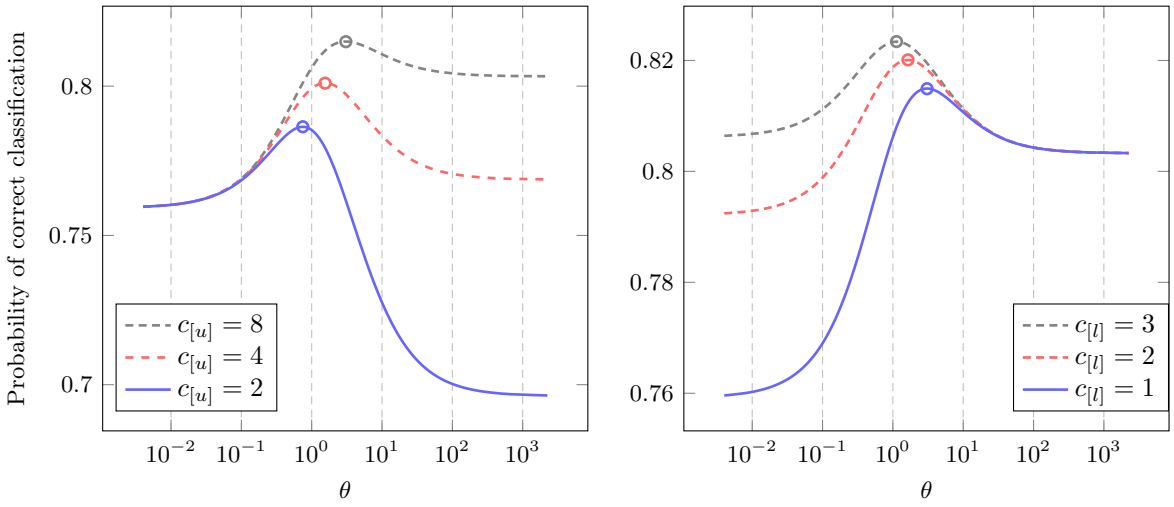

Figure 4.1: Asymptotic probability of correct classification as a function of θ with $\rho_1 = \rho_2$, $p = 100$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $\{C\}_{i,j} = .1^{|i-j|}$. Left: various $c_{[u]}$ with $c_{[l]} = 1$. Right: various $c_{[l]}$ with $c_{[u]} = 8$. Optimal values marked in circle.

data information (i.e., $\theta \to +\infty$) below the phase transition threshold. Indeed, we observe from (4.8) in the appendix that r_{ctr} has a well-defined positive value only when the following condition on θ is satisfied:

$$1 - \left(\frac{\theta}{1+\theta}\right)^2 \frac{q(\theta)}{(\rho_1\rho_2)^2 c_{[u]}} > 0. \tag{4.11}$$

Letting $\theta \to +\infty$ in the case of $\mathbf{C}_1 = \mathbf{C}_1 = \mathbf{C}$, for which $q(\theta) = 1/\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^4$ according to (4.9), we find the inequality condition (4.11) of $c_{[u]}$ to coincide with the phase transition threshold in (4.3), as expected. Generally speaking, a certain value θ' of θ is attainable through the adjustment of α if the inequality (4.11) is satisfied at $\theta = \theta'$. As such, we note importantly that *the attainable range of θ can only enlarge with greater $c_{[u]}$*.

It is obvious by looking at (4.8) that, at the same value of θ, r_{ctr} is *a strictly decreasing function of both $c_{[l]}$ and $c_{[u]}$*. Combining this observation with the remark that the attainable range of θ can only broaden with larger $c_{[u]}$ and is not affected by the value of $c_{[l]}$, we deduce straightforwardly that, with an appropriately chosen α, the performance of the proposed method consistently benefits from the addition of input data, *whether labelled or unlabelled*, as illustrated in Figure 4.1.

The following conclusion summarizes the main remarks obtained above.

Conclusion 1. *The proposed centered similarities regularization, implemented by Algorithm 1 with the hyperparameter α, allows one to*

1. *recover the high dimensional performance of Laplacian regularization at $\alpha \to +\infty$;*
2. *recover the high dimensional performance of spectral clustering at $\alpha \downarrow \|\hat{\mathbf{W}}_{[uu]}\|$;*

3. *accomplish a consistent high dimensional semi-supervised learning for α set between the two extremes, thus leading to an increasing performance gain over Laplacian regularization with greater amounts of unlabelled data.*

4.4 Experimentation

This section provides empirical evidence to support the proposition of centered similarities regularization, by comparing it with Laplacian regularization through simulations under and beyond the settings of the theoretical analysis.

4.4.1 Validation on Finite-Size Systems

We first validate the asymptotic results of the previous section on finite data sets of relatively small sizes ($n, p \sim 100$). Recall from Section 4.3 that the asymptotic performance of Laplacian regularization and spectral clustering are recovered by centered similarities regularization at extreme values of the hyperparameter θ. In other words, the high dimensional accuracies of Laplacian regularization and spectral clustering are given by Equation (4.8) of Theorem 4.3.1, respectively in the limit $\theta = 0$ and $\theta = +\infty$ (when spectral clustering yields non-trivial solutions); this is how the theoretical values of both methods are computed in Figure 4.2. The finite-sample results are given for the best (oracle) choice of the hyperparameter γ in the generalized Laplacian matrix $\mathbf{L}^{(\gamma)} = \mathbf{I} - \mathbf{D}^{-1-\gamma}\mathbf{W}\mathbf{D}^{\gamma}$ for Laplacian regularization and spectral clustering, and for the optimal (oracle) choice of the hyperparameter α for centered similarities regularization.

Under a non-trivial Gaussian mixture model setting (see caption) with $p = 100$, Figure 4.2 demonstrates a sharp prediction of the average empirical performance by the asymptotic analysis. As revealed by the theoretical results, the Laplacian regularization fails to learn effectively from unlabelled data, causing it to be outperformed by the purely unsupervised spectral clustering approach (for which the labelled data are treated as unlabelled ones) for sufficiently numerous unlabelled data. The performance curve of the proposed centered similarities regularization, on the other hand, is consistently above that of spectral clustering, with a growing advantage over Laplacian regularization as the number of unlabelled data increases.

Figure 4.2 also interestingly shows that the unsupervised performance of spectral clustering is noticeably reduced when the covariance matrix of the data distribution changes from the identity matrix to a slightly disrupted model (here for $\{\mathbf{C}\}_{i,j} = .1^{|i-j|}$). On the contrary, the Laplacian regularization, the high dimensional performance of which relies essentially on labelled data, is barely affected. This is explained by the different impacts labelled and unlabelled data have on the learning process, which can be understood from the theoretical results of the previous section.

4.4.2 Beyond the Model Assumptions

After verifying the advantage of the proposed centered similarities regularization in a finite (and not so large) sample setting, we are now interested in examining the extent of its superiority beyond the analysis framework.

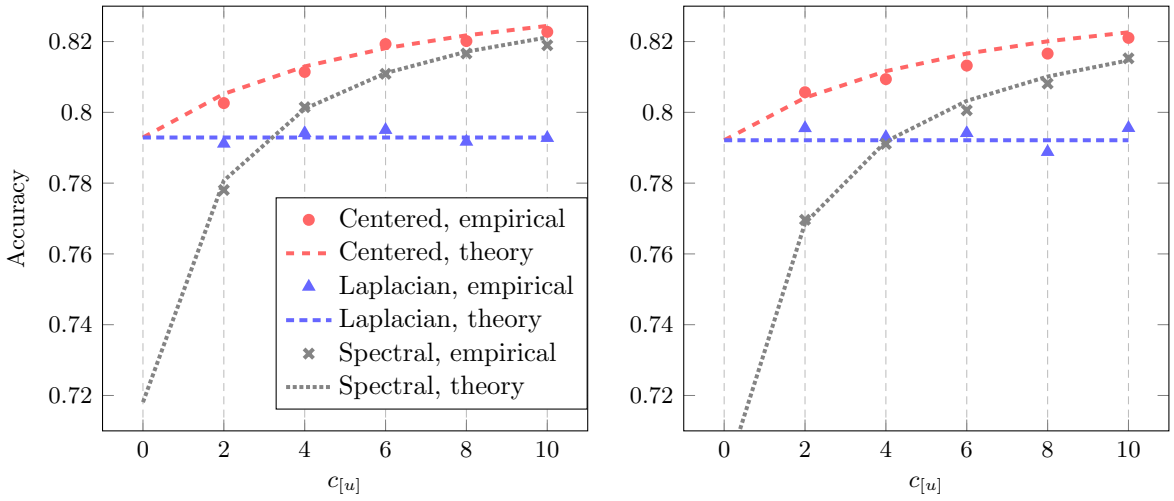

Figure 4.2: Empirical and theoretical accuracy as a function of $c_{[u]}$ with $c_{[l]} = 2$, $\rho_1 = \rho_2$, $p = 100$, $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [-1, 0, \ldots, 0]^\mathsf{T}$, $\mathbf{C} = \mathbf{I}_p$ (left) or $\{\mathbf{C}\}_{i,j} = .1^{|i-j|}$ (right). Graph constructed with $w_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/p}$. Averaged over $50000/n_{[u]}$ iterations.

As thoroughly discussed in Subsection 4.2, the key element causing the unlabelled data learning inefficiency of Laplacian regularization is the negligible distinction between inter-class and intra-class similarities, induced by the *distance concentration* of high dimensional data. It is important to understand that this concentration phenomenon is essentially independent of the Gaussianity of the data. Proposition 3.3.1 can indeed be generalized to a wider statistical model by a mere law of large numbers; this is the case for instance of all high dimensional data vectors $\mathbf{x}_i$ of the form $\mathbf{x}_i = \boldsymbol{\mu}_k + \mathbf{C}_k^{\frac{1}{2}} \mathbf{z}_i$, for $k \in \{1, 2\}$, where $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $\mathbf{C}_k \in \mathbb{R}^{p \times p}$ are means and covariance matrices as specified in Assumption 5.1 and $\mathbf{z}_i \in \mathbb{R}^p$ any random vector of independent elements with zero mean, unit variance and bounded fourth order moment.

As a side comment, it worth pointing out that the $k-$nearest neighbors (KNN) graphs, constructed by letting $w_{ij} = 1$ if data points $\mathbf{x}_i$ or $\mathbf{x}_j$ is among the k nearest (k being the parameter to be set beforehand) to the other data point, and $w_{ij} = 0$ if not, are not covered by the present analytic framework. Our study only deals with graphs where w_{ij} is exclusively determined by the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, while in the KNN graphs, w_{ij} is dependent of all pairwise distances of the whole data sets. Nonetheless, KNN graphs evidently suffer the same problem of distance concentration, for they are still based on the distances between data points. It is thus natural to expect that the proposed centering procedure may also be advantageous on KNN graphs.

Upon the above remarks, we expect the advantage of the proposed method to manifest itself on practical datasets, whenever a weak difference between inter-class and intra-class similarities is observed (and whenever the data themselves or the relevant features to classify are obviously not too far from a mixture model). The exact convergence of all distances to a common limit is of course an extreme mathematically ideal scenario; to gain an actual sense of how the Laplacian regularization and the proposed centered similarities approaches behave under different levels of distance concentration, we provide first, as a real-life example, simulations on datasets from the standard MNIST database of handwritten digits [52]. These are depicted in Figures 4.3–4.4.

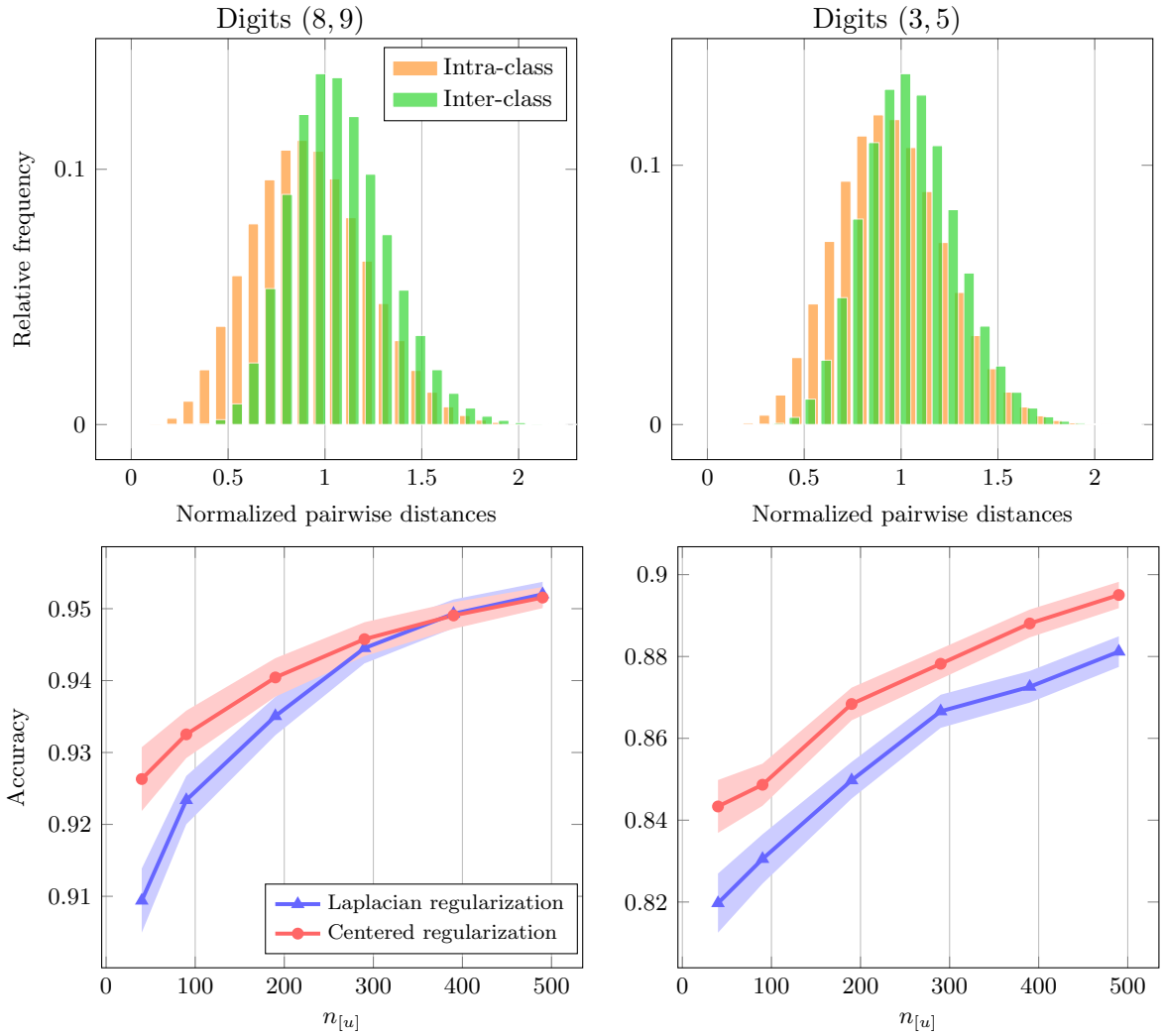

Figure 4.3: Top: distribution of normalized pairwise distances $\|\mathbf{x}_i - \mathbf{x}_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ for 2-class MNIST data. Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

As the performance of the methods tends to depend on the similarity graph, for a fair and extensive comparison of Laplacian and centered similarities regularizations, the results displayed here are obtained on their respective best performing graphs, selected among commonly used graphs including KNN graphs with various numbers of neighbors $k = \{2^1, \ldots, 2^q\}$, for q the largest integer such that $2^q < n$, and graphs constructed by Gaussian (also called RBF) kernels, i.e., $w_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2}$, with bandwidth σ set to the average data vector distance. The hyperparameters of the Laplacian and centered similarities regularization approaches are set optimally within the admissible range.[1]

Figure 4.3 shows that high classification accuracy is easily obtained on MNIST data, even with the classical Laplacian approach. However, it exhibits an unsatisfactory learning efficiency when compared to our proposed method. We also find that the benefit of the proposed algorithm is more perceptible on the classification task displayed in the right of Figure 4.3 (digits 3 versus 5) than on the left (digits 8 versus 9), for which the difference between inter-class and intra-class distances is more significant (and thus, in our setting, too "trivial"). To further evidence the impact of non-trivial classification, Figure 4.4 presents situations where the learning problem becomes more challenging in the presence of additive noise. Understandably, the distance concentration phenomenon is more acute in this noise-corrupted setting, and so is the performance gain generated by the centered similarities approach; this is indeed corroborated by Figure 4.4, demonstrating extremely large performance gains produced by the proposed method. In the right of Figure 4.4 where the similarity information is seriously disrupted by the noise, we observe the anticipated saturation effect when increasing $n_{[u]}$ for the Laplacian regularization, in contrast to the growing performance of the proposed approach. This suggests, in conclusion, that the centered similarities approach is a privileged solution in all situations, but is especially meaningful when the distinction between intra-class and inter-class similarities is quite subtle.

In order to further illustrate the advantage of the proposed method on more challenging datasets, we subsequently compare the Laplacian and centered similarities regularization methods on the popular Cifar10 database [61]. To obtain meaningful results, the data went through a feature extraction step using the standard pre-trained ResNet-50 network [62]. Other experimental settings are the same as for the above MNIST data. The simulations are reported in Figure 4.5, where the findings confirm again the superiority of the proposed centered similarities approach.

4.5 Concluding Remarks

The key to the proposed semi-supervised learning method lies in the replacement of conventional Laplacian regularizations by a centering operation on similarities. The motivation behind this operation is rooted in the large dimensional concentration of pairwise-data distances and thus likely to extend beyond the present graph-based semi-supervised learning schemes. It would in particular be interesting to know whether other advanced learning models involving Laplacian regularizations benefit from the same update. A specific example is Laplacian support vector machines (Laplacian SVMs) [63], which is another widespread semi-supervised learning algorithm. Answering this question about Laplacian SVMs is however not a straightforward

[1] Specifically, the hyperparameter γ of Laplacian regularization is searched among the values from -2 to 0 with a step of 0.02, and the hyperparameter α of centered similarities regularization within the grid $\alpha = (1+10^t)\|\hat{\mathbf{W}}_{[uu]}\|$ where t varies from -3 to 3 with a step of 0.1. The results outside these ranges are observed to be non-competitive.

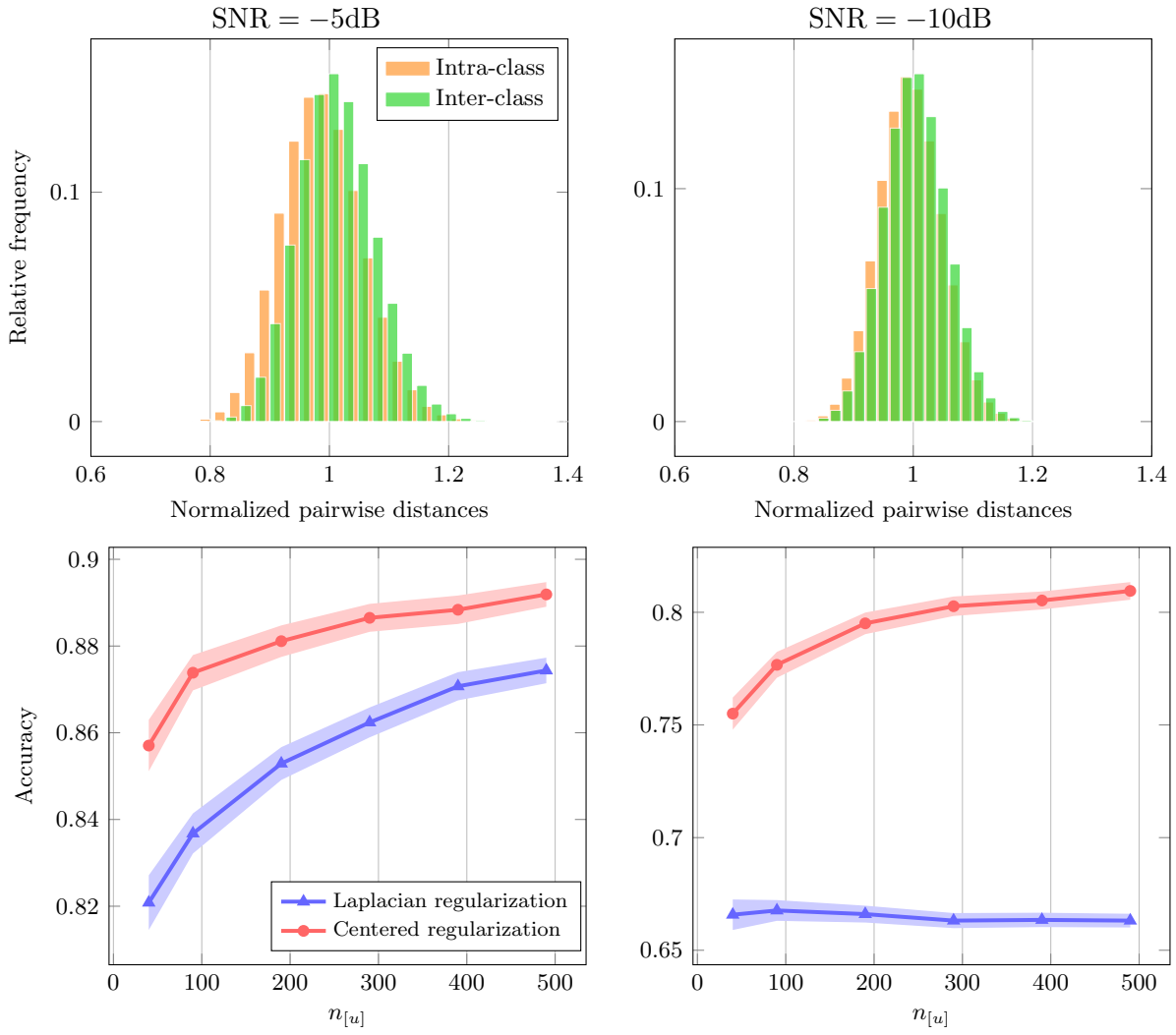

Figure 4.4: Top: distribution of normalized pairwise distances $\|\mathbf{x}_i - \mathbf{x}_j\|^2/\bar{\delta}$ $(i \neq j)$ with $\bar{\delta}$ the average of $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ for noised MNIST data (8,9). Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

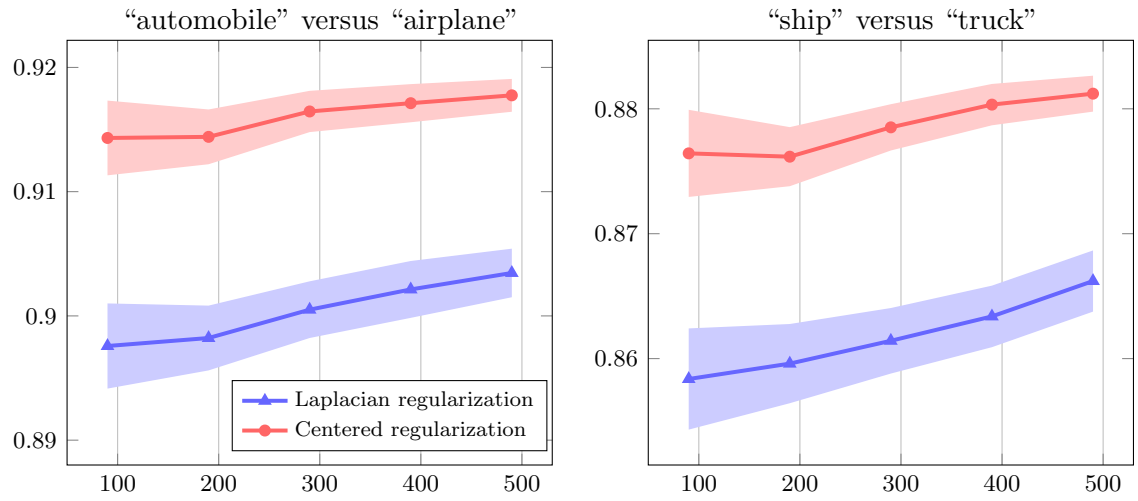

Figure 4.5: Average accuracy on two-class Cifar10 data as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations with 99% confidence intervals represented by shaded regions.

extension of the present analysis. Unlike the outcomes of Laplacian regularization, Laplacian SVMs are learned through an optimization problem without an explicit solution; additional technical tools, such as those recently devised in the work of [6] and in the subsequent chapters, to deal with implicit objects are required for analyzing their performance.

As already anticipated by the theoretical results, it is not surprising that the proposed centered similarities regularization empirically produces large performance gains over the standard Laplacian regularization when the aforementioned distance concentration problem is severe on the experimented data. However, it is quite illuminating to observe that even on datasets with weak distance concentration, for which the standard Laplacian approach exhibits a clear performance growth with respect to unlabelled data, the advantage of the proposed algorithm is still preserved. This attests to the general potential of such high dimensional studies for improving machine learning algorithms by identifying and settling some underlying issues compromising their learning performance, which would be difficult to spot if not through high dimensional analyses.

Part II

Statistical learning methods with no closed-form solution

Chapter 5

Statistical properties of high dimensional support vector machines

5.1 Introduction

Support vector machines (SVMs), originally proposed in [18], are one of the most popular classification tools in machine learning, thanks to their highly competitive performance in various real-world applications such as face recognition, image processing, and text mining, as well as their easy implementation and computational efficiency. The idea of SVMs is quite simple: given a set of input training data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ with class labels $y_i = \pm 1$, $i = \{1, \ldots, n\}$, SVMs aim to find a hyperplane defined by $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0 = 0$ which separate the two classes of input training vectors with largest distance between them, in order to maximize the generalization ability on unseen data (the optimization problem is as formulated in Equation 5.1 of Section 5.2). It was demonstrated that only a part of the training data, the so-called support vectors which determine the margin between the classes, have to be taken into account when constructing the decision function. As a matter of fact, the decision direction $\boldsymbol{\beta}$ is of the form $\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n} c_i y_i \mathbf{x}_i$ where the non-negative weights c_i are usually learned through a dual formulation of the original SVM optimization. We thus refer to them as the dual coefficients. Besides, by the properties of the dual optimization, c_i is zero unless $\mathbf{x}_i$ is a vector on the margin. The SVM method is also extended to data sets that are not linearly separable by allowing some data vectors to fall inside the margin, with a hyperparameter τ controlling the penalty of the data vectors inside the margin. The mathematical formulation of this extension is presented in (5.2). From the standpoint of empirical risk minimization [64], SVMs can be seen as a special case with a hinge loss function and a L_2 regularization term weighted by $1/\tau$ (see details in Section 5.4).

The present investigation is placed instead under a very generic form of mixture models, providing a more intuitive proof approach as well as additional consequences. Our study notably generalizes the work of [22], where the authors exploit a (non formally rigorous) statistical physics approach to evaluate the asymptotic SVM performance in a specific Gaussian mixture model with a spiked covariance matrix and constraint on the direction of the class signal. Remarkably, although it was found by [22] that the performance is maximized at trivial, overly regularized solutions, the general results in this chapter show that this conclusion is actually a consequence of the specific data model in the study of [22], which no longer holds in a broader setting.

Our main findings are summarized as follows:

- The decision direction $\boldsymbol{\beta}$ is asymptotically a multivariate normal vector whose mean $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be expressed as $[\boldsymbol{\nu}, \boldsymbol{\Sigma}] = G(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ for some function G dependent of the statistical distribution of the data vectors, the size ratio n/p and the SVM hyperparameters, which is defined in this chapter.
- To obtain more interpretable results, we find especially that under the Gaussian mixture data model with arbitrary means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\mathbf{C}_1, \mathbf{C}_2$ for the two classes, $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ have rather simple and insightful expressions controlled by five variables related to the statistical distribution of the dual coefficients c_i. From this result unfold the following remarks.
  - If $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$, or $\mathbf{C}_1 = \mathbf{I}_p + \mathbf{P}_1$ and $\mathbf{C}_1 = \mathbf{I}_p + \mathbf{P}_2$ with $\mathbf{P}_1, \mathbf{P}_2$ some low rank matrices and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ being one of the eigenvectors of both $\mathbf{P}_1$ and $\mathbf{P}_2$ (which is the case in the analysis of [22]), the SVM algorithm achieves its optimal performance at a trivial solution where $\boldsymbol{\beta}$ is just the average sum of the vectors $y_i \mathbf{x}_i$, $i \in \{1, \ldots, n\}$, obtained in the limit of $\tau \to 0$.
  - Still in the case $\mathbf{C}_1 = \mathbf{I}_p + \mathbf{P}_1$ and $\mathbf{C}_1 = \mathbf{I}_p + \mathbf{P}_2$, but with $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ being partially aligned with the eigenvectors of $\mathbf{P}_1, \mathbf{P}_2$, the hyperparameter τ allows to realize a trade-off between the bias and the variance of $\boldsymbol{\beta}$. The optimal solution is thus found at a non-trivial setting with non-negligible τ.
  - When $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ lives simultaneously in the eigenspaces of $\mathbf{C}_1$ and $\mathbf{C}_2$, the tuning of τ only affect the variance of $\boldsymbol{\beta}$. Interestingly, the variance does not necessarily changes monotonically with τ. As will be detailed in Section 5.4, depending on some condition on $\mathbf{C}_1$, $\mathbf{C}_2$ and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2$. the variance is minimized either at non-trivial τ or in the limit of $\tau \to 0$
- The statistical distribution of the dual coefficients c_i is asymptotically determined by the distribution of $\boldsymbol{\beta}$ and that of the data samples, through a simple, explicit relation given in this chapter.

5.2 Preliminaries

As in the common setting of supervised classification, we dispose of n observations $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in \mathbb{R}^p$ being the feature vectors and $y_i \in \{-1, 1\}$ their associated binary class labels, for $i = \{1, \ldots, n\}$. Support vector machines aim to construct a hyperplane defined by the subset of points $\mathbf{x} \in \mathbb{R}^p$ satisfying $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0 = 0$, which separates the set of training data by their class into two subsets with a maximal gap between them. The problem is formulated as

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\|^2 \\ &\text{s.t.} \, \forall i \in \{1, \ldots, n\}, \; y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) \geq 1. \end{aligned} \tag{5.1}$$

The n inequalities constrain the training data samples in two class-areas bounded by the two parallel hyperplanes defined by $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0 = -1$ (any $\mathbf{x}_i$ of the class labelled by -1 is on or "below" this boundary), and $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0 = 1$ (any $\mathbf{x}_i$ of the class labelled by 1 is on or "above"

this boundary). The region between these hyperplanes is conventionally referred to as the *margin* and vectors $\mathbf{x}$ falling *on* the two bordering hyperplanes as the *support vectors*.

It may occur that there exists no hyperplane that perfectly separates the training samples according to their classes, in which case the optimization problem (5.1) is not solvable. To extend SVMs to these cases, one usually resorts to a "soft-margin" alternative that allows training samples to lie inside the margin (unlike in the "hard-margin" problem (5.1)). The problem is then cast as

$$\begin{aligned} &\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{\tau}{n}\sum_{i=1}^{n}\epsilon_n \\ &\text{s.t.}\ \forall i \in \{1,\dots,n\},\ y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) \geq 1 - \epsilon_i,\ \xi_i \geq 0. \end{aligned} \tag{5.2}$$

with the hyperparameter τ controlling how much we penalize training samples falling inside the margin.

For both cases, a Lagrange multipliers approach gives the dual problem:

$$\begin{aligned} &\max_{c_1,\dots,c_n} \sum_{i=1}^{n} c_i - \frac{1}{2n}\sum_{i,j=1}^{n} c_i c_j y_i y_j \mathbf{x}_i^\mathsf{T}\mathbf{x}_j \\ &\text{s.t.}\ \forall i \in \{1,\dots,n\},\ 0 \leq c_i \leq \tau, \sum_{i=1}^{n} c_i y_i = 0 \end{aligned} \tag{5.3}$$

the solution (if it exists) for the hard-margin problem (5.1) being retrieved at $\tau = +\infty$.

With the dual solutions c_i, the hyperplane direction $\boldsymbol{\beta}$ is obtained as

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i c_i \mathbf{x}_i. \tag{5.4}$$

Additionally, by the Karush-Kuhn-Tucker conditions, we have the following relations

$$\begin{cases} c_i = 0 & \text{for}\ \ y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) > 1 \\ 0 < c_i < \tau & \text{for}\ \ y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) = 1 \\ c_i = \tau & \text{for}\ \ y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) < 1. \end{cases} \tag{5.5}$$

Evidently, $\boldsymbol{\beta}$ is determined solely by the data vectors $\mathbf{x}_i$ on the borders of margin (i.e., with $y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) = 1$) or inside the margin (i.e., with $y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) < 1$), the so-called support vectors. Note interestingly that

$$c_i = \phi_\tau\left(\frac{1 - y_i\frac{1}{n}\sum_{j\neq i} c_j y_j \mathbf{x}_j^\mathsf{T}\mathbf{x}_i - y_i\beta_0}{\|\mathbf{x}_i\|^2}\right)$$

with

$$\phi_\tau(t) = \begin{cases} 0 & \text{for}\ \ t < 0 \\ t & \text{for}\ \ 0 < t < \tau \\ \tau & \text{for}\ \ t > \tau \end{cases} \tag{5.6}$$

where τ is the penalty hyperparameter in (5.2).

After the separating hyperplane is determined by either the primal or dual optimization, the final step consists in classifying a new coming data $\mathbf{x}$ by the sign of the decision function $f(\mathbf{x}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_0$.

We consider here a general high dimensional data model presented in the following assumption.

Assumption 5.1. *The training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are independently drawn from a distribution $\mathcal{D}$ of a mixture model such that, for $k \in \{1, 2\}$, $\mathbb{P}(y_i = (-1)^k) = \rho_k$, and*

$$y_i = (-1)^k \Leftrightarrow$$
$$x_i = \boldsymbol{\mu}_k + \mathbf{C}_k^{\frac{1}{2}}\mathbf{z}_i$$

for $\boldsymbol{\mu}_k \in \mathbb{R}^p$, $\mathbf{C}_k \in \mathbb{R}^{p\times p}$ positive definite , and $\mathbf{z}_i \in \mathbb{R}^p$ some random vector of i.i.d. entries with zero mean, unit variance and bounded forth moment.

Besides, for arbitrarily large p, the ratio of training data number over dimensionality $c_0 \equiv \frac{n}{p}$ is uniformly bounded in $(0, +\infty)$, and we have the controls $\|\mathbf{C}_k\| = O(1)$, $\|\mathbf{C}_k^{-1}\| = O(1)$, $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\| = O(1)$ (with $O(\cdot)$ notation with respect to p).

For notational convenience, in the following we denote by $\mathcal{C}_k$, $k \in \{1, 2\}$ the set of $i \in \{1, \ldots, n\}$ with $y_i = (-1)^k$. It is worth pointing out that the growth-rate controls on the data means $\boldsymbol{\mu}_k$ and covariance matrices $\mathbf{C}_k$ in Assumption 5.1 are imposed for two reasons: 1) to ensure that the data vectors $\mathbf{x}_i$ are not intrinsically low dimensional by enforcing variations of similar magnitude in the entries of $\mathbf{x}_i$ through the conditions $\|\mathbf{C}_k\| = O(1)$ and $\|\mathbf{C}_k^{-1}\| = O(1)$, and 2) to enforce a "non-trivial" learning scenario by controlling the distance $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\| = O(1)$ between classes, so that the classification is neither overly difficult nor overly easy for data of large dimension p.

5.3 Statistical characterizations

As explained before, a special property of SVMs is that they are determined only by a part of the training data (the support vectors). Whether a training data $\mathbf{x}_i$ is a support vector and how much it affects the direction of theseparating hyperplane is reflected by its corresponding dual coefficient c_i determined by (5.3), since $\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n y_i c_i \mathbf{x}_i$. Also, the relative position of data vector $\mathbf{x}_i$ with respect to the separating hyperplane described by the training score

$$R_i = y_i(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i + \beta_0), \quad i \in \{1, \ldots, n\}, \tag{5.7}$$

has some relations with c_i as described in (5.5). Even though we can get easily from c_i to $\boldsymbol{\beta}$ by the equation $\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n y_i c_i \mathbf{x}_i$, not much can be said directly about the relation between their statistical behaviors since we do not have an explicit expression for c_i nor $\boldsymbol{\beta}$. Our first step is to relate the statistical distribution of c_i and R_i to that of the separating hyperplane parameters $(\boldsymbol{\beta}, \beta_0)$ in high dimensions, by using the results presented in the following proposition.

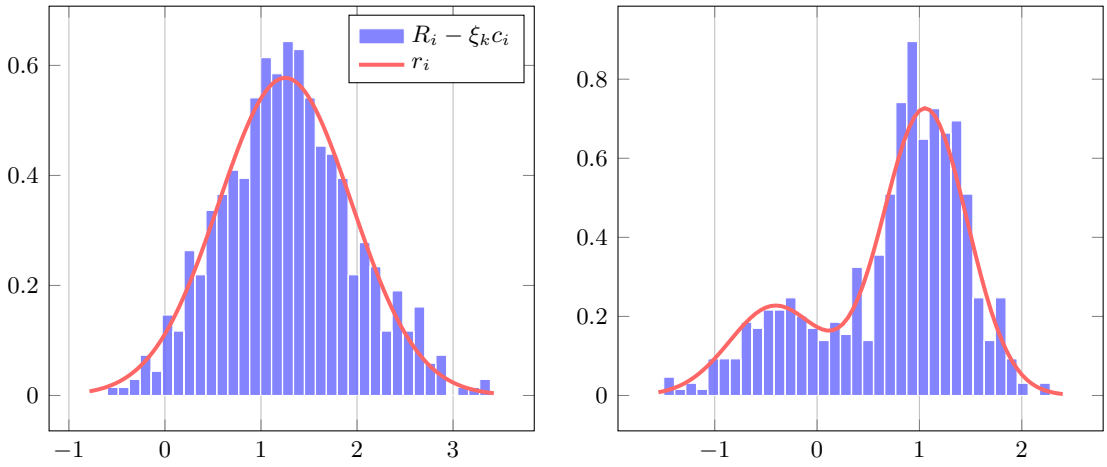

Figure 5.1: Comparison between the density histogram of $R_i - \xi_k c_i$ and the distribution of r_i as defined in Proposition 5.3.1 with $p = 200$, $n = 600$. Left: Bernouilli distributed data features with $\mathbf{C}_1 = \mathbf{C}_2 = \frac{(\sqrt{p}-2)(\sqrt{p}+2)}{4p}\mathbf{I}_p$ $\boldsymbol{\mu}_k = \frac{\sqrt{p}+(-1)^k 2}{2\sqrt{p}}[\mathbf{1}_{p/2};\ \mathbf{0}_{p/2}] + \frac{\sqrt{p}-(-1)^k 2}{2\sqrt{p}}[\mathbf{0}_{p/2};\ \mathbf{1}_{p/2}]$, $k = \{1, 2\}$, $\rho_1 = \rho_2$; $\tau = 10$. Right: normal distributed data features with $\mathbf{C}_1 = \mathbf{I}_p$, $\{\mathbf{C}_2\}_{ij} = .4^{|i-j|}$, $\boldsymbol{\mu}_k = (-1)^k[1;\ \mathbf{0}_{p-1}]$, $k = \{1, 2\}$, $\rho_1 = 3\rho_2$; $\tau = 1$.

Proposition 5.3.1. *Under Assumption 5.1, let*

$$\{\xi_1, \xi_2\} = F\left(\frac{1}{n}\sum_{i\in\mathcal{C}_1} \mathbf{1}_{(0,\tau)}(c_i), \frac{1}{n}\sum_{j\in\mathcal{C}_2} \mathbf{1}_{(0,\tau)}(c_j)\right)$$

where $F(t_1, t_2) = \{s_1, s_2\}$ is a mapping from $\mathbb{R}^* \times \mathbb{R}^*$ to $\mathbb{R}^* \times \mathbb{R}^*$ with $\{s_1, s_2\}$ the unique solution of

$$s_k = \frac{1}{n}\operatorname{tr}\mathbf{C}_k\left(\mathbf{I}_p + \sum_{a=1}^{2} s_a^{-1} t_a \mathbf{C}_a\right)^{-1},\ k \in \{1, 2\}. \tag{5.8}$$

Then, $\forall i \in \{1, \ldots, n\}$ with $i \in \mathcal{C}_k$, $k \in \{1, 2\}$,

$$(R_i - \xi_k c_i) - r_i = o_P(1), \quad \text{where} \quad r_i \overset{\mathcal{L}}{=} y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i' + \beta_0) \tag{5.9}$$

for $\mathbf{x}_1', \ldots, \mathbf{x}_n'$ independent copies of $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Note that the random variables r_i defined in Proposition 5.3.1 follow the same distribution as $y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i' + \beta_0)$, which are the prediction scores of some new coming data vectors $\mathbf{x}_i'$ independent of the training samples. Since the new data vectors $\mathbf{x}_i'$ are not involved in the training process, the r_i reflect the generalization performance of SVMs , in contrast to the training performance given by the R_i.

Importantly, the crucial difference between r_i and R_i resides in fact that the former is statistically independent of $\boldsymbol{\beta}$, while the latter is obviously not. Therefore, unlike R_i, the statistical distribution of which is inaccessible due to the implicit dependence between $\boldsymbol{\beta}$ and $\mathbf{x}_i$, the probability distribution of r_i is known as a function of that of $\boldsymbol{\beta}$ and data model.

As a numerical evidence to Proposition 5.3.1, we compare the density histogram of $R_i - \xi_k c_i$ and the distribution of r_i obtained upon one realization of the SVM algorithm. The close matches displayed in Fig 5.1 demonstrate the validity of the high dimensional results of Proposition 5.3.1 on data sets of moderate size ($p, n \sim 100$).

Remark also that it can be derived from (5.5) and (5.6) that

$$\phi_\tau \left(\frac{1 - (R_i - \xi_k c_i)}{\xi_k} \right) = c_i.$$

As a direct consequence of Proposition 5.3.1 and the above equation, we have the following corollary which establishes statistical relations between c_i and $\boldsymbol{\beta}$ in the high dimensional setting, by expressing c_i through a function of r_i.

Corollary 5.1. *Under the conditions and notations of Proposition 5.3.1, we have, $\forall i \in \{1, \ldots, n\}$ with $i \in \mathcal{C}_k$, $k \in \{1, 2\}$,*

$$c_i - \phi_\tau \left(\frac{1 - r_i}{\xi_k} \right) = o_P(1)$$

.

Proposition 5.3.2. *Let Assumption 5.1 hold. Under the notations of Proposition 5.3.1, we have*

$$\| g(\boldsymbol{\beta}) - \boldsymbol{\nu} \| = o_P(1)$$

where

$$g(\boldsymbol{\beta}) = \boldsymbol{\beta} + \sum_{a=1}^{2} \rho_a \mathbb{E}_{\mathbf{x}_i'} \left\{ \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i' - y_i \beta_0}{\xi_a} \right) y_i (\mathbf{x}_i' - \boldsymbol{\mu}_a) \middle| y_i = (-1)^a \right\} \tag{5.10}$$

$$\boldsymbol{\nu} \sim \mathcal{N} \left(\eta \boldsymbol{\mu}, \frac{\rho_1 \gamma_1 \mathbf{C}_1 + \rho_2 \gamma_2 \mathbf{C}_2}{n} \right)$$

with

$$\eta = \mathbb{E}\{c_i\}, \quad \gamma_k = \mathbb{E}\{c_i^2 | i \in \mathcal{C}_k\}, \quad k \in \{1, 2\}.$$

Proposition 5.3.2 states that in the large dimensional regime, there exists a transformation $g(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$, given by (5.10), that follows a multivariate normal distribution with statistical parameters given as functions of the first and second moments of c_i. As such, Proposition 5.3.2 provides in fact, in addition to Proposition 5.3.1, a second way to link the distribution of $\boldsymbol{\beta}$ back to that of c_i. Combining the results in Proposition 5.3.1-5.3.2 should allow to determine, from the statistical properties of the data model, the statistical distribution of the learned parameters for high dimensional SVMs. However, the expectation term in the expression (5.10) of $g(\boldsymbol{\beta})$ makes it computationally difficult to apply these results. Also, in terms of interpretability, it is hard to comment on the learning behavior and performance of SVMs from the results of Proposition 5.3.1-5.3.2.

Remarkably, we find that under the normality of data, $g(\boldsymbol{\beta})$ has a rather convenient form:

$$g(\boldsymbol{\beta}) = \left(\mathbf{I}_p + \sum_{a=1}^{2} \frac{\mathbb{E}_{\mathbf{x}_i'} \left\{ \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i' - y_i \beta_0}{\xi_a} \right) y_i (\mathbf{x}_i' - \boldsymbol{\mu}_a)^\mathsf{T} \boldsymbol{\beta} \middle| y_i = (-1)^a \right\}}{\boldsymbol{\beta}^\mathsf{T} \mathbf{C}_a \boldsymbol{\beta}} \mathbf{C}_a \right) \boldsymbol{\beta},$$

allowing us to derive, from Proposition 5.3.1 and 5.3.2, the high dimensional distribution of $\boldsymbol{\beta}$ parametrized by five variables, as presented in the following theorem.

Theorem 5.3.1. *Let Assumption 5.1 hold for multivariate normally distributed* $\mathbf{x}_i$*. Defining* $\boldsymbol{\mu} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/2$*, we have*

$$\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\| = o_P(1)$$

where

$$\left(\mathbf{I}_p + \rho_1\tilde{\theta}_1\mathbf{C}_1 + \rho_2\tilde{\theta}_2\mathbf{C}_2\right)\tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\tilde{\eta}\boldsymbol{\mu}, \frac{\rho_1\tilde{\gamma}_1\mathbf{C}_1 + \rho_2\tilde{\gamma}_2\mathbf{C}_2}{n}\right)$$

with $(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\eta}, \tilde{\gamma}_1, \tilde{\gamma}_2) \in \mathbb{R}_+^5$ *determined by the following system of five fixed-point equations*

$$\tilde{\eta} = \rho_1\mathbb{E}\{\tilde{c}_{[1]}\} + \rho_2\mathbb{E}\{\tilde{c}_{[2]}\}, \quad \tilde{\gamma}_k = \mathbb{E}\{\tilde{c}_{[k]}^2\}, \quad \tilde{\theta}_k = \frac{\mathbb{E}\{\mathbf{1}_{(0,\tau)}(\tilde{c}_{[k]})\}}{\tilde{\xi}_k}, \quad k \in \{1,2\}, \tag{5.11}$$

where

$$\tilde{c}_{[a]} \sim \phi_\tau\left(\frac{1 - y_i\tilde{\beta}_0 - y_i\tilde{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i'}{\tilde{\xi}_a}\right), \quad \textit{for} \quad i \in \mathcal{C}_a, \quad a \in \{1,2\}$$

with $\mathbf{x}_i' \sim \mathbf{x}_i$ *independent of* $\tilde{\boldsymbol{\beta}}$*, and* $\tilde{\beta}_0$*,* $\tilde{\xi}_1, \tilde{\xi}_2$ *jointly given by*

$$\begin{aligned}\{\tilde{\xi}_1, \tilde{\xi}_2\} &= F\Big(\rho_1\mathbb{E}\{\mathbf{1}_{(0,\tau)}(\tilde{c}_{[1]})\}, \rho_2\mathbb{E}\{\mathbf{1}_{(0,\tau)}(\tilde{c}_{[2]})\}\Big)\\ \rho_1\mathbb{E}\{\tilde{c}_{[1]}\} &= \rho_2\mathbb{E}\{\tilde{c}_{[2]}\}.\end{aligned}$$

Moreover, $c_i - \tilde{c}_i = o_P(1)$ *where* $\tilde{c}_i \sim \tilde{c}_{[k]}$*, for* $i \in \mathcal{C}_k$*,* $k \in \{1,2\}$*; and* $\beta_0 - \tilde{\beta}_0 = o_P(1)$*.*

The statistical equations in Theorem 5.3.1 allow one to obtain, from the data model, the sample number over dimension ratio n/p and the soft-margin hyperparameter τ, the asymptotic distribution of $\boldsymbol{\beta}$. To check its reliablity on finite dimensonal data sets, we contrast in Figure 5.2 the statistical distribution of $\boldsymbol{\beta}$ obtained on data sets of size $p = 60$ with its high dimensional equivalence $\bar{\boldsymbol{\beta}}$ as defined by Theorem 5.3.1. The statistical properties of $\boldsymbol{\beta}$ is empirically estimated from 500 independent realizations. The estimated expectation $\mathbb{E}\{\boldsymbol{\beta}\}$ of $\boldsymbol{\beta}$ is shown on the top of Figure 5.3.1 to coincide with $\mathbb{E}\{\bar{\boldsymbol{\beta}}\}$; and the displays on the bottom support the result that the random part $\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\}$ of $\boldsymbol{\beta}$ follows practically the same distribution as $\bar{\boldsymbol{\beta}} - \mathbb{E}\{\bar{\boldsymbol{\beta}}\}$ by illustrating the statistical closeness between the elements of $\mathrm{Cov}\{\bar{\boldsymbol{\beta}}\}^{-\frac{1}{2}}(\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\})$ and standard normal variables.

It is evident that, from the high dimensional distribution of the SVM parameters determined by Theorem 5.3.1, we directly obtain the probability of correct classification (i.e., the expected classification accuracy) on unseen data, given in the following corollary.

Corollary 5.2. *Under the conditions and notations of Theorem 5.3.1, we have that, for some* $(\mathbf{x}, y) \sim \mathcal{D}$ *independent of the training samples* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ *with* $y = (-1)^k$*,* $k \in \{1,2\}$*,*

$$\mathbb{P}\{y(\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0) > 0 | \boldsymbol{\beta}, \beta_0\} = Q\left(\frac{y(\tilde{\eta}\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\boldsymbol{\mu}_k + \beta_0)}{\sqrt{\tilde{\eta}^2\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\mathbf{C}_k\mathbf{G}\boldsymbol{\mu} + \frac{1}{n}\sum_{a=1}^2 \rho_a\tilde{\gamma}_a \operatorname{tr}\mathbf{C}_a\mathbf{G}\mathbf{C}_k\mathbf{G}}}\right) + o_P(1) \tag{5.12}$$

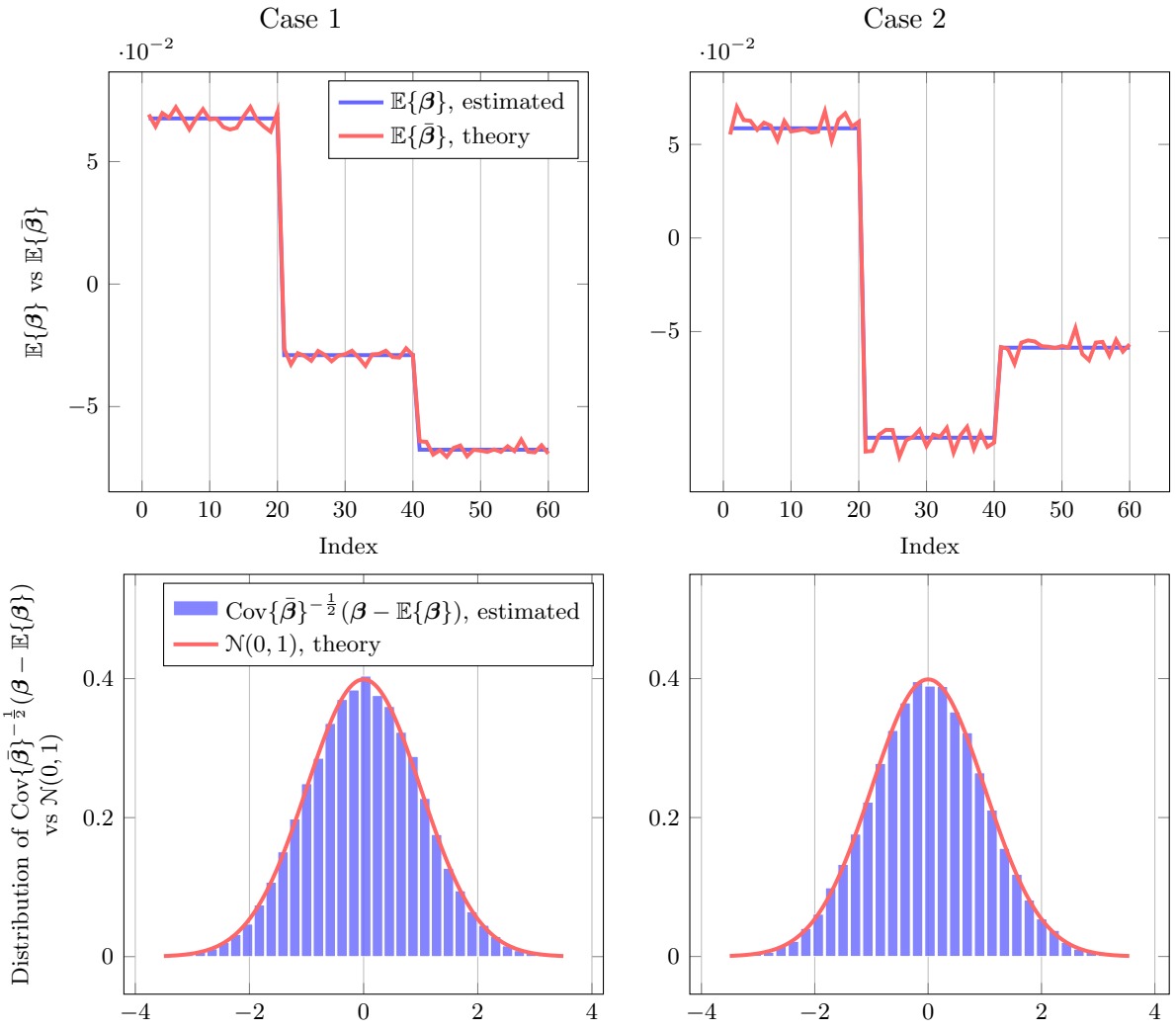

Figure 5.2: Comparison between empirical (estimated over 500 realizations) and theoretical (from Theorem 5.3.1) statistical distribution of $\boldsymbol{\beta}$ for multivariate normal distributed data with $p = 60$, $n = 120$. Case 1: $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \frac{3}{2\sqrt{p}}[\mathbf{1}_{p/3};\ -\mathbf{1}_{2p/3}]$; $\mathbf{C}_1 = \mathrm{diag}([\mathbf{1}_{p/3};\ 8\mathbf{1}_{p/3};\ 4\mathbf{1}_{p/3}])$, $\mathbf{C}_2 = \mathrm{diag}([4\mathbf{1}_{p/3};\ 8\mathbf{1}_{p/3};\ \mathbf{1}_{p/3}])$, $\rho_1 = \rho_2$, $\tau = 2$. Case 2: $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \frac{1}{\sqrt{p}}[\mathbf{1}_{p/3};\ -\mathbf{1}_{2p/3}]$, $\mathbf{C}_1 = \mathrm{diag}([\mathbf{1}_{2p/3};\ 4\mathbf{1}_{p/3}])$, $\mathbf{C}_2 = \mathrm{diag}([4\mathbf{1}_{p/3};\ \mathbf{1}_{2p/3}])$, $\rho_1 = \rho_2$, $\tau = 4$.

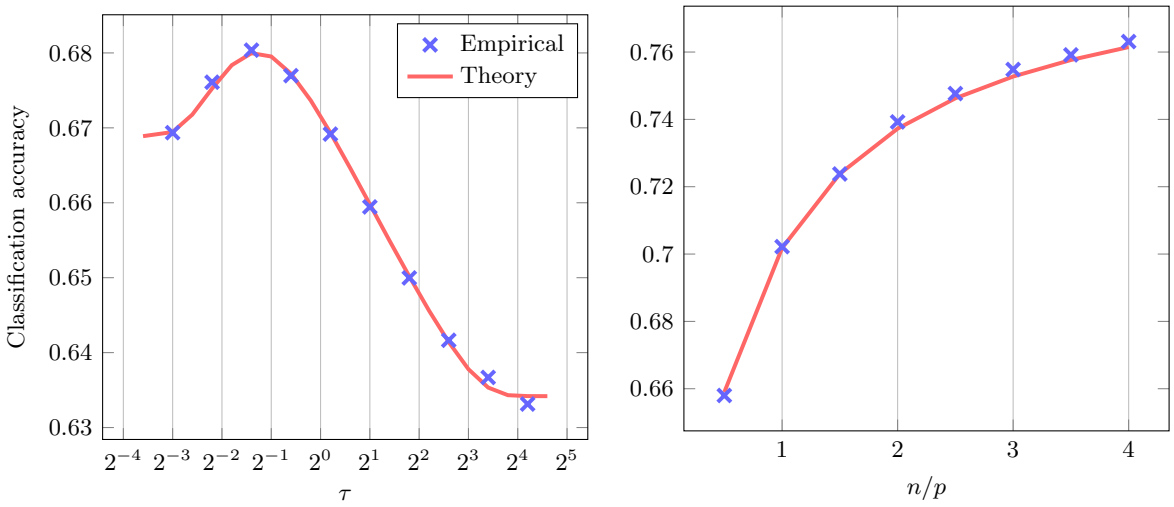

Figure 5.3: Comparison between empirical (averaged over 200 runs) and theoretical accuracy (given by Corollary 5.2) for multivariate normal distributed data of $p = 200$. Left: accuracy as a function of τ with $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \frac{1}{\sqrt{p}}\mathbf{1}_p$, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p + \frac{12}{p}[\mathbf{1}_{p/2}; \mathbf{0}_{p/2}][\mathbf{1}_{p/2}; \mathbf{0}_{p/2}]^\mathsf{T}$, $\rho_1 = \rho_2$, $n = 200$. Right: accuracy as a function of n/p with $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \frac{1}{\sqrt{p}}\mathbf{1}_p$, $\mathbf{C}_1 = \mathbf{I}_p$, $\{\mathbf{C}_2\}_{ij} = .4^{|i-j|}$, $\rho_1 = \rho_2$, $\tau = 1$.

where $\mathbf{G} = \left(\mathbf{I}_p + \rho_1\tilde{\theta}_1\mathbf{C}_1 + \rho_2\tilde{\theta}_2\mathbf{C}_2\right)^{-1}$, *and* $Q(t) \equiv \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$ *is the Q-function of the standard Gaussian distribution.*

As can be observed from (5.12), in the cases of finite n/p ratios, the classification accuracy of SVMs stabilizes around a deterministic constant at sufficiently large p, irrespective of the realization of the training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. The results in Figure 5.3 show that the asymptotic accuracy given by (5.12) predicts with great precision the average accuracy on data sets of relatively small n, p (~ 100), providing thus an adequate quantitative characterization of the learning performance which changes with the tuning of the hyperparameter τ (as shown on the left of Figure 5.3) or when more data samples are fed into the training process (as presented on the right of Figure 5.3).

5.4 Insights into the learning process: the bias-variance decomposition

We notice from Theorem 5.3.1 that under data normality, the high dimensional equivalence $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ has a quite interesting form of distribution, which can help shed light on the bias-variance trade-off controlled by the hyperparameter τ.

Before our discussion, it is worth pointing out that SVMs can be seen as a special case from the family of the empirical risk minimization (ERM) algorithms, the general idea of which is to find a mapping $h(\mathbf{x})$ of the data vector $\mathbf{x}$ that minimizes the sum of a certain loss $L(h(\mathbf{x}), y)$ between $h(\mathbf{x})$ and the desired output y over the training samples. Indeed, the SVM optimization

formulation (5.2) can be rewritten as

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{n}\sum_{i=1}^{n} L_{\text{hinge}}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i + \beta_0, y_i) + \frac{\tau^{-1}}{2}\|\boldsymbol{\beta}\|^2$$

where L_{hinge} denotes the hinge loss function $L_{\text{hinge}}(t, y) \equiv \max\{0, 1 - ty\}$. The SVM algorithm gives in fact a regularized solution of a empirical risk minimization with the hinge loss. Since adding regularization terms is commonly accepted as a means of reducing the variance at the cost of an increased bias, the hyperparameter τ, a small value of which reflects a high level of regularization, is expected to yield solutions of high bias and low variance at its large values and conversely at its small values. However, as we shall see in the following discussion, this statement is only partially correct.

Returning to our initial discussion, a preferable solution of $\boldsymbol{\beta}$ should create a better separation between classes, meaning that its projection $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}$ on some new coming data vector $\mathbf{x}$ (with underlying class label $y = \pm 1$) has a small ratio of the variance within classes to the variance between classes. We then consider a measure of noise-to-signal ratio for each class:

$$H_k = \frac{\text{Var}\{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}|y = (-1)^k\}}{\left(\mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_1 - \mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_2\right)^2}, \quad k = \{1, 2\}.$$

To conduct a bias-variance investigation of $\boldsymbol{\beta}$, we decompose $\boldsymbol{\beta}$ as its expectation $\mathbb{E}\{\boldsymbol{\beta}\}$ plus a non-informative random part $\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\} = \boldsymbol{\epsilon}$, and rewrite H_k as $H_k = B_k + V_k$ where we retrieve a bias penalty term for $\boldsymbol{\beta}$

$$B_k = \frac{\text{Var}\{\mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\mathbf{x}|y = (-1)^k\}}{\left(\mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_1 - \mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_2\right)^2}$$

depending only on the expectation of $\boldsymbol{\beta}$, and a measure of its variance

$$V_k = \frac{\text{Var}\{\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{x}|y = (-1)^k\}}{\left(\mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_1 - \mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}}\boldsymbol{\mu}_2\right)^2}.$$

which concerns the random part $\boldsymbol{\epsilon}$ of $\boldsymbol{\beta}$ and goes to zero in the limit of $n \gg p$. Under the notations of Theorem 5.3.1, we get

$$B_k = \frac{\boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\mathbf{C}_k\mathbf{G}\boldsymbol{\mu}}{(2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\boldsymbol{\mu})^2} + o_P(1), \quad V_k = \sum_{a=1}^{2}\frac{\tilde{\gamma}_a}{\tilde{\eta}^2}\frac{\text{tr}\,\mathbf{C}_a\mathbf{G}\mathbf{C}_k\mathbf{G}/n}{(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{G}\boldsymbol{\mu})^2} + o_P(1), \quad k\{1, 2\}.$$

where we recall that

$$\mathbf{G} = \left(\mathbf{I}_p + \rho_1\tilde{\theta}_1\mathbf{C}_1 + \rho_2\tilde{\theta}_2\mathbf{C}_2\right)^{-1}.$$

The bias and variance of $\boldsymbol{\beta}$ are thus controlled by $\tilde{\gamma}_1/\tilde{\eta}^2$, $\tilde{\gamma}_2/\tilde{\eta}^2$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$.

Remark first that as c_i follows essentially the same distribution as $\tilde{c}_i$ (as defined in Theorem 5.3.1) in high dimensions, these variables can be in fact estimated from the dual coefficients. Precisely, let

$$\hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad \hat{\gamma}_k = \frac{1}{n_k}\sum_{i\in\mathcal{C}_k} c_i^2, \quad \hat{\theta}_k = \frac{1}{n_k\xi_k}\sum_{i\in\mathcal{C}_k}\mathbf{1}_{(0,\tau)}(c_i), \quad k \in \{1, 2\}, \tag{5.13}$$

where $\{\xi_1, \xi_2\} = F\left(\frac{1}{n}\sum_{i\in\mathcal{C}_1}\mathbf{1}_{(0,\tau)}(c_i), \frac{1}{n}\sum_{j\in\mathcal{C}_2}\mathbf{1}_{(0,\tau)}(c_j)\right)$ with F as defined in Proposition 5.3.1, we have the following proposition establishing the above empirical quantities as consistent estimators of these variables.

Proposition 5.4.1. *Under the conditions and notations of Theorem 5.3.1,* $(\hat{\theta}_1, \hat{\theta}_2, \hat{\eta}, \hat{\gamma}_1, \hat{\gamma}_2)$ *given by* (5.13) *is a consistent estimator of* $(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\eta}, \tilde{\gamma}_1, \tilde{\gamma}_2)$.

We observe from (5.13) that $\hat{\gamma}_1/\hat{\eta}^2$ and $\hat{\gamma}_2/\hat{\eta}^2$ measure the variability of the dual coefficients c_i for each class. Note from (5.5) that when the training vectors are all inside the margin, all c_i have the same deterministic value $c_i = \tau$, in which case $\hat{\gamma}_1/\hat{\eta}^2$, $\hat{\gamma}_2/\hat{\eta}^2$ are both minimized to 1. Since smaller values of τ give rise to larger margins as can be understood from the optimization problem (5.2), $\hat{\gamma}_1/\hat{\eta}^2$, $\hat{\gamma}_2/\hat{\eta}^2$ attain their minimum at sufficiently small τ.

Let $n_{\mathcal{B}k}$, $k = \{1, 2\}$, denote the numbers of training data vectors $\mathbf{x}_i$ labelled with $y_i = (-1)^k$ which are on the border, i.e.,

$$n_{\mathcal{B}k} = \sum_{i=1}^{n}\mathbf{1}_{(0,\tau)}(c_i).$$

For $k = \{1, 2\}$, $\hat{\theta}_k$ is determined by $(n_{\mathcal{B}1}, n_{\mathcal{B}2})$, and has a value of zero if $n_{\mathcal{B}k} = 0$. Understandably, $(n_{\mathcal{B}1}, n_{\mathcal{B}2})$ are both zero in the limit of $\tau \to 0$, where the margin is so large that all the training vectors are inside it and there is no training vector that falls on the border.

Since $(\hat{\theta}_1, \hat{\theta}_2, \hat{\eta}, \hat{\gamma}_1, \hat{\gamma}_2)$ estimates consistently $(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\eta}, \tilde{\gamma}_1, \tilde{\gamma}_2)$, we obtain the following conclusions that will be useful in the subsequent discussion:

- In the limit of really small τ, $\tilde{\gamma}_1/\tilde{\eta}^2$ and $\tilde{\gamma}_2/\tilde{\eta}^2$ tend to their minimum 1.
- The variables $\tilde{\theta}_1, \tilde{\theta}_2$ are also minimized to 0 as $\tau \to 0$.

A first remark to be made about the bias terms of $\boldsymbol{\beta}$ is that they remain unchanged for high dimensional data if $\boldsymbol{\mu}$ is a common eigenvector of both $\mathbf{C}_1$ and $\mathbf{C}_2$, since

$$\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\mathbf{C}_k\mathbf{G}\boldsymbol{\mu}}{(2\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\boldsymbol{\mu})^2} = \frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_k\boldsymbol{\mu}}{4\|\boldsymbol{\mu}\|^2}, \quad k = \{1, 2\},$$

indicating a constant bias irrespective of $\tilde{\theta}_1$ and $\tilde{\theta}_2$.

When $\boldsymbol{\mu}$ does not live in the eigenspace of $\mathbf{C}_k$ for $k = \{1, 2\}$, we note importantly that

$$\min_{\mathbf{G}} \frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\mathbf{C}_k\mathbf{G}\boldsymbol{\mu}}{(2\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\boldsymbol{\mu})^2} = \frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_k^{-1}\boldsymbol{\mu}}{(2\boldsymbol{\mu}^\mathsf{T}\mathbf{C}_k^{-1}\boldsymbol{\mu})^2},$$

which is attained only at $\mathbf{G} \propto \mathbf{C}_k^{-1}$. Therefore, a larger value of $\tilde{\theta}_k$ tends to lower the bias term B_k as it pulls $\mathbf{G}$ closer towards a proportion of $\mathbf{C}_k^{-1}$. Particularly, in the homoscedastic case with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, for which we have $B_1 = B_2 = B$ and

$$\mathbf{G} = \left(\mathbf{I}_p + \tilde{\theta}\mathbf{C}\right)^{-1}, \quad \text{with} \quad \tilde{\theta} = \rho_1\tilde{\theta}_1 + \rho_2\tilde{\theta}_2.$$

The bias term B is thus a decreasing function of $\tilde{\theta}$, and it can be shown from the above results that $\tilde{\theta}$ increases with $n_b = \sum_{i=1}^{n} \mathbf{1}_{(0,\tau)}(c_i)$, which equals the total number of support vectors on the border. Moreover, $\tilde{\theta}$ tends to infinity as n_b approaches the dimension p, the bias term B hence goes to its minimum in this limit.

We move now to the discussion about the variance of $\boldsymbol{\beta}$. It is easy to see that smaller values of $\tilde{\gamma}_1/\tilde{\eta}^2$ and $\tilde{\gamma}_2/\tilde{\eta}^2$ are beneficial at all times as they always lead to reduced variances. Since $\tilde{\gamma}_1/\tilde{\eta}^2$ and $\tilde{\gamma}_2/\tilde{\eta}^2$ are minimized at sufficiently small τ, choosing a small τ has a positive effect for decreasing the variance of $\boldsymbol{\beta}$ on that account. However, it is easy to see that, unlike with $\tilde{\gamma}_1/\tilde{\eta}^2$ and $\tilde{\gamma}_2/\tilde{\eta}^2$, the variance of $\boldsymbol{\beta}$ can increse or decrease with $\tilde{\theta}_1$ and $\tilde{\theta}_2$.

An interesting scenario is when $\mathbf{C}_1$ and $\mathbf{C}_2$ are identity matrices plus some low-rank perturbations, i.e.,

$$\mathbf{C}_k = \mathbf{I} + \sum_{d=1}^{m} \lambda_{[k]m}^2 \mathbf{v}_{[k]m} \mathbf{v}_{[k]m}^{\mathsf{T}}, \quad k = \{1, 2\},$$

for some $m = O(1)$ and $\mathbf{v}_{[k]d}^{\mathsf{T}} \mathbf{v}_{[k]d'} = \delta_{dd'}$, in which case

$$\begin{aligned} &\operatorname{tr} \mathbf{C}_a \mathbf{G} \mathbf{C}_k \mathbf{G}/n = (1 + \rho_1 \tilde{\theta}_1 + \rho_2 \tilde{\theta}_2)^{-2} (p/n) + o_P(1) \\ &(\boldsymbol{\mu}^{\mathsf{T}} \mathbf{G} \boldsymbol{\mu})^2 \leq (1 + \rho_1 \tilde{\theta}_1 + \rho_2 \tilde{\theta}_2)^{-2} \|\boldsymbol{\mu}\|^2 \end{aligned}$$

where the equality of the second line is reached at $\tilde{\theta}_1 = \tilde{\theta}_2 = 0$. Since $\tilde{\theta}_1$, $\tilde{\theta}_2$ go to their minimum 0 as $\tau \to 0$, and so do $\tilde{\gamma}_1/\tilde{\eta}^2$, $\tilde{\gamma}_2/\tilde{\eta}^2$ to a minimal value of 1 , we conclude that under the spiked model for $\mathbf{C}_1$, $\mathbf{C}_2$, the variance terms V_1, V_2 are minimized in the extreme of low τ, conforming to the common belief that small τ leads to decreased variance.

Since the bias remains unchanged with respect to τ when $\boldsymbol{\mu}$ is an eigenvector of both $\mathbf{C}_1$ and $\mathbf{C}_2$, in this case, the performance only depends on the variance, which is optimized at small τ under the spiked model. However, if the condition of $\boldsymbol{\mu}$ being an eigenvector of both $\mathbf{C}_1, \mathbf{C}_2$ does not hold, the tuning of τ allows one to search for an optimal trade-off between the bias and the variance. Namely, when $\boldsymbol{\mu}$ is either aligned with or orthogonal to the eigenvectors of $\mathbf{C}_1, \mathbf{C}_2$, the classification accuracy goes to its maximum at $\tau \to 0$ where $\boldsymbol{\beta}$ is proportional to $\sum_{i=1}^{n} y_i \mathbf{x}_i$, a trivial solution. Otherwise, there exist a non-trivial value of τ at which we achieve optimal performance. These two remarks are illustrated in Figure 5.4, where we display the case of $\boldsymbol{\mu}$ being completely aligned with the eigenvector on the left side and that of $\boldsymbol{\mu}$ being only partially aligned with the eigenvector on the right side. Notice that our conclusions are consistent with the observation of [22] that the best performance is attained at small τ, as the analysis of [22] is restricted to the spiked model with $\boldsymbol{\mu}$ being either aligned with or orthogonal to the eigenvectors.

Evidently, the statement that insignificant values of τ induce minimized variance does not apply generally to arbitrary $\mathbf{C}_1$, $\mathbf{C}_2$ since the variance terms V_1, V_2 do not necessarily go to their minimum at $\tilde{\theta}_1, \tilde{\theta}_2 = 0$. To gain more insights on how V_1, V_2 change with $\theta_1, \theta_2 = 0$, consider that $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$, under which condition we have $V_1 = V_2 = V$. From the expression of the variance terms given above, we observe that in the limit of large p, V is an increasing function

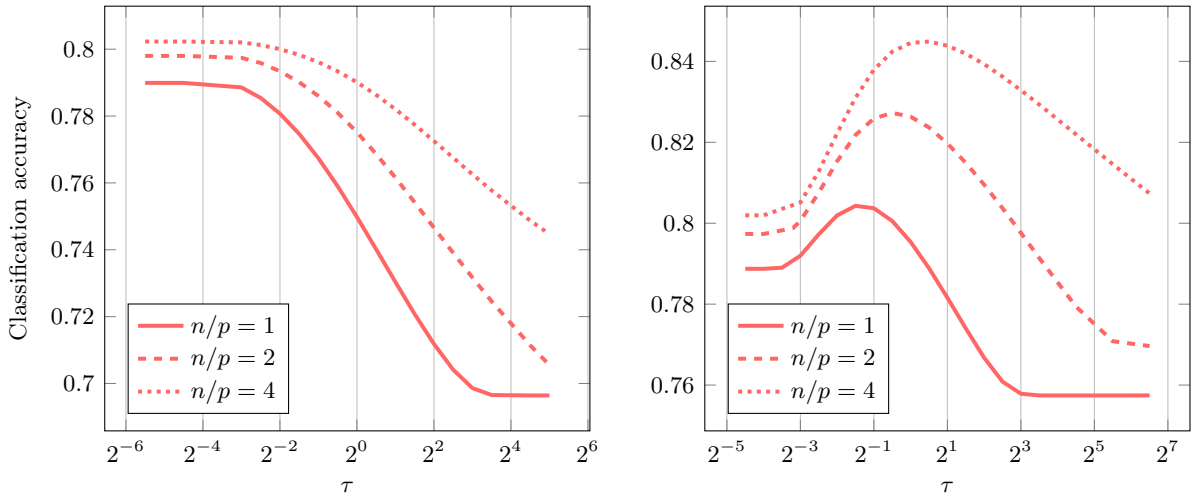

Figure 5.4: Classification accuracy (computed from Corollary 5.2) as a function of τ for multivariate normal distributed data of $p = 200$. Left: $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [3/2;\ \mathbf{0}_{p-1}]$, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p + 4[1;\ \mathbf{0}_{p-1}][1;\ \mathbf{0}_{p-1}]^\mathsf{T}$, $\rho_1 = \rho_2$. Right: accuracy as a function of n/p with $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [3/2;\ \mathbf{0}_{p-1}]$, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p + 2[1; 1;\ \mathbf{0}_{p-2}][1; 1;\ \mathbf{0}_{p-2}]^\mathsf{T}$, $\rho_1 = \rho_2$.

of $\tilde{\theta} = \rho_1\tilde{\theta}_1 + \rho_2\tilde{\theta}_2$ if

$$\frac{\operatorname{tr}\mathbf{C}\mathbf{G}\mathbf{C}\mathbf{G}/n}{(\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\boldsymbol{\mu})^2} = \frac{\frac{1}{n}\operatorname{tr}\left[\mathbf{C}\left(\mathbf{I}_p + \tilde{\theta}\mathbf{C}\right)^{-1}\right]^2}{\left[\boldsymbol{\mu}^\mathsf{T}\left(\mathbf{I}_p + \tilde{\theta}\mathbf{C}\right)^{-1}\boldsymbol{\mu}\right]^2}$$

increases with τ, or the other way around. Roughly speaking, the above term tends to decrease with respect to $\tilde{\theta}$ when $\boldsymbol{\mu}$ lives in the span of the eigenvectors of $\mathbf{C}$ associated with large eigenvalues, and conversely. As an example, we let $\mathbf{C} = \{.4^{|i-j|}\}_{i,j=1}^p$ and trace in Figure 5.5 $\frac{\operatorname{tr}\mathbf{C}\mathbf{G}\mathbf{C}\mathbf{G}/n}{(\boldsymbol{\mu}^\mathsf{T}\mathbf{G}\boldsymbol{\mu})^2}$ as a function θ for $\boldsymbol{\mu}$ being the eigenvector of $\mathbf{C}$ associated with small (shown on the left) or large (displayed on the right) eigenvalues. When the variance term V reduces as θ grows larger, it is minimized at a certain non-trivial value of τ where an optimal compromise between diminishing $\tilde{\gamma}_1/\tilde{\eta}^2, \tilde{\gamma}_2/\tilde{\eta}^2$ and increasing θ is found. Since the bias is constant when $\boldsymbol{\mu}$ is an eigenvector of $\mathbf{C}$, the classification performance is maximized when the variance is minimized. We hence observe the behavior of the variance with respect to τ from the curves of classification accuracy given in Figure 5.6.

5.5 Concluding remarks

The purpose of this chapter is to study high dimensional SVMs in the regime of finite n/p ratios. Under a general high dimensional model, we established two statistical equations linking the distribution of the dual coefficients c_i with that of the parameters $(\boldsymbol{\beta}, \beta_0)$ of the separating hyperplane. Combining these two statistical equations gives access to a full statistical characterization of SVMs. The interest of the regime with finite n/p is that the learned direction $\boldsymbol{\beta}$ of the separating hyperplane remains a random vector, the statistical distribution of which is

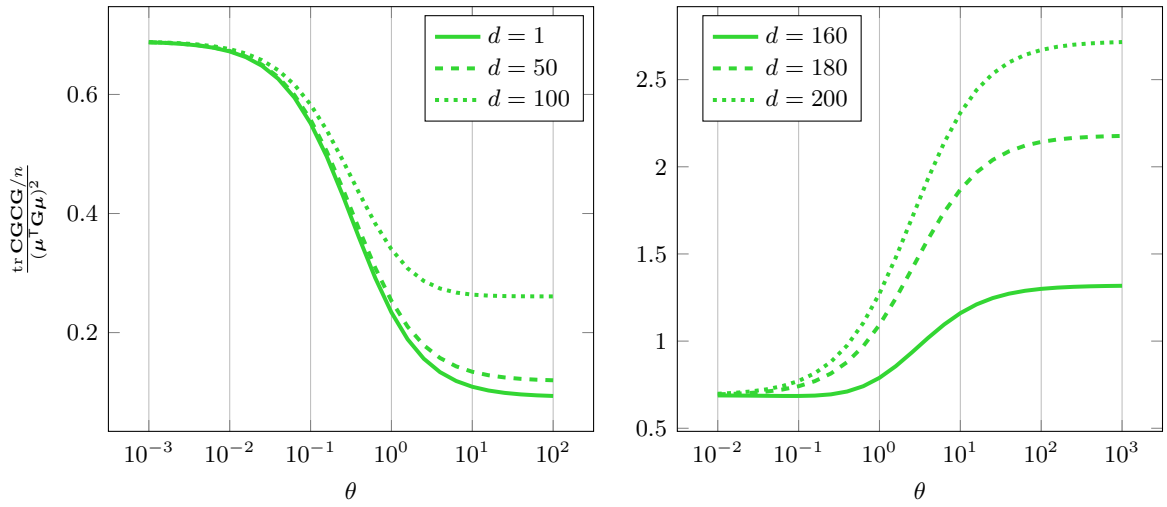

Figure 5.5: $\frac{\operatorname{tr}\mathbf{CGCG}/n}{(\boldsymbol{\mu}^\top\mathbf{G}\boldsymbol{\mu})^2}$ as a function of θ with $n = 400$, $p = 200$, $\mathbf{C} = \{.4^{|i-j|}\}_{i,j=1}^{p}$, and $\boldsymbol{\mu} = \mathbf{v}_d$ where $\mathbf{v}_d$ is the normalized eigenvector of $\mathbf{C}$ associated with its d-th smallest eigenvalue.

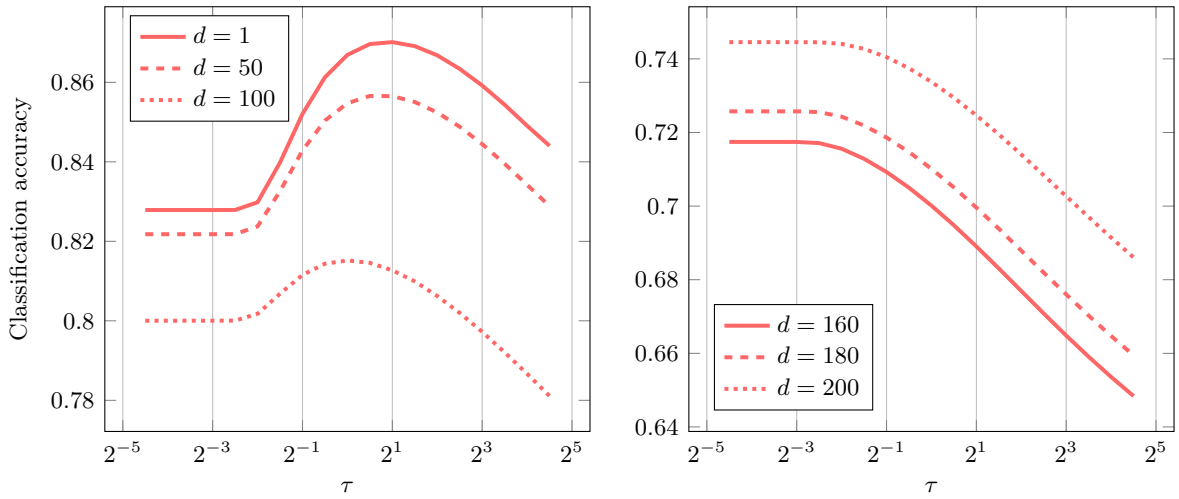

Figure 5.6: Classification accuracy (computed from Corollary 5.2) as a function of τ for multivariate normal distributed data with $p = 200$, $n = 400$, $\mathbf{C}_1 = \mathbf{C}_2 = \{.4^{|i-j|}\}_{i,j=1}^{p}$, and $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{v}_d$ where $\mathbf{v}_d$ is the normalized eigenvector of $\mathbf{C}$ associated with its d-th smallest eigenvalue.

sensitive to the addition of training samples and the tuning of the hyperparameter τ. This is in contrast to the regime $n \gg p$ where $\boldsymbol{\beta}$ converges to a deterministic value. Besides, under the normality of data, the distribution of $\boldsymbol{\beta}$ has a remarkably insightful form, linking the statistical parameters of $\boldsymbol{\beta}$ to those of the dual coefficients through simple relations.

By exploiting this quantitative analysis of SVMs, we conducted a detailed discussion on the bias-variance trade-off and discovered more precise consequences of choosing large or small τ beyond the general conclusions of the current literature. The application of our analysis to the study of the bias-variance decomposition is a telling example that demonstrates how our results can provide exhaustive details on the behavior of SVMs. Future investigations are envisioned to take advantage of the remarks drawn from this highly informative analysis for a more clever practical usage of the SVM method.

Chapter 6

A joint analytical framework for logistic regression and other empirical risk minimization algorithms

6.1 Introduction

In this chapter, we consider the following general classification problem: given a training set of n pre-labelled samples with feature vectors of dimension p, the objective is to predict the class label y (e.g., $y = \pm 1$) of a new observation $\mathbf{x}$ based on the knowledge of these training samples. The basic setup of a large number of classification algorithms is to obtain the class label y of a new instance by combining its feature vector $\mathbf{x}$ with a vector of weights $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $y = \text{sign}(\boldsymbol{\beta}^\mathsf{T}\mathbf{x})$. The weight vector $\boldsymbol{\beta}$ is usually learned by fitting the known class of training samples, for example, by minimizing the classification error (also known as the 0–1 loss) on the given training set. Despite being a natural choice, the minimization of the non-convex 0–1 loss is known to be NP-hard [65]. To address this issue, the empirical risk minimization (ERM) principle [64] suggests to obtain $\boldsymbol{\beta}$ by minimizing a certain *convex* surrogate of the 0–1 loss on the training set. Within this framework, the comparison between different designs of loss functions has been long discussed in the literature [64, 66, 67], mostly in the setting where the number of training data n largely exceeds their dimension p (i.e., p is considered small while n goes to infinity). Besides computational convenience, the usage of convex loss functions is also supported by their property of leading to the same Bayes optimal solution that minimizes the 0–1 loss in the limit of $n \gg p$ [66]. In spite of this remark, the classification accuracy can significantly depend on the choice of the loss function when n is not exceedingly larger than p. While it is crucial to know in practice which loss function to use for a given number of training samples, little is known in the regime of finite n/p.

In this chapter, we derive, in the regime of large n, p with finite n/p ratios, a unified stochastic description of the (generally implicit) optimization solution obtained from minimizing the empirical risk of any convex and smooth loss, under a high dimensional mixture model of multivariate normal feature vectors. Through this convenient and informative stochastic representation of

the learned parameters, our analysis allows notably one to discover the possibilities of improvement and comment on the optimality of this empirical risk minimization approach, as briefly summarized in the following paragraph.

To begin with, the maximal likelihood principle [68, 69, 70] states that the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ given by the negative log-likelihood loss function is a consistent estimator of the true parameter vector $\boldsymbol{\beta}_*$ underlying the conditional class probability $P(y|\mathbf{x})$, and often provides the best efficiency compared to other loss functions when $n \gg p$. However, it is empirically observed (in simulations that will be shown subsequently) that at finite n/p: 1) $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is a biased estimator of $\boldsymbol{\beta}_*$, up to a factor depending on n/p; 2) higher classification accuracy can be achieved with other losses, going against the natural use of maximal likelihood methods in high dimensions. These empirical evidences raise the questions on the possibility of bias-correcting as well as of an optimal choice of the loss function for finite n/p. From an ensemble learning perspective, it is also found that the classification accuracy can be improved by linearly combining solutions learned with different loss functions, as long as the weights assigned to the member solutions are properly chosen. It would thus be of interest to investigate on the conditions and the limitations of this improvement. Driven by these empirically motivated questions, our main findings are summarized as follows:

- Besides $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, all solutions $\hat{\boldsymbol{\beta}}$ within the present framework are aligned with $\boldsymbol{\beta}_*$ in expectation. The rescaling factor α that renders $\alpha\hat{\boldsymbol{\beta}}$ an unbiased estimator of $\boldsymbol{\beta}_*$ in high dimensions is given as an explicit function of $\hat{\boldsymbol{\beta}}$ and the training samples.

- The square loss, rather than the negative log-likelihood loss, is proved to yield the best classification accuracy for the high dimensional mixture model under study. This optimality holds universally for all n/p ratios and is irrespective of the model parameters.

- The performance gain from linearly combining different solutions can be achieved through an optimal strategy given in this chapter. Our further investigation leads however to an instructive remark that the classification accuracy attained by this ensemble learning approach is upper bounded by the accuracy that is produced solely by the solution of square loss.

In the remainder of the chapter, we introduce the objects of interest in Section 6.2. Our main technical results are presented in Section 6.3, based on which we propose the aforementioned high dimensional improvements. In Section 6.4 we discuss the optimal choice of loss function and the limitations of the ensemble method. To complete our theoretical results, we provide in Section 6.5 an asymptotic deterministic description of the system performance. The chapter closes with concluding remarks and envisioned extensions in Section 6.6.

6.2 Preliminaries

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

model:

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

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

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

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

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

$$\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}. \tag{6.1}$$

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

To ensure a non-trivial misclassification rate in the high dimensional setting (i.e., the misclassification probability is neither 0 nor 1 for large p), we shall (as in [71]) work under the following assumptions.

Assumption 6.1 (Growth rate)**.** *The sample ratio n/p is uniformly bounded in $(1, +\infty)$ for arbitrarily large p. Also, $\|\boldsymbol{\mu}\| = O(1)$, $\|\mathbf{C}\| = O(1)$ and $\|\mathbf{C}^{-1}\| = O(1)$ with respect to p.*

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

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}) \tag{6.2}$$

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

Assumption 6.2 (Loss function)**.** *The function ρ is convex and at least twice differentiable with $\rho''(t)$ bounded away from zero except at $t = \infty$.*

In particular, with the logistic loss $\rho(t) = \ln(1 + e^{-t})$ that gives the maximum likelihood estimate of $\boldsymbol{\beta}_*$, we obtain the logistic regression classifier. The least squares classifier is given by the square loss $\rho(t) = (t - 1)^2$. Another popular choice is the exponential loss $\rho(t) = e^{-t}$, widely used in boosting algorithms [72, 73].

It is worth noting that, in the high dimensional setting of Assumption 6.1, the existence of the unique solution to (6.2) is not guaranteed for all n, p. A simple example is the case $n < p$ (as excluded from Assumption 6.1), for which one can show that (6.2) has multiple solutions. Furthermore, it was shown in [74] that, in the case of logistic regression, $\|\hat{\boldsymbol{\beta}}\|$ is finite if and only if some dimensionality condition is met. The discussion on the existence condition is out of the scope of this work and we assume here that the learned classifier is "well-behaved" in the sense that the optimization problem (6.2) is well defined with a unique solution $\hat{\boldsymbol{\beta}}$ of finite norm and bounded prediction scores $\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}} = O(1)$ for training data vectors $\mathbf{x}_i$.

6.3 Main results and improvements

Before introducing the main theoretical results, we define some random elements that will appear in the theorem. By cancelling the derivative of the convex loss function ρ, we obtain from (6.2) that $\mathbf{Xc} = \mathbf{0}$ with

$$\mathbf{c} = [c_1, \ldots, c_n]^\mathsf{T} \equiv [y_1 \psi(y_1 \mathbf{x}_1^\mathsf{T} \hat{\boldsymbol{\beta}}), \ldots, y_n \psi(y_n \mathbf{x}_n^\mathsf{T} \hat{\boldsymbol{\beta}})]^\mathsf{T}, \tag{6.3}$$

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

$$\mathbf{r} = [r_1, \ldots, r_n]^\mathsf{T} \equiv [\mathbf{x}_1^\mathsf{T} \hat{\boldsymbol{\beta}} - \kappa c_1, \ldots, \mathbf{x}_n^\mathsf{T} \hat{\boldsymbol{\beta}} - \kappa c_n]^\mathsf{T}, \tag{6.4}$$

with

$$\kappa = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i^\mathsf{T} \mathbf{Q} \mathbf{x}_i / n}{1 + \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_i^\mathsf{T} \mathbf{Q} \mathbf{x}_i / n}, \tag{6.5}$$

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

$$\mathbf{r}_c = -\left(\mathbf{I}_n - \frac{1}{n} \mathbf{y} \mathbf{y}^\mathsf{T} \right) \mathbf{r}. \tag{6.6}$$

With the above notations, we are now in position to introduce the main technical result of this chapter, which concerns a stochastic description of the classifier $\hat{\boldsymbol{\beta}}$ defined in (6.2), in the following theorem.

Theorem 6.3.1. *Let Assumptions 6.1 and 6.2 hold. Then,*

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

for $\boldsymbol{\beta}_*$, $\mathbf{c}$, $\mathbf{r}_c$ defined respectively in (6.1), (6.3) *and* (6.6), *$\mathbf{u} \in \mathbb{R}^p$ a random vector uniformly distributed on the unit sphere and*

$$\alpha = \frac{2n \mathbf{c}^\mathsf{T} \mathbf{r}_c}{\mathbf{c}^\mathsf{T} \mathbf{y} \|\mathbf{r}_c\|^2}. \tag{6.7}$$

Theorem 6.3.1 gives a high dimensional equivalence $\tilde{\boldsymbol{\beta}}$ for the optimization solution $\hat{\boldsymbol{\beta}}$, so that the high dimensional performance of $\hat{\boldsymbol{\beta}}$ can be studied via $\tilde{\boldsymbol{\beta}}$. Indeed, consider the probability of misclassification

$$P(y \mathbf{x}^\mathsf{T} \boldsymbol{\beta} < 0 | \boldsymbol{\beta}) \equiv \mathcal{M}_{\mathrm{C}}(\boldsymbol{\beta}) \tag{6.8}$$

for some $(\mathbf{x}, y) \sim \mathcal{D}$ independent of $\boldsymbol{\beta}$; we deduce from Theorem 6.3.1 that

$$\mathcal{M}_{\mathrm{C}}(\hat{\boldsymbol{\beta}}) = Q\left(\frac{\boldsymbol{\mu}^\mathsf{T} \mathbf{C}^{-1} \boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^\mathsf{T} \mathbf{C}^{-1} \boldsymbol{\mu} + \frac{p\|\mathbf{c}\|^2}{(\mathbf{c}^\mathsf{T} \mathbf{y})^2}}} \right) + o_P(1), \tag{6.9}$$

where $Q(t) \equiv \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du$ denotes the Q-function of the standard Gaussian distribution. As shown in Figure 6.1, the approximation of classification performance $\mathcal{M}_{\mathrm{C}}(\hat{\boldsymbol{\beta}})$ given by (6.9) is of high precision for moderately large n, p. Note also from Theorem 6.3.1 that$\tilde{\boldsymbol{\beta}}$ is proportional to the true parameter $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$ in expectation, with an random "noise" term $\frac{2\sqrt{p}\|\mathbf{c}\|}{\mathbf{c}^\mathsf{T} \mathbf{y}} \mathbf{C}^{-\frac{1}{2}} \mathbf{u}$ that is of no use to the classification.[1] Clearly, one shall maximize the signal-to-noise ratio of $\tilde{\boldsymbol{\beta}}$ by minimizing $\frac{\|\mathbf{c}\|}{|\mathbf{c}^\mathsf{T} \mathbf{y}|}$. This conclusion can also be easily reached from (6.9).

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

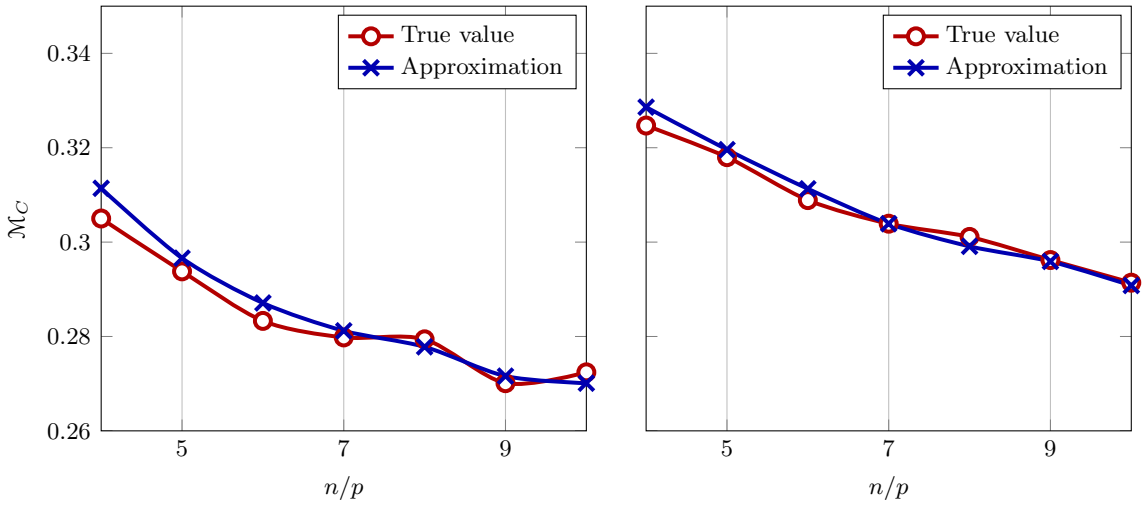

Figure 6.1: Comparison between the expected classification error $\mathcal{M}_{\mathrm{C}}$ and its approximation in (6.9) for $p = 256$, with $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$, $\rho(t) = \ln(1 + e^{-t})$ (left) and $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{2p}$, $\mathbf{C}_{ij} = 0.1^{|i-j|}$, $\rho(t) = (t-1)^2/2$ (right).

Even though the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ obtained from $\rho(t) = \ln(1+e^{-t})$ estimates exactly $\boldsymbol{\beta}_*$ in the limit of $n \gg p$, it is biased in the high dimensional setting with finite n/p. Indeed, by estimating the expectation of $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ with the empirical mean $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ obtained over 500 independent realizations, we observe in Figure 6.2 that $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ is proportional, and clearly not equal, to the true parameter vector $\boldsymbol{\beta}_*$. According to Theorem 6.3.1, we can render the high dimensional solution unbiased by multiplying α given in (6.7). As corroborating evidence, the estimated expectation $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^{\mathrm{avg}}$ of $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is observed in Figure 6.2 to coincide with $\boldsymbol{\beta}_*$.

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

Corollary 6.1. *With the assumptions and notations of Theorem 6.3.1, we have*

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

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

$$\mathcal{M}_{\mathrm{E}}(\boldsymbol{\beta}) = \mathbb{E}[s(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}) - s(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_*)|\boldsymbol{\beta}]^2 \tag{6.10}$$

where $(\mathbf{x}, y) \sim \mathcal{D}$ is independent of $\boldsymbol{\beta}$, and $s(t) = \frac{1}{1+e^{-t}}$. We demonstrate in Figure 6.3 the utility of the proposed rescaling strategy with the significant performance gains measured by $\mathcal{M}_{\mathrm{E}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) - \mathcal{M}_{\mathrm{E}}(\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$, which are especially large at small n/p ratios. Moreover, note that both $\mathbf{c}$, $\mathbf{r}_c$ (and thus α) are fast computed once $\hat{\boldsymbol{\beta}}$ is obtained from solving (6.2). Therefore, the proposed rescaling scheme is computationally efficient in the sense that it induces little extra cost to the training of the original classifier $\hat{\boldsymbol{\beta}}$.

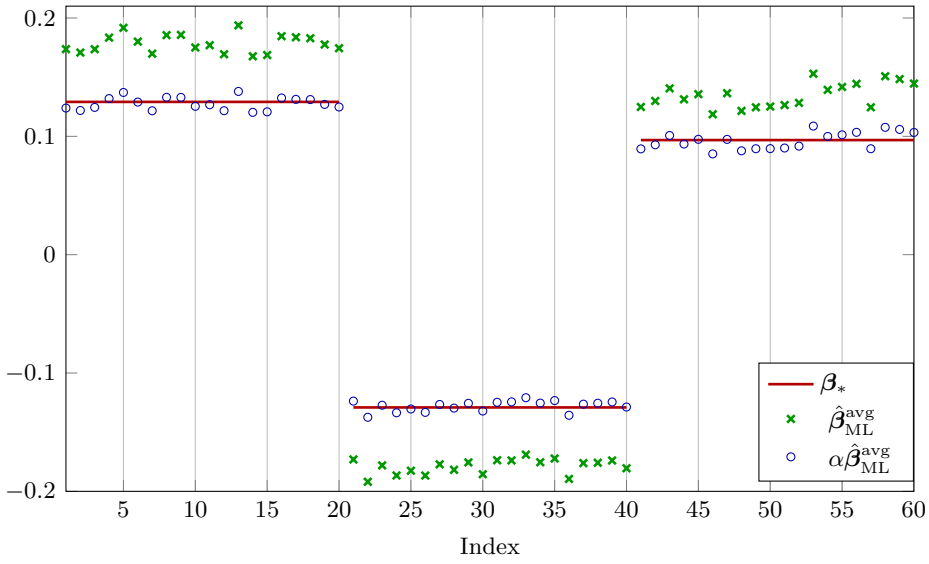

Figure 6.2: Comparison of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ (averaged over 500 realizations), the true parameter $\boldsymbol{\beta}_*$ and the rescaled classifier $\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ defined in Theorem 6.3.1 with $\boldsymbol{\mu} = [\mathbf{1}_{p/3}, -\mathbf{1}_{p/3}, \frac{3}{4}\mathbf{1}_{p/3}]/\sqrt{p}$, $\mathbf{C} = 2\mathbf{I}_p$, for $p = 60$ and $n = 300$.

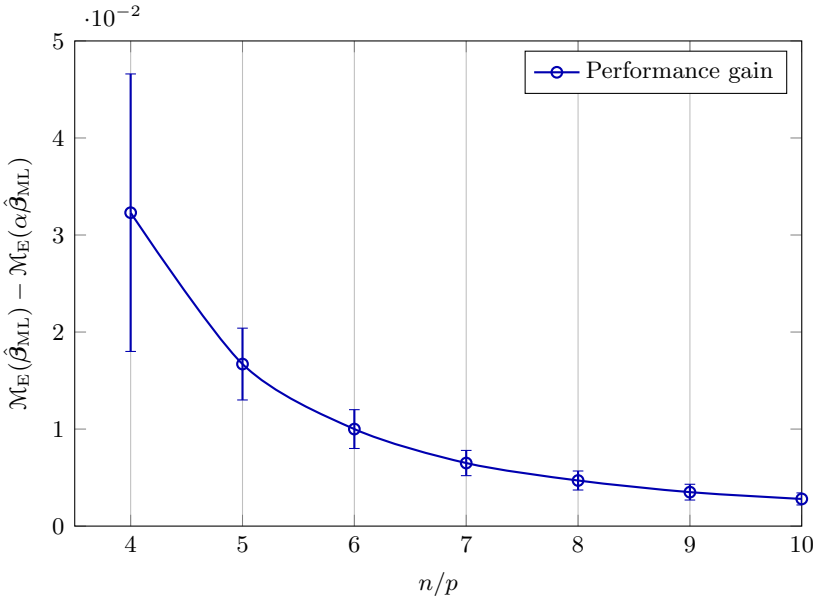

Figure 6.3: Performance gain $\mathcal{M}_{\mathrm{E}}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) - \mathcal{M}_{\mathrm{E}}(\alpha\hat{\boldsymbol{\beta}}_{\mathrm{ML}})$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

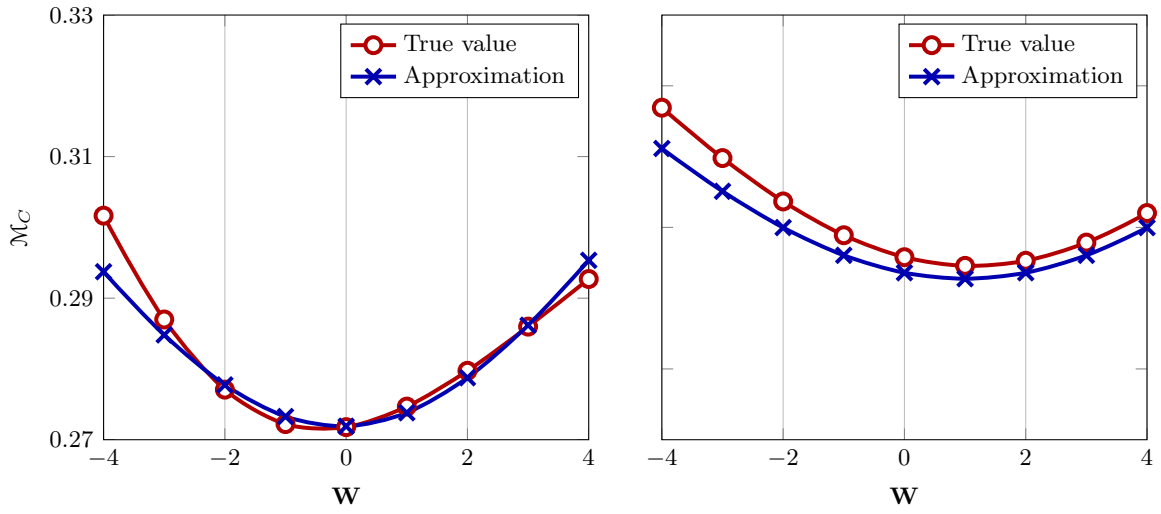

Figure 6.4: Comparison between the expected classification errors of the ensemble classifier $\mathcal{M}_{\mathrm{C}}(\tilde{\boldsymbol{\beta}}_{\mathrm{ES}})$ and its approximation $\mathcal{M}_{\mathrm{C}}(\tilde{\boldsymbol{\beta}}_{\mathrm{ES}})$ as a function of $\mathbf{W}$. For $\rho_1(t) = \ln(1+e^{-t})$, $\rho_2(t) = e^{-t}$, $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$ (left) and $\rho_1(t) = (t-1)^2/2$, $\rho_2(t) = \ln(1+e^{-t})$, $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{2p}$, $\mathbf{C}_{ij} = 0.1^{|i-j|}$ (right), $p = 256$, $n = 10p$.

Corollary 6.1 indicates that all rescaled classifiers $\alpha\hat{\boldsymbol{\beta}}$ are equally efficient in expectation. It is then pertinent to ask whether it is possible to reduce the variance. One of the basic strategies in this respect is to linearly combine several (rescaled) classifiers $\alpha_k\hat{\boldsymbol{\beta}}_k$ learned with different loss functions, to form an ensemble classifier [75]. In the following theorem we give a stochastic characterization of such ensemble classifier.

Theorem 6.3.2. *Let Assumptions 6.1 and 6.2 hold, and $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m$ stand respectively for classifiers learned with loss functions $\rho_1, \ldots, \rho_m$, m being some positive integer. For any set of m real-valued coefficients $\{w_1, \ldots, w_m\}$ such that $\sum_{k=1}^m w_k = 1$, define the ensemble classifier*

$$\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = \sum_{k=1}^{m} w_k \alpha_k \hat{\boldsymbol{\beta}}_k \tag{6.11}$$

with α_k the rescaling factor of $\hat{\boldsymbol{\beta}}_k$ given in (6.7)*. Then,*

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

with

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

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

$$\mathbf{c}_{\mathrm{ES}} = \sum_{k=1}^{m} \frac{w_k \mathbf{c}_k}{\mathbf{c}_k^{\mathsf{T}}\mathbf{y}} \tag{6.12}$$

for $\mathbf{c}_k$ defined in (6.3) *with respect to the loss function ρ_k and the training set $(\mathbf{X}, \mathbf{y})$.*

In Figure 6.4 we consider the ensemble classifier $\hat{\boldsymbol{\beta}}_{\mathrm{ES}} = w\alpha_1\hat{\boldsymbol{\beta}}_{\rho_1} + (1-w)\alpha_2\hat{\boldsymbol{\beta}}_{\rho_2}$ from different loss functions ρ_1, ρ_2, and compare its classification performance with its high dimensional equivalent $\tilde{\boldsymbol{\beta}}_{\mathrm{ES}}$ given in Theorem 6.3.2 as a function of the weight w. A close match is observed in both settings with different combinations of loss functions, suggesting that the optimal weights can be estimated with great precision from the vector $\mathbf{c}$ of its member classifiers. Indeed, Theorem 6.3.2 entails that the optimal weights w_k yielding the best performance can be obtained by minimizing $\|\mathbf{c}_{\mathrm{ES}}\|$. This remark is formally stated in the following corollary, where we also provide a necessary and sufficient condition under which $\hat{\boldsymbol{\beta}}_{\mathrm{ES}}$ is guaranteed to surpass all its (rescaled) member classifiers $\alpha_k\hat{\boldsymbol{\beta}}_k$ in terms of both classification accuracy and class probability estimation.

Corollary 6.2. *With the assumptions and notations in Theorem 6.3.2, the optimal ensemble classifier is given by*

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

with

$$\{w_1^{\mathrm{opt}}, \ldots, w_m^{\mathrm{opt}}\} = \mathrm{argmin}_{\{w_1,\ldots,w_m\}} \|\mathbf{c}_{\mathrm{ES}}\|, \tag{6.13}$$

then, with high probability,

$$\mathcal{M}(\hat{\boldsymbol{\beta}}_{\mathrm{ES}}^{\mathrm{opt}}) \geq \max_{k\in\{1,\ldots,m\}} \mathcal{M}(\alpha_k\hat{\boldsymbol{\beta}}_k) \tag{6.14}$$

for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$ as defined in (6.8) *and* (6.10)*, respectively. Furthermore, the inequality in* (6.14) *is strict if and only if, there exists $k \in \{1, \ldots, m\}$ such that,*

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

The simulations in Figure 6.5 confirm the benefits of this ensemble approach. Compared to its member classifiers, the ensemble classifier with the optimal weights w_k^{opt} given by (6.13) produces a similar effect as adding $p/5$ training samples in reducing classification error.

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

6.4 Optimality of the empirical risk minimization appraoch

It has been shown in Corollary 6.1 that, regardless of the choice of loss function ρ, the true parameter vector $\boldsymbol{\beta}_*$ is reached by the rescaled classifier $\alpha\hat{\boldsymbol{\beta}}$, for α given by (6.7), in the limit of $n \gg p$. Yet, it is still unclear which choice of ρ is optimal (if any) at finite n/p, which is a far more interesting question to provide guidance in practice.

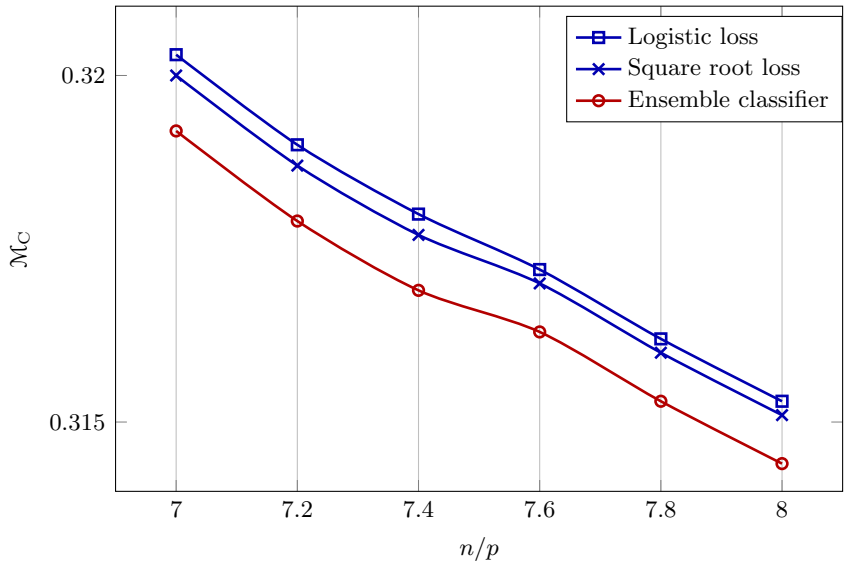

Figure 6.5: Comparison of classification error rate between the logistic loss $\rho(t) = \ln(1+e^{-t})$, the square root loss $\rho(t) = \sqrt{(t-1)^2+1}$ and the associated ensemble classifier given in Corollary 6.2 for $\boldsymbol{\mu} = [0.6, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 250$.

A default option, which is commonly believed to yield optimal learning results, would be to apply the maximal likelihood solution $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, obtained here with the logistic loss $\rho(t) = \ln(1+e^{-t})$. However, as can be observed in Figure 6.6 where the classification performance of the maximal likelihood solution (in blue) is provided along with the results produced by the square loss $\rho(t) = (t-1)^2/2$ (in red), the maximum likelihood classifier is *consistently* surpassed by the least squares classifier, for n/p ranging from 4 to 10. In light of this empirical evidence which contradicts the maximal likelihood principle for not too large n/p, one may ask whether this observed superiority of square loss over logistic loss holds for all n/p ratios, or more generally, whether there exists a loss function providing the best high dimensional classification results for any given size of training samples.

To answer these questions, note first that since $\alpha\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to $\alpha\tilde{\boldsymbol{\beta}}$ in high dimensions, it is straightforward to see (from the remarks following Theorem 6.3.1) that, with high probability,

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

where $\mathcal{M}$ can be either the classification error function $\mathcal{M}_{\mathrm{C}}$ given by (6.8) or the estimation error function $\mathcal{M}_{\mathrm{E}}$ in (6.10).

To put it differently, the search for the optimal loss function ρ can be reduced to the minimization of $\frac{\|\mathbf{c}\|}{|\mathbf{c}^\mathsf{T}\mathbf{y}|}$ with respect to ρ. Now notice that we always have $\mathbf{X}\mathbf{c} = \mathbf{0}$ from (6.2) and $\mathbf{X} \in \mathbb{R}^{p\times n}$ is of rank p for $n > p$ with probability one. Then consider the singular value decomposition

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

where $\mathbf{u} \in \mathbb{R}^{p\times p}$, $\mathbf{v} \in \mathbb{R}^{n\times n}$ are some unitary matrices such that $\boldsymbol{\Sigma} = \begin{bmatrix}\mathbf{S} & \mathbf{0}\end{bmatrix}$ with $\mathbf{S} \in \mathbb{R}^{p\times p}$

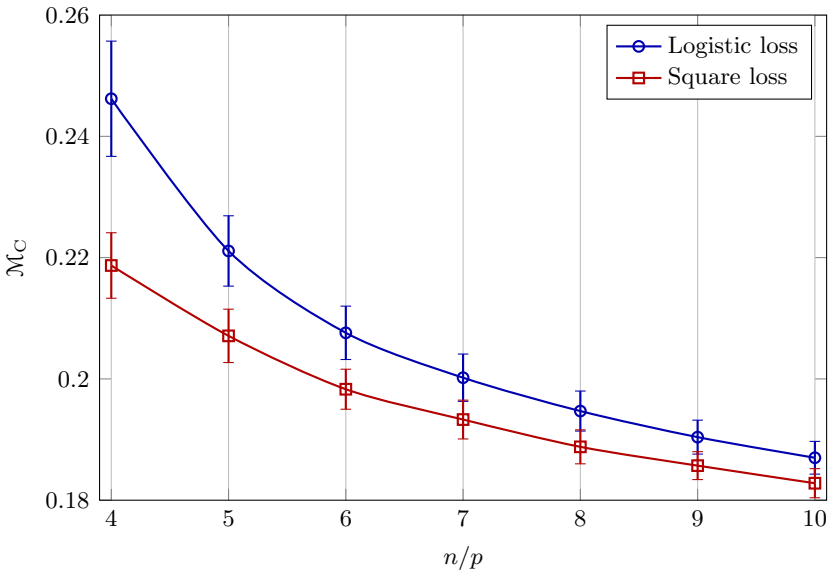

Figure 6.6: Comparison of the expected classification error rate between the logistic loss $\rho(t) = \ln(1 + e^{-t})$ and the square loss $\rho(t) = (t-1)^2/2$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

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

$$\mathbf{c}_\rho = \mathbf{v}_2\boldsymbol{\eta}_\rho. \tag{6.15}$$

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

$$\mathbf{c}_{\text{opt}} = a\mathbf{v}_2\mathbf{v}_2^\mathsf{T}\mathbf{y}, \tag{6.16}$$

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

As a matter of fact, with the square loss function[2] $\rho(t) = (t-1)^2/2$, the optimization problem in (6.2) is of explicit solution

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

which is the least squares classifier. We obtain from (6.3) that

$$\mathbf{c}_{\text{LS}} = \mathbf{y} - \mathbf{X}^\mathsf{T}\hat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{v}_2\mathbf{v}_2^\mathsf{T}\mathbf{y}, \tag{6.17}$$

meeting the optimality condition given in (6.16). This remark, combined with the above arguments, leads to the following proposition on the optimal choice of the loss function.

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

Proposition 6.4.1. *Let Assumptions 6.1 and 6.2 hold. Denote by* $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ *the solution of* (6.2) *with the square loss function* $\rho(t) = (t-1)^2/2$, $\boldsymbol{\beta}_{\rho'}$ *the solution with some loss function* ρ', *and* $\alpha_{\mathrm{LS}}, \alpha_{\rho'}$ *respectively the rescaling factor of* $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}, \hat{\boldsymbol{\beta}}_{\rho'}$ *given in* (6.7). *Then, for any given* $\boldsymbol{\mu}, \mathbf{C}$ *and* n/p *ratio, we have that*

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

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

As we recall, the true parameter vector $\boldsymbol{\beta}_*$ is given by $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$, which can also be consistently estimated by the linear discriminant analysis (LDA) classifier discussed in the introduction of Chapter 1

$$\hat{\boldsymbol{\beta}}_{\mathrm{LDA}} = 2\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}} \tag{6.18}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\mathbf{X}\mathbf{y}, \hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}$ are respectively consistent estimators of the true mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$. Actually, since

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

with $\hat{\boldsymbol{\mu}}^{\mathsf{T}}(\mathbf{X}\mathbf{X}^{\mathsf{T}}/n)^{-1}\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}^{\mathsf{T}}\hat{\mathbf{C}}^{-1}\hat{\boldsymbol{\mu}}}{1+\hat{\boldsymbol{\mu}}^{\mathsf{T}}\mathbf{C}^{-1}\hat{\boldsymbol{\mu}}} < 1$ by Sherman-Morrison formula, we observe that $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ is in fact proportional to $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. As such, $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ leads to the same classification results as $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. However, when it comes to the prediction of class probability $P(y = \pm 1|\mathbf{x})$, the estimation error can be significantly reduced by using $\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ instead of $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$, thanks to the bias-correcting effect (as stated in Corollary 6.1) of the rescaling factor α_{LS} for finite n/p. This remark is confirmed in Figure 6.7, where the performance gain of $\alpha_{\mathrm{LS}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ over $\hat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ in class probability estimation is reported.

After answering the question of optimality for individual classifiers, we move on to discuss the learning efficiency of the ensemble learning classifiers $\hat{\boldsymbol{\beta}}_{\mathrm{ES}}$ described in Theorem 6.3.2. As shown in Figure 6.5, the ensemble classifier yields superior results when compared to all of its member classifiers. However, since the learning process is always performed on the same training set, there exists certainly a limit for the performance gain achieved by this approach. To inquire into this limit, we develop the arguments below.

Similarly to the performance discussion on (rescaled) individual classifiers, it can be derived from Theorem 6.3.2 that

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

with high probability, for $\mathcal{M} = \mathcal{M}_{\mathrm{C}}$ or $\mathcal{M}_{\mathrm{E}}$. According to (6.12) and (6.15), we have

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

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

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

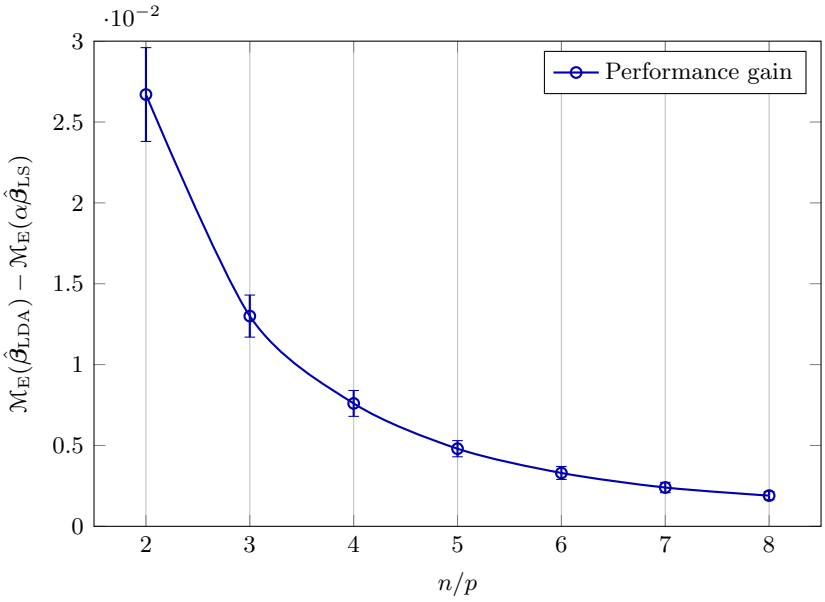

Figure 6.7: Performance gain $\mathcal{M}_{\rm E}(\hat{\boldsymbol{\beta}}_{\rm LDA}) - \mathcal{M}_{\rm E}(\alpha\hat{\boldsymbol{\beta}}_{\rm LS})$ with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

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

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

Moreover, since $\frac{1}{|\mathbf{v}_2^\mathsf{T}\mathbf{y}|} = \frac{\|\mathbf{c}_{\rm LS}\|}{|\mathbf{c}_{\rm LS}^\mathsf{T}\mathbf{y}|}$ with $\mathbf{c}_{\rm LS}$ given in (6.17), we deduce that the norm of $\mathbf{c}_{\rm ES}$ reaches its minimum at $\hat{\boldsymbol{\beta}}_{\rm ES} = \alpha_{\rm LS}\hat{\boldsymbol{\beta}}_{\rm LS}$, and thus conclude on the performance limit of ensemble learning classifier as follows.

Proposition 6.4.2. *Let Assumptions 6.1 and 6.2 hold, and $\hat{\boldsymbol{\beta}}_{\rm ES}$ be any ensemble learning classifier of the form* (6.11)*. Denote by $\hat{\boldsymbol{\beta}}_{\rm LS}$ the solution of* (6.2) *with the square loss function $\rho(t) = (t-1)^2/2$, and $\alpha_{\rm LS}$ its rescaling factor as defined in* (6.7)*. Then, for any given $\boldsymbol{\mu}, \mathbf{C}$ and n/p ratio, we have*

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

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

6.5 Asymptotic deterministic description of the learning performance

As discussed in Section 6.3, the high dimensional classification performance of $\hat{\boldsymbol{\beta}}$ can be computed with the associated random vector $\mathbf{c}$ via (6.9). The distribution of $\mathbf{c}$ thus provides a direct access to the classification performance of $\hat{\boldsymbol{\beta}}$ for any loss function. However, as $\mathbf{c}$ is a

function of the predicted scores $\mathbf{x}_1^\mathsf{T}\hat{\boldsymbol{\beta}}, \ldots, \mathbf{x}_n^\mathsf{T}\hat{\boldsymbol{\beta}}$ on all training samples, its statistical behavior is difficult to capture since $\hat{\boldsymbol{\beta}}$ depends on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a (generally) implicit manner through the optimization problem in (6.2). Nonetheless, by considering the regime of large n, p, one can link the distribution of $\mathbf{c}$ (and that of the random vector $\mathbf{r}$ defined in (6.4)) to the predicted score of new data as specified in the following theorem.

Theorem 6.5.1. *Let Assumptions 6.1 and 6.2 hold, then there exist two positive constants m, σ such that $y\mathbf{x}^\mathsf{T}\hat{\boldsymbol{\beta}} \xrightarrow{\mathcal{D}} \mathcal{N}(m, \sigma^2)$ for some $(\mathbf{x}, y) \sim \mathcal{D}$ independent of $\hat{\boldsymbol{\beta}}$. For random vectors $\mathbf{c}, \mathbf{r}$ defined in (6.3) and (6.4), we have that for all $i \in \{1, \ldots, n\}$,*

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

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

$$g_{\bar{\kappa}}(t) \equiv \psi\left(\text{prox}_{\bar{\kappa}}(t)\right) \tag{6.19}$$

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

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

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

$$m = \frac{\mathbb{E}[g_{\bar{\kappa}}(r)]\sigma^2}{m\mathbb{E}[g_{\bar{\kappa}}(r)] - \mathbb{E}[r g_{\bar{\kappa}}(r)]}\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}, \tag{6.20}$$

$$\sigma = \frac{m}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}}} + \sqrt{\frac{p}{n}}\frac{\sigma^2\sqrt{\mathbb{E}[g_{\bar{\kappa}}(r)^2]}}{m\mathbb{E}[g_{\bar{\kappa}}(r)] - \mathbb{E}[r g_{\bar{\kappa}}(r)]}. \tag{6.21}$$

Since m, σ introduced in Theorem 6.5.1 are given as the solutions of the two deterministic equations (6.20) and (6.21), we can obtain the high dimensional classification performance directly from the parameters of the data model and the n/p ratio without the actual training of the classifier.

Corollary 6.3. *Under the conditions and notations of Theorem 6.5.1, the expected classification error rate is given by*

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

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

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

In Figure 6.8 we compare the empirical distribution of $y_i r_i$ and $y_i c_i$ with the theoretical predictions in Theorem 6.5.1. A close match is observed for $p = 256$ and $n = 6p$ which confirms our theoretical results. In Figure 6.9 we plot, as a numerical validation of Corollary 6.3, the classification error rate $\mathcal{M}_{\text{C}}$ and the associated values of $Q\left(\frac{m}{\sigma}\right)$ as a function of the n/p ratio, for logistic and square losses.

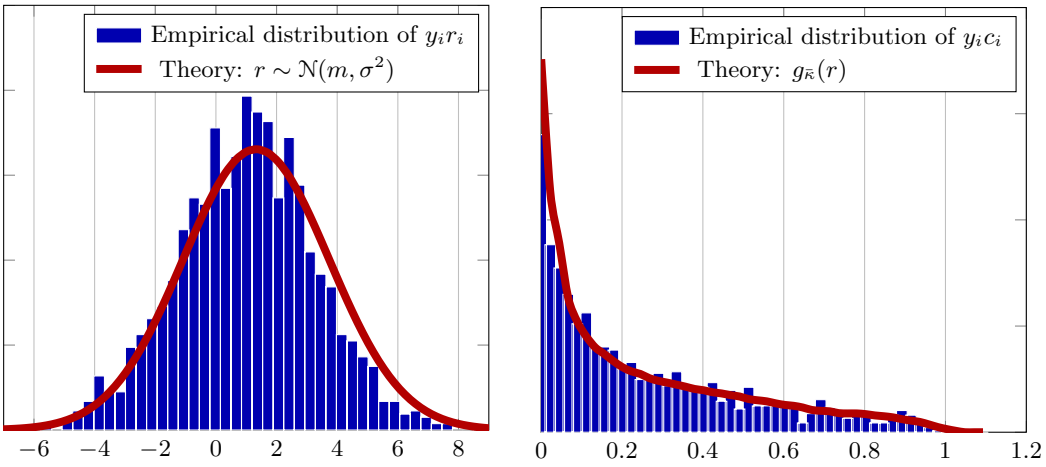

Figure 6.8: Comparison between the empirical distribution of $y_i r_i$ and $y_i c_i$ with their theoretical prediction given in Theorem 6.5.1. For logistic loss $\rho(t) = \ln(1 + e^{-t})$, $\boldsymbol{\mu} = [1, \mathbf{0}_{p-1}]$, $\mathbf{C} = 2\mathbf{I}_p$ and $p = 256$, $n = 6p$.

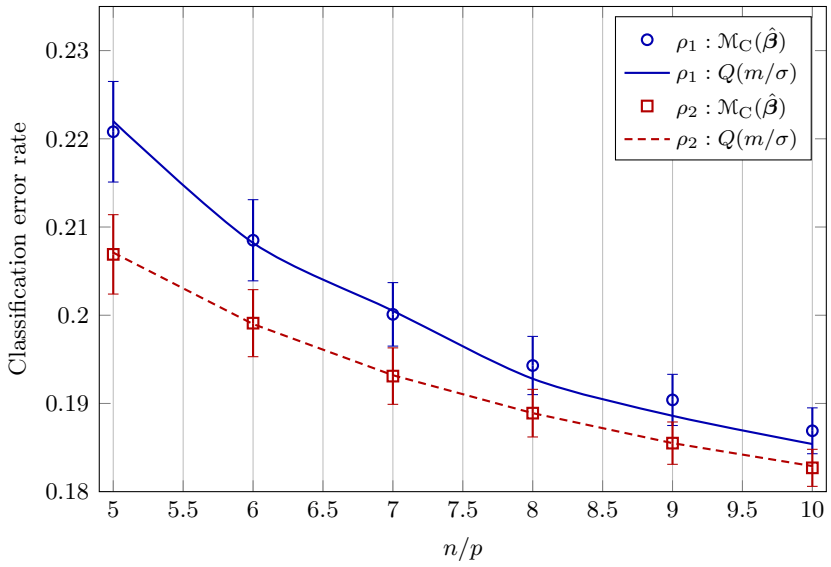

Figure 6.9: Comparison of classification error rate between the logistic loss $\rho_1(t) = \ln(1 + e^{-t})$, the square loss $\rho_2(t) = (t-1)^2/2$ and the theoretical results given in Corollary 6.3 with a width of ± 1 standard deviation (generated from 500 trials) for $\boldsymbol{\mu} = [\mathbf{1}_{p/2}, -\mathbf{1}_{p/2}]/\sqrt{p}$, $\mathbf{C} = \mathbf{I}_p$ and $p = 256$.

6.6 Concluding remarks

In this chapter, we investigated the problem of high dimensional classification within the general framework of empirical risk minimization. We showed that, for the high dimensional mixture model under consideration, all classifiers $\hat{\boldsymbol{\beta}}$ given in (6.2) are aligned in expectation to the oracle direction, however with different scaling factors that depend on the ratio n/p. Based on this result, we proposed a rescaling method to render high dimensional solutions unbiased for an enhanced class probability estimation. We demonstrated subsequently that the square loss solution, instead of the maximal likelihood solution given by the negative log-likelihood loss (i.e., the logistic loss), yields the best results in both classification and class probability estimation after being corrected by the proposed rescaling strategy. Our analysis furthermore served to statistically characterize linear combinations of classifiers learned with different loss functions, allowing us to conclude on the possibilities and limitations of this ensemble learning approach.

The proposed analysis framework is generalizable to more generic mixture models of non-Gaussian feature vectors (in a similar manner to the results of SVMs given in Chapter 5), however at the cost of the readability and interpretability of the theoretical results. Indeed, under the data model of this chapter, we were able to obtain a convenient statistical representation for the ERM solutions, which was notably useful for deriving the conclusions on the optimality of the empirical risk minimization approach. It is interesting to see if these insightful remarks generalize to other more elaborate settings in future investigations, which may be conducted by further developing the results and the mathematical approach of this chapter.

The extension to non-smooth and non-convex loss functions is on the other hand more technically challenging. The hinge loss function, arguably the most known of the non-smooth loss functions, is used in the SVM method, which is studied in the previous chapter and can be seen as a regularized version of the empirical risk minimization technique. The present study can also be further developed to encompass regularized solutions, as a means to explore the joint effect of loss functions and regularizations (e.g., ℓ_1 or ℓ_2 regularization).

Chapter 7

Conclusions and perspectives

Up to now, the field of machine learning has been largely driven by engineering techniques. The understanding of these techniques, or more essentially, the understanding of the learning process itself remains unsatisfactory. This lack of understanding is at the root of the fact that, so far, the most performing classifiers often owe their impressive, sometimes superhuman, performance to the power of massive amounts of input data. Precisely, when the methods of machine learning are said now to achieve superhuman performance in certain learning tasks, it usually refers to the regime $n \gg p$, which certainly does not imply that these methods can compare to the human level of learning efficiency with relatively less data samples. To reach the next stage in the development of artificial intelligence, we clearly need to turn our attention to learning scenarios with comparable n, p in search for improved learning efficiency. This is where the overarching ambition of this research project lies.

A deep analysis of the learning process is however far from being an easy task. In fact, empirical successes almost always exceed theoretical advances in the domain of machine learning. Motivated by the big data paradigm, we place ourselves in the setting of large and comparable n, p. The high dimensionality of modern data actually happens to be theoretically convenient for analyzing the regime of finite n/p. In this dissertation, we merged the advanced techniques of random matrix theory and the leave-one-out perturbation strategy which allow us to conduct, on structured data, more elaborate analyses than provided by the state of the art.

The potential of such high dimensional analyses is in particular demonstrated by our contribution in semi-supervised graph-based learning. By fully characterizing the learning outcomes as a function of the size ratio n/p in the limit of large p, we discovered that the currently used Laplacian regularization algorithms are flawed as they fail to effectively extract information from the unlabelled data. In view of this finding, we proposed a novel approach as a solution, which we proved to have a consistent performance growth when the size of the unlabeled set increases, while maintaining a labeled data learning efficiency lower-bounded by that of the Laplacian regularization approach. A fundamental observation about this proposed method is that its advantage over the Laplacian approach appears to extend beyond the high dimensional data model, evidencing the prospect of improving the general efficiency of machine learning approaches (even for small dimensional data) through large dimensional studies.

Another important contribution of this thesis is to show that, even in the well developed area of supervised learning, there still remains a lot to be discovered. The power of our new

technical approach allows one to address the issue of admitting no explicit solution, and enables us to analyze the widely used methods of SVMs and logistic regression, under realistic mixture models. Thanks to an insightful statistical description of the learned parameters, we were able to deduce surprising remarks about the bias-variance trade-off, the maximum likelihood principle and the limitations of ensemble learning, which are not covered in the literature, some even in contradiction with common belief.

It is worth noting that for the study of the kernel methods, we concentrated in this thesis on kernels of the form $h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ with some sufficiently smooth function h. Since, under our high dimensional assumption, the value of $h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ converges to the same limit $h(\tau)$ for any pair of data vectors $\mathbf{x}_i, \mathbf{x}_j$, we analyzed the performance of the kernel methods by developing the small fluctuations around this limit. It is should be pointed out that in the development of $h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ around $h(\tau)$, there are some smaller terms left out in our analyses that can be useful to the learning task. However, for these smaller terms to have a non-negligible impact on the learning performance, the larger terms (that are kept in our analyses) should be somehow "made smaller" in order of magnitude. This idea was for example investigated in [76], where the authors analyzed the performance of kernel spectral clustering for kernel functions h with $h'(\tau) = O(p^{-\frac{1}{2}})$, a condition that is required for the learning method to access information in smaller terms. We can envision the same extension to the analyses of this thesis in order to further understand the role of the kernel function in the methods under study.

This thesis opens up many future directions in the line of high dimensional research. One direction is to go deeper in the study of semi-supervised learning, where our first works turned out to be instrumental in understanding and improving this rather underdeveloped field of machine learning. Since our technical tools allow one to examine learning methods with implicit solutions, they can be applied to a large range of more complex semi-supervised learning algorithms than those studied in this thesis, which have explicit solutions. For instance, as an extension of SVMs in the semi-supervised learning setting, the method of transductive support vector machines (TSVMs) [77] aims to construct a separating hyperplane on a low-density region of the whole data set with labelled and unlabelled examples that divides the labelled instances according to their classes. Like SVMs, the TSVM algorithm admits no closed-form solution. More than that, the convexity of the optimization problem is no longer ensured as a result of including unlabeled data. Despite some successful applications, it is observed that TSVMs yield in certain scenarios roughly the same performance as SVMs or even worse, arguably due to the difficulty of finding the global optimum. With the help of high dimensional analyses, we can provide a complete description of all the local and global solutions to the TSVM optimization. A first question that the high dimensional results can help answer is how many labelled data are required or how much weight should be associated to them so that the TSVM optimization can lead to reasonable solutions. Based on the properties of the local and global solutions revealed by the high dimensional study of TSVMs, it would also be possible to design improved strategies to guide more efficiently the optimization algorithms towards the global minimum. Instead of implementing the low-density separation principle over the entire set of labelled and unlabelled data, Laplacian SVMs [63] incorporate the unlabelled information by adding a Laplacian regularization term into the original SVM optimization, and have the advantage of preserving the convexity of the optimization problem. Since Laplacian SVMs share the same Laplacian regularization terms as the semi-supervised Laplacian regularization algorithms studied in Chapter 3, it is natural to wonder if they suffer also the same high dimensional conse-

quences. By conducting large dimensional analyses, we can find out the answer to this question, and more generally understand how imposing the smoothness of class scores on graph through Laplacian regularization affects the original SVM solution, which might provide guidance on the choice of hyperparameters or even inspire superior alternatives.

A closely related research area to semi-supervised learning is active learning, which involves also labelled and unlabelled data, but is focused on improving the learning performance by finding the optimal set of unlabelled data to label [78]. Semi-supervised learning and active learning are different in what they care most about unlabelled data: the former is concerned with what can be learned from unlabelled data, whereas the latter is interested in the uncertainty of unlabelled instances. For example, self-training is a standard semi-supervised technique that trains first a classifier with the labelled data set, then uses it to classify unlabelled data before feeding the most confidently classified unlabelled samples into the training set, and repeats the process. In contrast, a basic concept in active learning is to select the least confident unlabelled instances to label. Active learning can be seen as a complementary technique to semi-supervised learning, and we can push further the high dimensional study on semi-supervised learning methods by investigating the possibility of enhancing their performance through active learning. Indeed, based on the statistical characterization of learning outcomes given by high dimensional analyses, we can derive the distribution of updated results after applying a certain active learning technique. By doing so, we can shed light on the impact of active learning algorithms, discuss their limitations as well as search for optimal strategies.

Another major direction of exploration to apply these high dimensional statistical tools concerns learning methods involving multiple sources of information like transfer learning (see, e.g., [79] for an overview). The idea of transfer learning is to improve the results of a target learning task by exploiting the knowledge gained from a source task where the data samples are much more abundant compared to the target task. A major complication about transfer learning is that since it concerns multiple learning tasks, data samples are no longer considered to be drawn from the same distribution as commonly assumed in machine learning. Understandably, how to take advantage of the relation between the source task and the target task is the key question in transfer learning. In fact, negative transfer can happen when using data samples of the source task decreases the performance of the target task, as empirically evidenced by [80]. To avoid negative transfer, one needs to study and characterize the relatedness between tasks. As high dimensional analyses give rise to quantitative descriptions of learning systems, they might be used to design a measure of transferability between tasks, which can help practitioners decide when and which transfer learning techniques to apply. On a more precise level, since lots of transfer learning methods admit some hyperparameters to control the emphasis on the data of the source task, the quantification of transferability on high dimensional data would allow one to determine the exact optimal weight to assign to the information from the source task. Although sharing some similarity with semi-supervised learning, transfer learning requires a more elaborate modeling of the data distribution for theoretical studies than those considered in a semi-supervised learning setting, as the latter are only meant to describe a single learning task.

Even though we focused in this dissertation on analyzing each learning method individually, the technical approach, even some of our results, can be straightforwardly used in the study of ensemble learning or distributed learning. Actually, in Chapter 6 we already applied the joint framework of the empirical risk minimization (ERM) algorithms to investigate the simple

ensemble learning technique that linearly combines several ERM classifiers with different loss functions, trained on the same data set. We showed how our high dimensional results can be employed to estimate the optimal weights to assign to each member classifier. More interestingly, we found that, under the Gaussian-mixture models with identical covariances considered in the analysis, the performance of this elementary ensemble learning technique is upper bounded by that of a single least squares classifier. This remark is illustrative of the importance of theoretical studies. Indeed, when looking only at empirical results, the practitioner might be under the impression that this ensemble learning method was beneficial to the improvement of the learning performance, although in reality this improvement is only observed when the optimal loss function (here the square loss) is not used. It is interesting to examine the extent of this remark by studying more sophisticated ensemble learning techniques under more involved data models. It is also possible to apply our high dimensional results to the design of better distributed learning strategies. Distributed learning is based on several learning models, each of them learning from subsets of the data samples, which can be of different sizes. Understandably, the optimal weights for combining the individual learning models are dependent, however not necessarily in an obvious way, of their input data size. Since the results of analyses with comparably large n, p are quantitative in the sense that they capture the exact effect of the data size, they can be used to design consistent estimators of the optimal weights that maximize the performance of distributed learning in the limit of very large dimensions.

In addition to investigating general learning methods, the approach of large dimensional analyses can be used to study any particular task in big data mining. After mathematically modeling the learning problem, we can examine, with the help of high dimensional statistical tools, the adequacy of current techniques for handling the data of this learning task, provide guidance on how to employ them in practice, and eventually search for superior alternatives based on the insights drawn from the performance analysis. Naturally, the key to this kind of research work is a proper mathematical modeling of the learning problem which captures the essential properties of the data type and reflects the real difficulties faced in practice. All this requires a solid grasp of the learning task, built upon extensive empirical observations. Once the high dimensional analysis is carried out under a presumably appropriate statistical model of the data samples, we can discover which statistical properties of data are captured by which learning strategies and with how much efficiency, then answer important questions about whether or not the intended purpose of the learning strategies is fulfilled and under which conditions. These remarks will in turn enable us to choose wisely the learning technique to use according to the specific data set at hand and the requirements of the learning task. Eventually, the findings of the high dimensional study can help devise better learning procedures, e.g., by smartly combining several learning techniques that capitalize on different sets of data statistical properties, or by inventing new learning tools exploiting the aspects in the data structure that are not put to use (or not with enough efficiency) by existing techniques.

By comparing the theoretical results with empirical outcomes obtained from executing the learning task, the high dimensional analysis will further allow one to check the validity of the presumed data model. It should be noted that a close match between the theoretical results and the empirical outcomes does not necessarily imply that the actual data distribution is very similar to the one assumed in the analysis. However, it does suggest that the learning algorithm under study treats data instances of this particular learning task as if they were drawn from the assumed data distribution. Recall for instance from Chapter 3 that the theoretical performance of semi-supervised Laplacian regularization obtained under high dimensional Gaussian mixture models

is remarkably close to the actual performance on the MNIST data sets, even though the MNIST data are nowhere near appropriately described by a Gaussian mixture model. This observation tells us that the semi-supervised Laplacian regularization algorithms extract information from MNIST data in a manner similar to Gaussian data vectors, the distribution of which is fully determined by its first and second moments. If we later found a method that yields a learning performance higher than the theoretical value obtained under high dimensional Gaussian mixture models, this would mean that this other method captures more complicated data structure than Gaussian mixture models. The process of fitting the performance of this method would help the refinement of the data model. Conversely, if a relatively simple data model gives essentially the same level of approximation for the actual performance as a more complicated model, it is preferable to use the simpler model that produces more straightforward results from which it is easier to extract meaningful messages. As such, the high dimensional analysis provides a way to deepen the understanding of data structures and also to improve its statistical modeling.

Publications

Tutorials in International Conferences

- R. Couillet, Z. Liao, X. Mai, "Random Matrix Advances in Machine Learning and Neural Nets", European Signal Processing Conference (EUSIPCO'18), Rome, Italy, 2018. [slides]

Articles in Journals

- X. Mai, Z. Liao, "High Dimensional Classification via Empirical Risk Minimization: Statistical Analysis and Optimality", in preparation, 2019.
- X. Mai, R. Couillet, "Statistical Behavior and Performance of Support Vector Machines for Large Dimensional Data", in preparation, 2019.
- X. Mai, R. Couillet, "Consistent Semi-Supervised Graph Regularization for High Dimensional Data", submitted to Journal of Machine Learning Research, 2019. [article]
- X. Mai, R. Couillet,"A Random Matrix Analysis and Improvement of Semi-Supervised Learning for Large Dimensional Data", Journal of Machine Learning Research, vol. 19, no. 79, pp. 1-27, 2018. [article]

Articles in International Conferences

- X. Mai, Z. Liao, R. Couillet, "A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019. [article]
- X. Mai, R. Couillet, "Revisiting and Improving Semi-Supervised Learning: A Large Dimensional Approach", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019. [article]
- X. Mai, R. Couillet, "Semi-Supervised Spectral Clustering", Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2018. [article]
- R. Couillet, Z. Liao, X. Mai, "Classification Asymptotics in the Random Matrix Regime", European Signal Processing Conference (EUSIPCO'18), Rome, Italy, 2018. [article]

- X. Mai, R. Couillet, "The Counterintuitive Mechanism of Graph-based Semi-Supervised Learning in the Big Data Regime", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017. [article]

Appendix A

Supplementary material of Chapter 4

A.1 Generalized theorem

We begin with some additional notations that will be useful in the proofs.

- For $\mathbf{x}_i \in \mathcal{C}_k$, $\boldsymbol{\omega}_i \equiv (\mathbf{x}_i - \mu_k)/\sqrt{p}$, and $\boldsymbol{\Omega} \equiv [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n]^\mathsf{T}$

- $\mathbf{j}_k \in \mathbb{R}^n$ is the canonical vector of $\mathcal{C}_k$, in the sense that its i-th element is 1 if $\mathbf{x}_i \in \mathcal{C}_k$ or 0 otherwise. $\mathbf{j}_{[l]k}$ and $\mathbf{j}_{[u]k}$ are respectively the canonical vectors for labelled and unlabelled data of $\mathcal{C}_k$.

- $\psi_i \equiv \|\boldsymbol{\omega}_i\|^2 - \mathrm{E}[\|\boldsymbol{\omega}_i\|^2]$, $\boldsymbol{\psi} \equiv [\psi_1, \cdots, \psi_n]^\mathsf{T}$ and $(\boldsymbol{\psi})^2 \equiv [(\psi_1)^2, \cdots, (\psi_n)^2]^\mathsf{T}$.

With these notations at hand, we introduce next the generalized version of Theorem 3.4.1 for all $\gamma = O(1)$ (rather than $\gamma = -1 + O(1/\sqrt{n})$).

Theorem A.1.1. *For $\mathbf{x}_i \in \mathcal{C}_b$ an unlabelled vector (i.e., $i > n_{[l]}$), let $\hat{f}_{ia}$ be given by* (3.7) *with* $\mathbf{F}$ *defined in* (3.2) *for $\gamma = O(1)$. Then, under Assumptions 3.1–3.2,*

$$p\{\hat{\mathbf{F}}\}_{i\cdot} = p(1 + z_i)\mathbf{1}_K + \mathbf{g}_i + o_P(1)$$
$$\mathbf{g}_i \sim \mathcal{N}(\mathbf{m}_b, \boldsymbol{\Sigma}_b)$$

where z_i is as in Theorem 3.4.1 and

(i) for $\{\mathbf{F}\}_{i\cdot}$ considered on the σ-field induced by the random variables $\mathbf{x}_{[l]+1}, \ldots, \mathbf{x}_n$, $p =$

$1, 2, \ldots,$

$$[\mathbf{m}_b]_a = H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} (\gamma n_d + n_{[u]d}) H_{ad} + (1+\gamma)\frac{n}{n_{[l]}} \left[\Delta_a + \frac{p}{n_{[l]a}} \frac{h'(\tau)}{h(\tau)} \boldsymbol{\psi}_{[l]}^T \mathbf{j}_{[l]a} - \gamma \frac{h'(\tau)^2}{h(\tau)^2} t_a t_b \right] \tag{A.1}$$

$$[\boldsymbol{\Sigma}_b]_{a_1 a_2} = \left(\frac{(-\gamma^2 - \gamma)n - n_{[l]}}{n_{[l]}} \frac{h'(\tau)^2}{h(\tau)^2} + \frac{h''(\tau)}{h(\tau)} \right)^2 T_{bb} t_{a_1} t_{a_2} + \boldsymbol{\delta}_{a_1}^{a_2} \frac{h'(\tau)^2}{h(\tau)^2} \frac{4c_0 T_{ba_1}}{c_{[l]a_1}} + \frac{4h'(\tau)^2}{h(\tau)^2} \mu_{a_1}^{\circ} \mathbf{C}_b \mu_{a_2}^{\circ} \tag{A.2}$$

where

$$H_{ab} = \frac{h'(\tau)}{h(\tau)} \|\boldsymbol{\mu}_b^{\circ} - \boldsymbol{\mu}_a^{\circ}\|^2 + \left(\frac{h''(\tau)}{h(\tau)} - \frac{h'(\tau)^2}{h(\tau)^2} \right) t_a t_b \tag{A.3}$$

$$\Delta_a = \frac{\sqrt{p} h'(\tau)}{h(\tau)} t_a + \frac{\gamma h'(\tau)^2 + h(\tau) h''(\tau)}{2h(\tau)^2} \left(2T_{aa} + t_a^2 \right) + \frac{1}{n_{[l]}} \left(\frac{h'(\tau)}{h(\tau)} \right)^2 \left(\sum_{d=1}^{K} n_{[u]d} t_d \right) t_a. \tag{A.4}$$

(ii) for $F_{i\cdot}$ considered on the σ-field induced by the random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$,

$$\{\mathbf{m}_b\}_a = H_{ab} + \frac{1}{n_{[l]}} \sum_{d=1}^{K} (\gamma n_d + n_{[u]d}) H_{ad} + (1+\gamma)\frac{n}{n_{[l]}} \left[\Delta_a - \gamma \frac{h'(\tau)^2}{h(\tau)^2} t_a t_b \right]$$

$$\{\boldsymbol{\Sigma}_b\}_{a_1 a_2} = \left(\frac{(-\gamma^2 - \gamma)n - n_{[l]}}{n_{[l]}} \frac{h'(\tau)^2}{h(\tau)^2} + \frac{h''(\tau)}{h(\tau)} \right)^2 T_{bb} t_{a_1} t_{a_2} + \boldsymbol{\delta}_{a_1}^{a_2} \frac{h'(\tau)^2}{h(\tau)^2} \left(\frac{(1+\gamma)^2 n^2}{n_{[l]}^2} \frac{2T_{aa}}{c_{[l]a_1}} + \frac{4T_{ba_1}}{c_{[l]a_1}} \right) + \frac{4h'(\tau)^2}{h(\tau)^2} \boldsymbol{\mu}_{a_1}^{\circ} \mathbf{C}_b \boldsymbol{\mu}_{a_2}^{\circ}$$

with H_{ab} given in (A.3) *and Δ_a in* (A.4)*.*

Let $\mathbb{P}(\mathbf{x}_i \to \mathcal{C}_b | \mathbf{x}_i \in \mathcal{C}_b, \mathbf{x}_1, \cdots, \mathbf{x}_{n_{[l]}})$ denote the probability of correct classification of $\mathbf{x}_i \in \mathcal{C}_b$ unlabelled, conditioned on $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]}}$, and $\mathbb{P}(\mathbf{x}_i \to \mathcal{C}_b | \mathbf{x}_i \in \mathcal{C}_b)$ the unconditional probability. Recall that the probability of correct classification of $\mathbf{x}_i \in \mathcal{C}_b$ is the same as the probability of $\{\hat{\mathbf{F}}\}_{ib} > \max_{a \neq b} \{\hat{\mathbf{F}}\}_{ib}$, which, according to the above theorem, is asymptotically the probability that $[\mathbf{g}_i]_b$ is the greatest element of $\mathbf{g}_i$. Particularly for $K = 2$, we have the following corollary.

Corollary A.1. *Under the conditions of Theorem 1, and with $K = 2$, we have, for $a \neq b \in \{1, 2\}$,*

(i) Conditionally on $\mathbf{x}_1, \cdots, \mathbf{x}_{n_{[l]}}$,

$$\mathbb{P}\left(\mathbf{x}_i \to \mathcal{C}_b | \mathbf{x}_i \in \mathcal{C}_b, \mathbf{x}_1, \cdots, \mathbf{x}_{n_{[l]}} \right) - \Phi(\theta_b^a) \to 0$$

$$\theta_b^a = \frac{[\mathbf{m}_b]_b - [\mathbf{m}_b]_a}{\sqrt{[\boldsymbol{\Sigma}_b]_{bb} + [\boldsymbol{\Sigma}_b]_{aa} - 2[\boldsymbol{\Sigma}_b]_{ab}}}$$

where $\Phi(u) = \frac{1}{2\pi} \int_{-\infty}^{u} \exp(-t^2/2) dt$ and $\mathbf{m}_b$, $\boldsymbol{\Sigma}_b$ are given in (i) of Theorem A.1.1.

(ii) Unconditionally,

$$\mathbb{P}\left(\mathbf{x}_i \to \mathcal{C}_b | \mathbf{x}_i \in \mathcal{C}_b\right) - \Phi(\theta_b^a) \to 0$$
$$\theta_b^a = \frac{[\mathbf{m}_b]_b - [\mathbf{m}_b]_a}{\sqrt{[\mathbf{\Sigma}_b]_{bb} + [\mathbf{\Sigma}_b]_{aa} - 2[\mathbf{\Sigma}_b]_{ab}}}$$

where here $\mathbf{m}_b$, $\mathbf{\Sigma}_b$ *are given in (ii) of Theorem A.1.1.*

The remainder of the appendix is dedicated to the proof of Theorem A.1.1 and Corollary A.1 from which the results of Section 3.4 directly unfold.

A.2 Proof of the generalized theorem

The proof of Theorem A.1.1 is divided into two steps: first, we Taylor-expand the normalized scores for unlabelled data $\hat{\mathbf{F}}_{[u]}$ using the convergence $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \xrightarrow{\text{a.s.}} \tau$ for all $i \neq j$; this expansion yields a random equivalent $\hat{\mathbf{F}}^{\text{eq}}_{[u]}$ in the sense that $p(\hat{\mathbf{F}}_{[u]} - \hat{\mathbf{F}}^{\text{eq}}_{[u]}) \xrightarrow{\text{a.s.}} 0$. Proposition 3.4.1 is directly obtained from $\hat{\mathbf{F}}^{\text{eq}}_{[u]}$. We then complete the proof by demonstrating the convergence to Gaussian variables of $\hat{\mathbf{F}}^{\text{eq}}_{[u]}$ by means of a central limit theorem argument.

A.2.1 Step 1: Taylor expansion

In the following, we provide a sketch of the development of $\mathbf{F}_{[u]}$; most unshown intermediary steps can be retrieved from simple, yet painstaking algebraic calculus.

Recall from (3.2) the expression of the unnormalized scores for unlabelled data

$$\mathbf{F}_{[u]} = (\mathbf{I}_{n_u} - \mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^{\gamma})^{-1}\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^{\gamma}\mathbf{F}_{[l]}.$$

We first proceed to the development of the terms $\mathbf{W}_{[ul]}$, $\mathbf{W}_{[uu]}$, subsequently to $\mathbf{D}_{[l]}, \mathbf{D}_{[u]}$, to then reach an expression for $\mathbf{F}_{[u]}$. To this end, owing to the convergence $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{\text{a.s.}} \tau$ for all $i \neq j$, we first Taylor-expand $\mathbf{W}_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ around $h(\tau)$ to obtain the following expansion for $\mathbf{W}$, already evaluated by [45],

$$\mathbf{W} = \mathbf{W}^{(n)} + \mathbf{W}^{(\sqrt{n})} + \mathbf{W}^{(1)} + O(n^{-\frac{1}{2}}) \tag{A.5}$$

where $\|\mathbf{W}^{(n)}\| = O(n)$, $\|\mathbf{W}^{(\sqrt{n})} = O(\sqrt{n})$ and $\|\mathbf{W}^{(1)}\| = O(1)$, with the definitions

$$\begin{aligned}
\mathbf{W}^{(n)} &= h(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} \\
\mathbf{W}^{(\sqrt{n})} &= h'(\tau)\left[\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T} + \left(\sum_{b=1}^K \frac{t_b}{\sqrt{p}}\mathbf{j}_b\right)\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\sum_{a=1}^K \frac{t_a}{\sqrt{p}}\mathbf{j}_a^\mathsf{T}\right] \\
\mathbf{W}^{(1)} &= h'(\tau)\Bigg[\sum_{a,b=1}^K \frac{\|\boldsymbol{\mu}_a^\circ - \boldsymbol{\mu}_b^\circ\|^2}{p}\mathbf{j}_b\mathbf{j}_a^\mathsf{T} - \frac{2}{\sqrt{p}}\boldsymbol{\Omega}\sum_{a=1}^K \boldsymbol{\mu}_a^\circ\mathbf{j}_a^\mathsf{T} + \frac{2}{\sqrt{p}}\sum_{b=1}^K \mathrm{diag}(\mathbf{j}_b)\boldsymbol{\Omega}\boldsymbol{\mu}_b^\circ\mathbf{1}_n^\mathsf{T} \\
&\quad - \frac{2}{\sqrt{p}}\sum_{b=1}^K \mathbf{j}_b\boldsymbol{\mu}_b^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T} + \frac{2}{\sqrt{p}}\mathbf{1}_n\sum_{a=1}^K \boldsymbol{\mu}_a^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T}\mathrm{diag}(\mathbf{j}_a) - 2\boldsymbol{\Omega}\boldsymbol{\Omega}^\mathsf{T}\Bigg] \\
&\quad + \frac{h''(\tau)}{2}\Bigg[(\boldsymbol{\psi})^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n[(\boldsymbol{\psi})^2]^\mathsf{T} + \sum_{b=1}^K \frac{t_b^2}{p}\mathbf{j}_b\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\sum_{a=1}^K \frac{t_a^2}{p}\mathbf{j}_a^\mathsf{T} \\
&\quad + 2\sum_{a,b=1}^K \frac{t_at_b}{p}\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + 2\sum_{b=1}^K \mathrm{diag}(\mathbf{j}_b)\frac{t_b}{\sqrt{p}}\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + 2\sum_{b=1}^K \frac{t_b}{\sqrt{p}}\mathbf{j}_b\boldsymbol{\psi}^\mathsf{T} + 2\sum_{a=1}^K \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T}\mathrm{diag}(\mathbf{j}_a)\frac{t_a}{\sqrt{p}} \\
&\quad + 2\boldsymbol{\psi}\sum_{a=1}^K \frac{t_a}{\sqrt{p}}\mathbf{j}_a^\mathsf{T} + 4\sum_{a,b=1}^K \frac{T_{ab}}{p}\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + 2\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\Bigg] + (f(0) - h(\tau) + \tau h'(\tau))\mathbf{I}_n.
\end{aligned}$$

As $\mathbf{W}_{[ul]}$, $\mathbf{W}_{[uu]}$ are sub-matrices of $\mathbf{W}$, their approximated expressions are obtained directly by extracting the corresponding subsets of (A.5). Applying then (A.5) in $\mathbf{D} = \mathrm{diag}(\mathbf{W}\mathbf{1}_n)$, we next find

$$\mathbf{D} = nh(\tau)\left[\mathbf{I}_n + \frac{1}{nh(\tau)}\mathrm{diag}(\mathbf{W}^{(\sqrt{n})}\mathbf{1}_n + \mathbf{W}^{(1)}\mathbf{1}_n)\right] + O(n^{-\frac{1}{2}}).$$

Thus, for any $\sigma \in \mathbb{R}$, $(n^{-1}\mathbf{D})^\sigma$ can be Taylor-expanded around $h(\tau)^\sigma\mathbf{I}_n$ as

$$\begin{aligned}
(n^{-1}\mathbf{D})^\sigma &= h(\tau)^\sigma\left[\mathbf{I}_n + \frac{\sigma 1}{nh(\tau)}\mathrm{diag}((\mathbf{W}^{(\sqrt{n})} + \mathbf{W}^{(1)})\mathbf{1}_n) + \frac{\sigma(\sigma-1)}{2n^2h(\tau)^2}\mathrm{diag}^2(\mathbf{W}^{(\sqrt{n})}\mathbf{1}_n)\right] \\
&\quad + O(n^{-\frac{3}{2}})
\end{aligned} \tag{A.6}$$

where $\mathrm{diag}^2(\cdot)$ stands for the squared diagonal matrix. The Taylor-expansions of $(n^{-1}\mathbf{D}_{[u]})^\gamma$ and $(n^{-1}\mathbf{D}_{[l]})^\gamma$ are then directly extracted from this expression for $\sigma = \gamma$, and similarly for $(n^{-1}\mathbf{D}_{[u]})^{-1-\gamma}$ with $\sigma = -1-\gamma$. Since

$$\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^\gamma = \frac{1}{n}(n^{-1}\mathbf{D}_{[u]})^{-1-\gamma}\mathbf{W}_{[ul]}(n^{-1}\mathbf{D}_{[l]})^\gamma$$

it then suffices to multiply the Taylor-expansions of $(n^{-1}\mathbf{D}_{[u]})^\gamma$, $(n^{-1}\mathbf{D}_{[l]})^\gamma$, and $\mathbf{W}_{[ul]}$, given respectively in (A.6) and (A.5), normalize by n and then organize the result in terms of order $O(1)$, $O(1/\sqrt{n})$, and $O(1/n)$.

The term $\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^\gamma$ is dealt with in the same way. In particular,

$$\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^\gamma = \frac{1}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[l]}} + O(n^{-\frac{1}{2}}).$$

Therefore, $(\mathbf{I}_{n_{[u]}} - \mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^{\gamma})^{-1}$ may be simply written as

$$\left(\mathbf{I}_{n_{[u]}} - \frac{1}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}} + O(n^{-\frac{1}{2}})\right)^{-1} = \mathbf{I}_{n_{[u]}} + \frac{1}{n_{[l]}}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}} + O(n^{-\frac{1}{2}}).$$

Combining all terms together completes the full linearization of $\hat{\mathbf{F}}_{[u]}$.

This last derivation, which we do not provide in full here, is simpler than it appears and is in fact quite instructive in the overall behavior of $\mathbf{F}_{[u]}$. Indeed, only product terms in the development of $(\mathbf{I}_{n_{[u]}} - \mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[uu]}\mathbf{D}_{[u]}^{\gamma})^{-1}$ and $\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^{\gamma}\mathbf{F}_{[l]}$ of order at least $O(1)$ shall remain, which discards already a few terms. Now, in addition, note that for any vector $\mathbf{v}$, $\mathbf{v}\mathbf{1}_{n_{[l]}}^{\mathsf{T}}\mathbf{F}_{[l]} = \mathbf{v}\mathbf{1}_k^{\mathsf{T}}$ so that such matrices are non informative for classification (they have identical score columns); these terms are all placed in the intermediary variable z, the entries z_i of which are irrelevant and thus left as is (these are the z_i's of Proposition 3.4.1 and Theorem 3.4.1). It is in particular noteworthy to see that *all* terms of $\mathbf{W}_{[uu]}^{(1)}$ that remain after taking the product with $\mathbf{D}_{[u]}^{-1-\gamma}\mathbf{W}_{[ul]}\mathbf{D}_{[l]}^{\gamma}\mathbf{F}_{[l]}$ are precisely those multiplied by $h(\tau)\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[l]}}^{\mathsf{T}}\mathbf{F}_{[l]}$ and thus become part of the vector z. Since most informative terms in the kernel matrix development are found in $\mathbf{W}^{(1)}$, this means that the algorithm under study shall make little use of the *unsupervised* information about the data (those found in $\mathbf{W}_{[uu]}^{(1)}$). This is an important remark which, as discussed in Section 6.6, opens up the path to further improvements of the semi-supervised learning algorithms which would use more efficiently the information in $\mathbf{W}_{[uu]}^{(1)}$.

All calculus made, this development finally leads to $\mathbf{F}_{[u]} = \mathbf{F}_{[u]}^{\text{eq}}$ with, for $a, b \in \{1, \ldots, K\}$ and $\mathbf{x}_i \in \mathcal{C}_b$, $i > n_{[l]}$,

$$\begin{aligned}\left\{\hat{\mathbf{F}}^{\text{eq}}\right\}_{ia} &= 1 + \frac{1}{p}\left[H_{ab} + \frac{1}{n_{[l]}}\sum_{d=1}^{K} H_{ad}(\gamma n_d + n_{[u]d})\right] + (1+\gamma)\frac{n}{pn_{[l]}}\left[\Delta_a - \gamma\frac{h'(\tau)^2}{h(\tau)^2}t_a t_b\right] \\ &+ \left(\frac{(-\gamma^2-\gamma)n - n_{[l]}}{n_{[l]}}\frac{h'(\tau)^2}{h(\tau)^2} + \frac{h''(\tau)}{h(\tau)}\right)\frac{t_a}{\sqrt{p}}\psi_i + \frac{2h'(\tau)}{h(\tau)\sqrt{p}}\boldsymbol{\mu}_a^{\circ}\boldsymbol{\omega}_i \\ &+ \frac{h'(\tau)}{h(\tau)}\left(\frac{(1+\gamma)n}{n_{[l]}n_{[l]a}}\boldsymbol{\psi}_{[l]}^{\mathsf{T}}\mathbf{j}_{[l]a} + \frac{4}{n_{[l]a}}\mathbf{j}_{[l]a}^{T}\boldsymbol{\Omega}_{[l]}\boldsymbol{\omega}_i\right) + z_i \end{aligned} \tag{A.7}$$

where H_{ab} is as specified in (A.3), Δ_a as in (A.4), and $z_i = O(\sqrt{p})$ is some residual random variable only dependent on $\mathbf{x}_i$. Gathering the terms in successive orders of magnitude, Proposition 3.4.1 is then straightforwardly proven from (A.7).

A.2.2 Step 2: Central limit theorem

The focus of this step is to examine $\tilde{\mathbf{g}}_i = p\left\{\hat{\mathbf{F}}^{\text{eq}}\right\}_{i\cdot} - (1+z_i)\mathbf{1}_K)$. Theorem A.1.1 can be proven by showing that $\tilde{\mathbf{g}}_i = \mathbf{g}_i + o_P(1)$.

First consider Item (i) of Theorem A.1.1, which describes the behavior of $\hat{\mathbf{F}}_{[u]}$ conditioned on $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]}}$. Recall that a necessary and sufficient condition for a vector $\mathbf{v}$ to be a Gaussian vector is that all linear combinations of the elements of $\mathbf{v}$ are Gaussian variables. Thus, for given $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]}}$ deterministic, according to (A.7), $\tilde{\mathbf{g}}_i$ is asymptotically Gaussian if, for all $s_1 \in \mathbb{R}$, $\mathbf{s}_2 \in \mathbb{R}^p$, $s_1\psi_i + \mathbf{s}_2^T\boldsymbol{\omega}_i$ has a central limit.

Letting $\boldsymbol{\omega}_i = \frac{\mathbf{C}_b^{\frac{1}{2}}}{\sqrt{p}}\mathbf{r}$, with $\mathbf{r} \sim \mathcal{N}(0, \mathbf{I}_p)$, $s_1\psi_i + \mathbf{s}_2\boldsymbol{\omega}_i$ can be rewritten as $\mathbf{r}^\mathsf{T}\mathbf{A}\mathbf{r} + \mathbf{b}^\mathsf{T}\mathbf{r} + c$ with $A = s_1\frac{\mathbf{C}_b}{p}$, $b = \frac{\mathbf{C}_b}{p}\mathbf{s}_2$, $c = -s_1\frac{\mathrm{tr}\mathbf{C}_b}{p}$. Since $\mathbf{A}$ is symmetric, there exists an orthonormal matrix $\mathbf{U}$ and a diagonal Λ such that $\mathbf{A} = \mathbf{U}^\mathsf{T}\Lambda\mathbf{U}$. We thus get

$$\mathbf{r}^\mathsf{T}\mathbf{A}\mathbf{r} + \mathbf{b}^\mathsf{T}\mathbf{r} + c = \mathbf{r}^\mathsf{T}\mathbf{U}^\mathsf{T}\Lambda\mathbf{U}\mathbf{r} + \mathbf{b}^\mathsf{T}\mathbf{U}^T\mathbf{U}\mathbf{r} + c = \tilde{\mathbf{r}}^T\Lambda\tilde{\mathbf{r}} + \tilde{\mathbf{b}}\tilde{\mathbf{r}} + c$$

with $\tilde{\mathbf{r}} = \mathbf{U}\mathbf{r}$ and $\tilde{\mathbf{b}} = \mathbf{b}\mathbf{U}^\mathsf{T}$. By unitary invariance, we have $\tilde{\mathbf{r}} \sim \mathcal{N}(0, \mathbf{I}_p)$ so that $s_1\psi_i + \mathbf{s}_2\boldsymbol{\omega}_i$ is thus the sum of the independent but not identically distributed random variables $q_j = \lambda_j[\tilde{\mathbf{r}}]_j^2 + [\tilde{\mathbf{b}}]_j[\tilde{\mathbf{r}}]_j$, $i = 1, \ldots, p$. From Lyapunov's central limit theorem [81, Theorem 27.3], it remains to find a $\delta > 0$ such that $\frac{\sum_j \mathbb{E}|q_j - \mathbb{E}[q_j]|^{2+\delta}}{\left(\sum_j \mathrm{Var}[q_j]\right)^{1+\delta/2}} \to 0$ to ensure the central limit theorem. For $\delta = 1$, we have $\mathbb{E}[q_j] = \lambda_j$, $\mathrm{Var}[q_j] = 2\lambda_j^2 + [\tilde{\mathbf{b}}]_j^2$ and $\mathbb{E}\left[(q_j - \mathbb{E}[q_j])^3\right] = 8\lambda_j^3 + 6\lambda_j[\tilde{\mathbf{b}}]_j^2$, so that $\frac{\sum_j \mathbb{E}\left[|q_j - \mathbb{E}[q_j]|^3\right]}{\left(\sum_j \mathrm{Var}[q_j]\right)^{3/2}} = O(n^{-\frac{1}{2}})$.

It thus remains to evaluate the expectation and covariance matrix of $\tilde{\mathbf{g}}_i$ conditioned on $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{[l]}}$ to obtain (i) of Theorem A.1.1. For $\mathbf{x}_i \in \mathcal{C}_b$, we have

$$\begin{aligned}
\mathbb{E}\{[\tilde{\mathbf{g}}_i]_a\} &= H_{ab} + \frac{1}{n_{[l]}}\sum_{d=1}^{K}(\gamma n_d + n_{[u]d})H_{ad} \\
&\quad + (1+\gamma)\frac{n}{n_{[l]}}\left[\Delta_a + \frac{p}{n_{[l]a}}\frac{h'(\tau)}{h(\tau)}\boldsymbol{\psi}_{[l]}^T\mathbf{j}_{[l]a} - \gamma\frac{h'(\tau)^2}{h(\tau)^2}t_a t_b\right] \\
\mathrm{Cov}\{[\tilde{\mathbf{g}}_i]_{a_1}[\tilde{\mathbf{g}}_i]_{a_2}\} &= \left(\frac{(-\gamma^2-\gamma)n - n_{[l]}}{n_{[l]}}\frac{h'(\tau)^2}{h(\tau)^2} + \frac{h''(\tau)}{h(\tau)}\right)^2 T_{bb}t_{a_1}t_{a_2} \\
&\quad + \boldsymbol{\delta}_{a_1}^{a_2}\frac{h'(\tau)^2}{h(\tau)^2}\frac{4c_0 T_{ba_1}}{c_{[l]a_1}} + \frac{4h'(\tau)^2}{h(\tau)^2}\boldsymbol{\mu}_{a_1}^\circ\mathbf{C}_b\boldsymbol{\mu}_{a_2}^\circ + o(1).
\end{aligned}$$

From the above equations, we retrieve the asymptotic expressions of $[\mathbf{m}_b]_a$ and $[\boldsymbol{\Sigma}_b]_{a_1a_2}$ given in (A.1) and (A.2). This completes the proof of Item (i) of Theorem A.1.1. Item (ii) is easily proved by following the same reasoning.

Appendix B

Supplementary material of Chapter 5

B.1 Generalization of the main theorem and proof

B.1.1 Generalized Theorem

We first present an extended version of Theorem 4.3.1 for the general setting where $\mathbf{C}_1$ may differ from $\mathbf{C}_2$.

Theorem B.1.1. *Let Assumption 5.1 hold, h be three-times continuously differentiable in a neighborhood of τ, and $\mathbf{f}_{[u]}$ be the solution of (4.4) with fixed norm $n_{[u]}e^2$. Then, for $n_{[l]}+1 \leq i \leq n$ (i.e., $\mathbf{x}_i$ unlabelled) and $\mathbf{x}_i \in \mathcal{C}_k$,*

$$f_i = g_i + o_P(1)$$

where

$$g_i \sim \mathcal{N}\left((-1)^k(1-\rho_k)m, \sigma_k\right)$$

for some $m, \sigma_k^2 > 0$. More precisely, defining

$$\theta = \frac{c_{[u]}m}{2c_{[l]}},$$

letting

$$\boldsymbol{\nu}_k = \begin{bmatrix}\sqrt{-2h'(\tau)}\boldsymbol{\mu}_k^\mathsf{T} & \sqrt{h''(\tau)}\,\mathrm{tr}\,\mathbf{C}_k/\sqrt{p}\end{bmatrix}^\mathsf{T}$$
$$\boldsymbol{\Sigma}_k = \begin{bmatrix}-2h'(\tau)\mathbf{C}_k & \mathbf{0}_{p\times 1} \\ \mathbf{0}_{1\times p} & 2h''(\tau)\,\mathrm{tr}\,{\mathbf{C}_k}^2/p\end{bmatrix}$$

and $s : (0, \|(\rho_1\boldsymbol{\Sigma}_1 + \rho_2\boldsymbol{\Sigma}_2) + \rho_1\rho_2(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\|) \to (0,+\infty)$ be the injective function

$$s(\xi) = \xi\rho_1\rho_2(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\left\{\mathbf{I}_{p+1} - \xi\left[(\rho_1\boldsymbol{\Sigma}_1+\rho_2\boldsymbol{\Sigma}_2) + \rho_1\rho_2(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\right]\right\}^{-1}(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2),$$

the values of m and σ_k^2 are determined by the equations

$$e^2 = \rho_1\rho_2 m^2 + \rho_1\sigma_1^2 + \rho_2\sigma_2^2$$
$$\frac{\sigma_k^2}{m^2} = \omega_k(\theta) + \left(\frac{\theta}{1+\theta}\right)^2\frac{q(\theta)}{\rho_1\rho_2 c_{[u]}}\left(1 + \frac{\rho_1\sigma_1^2+\rho_2\sigma_2^2}{\rho_1\rho_2 m^2}\right) + \left(\frac{1}{1+\theta}\right)^2\frac{q(\theta)}{\rho_1\rho_2 c_{[l]}}$$

where

$$q(\theta) = \frac{\operatorname{tr}\left(\mathbf{Q}(\theta)^{-1}\bar{\mathbf{\Sigma}}\right)^2}{p\left[(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\mathbf{Q}(\theta)^{-1}(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)\right]^2}$$
$$\omega_k(\theta) = \frac{(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\mathbf{Q}(\theta)^{-1}\mathbf{\Sigma}_k\mathbf{Q}(\theta)^{-1}(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)}{\left[(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)^\mathsf{T}\mathbf{Q}(\theta)^{-1}(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)\right]^2}$$

with $\bar{\mathbf{\Sigma}} = \rho_1\mathbf{\Sigma}_1 + \rho_2\mathbf{\Sigma}_2$, $\mathbf{Q}(\theta) = \mathbf{I}_{p+1} - s^{-1}(\theta)\bar{\mathbf{\Sigma}}$ *(s^{-1} being the functional inverse of s).*

B.1.2 Proof of the generalized theorem

The proof of Theorem B.1.1 relies on a leave-one-out approach, in the spirit of [6], along with arguments from previous related analyses [58, 10] based on random matrix theory .

Main idea

The main idea of the proof is to first demonstrate that for unlabelled data scores f_i (i.e., with $i > n_{[l]}$),

$$f_i = \gamma\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) + o_P(1) \tag{B.1}$$

where γ is a finite constant, ϕ_c a certain mapping from the data space that we shall define, and $\boldsymbol{\beta}^{(i)}$ a random vector independent of $\phi_c(\mathbf{x}_i)$. Additionally, we shall show that

$$\boldsymbol{\beta}^{(i)} = \frac{1}{p}\sum_{i=1}^{n} f_i\phi_c(\mathbf{x}_i) + \boldsymbol{\epsilon} \tag{B.2}$$

with $\|\boldsymbol{\epsilon}\|/\|\boldsymbol{\beta}^{(i)}\| = o_P(1)$.

As a consequence of (B.1), the statistical behavior of the unlabelled data scores can be understood through that of $\boldsymbol{\beta}^{(i)}$, which itself depends on the unlabelled data scores as described by (B.2). By combining (B.1) and (B.2), we thus establish the equations ruling the asymptotic statistical behavior (i.e., mean and variance) of the unlabelled data scores f_i.

Detailed arguments

Here the big O notation $O(u_n)$ is understood in probability. We specify that when multidimensional objects are concerned, $O(u_n)$ is understood entry-wise. The notation $O_{\|\cdot\|}$ is understood as follows: for a vector $\mathbf{v}$, $\mathbf{v} = O_{\|\cdot\|}(u_n)$ means its Euclidean norm is $O(u_n)$ and for a square matrix $\mathbf{M}$, $\mathbf{M} = O_{\|\cdot\|}(u_n)$ means that the operator norm of $\mathbf{M}$ is $O(u_n)$.

First note that, as $w_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p) = h(\tau) + O(p^{-\frac{1}{2}})$, Taylor-expanding w_{ij} around $h(\tau)$ gives (see Appendix B.3 for a detailed proof) $\hat{W} = O_{\|\cdot\|}(1)$ and

$$\hat{W} = \frac{1}{p}\hat{\mathbf{\Phi}}^\mathsf{T}\hat{\mathbf{\Phi}} + [h(0) - h(\tau) + \tau h'(\tau)]\mathbf{P}_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{B.3}$$

where $\mathbf{P}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, and $\hat{\mathbf{\Phi}} = [\hat{\phi}(\mathbf{x}_1), \ldots, \hat{\phi}(\mathbf{x}_n)] = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]\mathbf{P}_n$ with

$$\phi(\mathbf{x}_i) = \begin{bmatrix}\sqrt{-2h'(\tau)}\mathbf{x}_i^\mathsf{T} & \sqrt{h''(\tau)}\|\mathbf{x}_i\|^2/\sqrt{p}\end{bmatrix}^\mathsf{T}.$$

Define $\boldsymbol{\nu}_k = \mathbb{E}\{\phi(\mathbf{x}_i)\}$, $\mathbf{\Sigma}_k = \mathrm{cov}\{\phi(\mathbf{x}_i)\}$ for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1, 2\}$, and let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ with $\mathbf{z}_i = \phi(\mathbf{x}_i) - \boldsymbol{\nu}_k$ (i.e., $\mathbb{E}\{\mathbf{z}_i\} = 0$). We also write the labelled versus unlabelled divisions $\mathbf{\Phi} = \begin{bmatrix}\mathbf{\Phi}_{[l]} & \mathbf{\Phi}_{[u]}\end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_{[l]} & \mathbf{Z}_{[u]}\end{bmatrix}$ and $\hat{\mathbf{\Phi}} = \begin{bmatrix}\hat{\mathbf{\Phi}}_{[l]} & \hat{\mathbf{\Phi}}_{[u]}\end{bmatrix}$.

Recall that $\mathbf{f}_{[u]} = \left(\alpha\mathbf{I}_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1}\hat{W}_{[ul]}\mathbf{f}_{[l]}$. To proceed, we need to show that $\frac{1}{n}\mathbf{1}_{n_{[u]}}^\mathsf{T}\mathbf{f}_{[u]} = O(p^{-\frac{1}{2}})$. This follows from (B.3) and the results in [58]. Specifically, applying (B.3), we can express $\mathbf{f}_{[u]}$ as

$$\mathbf{f}_{[u]} = \left(\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\hat{\mathbf{\Phi}}_{[u]} + \frac{r}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}}^\mathsf{T}\right)^{-1}\left(\frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\hat{\mathbf{\Phi}}_{[l]} - \frac{r}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[l]}}^\mathsf{T}\right)\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}})$$

where $\tilde{\alpha} = \alpha - h(0) + h(\tau) - \tau h'(\tau)$, $r = h(0) - h(\tau) + \tau h'(\tau)$. Since $\mathbf{1}_{[l]}^\mathsf{T}\mathbf{f}_{[l]} = 0$ from its definition given in (4.1),

$$\mathbf{f}_{[u]} = \left(\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\hat{\mathbf{\Phi}}_{[u]} + \frac{r}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}}^\mathsf{T}\right)^{-1}\frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\mathbf{\Phi}_{[l]}\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}}). \tag{B.4}$$

Write $\hat{\mathbf{\Phi}}_{[u]} = \mathbb{E}\{\hat{\mathbf{\Phi}}_{[u]}\} + \mathbf{Z}_{[u]} - (\mathbf{Z}\mathbf{1}_n/n)\mathbf{1}_{n_{[u]}}^\mathsf{T}$. Evidently, $\mathbb{E}\{\hat{\mathbf{\Phi}}_{[u]}\} = (\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2)s^\mathsf{T}$ where $s \in \mathbb{R}^{n_{[u]}}$ with $s_i = (-1)^k(n - n_k)/n$ for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1, 2\}$. By the large number law, $s = e + O(p^{-\frac{1}{2}})$ where $e \in \mathbb{R}^{n_{[u]}}$ with $e_i = (-1)^k(1 - \rho_k)$ for $\mathbf{x}_i \in \mathcal{C}_k$, therefore

$$\begin{aligned}\frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\hat{\mathbf{\Phi}}_{[u]} =& \frac{1}{p}\Big\{\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|^2 ee^\mathsf{T} + \mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}_{[u]} + (\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbf{1}_n/n^2)\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}}^\mathsf{T} + [\mathbf{Z}_{[u]}^\mathsf{T}(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2)]e^\mathsf{T} \\ &+ e[\mathbf{Z}_{[u]}^\mathsf{T}(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2)]^\mathsf{T} - (\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}\mathbf{1}_n/n)\mathbf{1}_{n_{[u]}}^\mathsf{T} - \mathbf{1}_{n_{[u]}}(\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}\mathbf{1}_n/n)^\mathsf{T}\Big\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned}$$

Set

$$\mathbf{U} = \frac{1}{\sqrt{p}}\begin{bmatrix}e & \mathbf{Z}_{[u]}^\mathsf{T}(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2) & \mathbf{1}_{n_{[u]}} & \mathbf{Z}_{[u]}^\mathsf{T}Z\mathbf{1}_n/n\end{bmatrix}$$

$$\mathbf{N} = \begin{bmatrix}\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|^2 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & (\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbf{1}_n/n^2) - \frac{r}{c_0} & -1 \\ 0 & 0 & -1 & 0\end{bmatrix}.$$

Invoking Woodbury's identity[82], we get

$$\begin{aligned}\left(\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\hat{\mathbf{\Phi}}_{[u]} + \frac{r}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}}^\mathsf{T}\right)^{-1} &= \left(\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}_{[u]} - \mathbf{U}\mathbf{N}\mathbf{U}^\mathsf{T}\right)^{-1} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \\ &= \mathbf{R} + \mathbf{R}\mathbf{U}(\mathbf{N}^{-1} - \mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U})^{-1}U^\mathsf{T}\mathbf{R} + O_{\|\cdot\|}(p^{-\frac{1}{2}})\end{aligned} \tag{B.5}$$

where $\mathbf{R} = \left(\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}_{[u]}\right)^{-1}$. Note also that

$$\frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^\mathsf{T}\mathbf{\Phi}_{[l]}\mathbf{f}_{[l]} = \sqrt{p}\mathbf{U}\begin{bmatrix}(\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^\mathsf{T}\frac{1}{p}\mathbf{\Phi}_{[l]}\mathbf{f}_{[l]} \\ 2c_{[l]}\rho_1\rho_2 \\ 0 \\ 0\end{bmatrix} + \frac{1}{p}Z_{[u]}^\mathsf{T}Z_{[l]}\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}}). \tag{B.6}$$

Similarly to the results of [58, Equation 7.6], $\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U}$ is of the form

$$\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U} = \begin{bmatrix} \mathbf{A} & 0_{2\times 2} \\ 0_{2\times 2} & \mathbf{B} \end{bmatrix} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{B.7}$$

for some matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{2\times 2}$ of $O(1)$-operator norm and $p^{-\frac{3}{2}}\|\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}_{[l]}\mathbf{f}_{[l]}\| = O(p^{-\frac{1}{2}})$. Substituting (B.5) and (B.6) into (B.4) and using the fact that $p^{-\frac{3}{2}}\|\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{Z}_{[u]}^\mathsf{T}\mathbf{Z}_{[l]}\mathbf{f}_{[l]}\| = O(p^{-\frac{1}{2}})$ allows us to obtain

$$\frac{1}{n}\mathbf{1}_{n_{[u]}}^\mathsf{T}\mathbf{f}_{[u]} = c_0^{-1}\begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix} K \begin{bmatrix} (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^\mathsf{T}\frac{1}{p}\boldsymbol{\Phi}_{[l]}\mathbf{f}_{[l]} \\ 2c_{[l]}\rho_1\rho_2 \\ 0 \\ 0 \end{bmatrix} + O(p^{-\frac{1}{2}}) \tag{B.8}$$

with

$$\mathbf{K} = \mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U} + \mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U}(\mathbf{N}^{-1} - \mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U}.$$

Since $\mathbf{U}^\mathsf{T}\mathbf{R}\mathbf{U}$ is of the form (B.7), we find from classical algebraic arguments that $\mathbf{K}$ is also of the same diagonal block matrix form. We thus finally get from (B.8) that $\frac{1}{n}\mathbf{1}_{n_{[u]}}^\mathsf{T}\mathbf{f}_{[u]} = O(p^{-\frac{1}{2}})$.

Now that we have shown that $\frac{1}{n}\mathbf{1}_{n_{[u]}}^\mathsf{T}\mathbf{f}_{[u]} = O(p^{-\frac{1}{2}})$, multiplying both sides of (B.4) with $\tilde{\alpha}\mathbf{I}_{n_{[u]}} - \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^\mathsf{T}\hat{\boldsymbol{\Phi}}_{[u]} + \frac{r}{n}\mathbf{1}_{n_{[u]}}\mathbf{1}_{n_{[u]}}^\mathsf{T}$ from the left gives

$$\tilde{\alpha}\mathbf{f}_{[u]} = \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^\mathsf{T}\hat{\boldsymbol{\Phi}}_{[u]}\mathbf{f}_{[u]} + \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^\mathsf{T}\hat{\boldsymbol{\Phi}}_{[l]}\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}}).$$

Decomposing this equation for any $i > n_{[l]}$ (i.e., $\mathbf{x}_i$ unlabelled) leads to

$$\tilde{\alpha}f_i = \frac{1}{p}\hat{\phi}(\mathbf{x}_i)^\mathsf{T}\hat{\boldsymbol{\Phi}}f + O(p^{-\frac{1}{2}}) \tag{B.9}$$

$$\tilde{\alpha}\mathbf{f}_{[u]}^{\{i\}} = \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^{\{i\}\mathsf{T}}\hat{\phi}(\mathbf{x}_i)f_i + \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^{\{i\}\mathsf{T}}\hat{\boldsymbol{\Phi}}_{[u]}^{\{i\}}\mathbf{f}_{[u]}^{\{i\}} + \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^{\{i\}\mathsf{T}}\hat{\boldsymbol{\Phi}}_{[l]}\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}}) \tag{B.10}$$

with $\mathbf{f}_{[u]}^{\{i\}}$ standing for the vector obtained by removing f_i from $\mathbf{f}_{[u]}$, $\hat{\boldsymbol{\Phi}}_{[u]}^{\{i\}}$ for the matrix obtained by removing $\hat{\phi}(\mathbf{x}_i)$ from $\hat{\boldsymbol{\Phi}}_{[u]}$.

Our objective is to compare the behavior of the vector $\mathbf{f}_{[u]}$ decomposed as $\{f_i, \mathbf{f}_{[u]}^{\{i\}}\}$ to the "leave-$\mathbf{x}_i$-out" version $\mathbf{f}_{[u]}^{(i)}$ to be introduced next. To this end, define the leave-one-out dataset $X^{(i)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{(n-1)\times p}$ for any $i > n_{[l]}$ (i.e., $\mathbf{x}_i$ unlabelled), and $\hat{\mathbf{W}}^{(i)} \in \mathbb{R}^{(n-1)\times(n-1)}$ the corresponding centered similarity matrix, for which we have, similarly to $\hat{\mathbf{W}}$,

$$\hat{\mathbf{W}}^{(i)} = \frac{1}{p}\hat{\boldsymbol{\Phi}}^{(i)\mathsf{T}}\hat{\boldsymbol{\Phi}}^{(i)} + [h(0) - h(\tau) + \tau h'(\tau)]\mathbf{P}_{n-1} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{B.11}$$

where $\hat{\boldsymbol{\Phi}}^{(i)} = [\hat{\phi}^{(i)}(\mathbf{x}_1), \ldots, \hat{\phi}^{(i)}(\mathbf{x}_{i-1}), \hat{\phi}^{(i)}(\mathbf{x}_{i+1}), \ldots, \hat{\phi}^{(i)}(\mathbf{x}_n)] = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_{i-1}), \phi(\mathbf{x}_{i+1}), \ldots, \phi(\mathbf{x}_n)]\mathbf{P}_{n-1}$. Denote by $\mathbf{f}_{[u]}^{(i)}$ the solution of the centered similarities regularization on the "leave-one-out" dataset $X_{(i)}$, i.e.,

$$\mathbf{f}_{[u]}^{(i)} = \left(\alpha\mathbf{I}_{n_{[u]}-1} - \hat{\mathbf{W}}_{[uu]}^{(i)}\right)^{-1}\hat{\mathbf{W}}_{[ul]}^{(i)}\mathbf{f}_{[l]}. \tag{B.12}$$

Substituting (B.11) into (B.12) leads to

$$\tilde{\alpha}\mathbf{f}_{[u]}^{(i)} = \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\hat{\mathbf{\Phi}}_{[u]}^{(i)}\mathbf{f}_{[u]}^{(i)} + \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\hat{\mathbf{\Phi}}_{[l]}\mathbf{f}_{[l]} + O(p^{-\frac{1}{2}}) \tag{B.13}$$

where $\hat{\mathbf{\Phi}}^{(i)} = \begin{bmatrix}\hat{\mathbf{\Phi}}_{[l]}^{(i)} & \hat{\mathbf{\Phi}}_{[u]}^{(i)}\end{bmatrix}$. From the definitions of $\hat{\mathbf{\Phi}}_{[u]}^{(i)}$ and $\hat{\mathbf{\Phi}}_{[u]}^{\{i\}}$, which essentially differ by the addition of the $O(1/p)$-norm term $\phi(\mathbf{x}_i)/n$ to every column, we easily have

$$\frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\hat{\mathbf{\Phi}}_{[u]}^{(i)} - \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{\{i\}\mathsf{T}}\hat{\mathbf{\Phi}}_{[u]}^{\{i\}} = O_{\|\cdot\|}(p^{-1}). \tag{B.14}$$

Thus, subtracting (B.13) from (B.10) gives

$$M^{(i)}\left(\mathbf{f}_{[u]}^{\{i\}} - \mathbf{f}_{[u]}^{(i)}\right) = \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\hat{\phi}(\mathbf{x}_i)\mathbf{f}_i + O(p^{-\frac{1}{2}}) \tag{B.15}$$

with

$$M^{(i)} = \tilde{\alpha}\mathbf{I}_{(n_{[u]}-1)} - \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\hat{\mathbf{\Phi}}_{[u]}^{(i)}.$$

Set $\boldsymbol{\beta} = \frac{1}{p}\hat{\mathbf{\Phi}}f = O_{\|\cdot\|}(1)$, the unlabelled data "regression vector", and its "leave-one-out" version $\boldsymbol{\beta}^{(i)} = \frac{1}{p}\hat{\mathbf{\Phi}}^{(i)}\mathbf{F}_{(i)}$ with $\mathbf{F}_{(i)} = \begin{bmatrix}\mathbf{f}_{[l]} & \mathbf{f}_{[u]}^{(i)}\end{bmatrix}$. Applying (B.14) and (B.15), we get that

$$\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)} = \left(\mathbf{I}_p + \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)}\left(M^{(i)}\right)^{-1}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\right)\frac{1}{p}f_i\hat{\phi}(\mathbf{x}_i) + O_{\|\cdot\|}(p^{-1}) = O_{\|\cdot\|}(p^{-\frac{1}{2}}). \tag{B.16}$$

By the above result, Equation (B.9) can be expanded as

$$\tilde{\alpha}f_i = \boldsymbol{\beta}^{(i)\mathsf{T}}\hat{\phi}(\mathbf{x}_i) + \frac{1}{p}\hat{\phi}(\mathbf{x}_i)^{\mathsf{T}}\left(\mathbf{I}_p + \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)}\left(M^{(i)}\right)^{-1}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}\right)\hat{\phi}(\mathbf{x}_i)f_i + O(p^{-\frac{1}{2}}). \tag{B.17}$$

To go further in the development of (B.17), we first need to evaluate the quadratic form

$$\kappa_i \equiv \frac{1}{p}\hat{\phi}(\mathbf{x}_i)^{\mathsf{T}}T^{(i)}\hat{\phi}(\mathbf{x}_i)$$

where

$$T^{(i)} = \mathbf{I}_p + \frac{1}{p}\hat{\mathbf{\Phi}}_{[u]}^{(i)}\left(M^{(i)}\right)^{-1}\hat{\mathbf{\Phi}}_{[u]}^{(i)\mathsf{T}}.$$

Since $T^{(i)} = O_{\|\cdot\|}(1)$ [27, Theorem 7.1] and $\hat{\phi}(\mathbf{x}_i)$ is independent of $T^{(i)}$, it unfolds from the "trace lemma" [27, Lemma 14.2] that $\kappa_i = O(1)$ and that κ_i converges almost surely to a deterministic limit κ independent of i at large n, p. Equation (B.17) then becomes

$$f_i = \gamma\boldsymbol{\beta}^{(i)\mathsf{T}}\hat{\phi}(\mathbf{x}_i) + O(p^{-\frac{1}{2}}). \tag{B.18}$$

where $\gamma = (\tilde{\alpha} - \kappa)^{-1}$.

We focus now on the term $\boldsymbol{\beta}^{(i)\mathsf{T}}\hat{\phi}(\mathbf{x}_i)$ in (B.18). To discard the "weak" dependence between $\boldsymbol{\beta}^{(i)\mathsf{T}}$ and $\hat{\phi}(\mathbf{x}_i)$, let us define

$$\phi_c(\mathbf{x}_i) = (-1)^k(1-\rho_k)(\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) + \mathbf{z}_i.$$

As $n_k/n = \rho_k + O(n^{-\frac{1}{2}})$, by the law of large numbers, $\mathbb{E}\{\hat{\phi}(\mathbf{x}_i)\} = (-1)^k[(n-n_k)/n](\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) = \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$. Remark that, unlike $\hat{\phi}(\mathbf{x}_i)$, $\phi_c(\mathbf{x}_i)$ is independent of all $\mathbf{x}_j$ with $j \neq i$, and therefore independent of $\boldsymbol{\beta}^{(i)}$. We thus now have

$$\boldsymbol{\beta}^{(i)\mathsf{T}}\hat{\phi}(\mathbf{x}_i) = \boldsymbol{\beta}^{(i)\mathsf{T}}\left(\mathbb{E}\{\hat{\phi}(\mathbf{x}_i)\} + \mathbf{z}_i - \frac{1}{n}\sum_{m=1}^{n}\mathbf{z}_m\right) = \boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) + \frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n + O(p^{-\frac{1}{2}}).$$

We get from (B.16) that $\frac{1}{n}\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{Z}\mathbf{1}_n = \frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n + O(p^{-\frac{1}{2}})$, leading to

$$f_i = \gamma\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) + \frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n + O(p^{-\frac{1}{2}}). \tag{B.19}$$

Since $\phi_c(\mathbf{x}_i)$ is independent of $\boldsymbol{\beta}^{(i)}$, according to the central limit theorem, $\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)$ asymptotically follows a Gaussian distribution.

To demonstrate that $\frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n$ is negligibly small, notice fist that, by summing (B.19) for all $i > n_{[u]}$, we have

$$\frac{1}{n}\mathbf{1}_{n_{[u]}}^{\mathsf{T}}\mathbf{f}_{[u]} = \frac{1}{n}\sum_{i=n_{[l]}+1}^{n}\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) + c_{[u]}(\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{Z}\mathbf{1}_n/n) + O(p^{-\frac{1}{2}}).$$

Since $\frac{1}{n}\mathbf{1}_{n_{[u]}}^{\mathsf{T}}\mathbf{f}_{[u]} = O(p^{-\frac{1}{2}})$, it suffices to prove $\frac{1}{n}\sum_{i=n_{[l]}+1}^{n}\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) = O(p^{-\frac{1}{2}})$ to consequently show that $\frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n = O(p^{-\frac{1}{2}})$ from the above equation. To this end, we shall examine the correlation between $\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)$ and $\boldsymbol{\beta}^{(j)\mathsf{T}}\phi_c(\mathbf{x}_j)$ for $i \neq j > n_{[l]}$. Consider $\boldsymbol{\beta}^{(ij)}, \hat{\boldsymbol{\Phi}}_{[u]}^{(ij)}, M^{(ij)}$ obtained in the same way as $\boldsymbol{\beta}^{(i)}, \hat{\boldsymbol{\Phi}}_{[u]}^{(i)}, M^{(i)}$, but this time by leaving out the two unlabelled samples $\mathbf{x}_i, \mathbf{x}_j$. Similarly to (B.16), we have

$$\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(ij)} = \left(\mathbf{I}_p + \frac{1}{p}\hat{\boldsymbol{\Phi}}_{[u]}^{(ij)}\left(M^{(ij)}\right)^{-1}\hat{\boldsymbol{\Phi}}_{[u]}^{(ij)\mathsf{T}}\right)\frac{1}{p}f_j\hat{\phi}(\mathbf{x}_j) + O_{\|\cdot\|}(p^{-1}) = O_{\|\cdot\|}(p^{-\frac{1}{2}}). \tag{B.20}$$

It follows from the above equation that, for $i \neq j > n_{[l]}$,

$$\begin{aligned}
&\mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_j)\} - \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\}\mathbb{E}\{\boldsymbol{\beta}^{(j)\mathsf{T}}\phi_c(\mathbf{x}_j)\} \\
&= \mathbb{E}\{\boldsymbol{\beta}^{(ij)\mathsf{T}}\phi_c(\mathbf{x}_i)\boldsymbol{\beta}^{(ij)\mathsf{T}}\phi_c(\mathbf{x}_j)\} - \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\}\mathbb{E}\{\boldsymbol{\beta}^{(j)\mathsf{T}}\phi_c(\mathbf{x}_j)\} + O(p^{-1}) \\
&= \mathbb{E}\{\boldsymbol{\beta}^{(ij)\mathsf{T}}\phi_c(\mathbf{x}_i)\}\mathbb{E}\{\boldsymbol{\beta}^{(ij)\mathsf{T}}\phi_c(\mathbf{x}_j)\} - \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\}\mathbb{E}\{\boldsymbol{\beta}^{(j)\mathsf{T}}\phi_c(\mathbf{x}_j)\} + O(p^{-1}) \\
&= O(p^{-1}),
\end{aligned} \tag{B.21}$$

leading to the conclusion that $\frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n}\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) = \frac{1}{n_{[u]}}\sum_{i=n_{[l]}+1}^{n}\mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\} + O(p^{-\frac{1}{2}}) = O(p^{-\frac{1}{2}})$. Hence, $\frac{1}{n}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\mathbf{1}_n = O(p^{-\frac{1}{2}})$. Finally, we have that, for $i > n_{[l]}$,

$$f_i = \gamma\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i) + O(p^{-\frac{1}{2}}), \tag{B.22}$$

indicating that, up to the constant γ, f_i asymptotically follows the same Gaussian distribution as $\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)$.

Moreover, taking the expectation and the variance of the both sides of (B.22) for $\mathbf{x}_i \in \mathcal{C}_k$ yields

$$\begin{aligned}\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} &= \gamma \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\}(-1)^k(1-\rho_k)(\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) + O(p^{-\frac{1}{2}}) \\ \mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} &= \gamma^2 \mathrm{tr}\big[\mathrm{cov}\{\boldsymbol{\beta}^{(i)}\}\boldsymbol{\Sigma}_k\big] + \gamma^2 \mathbb{E}\{\boldsymbol{\beta}^{(i)}\}^\mathsf{T}\boldsymbol{\Sigma}_k \mathbb{E}\{\boldsymbol{\beta}^{(i)}\} + O(p^{-\frac{1}{2}}).\end{aligned}$$

Since $\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$ as per (B.16), we obtain

$$\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = \gamma \mathbb{E}\{\boldsymbol{\beta}^\mathsf{T}\}(-1)^k(1-\rho_k)(\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) + O(p^{-\frac{1}{2}}) \tag{B.23}$$

$$\mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = \gamma^2 \mathrm{tr}\big[\mathrm{cov}\{\boldsymbol{\beta}\}\boldsymbol{\Sigma}_k\big] + \gamma^2 \mathbb{E}\{\boldsymbol{\beta}\}^\mathsf{T}\boldsymbol{\Sigma}_k \mathbb{E}\{\boldsymbol{\beta}\} + O(p^{-\frac{1}{2}}). \tag{B.24}$$

After linking the distribution parameters of unlabelled scores to those of $\boldsymbol{\beta}$ with Equation (B.23) and Equation (B.24), we now turn our attention to the statistical behaviour of $\boldsymbol{\beta}$. Substituting (B.22) into $\boldsymbol{\beta} = \frac{1}{p}\hat{\boldsymbol{\Phi}}\mathbf{f}$ yields

$$\begin{aligned}\boldsymbol{\beta} &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \hat{\phi}(\mathbf{x}_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \boldsymbol{\beta}^{(i)\mathsf{T}} \phi_c(\mathbf{x}_i)\hat{\phi}(\mathbf{x}_i) + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \\ &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \phi_c(\mathbf{x}_i) + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \boldsymbol{\beta}^{(i)\mathsf{T}} \phi_c(\mathbf{x}_i)\phi_c(\mathbf{x}_i) + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned} \tag{B.25}$$

For $i > n_{[l]}$ and $\mathbf{x}_i \in \mathcal{C}_k$, we decompose $\phi_c(\mathbf{x}_i)$ as

$$\phi_c(\mathbf{x}_i) = \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + \frac{\boldsymbol{\Sigma}_k \boldsymbol{\beta}^{(i)}}{\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{z}_i} + \tilde{\mathbf{z}}_i \tag{B.26}$$

where

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i - \frac{\boldsymbol{\Sigma}_k \boldsymbol{\beta}^{(i)}}{\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{z}_i}.$$

By substituting the expression (B.26) of $\phi_c(\mathbf{x}_i)$ into (B.25) and using the fact that $\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$, we obtain

$$\begin{aligned}\left(\mathbf{I}_p - \gamma c_{[u]} \sum_{a=1}^{2} \rho_a \boldsymbol{\Sigma}_a\right)\boldsymbol{\beta} &= \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \boldsymbol{\beta}^{(i)\mathsf{T}} \phi_c(\mathbf{x}_i)\mathbb{E}\{\phi_c(\mathbf{x}_i)\} \\ &\quad + \frac{1}{p}\sum_{i=1}^{n_{[l]}} f_i \mathbf{z}_i + \frac{1}{p}\sum_{i=n_{[l]}+1}^{n} \gamma \boldsymbol{\beta}^{(i)\mathsf{T}} \phi_c(\mathbf{x}_i)\tilde{\mathbf{z}}_i + O_{\|\cdot\|}(p^{-\frac{1}{2}}).\end{aligned} \tag{B.27}$$

Recall that $\mathbf{f}_{[l]}$ is a deterministic vector (given in (4.1)) and note that

$$\mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\phi_c(\mathbf{x}_i)\tilde{\mathbf{z}}_i\} = \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{z}_i[\mathbf{z}_i - \boldsymbol{\Sigma}_k\boldsymbol{\beta}^{(i)}/(\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{z}_i)]\} = \mathbb{E}\{\boldsymbol{\beta}^{(i)\mathsf{T}}\mathbf{z}_i\mathbf{z}_i\} - \boldsymbol{\Sigma}_k\mathbb{E}\{\boldsymbol{\beta}^{(i)}\} = 0.$$

Taking the expectation of both sides of (B.27) thus gives

$$\begin{aligned}
&\Big(\mathbf{I}_p - \gamma c_{[u]} \sum_{a=1}^{2} \rho_a \mathbf{\Sigma}_a \Big) \mathbb{E}\{\boldsymbol{\beta}\} \\
&= \frac{1}{p} \sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + \frac{1}{p} \sum_{i=n_{[l]}+1}^{n} \gamma \mathbb{E}\{\boldsymbol{\beta}^{(i)}\}^{\mathsf{T}} \mathbb{E}\{\phi_c(\mathbf{x}_i)\} \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \\
&= \frac{1}{p} \sum_{i=1}^{n_{[l]}} f_i \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + \frac{1}{p} \sum_{i=n_{[l]}+1}^{n} \gamma \mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}} \mathbb{E}\{\phi_c(\mathbf{x}_i)\} \mathbb{E}\{\phi_c(\mathbf{x}_i)\} + O_{\|\cdot\|}(p^{-\frac{1}{2}}). \qquad \text{(B.28)}
\end{aligned}$$

Let $\mathbf{Q} = \mathbf{I}_p - \gamma c_{[u]} \bar{\mathbf{\Sigma}}$ with $\bar{\mathbf{\Sigma}} = \rho_1 \mathbf{\Sigma}_1 + \rho_2 \mathbf{\Sigma}_2$ and denote $m \equiv \gamma (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^{\mathsf{T}} \mathbb{E}\{\boldsymbol{\beta}\}$. With these notations, we get directly from the above equation that

$$m = \gamma \rho_1 \rho_2 (2c_{[l]} + m c_{[u]}) (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^{\mathsf{T}} \mathbf{Q}^{-1} (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) + o_P(1). \qquad \text{(B.29)}$$

With the notation m, (B.23) notably becomes

$$\mathbb{E}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\} = (-1)^k (1 - \rho_k) m + O(p^{-\frac{1}{2}}).$$

In addition, we get from (B.28) that

$$\gamma^2 \mathbb{E}\{\boldsymbol{\beta}\}^{\mathsf{T}} \mathbf{\Sigma}_k \mathbb{E}\{\boldsymbol{\beta}\} = \big[\gamma \rho_1 \rho_2 (2c_{[l]} + m c_{[u]})\big]^2 (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1). \qquad \text{(B.30)}$$

Furthermore, we have from (B.27) and (B.28)

$$\begin{aligned}
\mathrm{tr}[\mathrm{cov}\{\boldsymbol{\beta}\} \mathbf{\Sigma}_k] &= \mathbb{E}\left\{ (\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\})^{\mathsf{T}} \mathbf{\Sigma}_k (\boldsymbol{\beta} - \mathbb{E}\{\boldsymbol{\beta}\}) \right\} \\
&= \frac{1}{p^2} \sum_{i=1}^{n_{[l]}} f_i^2 \mathbb{E}\{\mathbf{z}_i^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} \mathbf{z}_i\} + \frac{1}{p^2} \sum_{i=n_{[l]}+1}^{n} \gamma^2 \mathbb{E}\{(\boldsymbol{\beta}^{(i)\mathsf{T}} \phi_c(\mathbf{x}_i))^2 \tilde{\mathbf{z}}_i^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} \tilde{\mathbf{z}}_i\} \\
&\quad + O(p^{-\frac{1}{2}}).
\end{aligned}$$

Since $\frac{1}{p} \mathbf{z}_i^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} \mathbf{z}_i = \frac{1}{p} \mathrm{tr}(\mathbf{Q}^{-1} \bar{\mathbf{\Sigma}})^2 + O(p^{-\frac{1}{2}})$ and $\frac{1}{p} \tilde{\mathbf{z}}_i^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} \tilde{\mathbf{z}}_i = \frac{1}{p} \mathrm{tr}(\mathbf{Q}^{-1} \bar{\mathbf{\Sigma}})^2 + O(p^{-\frac{1}{2}})$, by the trace lemma [27, Lemma 14.2] and Assumption 5.1,

$$\begin{aligned}
\gamma^2 \mathrm{tr}[\mathrm{cov}\{\boldsymbol{\beta}\} \mathbf{\Sigma}_k] =& \gamma^2 \big[\rho_1 \rho_2 (4c_{[l]} + m^2 c_{[u]}) + c_{[u]} \sum_{a=1}^{2} \rho_a \mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_a\}\big] \frac{1}{p} \mathrm{tr}(\mathbf{Q}^{-1} \bar{\mathbf{\Sigma}})^2 \\
&+ O(p^{-\frac{1}{2}}). \qquad \text{(B.31)}
\end{aligned}$$

Using the shortcut notation $\sigma_k^2 \equiv \mathrm{var}\{f_i | i > n_{[l]}, x \in \mathcal{C}_k\}$ for $k \in \{1, 2\}$, we get by substituting (B.30) and (B.31) into (B.24) that

$$\begin{aligned}
\sigma_k^2 =& \big[\rho_1 \rho_2 (2c_{[l]} + m c_{[u]})\big]^2 (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1)^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{\Sigma}_k \mathbf{Q}^{-1} (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) \\
&+ \gamma^2 \big[\rho_1 \rho_2 (4c_{[l]} + m^2 c_{[u]}) + c_{[u]} \sum_{a=1}^{2} \rho_a \boldsymbol{\beta}_a^2\big] \frac{1}{p} \mathrm{tr}(\mathbf{Q}^{-1} \bar{\mathbf{\Sigma}})^2 + o_P(1). \qquad \text{(B.32)}
\end{aligned}$$

Additionally, letting $\xi \equiv c_{[u]}\gamma$, we get from (B.28)

$$\mathbb{E}\{\boldsymbol{\beta}\} = 2c_{[l]}\rho_1\rho_2 \left[\mathbf{I}_p - \xi\left(\sum_{a=1}^{2}\rho_a\boldsymbol{\Sigma}_a + \rho_1\rho_2(\boldsymbol{\nu}_2-\boldsymbol{\nu}_1)(\boldsymbol{\nu}_2-\boldsymbol{\nu}_1)^\mathsf{T}\right)\right](\boldsymbol{\nu}_2-\boldsymbol{\nu}_1) + O(p^{-\frac{1}{2}}),$$

leading directly to

$$\theta = \xi\rho_1\rho_2(\boldsymbol{\nu}_2-\boldsymbol{\nu}_1)^\mathsf{T}\left[\mathbf{I}_p - \xi\left(\sum_{a=1}^{2}\rho_a\boldsymbol{\Sigma}_a + \rho_1\rho_2(\boldsymbol{\nu}_2-\boldsymbol{\nu}_1)(\boldsymbol{\nu}_2-\boldsymbol{\nu}_1)^\mathsf{T}\right)\right](\boldsymbol{\nu}_2-\boldsymbol{\nu}_1) + o_P(1) \quad \text{(B.33)}$$

where $\theta = c_{[u]}m/2c_{[l]}$.

Finally, the equations of Theorem B.1.1 are retrieved by gathering (B.29), (B.32) and (B.33) and by ignoring the vanishing terms. This completes the proof.

B.2 Proof of Proposition 4.3.1

As the eigenvector of $\mathbf{L}_s$ associated with the smallest eigenvalue is $\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n$, we consider

$$\mathbf{L}'_s = n\mathbf{D}^{-\frac{1}{2}}W\mathbf{D}^{-\frac{1}{2}} - n\frac{\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{D}^{\frac{1}{2}}}{\mathbf{1}_n^\mathsf{T}\mathbf{D}\mathbf{1}_n}.$$

Note that $\|\mathbf{L}'_s\| = O(1)$ as demonstrated by [58], and if $\mathbf{v}$ is an eigenvector of $\mathbf{L}_s$ associated with the eigenvalue u, then it is also an eigenvector of $\mathbf{L}'_s$ associated with the eigenvalue $-u+1$, except for the eigenvalue-eigenvector pair $(n, \mathbf{D}^{\frac{1}{2}}\mathbf{1}_n)$ of $\mathbf{L}_s$ turned into $(0, \mathbf{D}^{\frac{1}{2}}\mathbf{1}_n)$ for $\mathbf{L}'_s$. The second smallest eigenvector $\mathbf{v}_{\text{lap}}$ of $\mathbf{L}_s$ is the same as the largest eigenvector of $\mathbf{L}'_s$.

From the random matrix equivalent of $\mathbf{L}'_s$ given by [58, Theorem 1] and that of $\hat{\mathbf{W}}$ expressed in (B.3), we have

$$\hat{\mathbf{W}} = h(\tau)\mathbf{L}'_s + \frac{5h'(\tau)^2}{4}\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T} + O(p^{-\frac{1}{2}})$$

where $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_n]^\mathsf{T}$ with $\psi_i = \|\mathbf{x}_i\|^2 - \mathbb{E}[\|\mathbf{x}_i\|^2]$.

For $k \in \{1,2\}$, define $\mathbf{j}_k \in \mathbb{R}^n$ the indicator vector of class k with $[\mathbf{j}_k]_i = 1$ if $\mathbf{x}_i \in \mathcal{C}_k$, otherwise $[\mathbf{j}_k]_i = 0$. Then, we have

$$\begin{aligned} d_{\text{inter}}(\mathbf{v}) &= |\mathbf{j}_1^\mathsf{T}\mathbf{v}/n_1 - \mathbf{j}_2^\mathsf{T}\mathbf{v}/n_2| \\ d_{\text{intra}}(\mathbf{v}) &= \|\mathbf{v} - (\mathbf{j}_1^\mathsf{T}\mathbf{v}/n_1)\mathbf{j}_1 - (\mathbf{j}_2^\mathsf{T}\mathbf{v}/n_2)\mathbf{j}_2\|/\sqrt{n} \end{aligned}$$

for some $\mathbf{v} \in \mathbb{R}^n$.

Denote by λ_{lap} the eigenvalue of $h(\tau)\mathbf{L}'_s$ associated with $\mathbf{v}_{\text{lap}}$, and λ_{ctr} the eigenvalue of $\hat{\mathbf{W}}$ associated with $\mathbf{v}_{\text{ctr}}$. Under the condition of non-trivial clustering upon $\mathbf{v}_{\text{lap}}$ with $d_{\text{inter}}(\mathbf{v}_{\text{lap}})/d_{\text{intra}}(\mathbf{v}_{\text{lap}}) = O(1)$, we have $\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{lap}}/\sqrt{n_k} = O(1)$ from the above expressions of $d_{\text{inter}}(\mathbf{v})$ and $d_{\text{intra}}(\mathbf{v})$. The fact that $\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{lap}}/\sqrt{n_k} = O(1)$ implies that the eigenvalue λ_{lap} of $h(\tau)\mathbf{L}'_s$ remains at a non vanishing distance from other eigenvalues of $h(\tau)\mathbf{L}'_s$ [58]. The same can be said about $\hat{\mathbf{W}}$ and its eigenvalue λ_{ctr}.

Let γ be a positively oriented complex closed path circling only around λ_{lap} and λ_{ctr}. Since there can be only one eigenvector of $\mathbf{L}'_s$ ($\hat{\mathbf{W}}$, resp.) whose limiting scalar product with $\mathbf{j}_k$ for $k \in \{1,2\}$ is bounded away from zero [58, Theorem 4], which is $\mathbf{v}_{\text{lap}}$ ($resp.$, $\mathbf{v}_{\text{ctr}}$), we have, by Cauchy's formula [83, Theorem 10.15],

$$\frac{1}{n_k}(\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{lap}})^2 = -\frac{1}{2\pi i}\oint_\gamma \frac{1}{n_k}\mathbf{j}_k^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\mathbf{j}_k dz + o_P(1)$$
$$\frac{1}{n_k}(\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{ctr}})^2 = -\frac{1}{2\pi i}\oint_\gamma \frac{1}{n_k}\mathbf{j}_k^\mathsf{T}(\hat{\mathbf{W}} - z\mathbf{I}_n)^{-1}\mathbf{j}_k dz + o_P(1)$$

for $k \in \{1,2\}$. Since $\hat{\mathbf{W}}$ is a low-rank perturbation of $\hat{L}$, invoking Sherman-Morrison's formula [84], we further have

$$\mathbf{j}_k^\mathsf{T}(\hat{\mathbf{W}} - z\mathbf{I}_n)^{-1}\mathbf{j}_k = \mathbf{j}_k^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\mathbf{j}_k - \frac{(5h'(\tau)^2/4)\left(\mathbf{j}_k^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\boldsymbol{\psi}\right)^2}{1 + (5h'(\tau)^2/4)\boldsymbol{\psi}^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\boldsymbol{\psi}} + o_P(n_k).$$

As $\frac{1}{\sqrt{n_k}}\mathbf{j}_k^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\boldsymbol{\psi} = o_P(1)$ [58, Equation 7.6], we get

$$\frac{1}{n_k}\mathbf{j}_k^\mathsf{T}(\hat{\mathbf{W}} - z\mathbf{I}_n)^{-1}\mathbf{j}_k = \frac{1}{n_k}\mathbf{j}_k^\mathsf{T}(h(\tau)\mathbf{L}'_s - z\mathbf{I}_n)^{-1}\mathbf{j}_k + o_P(1),$$

and thus

$$\frac{1}{n_k}(\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{lap}})^2 = \frac{1}{n_k}(\mathbf{j}_k^\mathsf{T}\mathbf{v}_{\text{ctr}})^2 + o_P(1),$$

which concludes the proof of Proposition 4.3.1.

B.3 Asymptotic Matrix Equivalent for $\hat{\mathbf{W}}$

The objective of this section is to prove the asymptotic matrix equivalent for $\hat{\mathbf{W}}$ expressed in (B.3). Some additional notations that will be useful in the proof:

- for $\mathbf{x}_i \in \mathcal{C}_k$, $k \in \{1,2\}$, $\boldsymbol{\omega}_i \equiv \mathbf{x}_i - \boldsymbol{\mu}_k$, and $\boldsymbol{\Omega} \equiv [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n]^\mathsf{T}$;
- $\boldsymbol{\mu}_k^\circ = \boldsymbol{\mu}_k - \frac{1}{n}\sum_{k'=1}^2 n_{k'}\boldsymbol{\mu}_{k'}$, $t_k = \left(\text{tr}\mathbf{C}_k - \frac{1}{n}\sum_{k'=1}^2 n_{k'}\text{tr}\mathbf{C}_{k'}\right)/\sqrt{p}$;
- $\mathbf{j}_k \in \mathbb{R}^n$ is the canonical vector of $\mathcal{C}_k$, i.e., $[\mathbf{j}_k]_i = 1$ if $\mathbf{x}_i \in \mathcal{C}_k$ and $[\mathbf{j}_k]_i = 0$ otherwise;
- $\psi_i \equiv \left(\|\boldsymbol{\omega}_i\|^2 - \text{E}[\|\boldsymbol{\omega}_i\|^2]\right)/\sqrt{p}$, $\boldsymbol{\psi} \equiv [\psi_1, \cdots, \psi_n]^\mathsf{T}$ and $(\boldsymbol{\psi})^2 \equiv [(\psi_1)^2, \cdots, (\psi_n)^2]^\mathsf{T}$.

As $w_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p = h(\tau) + O(p^{-\frac{1}{2}})$ for all $i \neq j$, we can Taylor-expand $w_{ij} = h(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ around $h(\tau)$ to obtain the following expansion for $\mathbf{W}$, which can be found in [58]:

$$
\begin{aligned}
\mathbf{W} &= h(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{h'(\tau)}{\sqrt{p}}\left[\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T} + \sum_{b=1}^{2} t_b\mathbf{j}_b\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\sum_{a=1}^{2} t_a\mathbf{j}_a^\mathsf{T}\right] \\
&\quad + \frac{h'(\tau)}{p}\left[\sum_{a,b=1}^{2}\|\boldsymbol{\mu}_a^\circ - \boldsymbol{\mu}_b^\circ\|^2\mathbf{j}_b\mathbf{j}_a^\mathsf{T} - 2\boldsymbol{\Omega}\sum_{a=1}^{2}\boldsymbol{\mu}_a^\circ\mathbf{j}_a^\mathsf{T} + 2\sum_{b=1}^{2}\mathrm{diag}(\mathbf{j}_b)\boldsymbol{\Omega}\boldsymbol{\mu}_b^\circ\mathbf{1}_n^\mathsf{T}\right. \\
&\quad \left. - 2\sum_{b=1}^{2}\mathbf{j}_b\boldsymbol{\mu}_b^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T} + 2\mathbf{1}_n\sum_{a=1}^{2}\boldsymbol{\mu}_a^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T}\mathrm{diag}(\mathbf{j}_a) - 2\boldsymbol{\Omega}\boldsymbol{\Omega}^\mathsf{T}\right] \\
&\quad + \frac{h''(\tau)}{2p}\left[(\boldsymbol{\psi})^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n[(\boldsymbol{\psi})^2]^\mathsf{T} + \sum_{b=1}^{2} t_b^2\mathbf{j}_b\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\sum_{a=1}^{2} t_a^2\mathbf{j}_a^\mathsf{T}\right. \\
&\quad + 2\sum_{a,b=1}^{2} t_at_b\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + 2\sum_{b=1}^{2}\mathrm{diag}(\mathbf{j}_b)t_b\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + 2\sum_{b=1}^{2} t_b\mathbf{j}_b\boldsymbol{\psi}^\mathsf{T} + 2\sum_{a=1}^{2}\mathbf{1}_n\boldsymbol{\psi}^\mathsf{T}\mathrm{diag}(\mathbf{j}_a)t_a \\
&\quad \left. + 2\boldsymbol{\psi}\sum_{a=1}^{2} t_a\mathbf{j}_a^\mathsf{T} + 2\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\right] + (h(0) - h(\tau) + \tau h'(\tau))\mathbf{I}_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}).
\end{aligned}
$$

Applying $\mathbf{P}_n = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\right)$ on both sides of the above equation, we get

$$
\begin{aligned}
\hat{\mathbf{W}} &= \mathbf{P}_n\mathbf{W}\mathbf{P}_n \\
&= \frac{-2h'(\tau)}{p}\left[\sum_{a,b=1}^{2}(\boldsymbol{\mu}_a^{\circ\mathsf{T}}\boldsymbol{\mu}_b^\circ)\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + \mathbf{P}_n\boldsymbol{\Omega}\sum_{a=1}^{2}\boldsymbol{\mu}_a^\circ\mathbf{j}_a^\mathsf{T} + \sum_{b=1}^{2}\mathbf{j}_b\boldsymbol{\mu}_b^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T}\mathbf{P}_n + \mathbf{P}_n\boldsymbol{\Omega}\boldsymbol{\Omega}^\mathsf{T}\mathbf{P}_n\right] \\
&\quad + \frac{h''(\tau)}{p}\left[\sum_{a,b=1}^{2} t_at_b\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + \sum_{b=1}^{2} t_b\mathbf{j}_b\boldsymbol{\psi}^\mathsf{T}\mathbf{P}_n + \mathbf{P}_n\boldsymbol{\psi}\sum_{a=1}^{2} t_a\mathbf{j}_a^\mathsf{T} + \mathbf{P}_n\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\mathbf{P}_n\right] \\
&\quad + (h(0) - h'(\tau) + \tau h''(\tau))\mathbf{P}_n + O(p^{-\frac{1}{2}}) \\
&= \frac{1}{p}\hat{\boldsymbol{\Phi}}^\mathsf{T}\hat{\boldsymbol{\Phi}} + (h(0) - h(\tau) + \tau h'(\tau))\mathbf{P}_n + O_{\|\cdot\|}(p^{-\frac{1}{2}})
\end{aligned}
$$

where the last equality is justified by

$$
\begin{aligned}
\frac{1}{p}\hat{\boldsymbol{\Phi}}^\mathsf{T}\hat{\boldsymbol{\Phi}} &= \frac{-2h'(\tau)}{p}\left[\sum_{a,b=1}^{2}(\boldsymbol{\mu}_a^{\circ\mathsf{T}}\boldsymbol{\mu}_b^\circ)\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + \mathbf{P}_n\boldsymbol{\Omega}\sum_{a=1}^{2}\boldsymbol{\mu}_a^\circ\mathbf{j}_a^\mathsf{T} + \sum_{b=1}^{2}\mathbf{j}_b\boldsymbol{\mu}_b^{\circ\mathsf{T}}\boldsymbol{\Omega}^\mathsf{T}\mathbf{P}_n + \mathbf{P}_n\boldsymbol{\Omega}\boldsymbol{\Omega}^\mathsf{T}\mathbf{P}_n\right] \\
&\quad + \frac{h''(\tau)}{p}\left[\sum_{a,b=1}^{2} t_at_b\mathbf{j}_b\mathbf{j}_a^\mathsf{T} + \sum_{b=1}^{2} t_b\mathbf{j}_b\boldsymbol{\psi}^\mathsf{T}\mathbf{P}_n + \mathbf{P}_n\boldsymbol{\psi}\sum_{a=1}^{2} t_a\mathbf{j}_a^\mathsf{T} + \mathbf{P}_n\boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{T}\mathbf{P}_n\right].
\end{aligned}
$$

Equation (B.3) is thus proved.

Appendix C

Supplementary material of Chapters 6-7

Here the big O notation $O(u_n)$ and the small O notation $o(u_n)$ are understood in probability when speaking about random variables. Additionally, when multidimensional objects are concerned, $O(u_n)$ and $o(u_n)$ are understood entry-wise. The notation $O_{\|\cdot\|}$ is understood as follows: for a vector $\mathbf{v}$, $\mathbf{v} = O_{\|\cdot\|}(u_n)$ means its Euclidean norm is $O(u_n)$ and for a square matrix $\mathbf{M}$, $\mathbf{M} = O_{\|\cdot\|}(u_n)$ means that the operator norm of $\mathbf{M}$ is $O(u_n)$. We follow the same rules for the notation $o_{\|\cdot\|}$.

C.1 Proofs of the theoretical results in Chapter 6

C.1.1 Proof of Proposition 5.3.1

Let

$$\boldsymbol{\eta}_{(-i)} = \frac{1}{n} \sum_{j \neq i} y_j c_j \mathbf{x}_j, \tag{C.1}$$

we have

$$c_i = \phi_\tau \left((1 - y_i \boldsymbol{\eta}_{(-i)}^\mathsf{T} \mathbf{x}_i - y_i \beta_0) / (n^{-1} \|\mathbf{x}_i\|^2) \right) \tag{C.2}$$

for $\phi_\tau(t) = \max\{0, \min\{t, \tau\}\}$.

First note for future reference that, through concentration inequality arguments, we immediately have $n^{-1}\|\mathbf{x}_i\|^2 = n^{-1} \operatorname{tr} \mathbf{C}_k + O(p^{-\frac{1}{2}})$ for all $i \in \{1, \ldots, n\}$ with $y_i = (-1)^k$, where $k \in \{1, 2\}$.

According to (C.2), the key to understanding the statistical behavior of c_i in the large dimensional regime lies in the characterization of $\boldsymbol{\eta}_{(-i)}^\mathsf{T} \mathbf{x}_i$. Since $\boldsymbol{\eta}_{(-i)}$ depends on $\mathbf{x}_i$ in an intricate, implicit manner through the dual optimization problem (5.3), not much can be said about c_i directly from (C.2) at this stage.

Let $(\boldsymbol{\beta}_{(-i)}, \beta_{(-i)0})$ be the SVM solution obtained from (5.2) with all data except $\mathbf{x}_i$, and

$c_{(-i)j}$ as the dual problem coefficients obtained from (5.3) with all data except $\mathbf{x}_i$. Then

$$\boldsymbol{\beta}_{(-i)} = \frac{1}{n}\sum_{j\neq i} y_j c_{(-i)j}\mathbf{x}_j. \tag{C.3}$$

Although the expression (C.1) of $\boldsymbol{\eta}_{(-i)}$ is very similarly to that of $\boldsymbol{\beta}_{(-i)}$ given in (C.3), the two vectors are critically different in the fact that $\boldsymbol{\beta}_{(-i)}$ is independent of $\mathbf{x}_i$, while $\boldsymbol{\eta}_{(-i)}$ is not. The goal of the *leave-one-observation-out* step is to establish a relation between $\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ and $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$.

To this end, first note from (C.2) that, for all $j \neq i$,

$$\begin{aligned} c_j - c_{(-i)j} = \phi_\tau\left(\frac{1 - n^{-1}\sum_{l\neq i,j} y_j y_l c_l \mathbf{x}_l^\mathsf{T}\mathbf{x}_j - n^{-1} y_j y_i c_i \mathbf{x}_i^\mathsf{T}\mathbf{x}_j - y_j\beta_0}{n^{-1}\|\mathbf{x}_j\|^2}\right) \\ - \phi_\tau\left(\frac{1 - n^{-1}\sum_{l\neq i,j} y_j y_l c_{(-i)l} \mathbf{x}_l^\mathsf{T}\mathbf{x}_j - y_j\beta_{(-i)0}}{n^{-1}\|\mathbf{x}_j\|^2}\right). \end{aligned} \tag{C.4}$$

Evidently, for any $t_1, t_2 \in \mathbb{R}$, there exists a constant $d \in [0,1]$ such that $\phi_\tau(t_1) - \phi_\tau(t_2) = d(t_1 - t_2)$. We denote then by $d_{(-i)j}$ the constant within the interval $[0,1]$ that satisfies

$$c_j - c_{(-i)j} = \frac{d_{(-i)j}\big[- n^{-1}\sum_{l\neq i,j} y_j y_l (c_l - c_{(-i)l}) \mathbf{x}_l^\mathsf{T}\mathbf{x}_j - n^{-1} y_j y_i c_i \mathbf{x}_i^\mathsf{T}\mathbf{x}_j - y_j(\beta_0 - \beta_{(-i)0})\big]}{n^{-1}\|\mathbf{x}_j\|^2}.$$

Set $\bar{\mathbf{x}}_i = y_i\mathbf{x}_i/\sqrt{n}$. For all $j \neq i$ such that $d_{(-i)j} \neq 0$, denote by $\bar{\mathbf{X}}_{(-i)}$ the data matrix with $y_j\mathbf{x}_j/\sqrt{n}$ as column vectors, $\boldsymbol{\Delta}c_{(-i)}$ the vector composed of the differences $c_j - c_{(-i)j}$, $\mathbf{y}_{(-i)}$ the vector of labels y_j and $\mathbf{D}_{(-i)}$ the diagonal matrix of the non-zero $d_{(-i)j}$. Define $\mathrm{diag}(\mathbf{A})$ as the operator that returns the diagonal matrix having the same diagonal elements as the input square matrix $\mathbf{A}$. Then, the previous expression entails

$$\begin{aligned} &\left[\mathbf{D}_{(-i)}\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)} + \mathrm{diag}\left(\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)} - \mathbf{D}_{(-i)}\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)}\right)\right]\boldsymbol{\Delta}c_{(-i)} \\ &= -c_i\mathbf{D}_{(-i)}\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{x}}_i - \left(\beta_0 - \beta_{(-i)0}\right)\mathbf{D}_{(-i)}\mathbf{y}_{(-i)}, \end{aligned} \tag{C.5}$$

where

$$\boldsymbol{\Delta}c_{(-i)} = -\mathbf{M}_{(-i)}^{-1}\left[c_i\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{x}}_i + \left(\beta_0 - \beta_{(-i)0}\right)\mathbf{y}_{(-i)}\right]$$

with

$$\mathbf{M}_{(-i)} = \bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)} + \mathrm{diag}\left(\mathbf{D}_{(-i)}^{-1}\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)} - \bar{\mathbf{X}}_{(-i)}^\mathsf{T}\bar{\mathbf{X}}_{(-i)}\right).$$

Note importantly that $\mathbf{M}_{(-i)}$ is indeed invertible and that $\mathbf{M}_{(-i)}^{-1} = O_{\|\cdot\|}(1)$, demonstrated as follows. Equations (5.5) and (C.4) imply that $d_{(-i)j} = 1$ if and only if $y_j(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_j + \beta_0) = 1$ and $y_j(\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_j + \beta_{(-i)0}) = 1$; in other words, $\mathbf{x}_j$ is on the hyperplane with or without the i-th training sample left out. Since training samples are generated independently, we remark that any $p+1$ data samples define uniquely (if possible) two parallel hyperplanes in the space of $\mathbb{R}^p$. As the optimization (5.2) of SVMs implements a condition of maximal margin, the number of the samples on (or infinitesimally close to) the boundary is smaller than $p+1$, otherwise the probability of satisfying simultaneously (5.5) and $\sum_{i=1}^n c_i y_i = 0$ is zero. Without loss of

generality, let us now write $\bar{\mathbf{X}}_{(-i)} = \begin{bmatrix}\bar{\mathbf{X}}_{\mathcal{B}(-i)} & \bar{\mathbf{X}}_{\mathcal{C}(-i)}\end{bmatrix}$ where $\bar{\mathbf{X}}_{\mathcal{C}(-i)}$ is composed of the $\bar{\mathbf{x}}_j$ for which $d_{(-i)j}$ bounded away from 1. From our previous reasoning, the dimension of the square matrix $\bar{\mathbf{X}}^{\mathsf{T}}_{\mathcal{B}(-i)}\bar{\mathbf{X}}_{\mathcal{B}(-i)}$ is no greater than p with probability 1 and $\bar{\mathbf{X}}^{\mathsf{T}}_{\mathcal{B}(-i)}\bar{\mathbf{X}}_{\mathcal{B}(-i)}$ is thus full rank with eigenvalues bounded away from zero (from [85, 86]). Combining this fact with standard algebraic arguments, we deduce that $\|\mathbf{v}\|^2/\mathbf{v}^T\mathbf{M}_{(-i)}\mathbf{v} = O(1)$ for any vector $\mathbf{v}$. Therefore, $\mathbf{M}^{-1}_{(-i)} = O_{\|\cdot\|}(1)$.

Define $\mathcal{D}_{(-i)}$ the set of indices $j \in \{1, \ldots, i-1, i+1, \ldots, n\}$ such that $d_{(-i)j} \neq 0$, and let $n_{\mathcal{D}_{(-i)}} = |\mathcal{D}_{(-i)}|$, which is obviously an integer between 1 and n (with high probability). It is also clear that $n_{\mathcal{D}_{(-i)}}$ is the dimension of the square matrix $\mathbf{M}_{(-i)}$. Since $\sum_i y_i c_i = 0$ and $\sum_{j\neq i} y_j c_{(-i)j} = 0$ by the optimization constraints of (5.3),

$$\mathbf{y}^{\mathsf{T}}_{(-i)}\boldsymbol{\Delta}c_{(-i)} + y_i c_i = -c_i\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i - (\beta_0 - \beta_{(-i)0})\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)} + y_i c_i = 0.$$

Seeing that $\bar{\mathbf{x}}_i = y_i\mathbf{x}_i/\sqrt{n}$ is independent of $\bar{\mathbf{X}}_{(-i)}$, it is easily shown again by concentration inequalities) that $c_i\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i = O(\sqrt{n_{\mathcal{D}_{(-i)}}/n})$. Hence,

$$(\beta_0 - \beta_{(-i)0}) = \frac{c_i\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i - y_i c_i}{\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}\mathbf{y}_{(-i)}} = O(1/n_{\mathcal{D}_{(-i)}}).$$

as the above display reduces to $(\beta_0 - \beta_{(-i)0})\mathbf{y}^{\mathsf{T}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)} + O(1) = 0$. Notice then

$$\begin{aligned}\|\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i\|^2 &= O(n_{\mathcal{D}_{(-i)}}/n)\\ \|\left(\beta_0 - \beta_{(-i)0}\right)\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)}\|^2 &= O(1/n_{\mathcal{D}_{(-i)}}),\end{aligned}$$

leading to

$$\begin{aligned}\boldsymbol{\Delta}c_{(-i)} = -c_i\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i - \left(\beta_0 - \beta_{(-i)0}\right)\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)} &= O_{\|\cdot\|}(\sqrt{\max\{n_{\mathcal{D}_{(-i)}}/n, 1/n_{\mathcal{D}_{(-i)}}\}})\\ &= O_{\|\cdot\|}(1).\end{aligned}$$

Furthermore,

$$\begin{aligned}\boldsymbol{\eta}_{(-i)} - \boldsymbol{\beta}_{(-i)} = \frac{1}{\sqrt{n}}\bar{\mathbf{X}}_{(-i)}\boldsymbol{\Delta}c_{(-i)} &= -\frac{1}{\sqrt{n}}c_i\bar{\mathbf{X}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i - \frac{1}{\sqrt{n}}\left(\beta_0 - \beta_{(-i)0}\right)\bar{\mathbf{X}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)}\\ &= O_{\|\cdot\|}(\sqrt{\max\{n_{\mathcal{D}_{(-i)}}/n^2, 1/nn_{\mathcal{D}_{(-i)}}\}})\\ &= O_{\|\cdot\|}(n^{-\frac{1}{2}})\end{aligned}$$

and

$$\begin{aligned}y_i\mathbf{x}_i^{\mathsf{T}}(\boldsymbol{\eta}_{(-i)} - \boldsymbol{\beta}_{(-i)}) &= -c_i\bar{\mathbf{x}}_i^{\mathsf{T}}\bar{\mathbf{X}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i - \left(\beta_0 - \beta_{(-i)0}\right)\bar{\mathbf{x}}_i^{\mathsf{T}}\bar{\mathbf{X}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\mathbf{y}_{(-i)}\\ &= -c_i\bar{\mathbf{x}}_i^{\mathsf{T}}\bar{\mathbf{X}}_{(-i)}\mathbf{M}^{-1}_{(-i)}\bar{\mathbf{X}}^{\mathsf{T}}_{(-i)}\bar{\mathbf{x}}_i + O(1/\sqrt{nn_{\mathcal{D}_{(-i)}}}).\end{aligned}$$

The above approximation of $y_i\boldsymbol{\eta}^{\mathsf{T}}_{(-i)}\mathbf{x}_i$ allows us to replace $y_i\boldsymbol{\eta}^{\mathsf{T}}_{(-i)}\mathbf{x}_i$ in (C.2) with $y_i\boldsymbol{\beta}^{\mathsf{T}}_{(-i)}\mathbf{x}_i$, giving rise to

$$\begin{aligned}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i - \xi_i^r c_i &= \boldsymbol{\beta}^{\mathsf{T}}_{(-i)}\mathbf{x}_i + O(1/\sqrt{nn_{\mathcal{D}_{(-i)}}})\\ c_i &= \phi_\tau\left((1 - y_i\boldsymbol{\beta}^{\mathsf{T}}_{(-i)}\mathbf{x}_i - y_i\beta_0)/\xi_i^r\right) + O(1/\sqrt{nn_{\mathcal{D}_{(-i)}}})\end{aligned}$$

where

$$\xi_i^r = \bar{\mathbf{x}}_i^\mathsf{T}\left(\mathbf{I}_p - \bar{\mathbf{X}}_{(-i)}\mathbf{M}_{(-i)}^{-1}\bar{\mathbf{X}}_{(-i)}^\mathsf{T}\right)\bar{\mathbf{x}}_i = O(1).$$

The concentration arguments suggest that $\xi_{(-i)j}^r$ goes to a value independent of $\mathbf{x}_i$ in the limit of large n, p. Before further investigation, we keep it for the moment under this random form.

As such, we managed to "break" the non-trivial dependence within the term $\boldsymbol{\eta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ by a convenient replacement with the inner product $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ of independent vectors $\boldsymbol{\beta}_{(-i)}$ and $\mathbf{x}_i$. However, we introduced ξ_i^r in the process, which remains to be investigated. To this end, we first need to better characterize the values $d_{(-i)j}$ for the case where the number of non-zero $d_{(-i)j}$ is comparable to n (otherwise we have simply $\xi_i^r = n^{-1}\|\mathbf{x}_i\| + o_(1)$). Note that, by a Taylor expansion, for $t \in \mathbb{R}\setminus\{0,\tau\}$, $\phi_\tau(t+\delta t) - \phi_\tau(t) = \phi_\tau'(t)\delta t + O(\delta t^2)$ with $\phi_\tau'(t) = 0$ for $t \in (-\infty, 0)\cup(\tau, +\infty)$ and $\phi_\tau'(t) = 1$ for $t \in (0,\tau)$. Recall from the above discussion that

$$\begin{aligned} c_{(-i)j} &= \phi_\tau\left(\frac{1 - n^{-1}\sum_{l\neq i,j} y_j y_l c_{(-i)l}\mathbf{x}_l^\mathsf{T}\mathbf{x}_j - y_j\beta_{(-i)0}}{n^{-1}\|\mathbf{x}_j\|^2}\right) \\ &= \phi_\tau\left((1 - y_j\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j - y_j\beta_{(-ij)0})/\xi_{(-i)j}^r\right) + O(1/\sqrt{n n_{\mathcal{D}_{(-ij)}}}). \end{aligned}$$

where $\xi_{(-i)j}^r = \bar{\mathbf{x}}_j^\mathsf{T}\left(\mathbf{I}_p - \bar{\mathbf{X}}_{(-ij)}\mathbf{M}_{(-ij)}^{-1}\bar{\mathbf{X}}_{(-ij)}^\mathsf{T}\right)\bar{\mathbf{x}}_j$. Note here and in the following that the notation $(-ij)$, similarly to the notation $(-i)$, refers to mathematical objects obtained by leaving out the i-th and the j-th data samples. Thus, by letting

$$t_{(-i)j} = (1 - y_j\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j - y_j\beta_0)/\xi_{(-i)j}^r,$$

we have $d_{(-i)j} = \phi_\tau'(t_{(-i)j}) + O(n^{-\frac{1}{2}})$ for $t_{(-i)j} \in \mathbb{R}\setminus\{0,\tau\}$.

Remark that the probability of $t_{(-i)j} = 0$ or $t_{(-i)j} = \tau$ is zero. To see this, notice first that as $c_i \leq \tau = O(1)$ and $\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^n c_i\mathbf{x}_i$, $\boldsymbol{\beta} = O_{\|\cdot\|}(1)$ has its energy "evenly" distributed among its elements, in the sense that for any subset $\mathcal{A} \subseteq \{1,\ldots,p\}$ with its cardinality comparable to p, $\|\boldsymbol{\beta}\|^2/\sum_{d=\mathcal{A}}\{\boldsymbol{\beta}\}_d^2 = O(1)$. Therefore, as $\boldsymbol{\beta}_{(-ij)}$ is independent of x_j, by the principle of the central limit theorem, we can write $\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j = w_j + w_j'$ where w_j follows asymptotically a continuous (normal) distribution, independent of (and non-negligible when compared to) w_j'. We conclude thus that the probability of $t_{(-i)j} = 0$ or $t_{(-i)j} = \tau$ is zero.

On account of these arguments, we obtain that

$$\xi_i^r = \bar{\mathbf{x}}_i^\mathsf{T}\left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}(\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^\mathsf{T}\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}})^{-1}\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^\mathsf{T}\right]\bar{\mathbf{x}}_i + O(n^{-\frac{1}{2}}).$$

with $\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}$ composed of the $y_j\mathbf{x}_j/\sqrt{n}$, $j \neq i$, for which $c_{(-i)j} \in (0,\tau)$ (i.e., $t_{(-i)j} \in (0,\tau)$). We define additionally $\mathcal{B}_{(-i)}$ the index set of j such that $c_{(-i)j} \in (0,\tau)$ and $n_{\mathcal{B}_{(-i)}}$ the cardinality of $\mathcal{B}_{(-i)}$. The notations $\mathcal{B}$ and $\mathcal{B}(-ij)$ are understood in the same way as $\mathcal{B}_{(-i)}$, for respectively the whole data set and the data set without the i-th and the j-th data samples.

Now we move to discuss the convergence of ξ_i^r. By the "trace lemma" [27, Lemma 14.2], we have that, for $y_i = (-1)^k$,

$$\begin{aligned} \xi_i^r &= \frac{1}{n}\mathrm{tr}\mathbf{C}_k\left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}(\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^\mathsf{T}\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}})^{-1}\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^\mathsf{T}\right] + o(1) \\ &= \frac{1}{n}\mathrm{tr}\mathbf{C}_k\left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}}(\bar{\mathbf{X}}_{\mathcal{B}}^\mathsf{T}\bar{\mathbf{X}}_{\mathcal{B}})^{-1}\bar{\mathbf{X}}_{\mathcal{B}}^\mathsf{T}\right] + o(1) \end{aligned}$$

where the second equality is justified by the definition of $\bar{\mathbf{X}}_{\mathcal{B}}$ and $\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}$ and the fact that $\boldsymbol{\Delta} c_{(-i)} = O_{\|\cdot\|}(1)$.

Let

$$\mathbf{Q} = \mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}}(\bar{\mathbf{X}}_{\mathcal{B}}^{\mathsf{T}}\bar{\mathbf{X}}_{\mathcal{B}})^{-1}\bar{\mathbf{X}}_{\mathcal{B}}^{\mathsf{T}},$$

we notice that

$$\begin{aligned}\mathbf{Q} &= \mathbf{I}_p - \lim_{z\to 0^+} \bar{\mathbf{X}}_{\mathcal{B}}(z\mathbf{I}_{n_b} + \bar{\mathbf{X}}_{\mathcal{B}}^{\mathsf{T}}\bar{\mathbf{X}}_{\mathcal{B}})^{-1}\bar{\mathbf{X}}_{\mathcal{B}}^{\mathsf{T}} \\ &= \lim_{z\to 0^+} z(z\mathbf{I}_p + \bar{\mathbf{X}}_{\mathcal{B}}\bar{\mathbf{X}}_{\mathcal{B}}^{\mathsf{T}})^{-1} \\ &= \lim_{z\to 0^+} z\left(z\mathbf{I}_p + \frac{1}{n}\sum_{i\in\mathcal{B}} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1},\end{aligned}$$

leading to

$$\xi_i^r = \lim_{z\to 0^+} z\frac{1}{n}\operatorname{tr}\mathbf{C}_k\mathbf{Q}(z) + o(1)$$

for

$$\mathbf{Q}(z) = \left(z\mathbf{I}_p + \frac{1}{n}\sum_{i\in\mathcal{B}} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}.$$

The matrix $\left(z\mathbf{I}_p + \frac{1}{n}\sum_{i\in\mathcal{B}} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}$ is in fact a standard object of study in random matrix theory, referred to as the resolvent of $\frac{1}{n}\sum_{i\in\mathcal{B}} \mathbf{x}_i\mathbf{x}_i^T$. Remark that $\frac{1}{n}\sum_{i\in\mathcal{B}} \mathbf{x}_i\mathbf{x}_i^T$ can be seen as a sort of empirical covariance matrix, the limiting form of its resolvent was investigated in [26]. Based on the results of [26], we get that,

$$\frac{1}{n}\operatorname{tr}\mathbf{C}_k\mathbf{Q}(z) \to \frac{1}{n}\operatorname{tr}\mathbf{C}_k\tilde{\mathbf{Q}}(z) \equiv \xi_k(z)$$

where

$$\tilde{\mathbf{Q}}(z) = \left[\mathbf{I}_p + \left(\frac{1}{z+\xi_1(z)}\frac{n_{\mathcal{B}_1}}{n}\mathbf{C}_1 + \frac{1}{z+\xi_2(z)}\frac{n_{\mathcal{B}_2}}{n}\mathbf{C}_2\right)\right]^{-1}$$

with $n_{\mathcal{B}_k}$ the count of $i \in \{1,\ldots,n\}$ such that $c_i \in (0,\tau)$ and $y_i = (-1)^k$, for $k \in \{1,2\}$. Taking the limit $z \to 0^+$, we obtain

$$\frac{1}{n}\operatorname{tr}\mathbf{C}_k\mathbf{Q} \to \frac{1}{n}\operatorname{tr}\mathbf{C}_k\tilde{\mathbf{Q}} \equiv \xi_k$$

where

$$\tilde{\mathbf{Q}} = \left[\mathbf{I}_p + \left(\frac{1}{\xi_1}\frac{n_{\mathcal{B}_1}}{n}\mathbf{C}_1 + \frac{1}{\xi_2}\frac{n_{\mathcal{B}_2}}{n}\mathbf{C}_2\right)\right]^{-1}.$$

We demonstrate thus that

$$F\left(\frac{n_{\mathcal{B}_1}}{n}, \frac{n_{\mathcal{B}_2}}{n}\right) = \{\xi_1, \xi_2\}$$

for F a mapping from $\mathbb{R}^* \times \mathbb{R}^*$ to $\mathbb{R}^* \times \mathbb{R}^*$ defined in (5.8), and

$$\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i - \xi_k c_i = \boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i + o(1)$$

for $y_i = (-1)^k$. Since $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{(-i)}\|^2 = o(1)$, $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ follows asymptotically the same law as $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i'$ for some random vector $\mathbf{x}_i' \overset{\mathcal{L}}{=} \mathbf{x}_i$, independent of $\boldsymbol{\beta}$. Proposition 5.3.1 is thus proven. And so is Corollary 5.1, which is demonstrated by

$$c_i = \phi_\tau\left(\frac{1 - y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i - y_i\beta_0}{\xi_k}\right) + o(1) \tag{C.6}$$

Additionally, it is interesting to points out that the above equation implies that $n_{\mathcal{D}_{(-i)}}$ is in fact comparable to n. To show this, remark first that, for any $a, b \in \mathbb{R}$ with $b > a$, and some $(\mathbf{x}, y) \overset{\mathcal{L}}{=} (\mathbf{x}_i, y_i)$, independent of $\boldsymbol{\beta}$.

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i) \to \mathbb{E}\{\mathbf{1}_{(a,b)}(y\boldsymbol{\beta}^\mathsf{T}\mathbf{x})\} = \mathbb{P}\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{x} \in (a,b)\}$$

where we recall from the previous discussion on the statistical behavior $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i$ that $\mathbb{P}\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{x} \in (a,b)\}$ is comparable to 1 if $b - a = O(1)$. The convergence is proven by

$$\frac{1}{n}\mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\} = \mathbb{P}\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{x} \in (a,b)\} + o(1)$$

and

$$\begin{aligned}
&\mathrm{Var}\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\} \\
=&\frac{1}{n^2}\sum_{i,j=1}^n \left[\mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-j)}^\mathsf{T}\mathbf{x}_j)\} - \mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\}\mathbb{E}\{\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-j)}^\mathsf{T}\mathbf{x}_j)\}\right] \\
=&\frac{1}{n^2}\sum_{i,j=1}^n \left[\mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i)\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j)\} - \mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\}\mathbb{E}\{\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-j)}^\mathsf{T}\mathbf{x}_j)\}\right] + o(1) \\
=&\frac{1}{n^2}\sum_{i,j=1}^n \left[\mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i)\}\mathbb{E}\{\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j)\} - \mathbb{E}\{\mathbf{1}_{(a,b)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i)\}\mathbb{E}\{\mathbf{1}_{(a,b)}(y_j\boldsymbol{\beta}_{(-j)}^\mathsf{T}\mathbf{x}_j)\}\right] + o(1) \\
=&o(1).
\end{aligned}$$

As a result of this convergence, we obtain

$$\begin{aligned}
\frac{n_{\mathcal{B}}}{n} =&\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(0,\tau)}(c_i) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(0,\tau)}\left[\phi_\tau\left(\frac{1 - y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i - y_i\beta_0}{\xi_k}\right)\right] + o(1) \\
=&\frac{1}{n}\sum_{i\in\mathcal{C}_1} \mathbf{1}_{(1+\beta_0-\xi_1\tau,1+\beta_0)}(y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i) + \frac{1}{n}\sum_{j\in\mathcal{C}_2} \mathbf{1}_{(1-\beta_0-\xi_2\tau,1-\beta_0)}(y_j\boldsymbol{\beta}_{(-j)}^\mathsf{T}\mathbf{x}_j) + o(1) \\
=&\frac{n_1}{n}\mathbb{P}\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{x} \in (1+\beta_0-\xi_1\tau, 1+\beta_0)|y=-1\} + \frac{n_2}{n}\mathbb{P}\{y\boldsymbol{\beta}^\mathsf{T}\mathbf{x} \in (1+\beta_0-\xi_2\tau, 1+\beta_0)|y=1\} \\
&+ o(1)
\end{aligned}$$

for some $(\mathbf{x}, y) \overset{\mathcal{L}}{=} (\mathbf{x}_i, y_i)$, independent of $\boldsymbol{\beta}$. Hence, $n_{\mathcal{B}}/n$ is comparable to 1. Since $n_{\mathcal{B}_{(-i)}} = n_{\mathcal{B}} + o(n)$ and $n_{\mathcal{D}_{(-i)}} \geq n_{\mathcal{B}_{(-i)}}$, $n_{\mathcal{D}_{(-i)}}$ is comparable to n.

In summary, we have the following approximation results (which will be useful for the derivation in the next section):

$$\beta_0 - \beta_{(-i)0} = O(n^{-1}) \tag{C.7}$$

$$\boldsymbol{\Delta} c_{(-i)} = -c_i (\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^{\mathsf{T}} \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}})^{-1} \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^{\mathsf{T}} \bar{\mathbf{x}}_i + o(n^{-\frac{1}{2}}) = O(n^{-\frac{1}{2}}) \tag{C.8}$$

$$\boldsymbol{\beta} - \boldsymbol{\beta}_{(-i)} = \frac{c_i}{\sqrt{n}} \left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}} (\bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^{\mathsf{T}} \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}})^{-1} \bar{\mathbf{X}}_{\mathcal{B}_{(-i)}}^{\mathsf{T}} \right] \bar{\mathbf{x}}_i + o_{\|\cdot\|}(n^{-\frac{1}{2}}) = O_{\|\cdot\|}(n^{-\frac{1}{2}}). \tag{C.9}$$

Moreover, letting

$$F\left(\frac{n_{\mathcal{B}_{(-i)1}}}{n}, \frac{n_{\mathcal{B}_{(-i)2}}}{n} \right) = \{\xi_{(-i)1}, \xi_{(-i)2}\},$$

we have then

$$\xi_{(-i)k} = \xi_k + O(n^{-\frac{1}{2}}), \quad k \in \{1, 2\} \tag{C.10}$$

as a consequence of $n_{\mathcal{B}_{(-i)k}}/n = n_{\mathcal{B}_k}/n + O(n^{-\frac{1}{2}})$, $k \in \{1, 2\}$, which is given by $c_i - c_{(-j)i} = O(n^{-\frac{1}{2}})$.

C.1.2 Proof of Proposition 5.3.2 and Theorem 5.3.1

In this proof, we make use of the results in Proposition 5.3.1 and Corollary 5.1, . Some arguments in the proof of Proposition 5.3.1 are also employed. In the following, we denote $\mathcal{C}_k$ with $k \in \{1, 2\}$ the set of indices $i \in \{1, \ldots, n\}$ such that $y_i = (-1)^k$.

Let us begin by writing $\boldsymbol{\beta}$ as

$$\begin{aligned}
\boldsymbol{\beta} &= \frac{1}{n} \sum_{i=1}^{n} c_i \mathbf{x}_i = \frac{1}{n} c_i \boldsymbol{\mu}_{(i)} + \frac{1}{n} c_i \mathbf{C}_{(i)} \mathbf{z}_i \\
&= \left(\frac{1}{n} \sum_{i \in \mathcal{C}_1} c_i \right) \boldsymbol{\mu}_1 + \left(\frac{1}{n} \sum_{j \in \mathcal{C}_2} c_j \right) \boldsymbol{\mu}_2 + \mathbf{C}_1 \left(\frac{1}{n} \sum_{i \in \mathcal{C}_1} c_i \mathbf{z}_i \right) + \mathbf{C}_2 \left(\frac{1}{n} \sum_{j \in \mathcal{C}_2} c_j \mathbf{z}_j \right).
\end{aligned}$$

We focus thus on the terms $\frac{1}{n} \sum_{i \in \mathcal{C}_k} c_i$, $\frac{1}{n} \sum_{i \in \mathcal{C}_k} c_i \mathbf{z}_i$, $k \in \{1, 2\}$.

Recall from (C.6) and (C.8) in the previous subsection that, for $i \in \mathcal{C}_1$, $k \in \{1, 2\}$,

$$\begin{aligned}
c_i &= \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}_{(-i)}^{\mathsf{T}} \mathbf{x}_i - y_i \beta_0}{\xi_k} \right) + o(1) \\
c_{(-j)i} &= \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}_{(-ij)}^{\mathsf{T}} \mathbf{x}_i - y_i \beta_0}{\xi_k} \right) + o(1) \\
c_i - c_{(-j)i} &= O(n^{-\frac{1}{2}}).
\end{aligned}$$

Moreover, we obtain from (C.7) and (C.10) that

$$\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i-y_i\beta_0}{\xi_k}\right)=\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i-y_i\beta_{(-ij)0}}{\xi_{(-ij)k}}\right)+o(1)$$

It follows then that, for $y_i=(-1)^k$, $y_j=(-1)^{k'}$ with $k,k'\in\{1,2\}$,

$$\begin{aligned}
&\mathrm{Cov}\{c_i,c_j\}\\
=&\mathbb{E}\{c_ic_j\}-\mathbb{E}\{c_j\}\mathbb{E}\{c_i\}=\mathbb{E}\{c_{(-j)i}c_{(-i)j}\}-\mathbb{E}\{c_i\}\mathbb{E}\{c_j\}+o(1)\\
=&\mathbb{E}\left\{\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i-y_i\beta_{(-ij)0}}{\xi_{(-ij)k}}\right)\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j-y_i\beta_{(-ij)0}}{\xi_{(-ij)k'}}\right)\right\}-\mathbb{E}\{c_j\}\mathbb{E}\{c_i\}+o(1)\\
=&\mathbb{E}\left\{\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_i-y_i\beta_{(-ij)0}}{\xi_{(-ij)k}}\right)\right\}\mathbb{E}\left\{\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-ij)}^\mathsf{T}\mathbf{x}_j-y_i\beta_{(-ij)0}}{\xi_{(-ij)k'}}\right)\right\}-\mathbb{E}\{c_j\}\mathbb{E}\{c_i\}+o(1)\\
=&o(1)
\end{aligned}$$

where the jump from the second line to the third is supported by the fact that $\boldsymbol{\beta}_{(-ij)}$, $\beta_{(-ij)0}$ and $\xi_{(-ij)k}$ are independent of $\mathbf{x}_i$ and $\mathbf{x}_j$. Therefore,

$$\mathrm{Var}\left\{\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_i\right\}=\frac{1}{n^2}\sum_{i\in\mathcal{C}_k}\mathrm{Var}\{c_i^2\}+\frac{1}{n^2}\sum_{i\neq j\in\mathcal{C}_k}\mathrm{Cov}\{c_i,c_j\}=o(1),$$

we get thus

$$\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_i=\frac{n_k}{n}\mathbb{E}\{c_i|y_i=(-1)^k\}+o(1),\quad k\in\{1,2\}.$$

We proceed now to study $\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_i\mathbf{z}_i$. To begin with, notice that

$$\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_i\mathbf{z}_i=\frac{1}{n}\sum_{i\in\mathcal{C}_k}\tilde{c}_i\mathbf{z}_i+o_{\|\cdot\|},\quad k\in\{1,2\}.$$

for

$$\tilde{c}_i=\phi_\tau\left(\frac{1-y_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{x}_i-y_i\beta_{(-i)0}}{\xi_{(-i)k}}\right)=c_i+o(1).$$

Let us define

$$\mathbf{e}_{[k]}=[e_{[k]1},\ldots,e_{[k]p}]=\frac{1}{n}\sum_{i\in\mathcal{C}_k}\left(\tilde{c}_i\mathbf{z}_i-\mathbb{E}_{\mathbf{x}_i}\{\tilde{c}_i\mathbf{z}_i\}\right),\quad k\in\{1,2\}$$

Obviously,

$$\mathbb{E}\{\mathbf{e}_{[k]}\}=\mathbf{0}_p.$$

Denoting

$$\tilde{c}_{(-j)i} = \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}_{(-ij)}^\mathsf{T} \mathbf{x}_i - y_i \beta_{(-ij)0}}{\xi_{(-ij)k}} \right),$$

for which we have from the earlier arguments that

$$\tilde{c}_{(-j)i} = \tilde{c}_i + O(n^{-\frac{1}{2}}) \text{ and } \mathrm{Cov}_{(\mathbf{x}_i, \mathbf{x}_j)}\{\tilde{c}_{(-j)i}[\mathbf{z}_i]_d, \tilde{c}_{(-i)j}[\mathbf{z}_j]_d\} = 0,$$

we show then $\mathrm{Var}\{e_{[k]d}\} = O(n^{-1})$ with the following development:

$$\begin{aligned}
\mathrm{Var}\{e_{[k]d}\} &= \frac{1}{n^2} \sum_{i \in \mathcal{C}_k} \mathrm{Var}_{\mathbf{x}_i}\{(\tilde{c}_i[\mathbf{z}_i]_d)^2\} + \frac{1}{n^2} \sum_{i \neq j \in \mathcal{C}_k} \mathrm{Cov}_{(\mathbf{x}_i, \mathbf{x}_j)}\{\tilde{c}_i[\mathbf{z}_i]_d, \tilde{c}_j[\mathbf{z}_j]_d\} \\
&= \frac{1}{n^2} \sum_{i \in \mathcal{C}_k} \mathrm{Var}_{\mathbf{x}_i}\{(\tilde{c}_i[\mathbf{z}_i]_d)^2\} + \frac{1}{n^2} \sum_{i \neq j \in \mathcal{C}_k} \mathrm{Cov}_{(\mathbf{x}_i, \mathbf{x}_j)}\{\tilde{c}_{(-j)i}[\mathbf{z}_i]_d, \tilde{c}_{(-i)j}[\mathbf{z}_j]_d\} + O(n^{-1}) \\
&= O(n^{-1}).
\end{aligned}$$

For $k \in \{1, 2\}$, we define $\mathcal{A}_k$ a subset of $\{1, \ldots, p\}$ such that $\forall d \in \mathcal{A}_k$, $[\mathbf{C}_k \boldsymbol{\beta}]_d = O(n^{-\frac{1}{2}})$, and $\mathcal{A}_k^c = \{1, \ldots, p\} \setminus \mathcal{A}_k$ the complement of $\mathcal{A}_k$. Since $\|\mathbf{C}_k\| = O(1)$ and $\|\mathbf{C}_k^{-1}\| = O(1)$, we have $|\mathcal{A}_k^c| = O(1)$. As $\mathrm{Var}\{e_{[k]d}\} = O(n^{-1})$, the overall behavior of $\mathbf{e}_{[k]}$ can be studied by focusing on the entries $e_{[k]d}$ with $d \in \mathcal{A}_k$.

Note that as $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{(-i)}\| = O(n^{-\frac{1}{2}})$, for all $d \in \mathcal{A}_k$ with $k \in \{1, 2\}$, $\left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d = O(n^{-\frac{1}{2}})$ if $[\mathbf{C}_k \boldsymbol{\beta}]_d = O(n^{-\frac{1}{2}})$. Therefore, for $d \in \mathcal{A}_k$, $i \in \mathcal{C}_k$,

$$\begin{aligned}
\tilde{c}_i &= \phi_\tau \left(\frac{1 - y_i \boldsymbol{\beta}_{(-i)}^\mathsf{T} \boldsymbol{\mu}_k - y_i \boldsymbol{\beta}_{(-i)}^\mathsf{T} \mathbf{C}_k \mathbf{z}_i - y_i \beta_{(-i)0}}{\xi_{(-i)k}} \right) \\
&= \phi_\tau \left(t_i^{\{d\}} \right) + \phi_\tau' \left(t_i^{\{d\}} \right) \frac{\left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d [\mathbf{z}_i]_d}{\xi_{(-i)k}} + O(n^{-1})
\end{aligned}$$

where

$$t_i^{\{d\}} = \frac{1 - y_i \boldsymbol{\beta}_{(-i)}^\mathsf{T} \boldsymbol{\mu}_k - y_i \sum_{d' \neq d} \left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_{d'} [\mathbf{z}_i]_{d'} - y_i \beta_{(-i)0}}{\xi_{(-i)k}}.$$

It is important to remark that $t_i^{\{d\}}$ is independent of $[\mathbf{z}_i]_d$. Moreover, for $j \neq i$, $t_i^{\{d\}}$ can be divided into two parts: one independent of $[\mathbf{z}_j]_d$ and one of order $o(n^{-\frac{1}{2}})$. To see this, observe from the results in the previous subsection that, for $j \in \mathcal{C}_k$,

$$\begin{aligned}
\boldsymbol{\beta}_{(-i)} =& \boldsymbol{\beta}_{(-ij)} + \frac{c_{(-i)j}}{n} \left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}} (\bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}}^\mathsf{T} \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}})^{-1} \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}}^\mathsf{T} \right] \mathbf{x}_j + o\| \cdot \|(n^{-\frac{1}{2}}) \\
=& \boldsymbol{\beta}_{(-ij)} + \frac{\phi_\tau \left(t_{(-i)j}^{\{d\}} \right)}{n} \left[\mathbf{I}_p - \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}} (\bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}}^\mathsf{T} \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}})^{-1} \bar{\mathbf{X}}_{\mathcal{B}_{(-ij)}}^\mathsf{T} \right] \mathbf{C}_k [\mathbf{z}_j(1), \cdots, \mathbf{z}_j(d-1), 0, \\
& \mathbf{z}_j(d+1), \cdots, \mathbf{z}_j(p)]^\mathsf{T} + o\| \cdot \|(n^{-\frac{1}{2}})
\end{aligned}$$

where

$$t_{(-i)j}^{\{d\}} = \frac{1 - y_j \boldsymbol{\beta}_{(-ij)}^{\mathsf{T}} \boldsymbol{\mu}_k - y_j \sum_{d' \neq d} \left[\mathbf{C}_k \boldsymbol{\beta}_{(-ij)}\right]_{d'} [\mathbf{z}_j]_{d'} - y_j \beta_{(-ij)0}}{\xi_{(-ij)k}}.$$

The aforementioned conclusion is thus deduced by considering an approximation of $\xi_{(-i)k}$ defined with respect to the approximation of $\boldsymbol{\beta}_{(-i)}$ given in the second line of the above equation, which is thus independent of $[\mathbf{z}_j]_d$ and at a distance of $o(n^{-\frac{1}{2}})$ from $\xi_{(-i)k}$, and recalling that $\beta_{(-i)0} - \beta_{(-ij)0} = O(n^{-1})$.

Notice hence that

$$\frac{1}{n} \sum_{i \in \mathcal{C}_k} \tilde{c}_i [\mathbf{z}_i]_d = \frac{1}{n} \sum_{i \in \mathcal{C}_k} \phi_\tau \left(t_i^{\{d\}}\right) [\mathbf{z}_i]_d + \frac{1}{n} \sum_{i \in \mathcal{C}_k} \phi_\tau' \left(t_i^{\{d\}}\right) \frac{\left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d [\mathbf{z}_i]_d^2}{\xi_{(-i)k}} + O(n^{-1})$$

where

$$\frac{1}{n} \sum_{i \in \mathcal{C}_k} \phi_\tau' \left(t_i^{\{d\}}\right) \frac{\left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d [\mathbf{z}_i]_d^2}{\xi_{(-i)k}} = \frac{1}{n} \sum_{i \in \mathcal{C}_k} \left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d \mathbb{E}\left\{ \frac{\phi_\tau' \left(t_i^{\{d\}}\right) [\mathbf{z}_i]_d^2}{\xi_{(-i)k}} \right\} + o(n^{-\frac{1}{2}})$$

according to the discussion in the above paragraph on statistical independences between $t_i^{\{d\}}$ and $[\mathbf{z}_j]_d$. This convergence entails that

$$e_{[k]d} = \frac{1}{n} \sum_{i \in \mathcal{C}_k} \phi_\tau \left(t_i^{\{d\}}\right) [\mathbf{z}_i]_d + o(n^{-\frac{1}{2}})$$

as

$$\frac{1}{n} \sum_{i \in \mathcal{C}_k} \mathbb{E}_{\mathbf{x}_i}\{\tilde{c}_i [\mathbf{z}_i]_d\} = \frac{1}{n} \sum_{i \in \mathcal{C}_k} \left[\mathbf{C}_k \boldsymbol{\beta}_{(-i)}\right]_d \mathbb{E}\left\{ \frac{\phi_\tau' \left(t_i^{\{d\}}\right) [\mathbf{z}_i]_d^2}{\xi_{(-i)k}} \right\} + o(n^{-\frac{1}{2}}).$$

We get then

$$\begin{aligned} \mathrm{Var}\{e_{[k]d}\} &= \frac{1}{n^2} \sum_{i \in \mathcal{C}_k} \mathbb{E}\left\{ \phi_\tau \left(t_i^{\{d\}}\right)^2 \right\} + o(n^{-1}) \\ &= \frac{1}{n^2} \sum_{i \in \mathcal{C}_k} \mathbb{E}\left\{c_i^2\right\} + o(n^{-1}). \end{aligned}$$

To demonstrate that $e_{[k]d}$ follows asymptotically a normal distribution, we consider a solution $\boldsymbol{\beta}_{\{-d\}}$ that is independent of all $[\mathbf{z}_i]_d$, $i \in \{1, \ldots, n\}$, obtained on the data set where we replace all the $[\mathbf{z}_i]_d$ with $i \in \mathcal{C}_k$ with some independent copies $[\mathbf{z}_i']_d \overset{\mathcal{L}}{=} [\mathbf{z}_i]_d$. Let $c_{\{-d\}i}$, $i \in \{1, \ldots, n\}$ stand for the corresponding dual solutions. Similarly to the discussion between c_i and $c_{(-j)i}$, it can be derived that $c_{\{-d\}i} - c_i = O(n^{-\frac{1}{2}})$. Since $c_{\{-d\}i}$ are independent of all $[\mathbf{z}_i]_d$ with $i \in \mathcal{C}_k$,

we obtain that

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i\in\mathcal{C}_k}\phi_\tau\left(t_i^{\{d\}}\right)[\mathbf{z}_i]_d-\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_{\{-d\}i}[\mathbf{z}_i]_d\right)^2\right\}\\
=&\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i\in\mathcal{C}_k}\phi_\tau\left(t_i^{\{d\}}\right)[\mathbf{z}_i]_d\right)^2\right\}+\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_{\{-d\}i}[\mathbf{z}_i]_d\right)^2\right\}\\
&-2\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i\in\mathcal{C}_k}\phi_\tau\left(t_i^{\{d\}}\right)[\mathbf{z}_i]_d\right)\left(\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_{\{-d\}i}[\mathbf{z}_i]_d\right)\right\}\\
=&\frac{1}{n^2}\sum_{i\in\mathcal{C}_k}\left[\mathbb{E}\left\{\phi_\tau\left(t_i^{\{d\}}\right)^2\right\}+\mathbb{E}\left\{c_{\{-d\}i}^2\right\}-2\mathbb{E}\left\{\phi_\tau\left(t_i^{\{d\}}\right)c_{\{-d\}i}\right\}\right]\\
&-\frac{2}{n^2}\sum_{i\neq j\in\mathcal{C}_k}\mathbb{E}\left\{\phi_\tau\left(t_i^{\{d\}}\right)[\mathbf{z}_j]_d\right\}\mathbb{E}\left\{c_{\{-d\}i}\right\}\mathbb{E}\left\{[\mathbf{z}_i]_d\right\}\\
=&O(n^{-\frac{3}{2}}),
\end{aligned}$$

leading to

$$e_{[k]d}=\frac{1}{n}\sum_{i\in\mathcal{C}_k}c_{\{-d\}i}[\mathbf{z}_i]_d+o(n^{-\frac{1}{2}}).$$

As $c_{\{-d\}i}[\mathbf{z}_i]_d$, $i\in\mathcal{C}_k$, are i.i.d random variables, we conclude by the central limit theorem that

$$e_{[k]d}=\tilde{e}_{[k]d}+o(n^{-\frac{1}{2}})$$

where

$$\tilde{e}_{[k]d}\sim\mathcal{N}\left(0,\frac{\rho_k}{n}\mathbb{E}\{c_i^2|i\in\mathcal{C}_k\}\right).$$

We retrieve thus the results of Proposition 5.3.2.

To obtain Theorem 5.3.1, it suffices to use Lemma 2.3, according to which we have that, conditioned on $\boldsymbol{\beta}_{(-i)}$,

$$\mathbf{z}_i-\frac{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}$$

is independent of $\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i$ under the Gaussianity of $\mathbf{z}_i$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_i}\{\tilde{c}_i\mathbf{z}_i\}&=\mathbb{E}_{\mathbf{x}_i}\left\{\frac{\tilde{c}_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\right\}+\mathbb{E}_{\mathbf{x}_i}\{\tilde{c}_i\}\mathbb{E}_{\mathbf{x}_i}\left\{\mathbf{z}_i-\frac{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\right\}\\
&=\mathbb{E}_{\mathbf{x}_i}\left\{\frac{\tilde{c}_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\right\}=\mathbb{E}_{\mathbf{x}_i}\left\{\frac{\tilde{c}_i\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{z}_i}{\boldsymbol{\beta}_{(-i)}^\mathsf{T}\mathbf{C}_{(i)}\boldsymbol{\beta}_{(-i)}}\right\},
\end{aligned}$$

which concludes the proof of Theorem 5.3.1.

C.2 Sketch of proofs for Chapter 7

The theoretical results in Chapter 6 can in fact be proven by following a almost exact reasoning employed in the proofs for Chapter 5. Here we build the derivation upon the mathematical arguments in the previous section.

Recall that we are interested in the presumed case where $\hat{\boldsymbol{\beta}}$ is unique with finite norm and the classification scores $\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i$ of training samples are bounded, allowing for simplified proofs. As in the development of the previous section, we shall connect $\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}$ to c_i by establishing a "leave-one-out" version of $\hat{\boldsymbol{\beta}}$ that is independent of $\mathbf{x}_i, y_i$. To this end, we denote $\hat{\boldsymbol{\beta}}_{(-i)}$ the solution of the original optimization problem in (6.2) for $\mathbf{X}_{(-i)}\mathbf{y}_{(-i)} \equiv [\mathbf{x}_1 y_1, \ldots, \mathbf{x}_{i-1}y_{i-1}, \mathbf{x}_{i+1}y_{i+1}, \ldots, \mathbf{x}_n y_n] \in \mathbb{R}^{p\times(n-1)}$, all training data except the pair $(\mathbf{x}_i, y_i)$, such that by cancelling the derivative we obtain

$$\frac{1}{n}\sum_{j\neq i} y_j \psi(y_j \mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{(-i)})\mathbf{x}_j = 0. \tag{C.11}$$

Recall the definition of $\mathbf{c}$ in (6.3) and the fact that $\frac{1}{n}\sum_{i=1}^n c_i\mathbf{x}_i = 0$, a simple subtraction from (C.11) yields

$$\frac{1}{n}\sum_{j\neq i}\left(c_j - y_j\psi(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{(-i)})\right)\mathbf{x}_j + \frac{1}{n}c_i\mathbf{x}_i = 0. \tag{C.12}$$

Since ρ is a convex function, there exists a value $d_{(-i)j} < 0$ between $\psi'(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{(-i)})$ and $\psi'(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}})$ such that

$$c_j - c_{(-i)j} = d_{(-i)j}y_j\mathbf{x}_j^\mathsf{T}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})$$

where we denote $c_{(-i)j} \equiv y_j\psi(y_j\mathbf{x}_j^\mathsf{T}\hat{\boldsymbol{\beta}}_{(-i)})$. Plugging the above estimate back into (C.12) we deduce

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

with $\mathbf{D}_{(-i)} \in \mathbb{R}^{n-1}$ a diagonal matrix with $d_{(-i)j}$ being its diagonal entries. As $d_{(-i)j}$ is bounded away from infinity or zero, $\frac{1}{n}\mathbf{X}_{(-i)}\mathbf{D}_{(-i)}\mathbf{X}_{(-i)}$ is indeed invertible for $n/p > 1$ and

$$\left(-\frac{1}{n}\mathbf{X}_{(-i)}\mathbf{D}_{(-i)}\mathbf{X}_{(-i)}^\mathsf{T}\right)^{-1} = O_{\|\cdot\|}(1).$$

We observe thus that

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = O_{\|\cdot\|}(n^{-\frac{1}{2}}).$$

Consequently,

$$y_i\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i - \kappa_i y_i c_i = y_i\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i - \kappa_i\psi(y_i\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}_i) = y_i\hat{\boldsymbol{\beta}}_{(-i)}^\mathsf{T}\mathbf{x}_i \tag{C.13}$$

with

$$\kappa_i = \frac{1}{n}\mathbf{x}_i^\mathsf{T}\left(-\frac{1}{n}\mathbf{X}_{(-i)}\mathbf{D}_{(-i)}\mathbf{X}_{(-i)}^\mathsf{T}\right)^{-1}\mathbf{x}_i.$$

Hence,

$$y_i \hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x}_i = \mathrm{prox}_{\kappa_i}(y_i \hat{\boldsymbol{\beta}}_{(-i)}^\mathsf{T} \mathbf{x}_i)$$

where $\mathrm{prox}_{\kappa_i}(t) \equiv \mathrm{argmin}_{z \in \mathbb{R}} \left(\kappa_i \rho(z) + \frac{1}{2}(z-t)^2 \right)$. Letting

$$\phi_\tau(t) = \frac{\mathrm{prox}_\tau(t) - t}{\tau},$$

we get that

$$c_i = y_i \phi_{\kappa_i}(y_i \hat{\boldsymbol{\beta}}_{(-i)}^\mathsf{T} \mathbf{x}_i).$$

Notice also that, as

$$\mathbf{x}_j^\mathsf{T}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}) = \frac{1}{n} c_i \mathbf{x}_j^\mathsf{T} \mathbf{Q}_{(-i)} \mathbf{x}_i = O(n^{-\frac{1}{2}})$$

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

implying that

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

We note thus that

$$\kappa_i = \frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{(-i)} \mathbf{x}_i + O(n^{-\frac{1}{2}})$$

with

$$\mathbf{Q}_{(-i)} = \left(-\frac{1}{n} \sum_{j \neq i} \psi'(y_j \mathbf{x}_j^\mathsf{T} \hat{\boldsymbol{\beta}}_{(-i)}) \mathbf{x}_j \mathbf{x}_j^\mathsf{T} \right)^{-1}.$$

As $\mathbf{Q}_{(-i)}$ is independent of $\mathbf{x}_i$, we obtain from the trace lemma ([27, Lemma 14.2]) that

$$\frac{1}{n} \mathbf{x}_i^\mathsf{T} \mathbf{Q}_{(-i)} \mathbf{x}_i = \frac{1}{n} \operatorname{tr} \mathbf{C}\mathbf{Q}_{(-i)} + O(n^{-1}) = \frac{1}{n} \operatorname{tr} \mathbf{C}\mathbf{Q} + O(n^{-1})$$

where we denote

$$\mathbf{Q} \equiv \left(-\frac{1}{n} \sum_{i=1}^{n} \psi'(y_i \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}) \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right)^{-1}.$$

From this we remark that all κ_i, $i\{1, \ldots, n\}$, have asymptotically the same value $\kappa = frac1n \operatorname{tr} \mathbf{C}\mathbf{Q}$.

With the above arguments, Theorem 6.3.1 and Theorem 6.3.2 can be demonstrated with the same manipulation as in Subsection C.1.2. To derive further the results of Theorem 6.5.1, it suffices to find the deterministic limit $\bar{\kappa}$ of κ. To this end, we look now for a deterministic matrix $\bar{\mathbf{Q}} = \mathbf{R}^{-1}$ such that

$$\frac{1}{n}(\operatorname{tr} \mathbf{C}\mathbf{Q} - \operatorname{tr} \mathbf{C}\bar{\mathbf{Q}}) = o(1).$$

As $\mathbf{A}^{-1}-\mathbf{B}^{-1}=\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ for any two square matrices $\mathbf{A},\mathbf{B}$ of the same dimension, we observe that

$$\begin{aligned}
\frac{1}{n}(\operatorname{tr}\mathbf{CQ}-\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}) &= \frac{1}{n}\operatorname{tr}\mathbf{CQ}\left(\mathbf{R}+\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)\bar{\mathbf{Q}} \\
&= \frac{1}{n}\operatorname{tr}\mathbf{CQR}\bar{\mathbf{Q}}+\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{tr}\mathbf{CQ}\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}} \\
&= \frac{1}{n}\operatorname{tr}\mathbf{CQR}\bar{\mathbf{Q}}+\frac{1}{n^2}\sum_{i=1}^{n}\operatorname{tr}\mathbf{C}\frac{\mathbf{Q}_{-i}\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1-\frac{\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{(-i)}\mathbf{x}_i}\bar{\mathbf{Q}}+o(1) \\
&= \frac{1}{n}\operatorname{tr}\mathbf{CQR}\bar{\mathbf{Q}}+\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})}{1-\frac{\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})}{n}\operatorname{tr}\mathbf{CQ}_{(-i)}}\operatorname{tr}\mathbf{CQ}_{(-i)}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}+o(1) \\
&= \frac{1}{n}\operatorname{tr}\mathbf{CQR}\bar{\mathbf{Q}}+\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})}{1-\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\kappa}\operatorname{tr}\mathbf{CQC}\bar{\mathbf{Q}}+o(1). \qquad \text{(C.14)}
\end{aligned}$$

Since

$$\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})=\psi'\left(\operatorname{prox}_\kappa(y_i\hat{\boldsymbol{\beta}}_{(-i)}^\mathsf{T}\mathbf{x}_i)\right)+o(1),$$

by the same concentration arguments in the proofs of Chapter 5, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})}{1-\psi'(y_i\mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})\kappa}=\mathbb{E}_r\left\{\frac{\psi'\left(\operatorname{prox}_\kappa(r)\right)}{1-\psi'\left(\operatorname{prox}_\kappa(r)\right)\kappa}\right\}+o(1)$$

for some random variable r independent of κ and $r\overset{\mathcal{L}}{=}\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{x}$ with $\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{C})$ independent of $\hat{\boldsymbol{\beta}}$. It then follows from (C.14) that

$$\frac{1}{n}(\operatorname{tr}\mathbf{CQ}-\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}})=o(1)\Leftrightarrow\frac{1}{n}\operatorname{tr}\mathbf{CQ}\left(\mathbf{R}-\mathbb{E}_r\left\{\frac{\psi'\left(\operatorname{prox}_\kappa(r)\right)}{1-\psi'\left(\operatorname{prox}_\kappa(r)\right)\kappa}\mathbf{C}\right\}\bar{\mathbf{Q}}\right)=o(1).$$

In view of this result, we obtain that

$$\mathbf{R}=\mathbb{E}\left\{\frac{-\psi'\left(\operatorname{prox}_{\bar{\kappa}}(r)\right)}{1-\psi'\left(\operatorname{prox}_{\bar{\kappa}}(r)\right)\bar{\kappa}}\right\}\mathbf{C}$$

with $\bar{\kappa}>0$ uniquely given by

$$\bar{\kappa}=\frac{1}{n}\operatorname{tr}\mathbf{R}^{-1}\mathbf{C}.$$

With the above equation of $\bar{\kappa}$ at hand, we thus deduce Theorem 6.5.1.

Appendix D

Résumé (Français)

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle centré sur le traitement automatique des données. À partir d'un ensemble d'échantillons de données et d'une tâche d'apprentissage, les algorithmes d'apprentissage automatique extraient des informations pertinentes pour la tâche à partir de l'ensemble de données sans instructions explicites. Naturellement, les performances des algorithmes d'apprentissage automatique sont limitées par la taille du jeu de données en entrée. L'augmentation rapide de la capacité de calcul a permis de collecter et de manipuler des ensembles de données volumineux dotés de nombreuses fonctionnalités, ce qui a permis le succès de nombreuses applications de méthodes d'apprentissage automatique, telles que la classification d'images, la reconnaissance de la parole et la prédiction de gènes. Même si des performances surhumaines ont été obtenues sur certaines tâches grâce à la puissance du big data, elles sont principalement réalisées à l'aide de modèles d'apprentissage supervisés et nécessitent une quantité considérable d'échantillons étiquetés. Le processus d'étiquetage coûteux et l'accès limité aux données dans de nombreux domaines appellent des approches d'apprentissage plus efficaces et plus flexibles. Pour améliorer les méthodes d'apprentissage actuelles, il est nécessaire de les comprendre à un niveau profond. Cependant, la nature non linéaire des algorithmes d'apprentissage, qui est à l'origine de leur succès empirique, les rend également théoriquement difficiles à étudier. En effet, la plupart des algorithmes d'apprentissage automatique, même les plus populaires, ont été motivés par un raisonnement intuitif et justifiés par des arguments heuristiques.

Depuis longtemps, on s'aperçoit que l'apprentissage sur de grandes données présente des défis uniques pour lesquels le terme *malédiction de la dimensionnalité* a été utilisé. Les arguments intuitifs qui sous-tendent la proposition de nombreux algorithmes d'apprentissage ne sont valables que pour des données de petites dimensions. Un phénomène important de la malédiction de la dimensionnalité est la *concentration des distances*, qui fait référence à la tendance des "distances" paire-à-paire entre les vecteurs de données à devenir indiscernables dans la limite des grandes dimensions. Étant donné que de nombreuses techniques d'apprentissage reposent sur la relation entre la proximité géométrique et l'affinité entre les données, leur validité est remise en question par ce phénomène de concentration des distances. Par conséquent, de nombreux phénomènes contre-intuitifs peuvent se produire, dont l'explication appelle une compréhension plus profonde de l'apprentissage en grande dimension. Malgré cette nécessité impérieuse de dévoiler le processus d'apprentissage des grandes données, la recherche théorique à cet égard est plutôt sous-développée dans la littérature. La plupart des analyses existantes de

techniques d'apprentissage suppose notamment que le nombre n d'échantillons de données est infiniment grand par rapport à leur dimension p, c'est-à-dire $n/p \to \infty$, une hypothèse qui ne convient guère lorsque la dimension est elle-même trop importante pour être considérée comme négligeable par rapport au nombre d'échantillons de données. L'objectif de cette thèse est d'analyser et d'améliorer les méthodes d'apprentissage dans le régime moderne des n, p grands et comparables.

Puisque les résultats d'apprentissage sont des variables aléatoires dépendant des données d'entrée, qui ne convergent vers des valeurs déterministes que lorsque $n \gg p$, l'analyse des algorithmes d'apprentissage pour des n, p comparables nécessite la tâche non triviale de caractériser leur caractère aléatoire. Prenons l'exemple de l'analyse discriminante linéaire (LDA), une méthode d'apprentissage simple et standard. La méthode LDA aborde le problème de l'apprentissage en supposant que les instances de données $(\mathbf{x}, y)$, avec $\mathbf{x} \in \mathbb{R}^p$ les vecteurs de caractéristiques et $y = \pm 1$ les étiquettes de classe, suivent un modèle de mélange gaussien avec des covariances identiques, c'est-à-dire pour $y = (-1)^k$ avec $k = \{1, 2\}$, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{C})$ (où nous supposons que $\mathbf{C}$ est de rang complet). Sous cette hypothèse, la solution Bayes-optimale consiste à affecter une observation $\mathbf{x}$ à la classe ± 1 par le signe de $\boldsymbol{\beta}^\mathsf{T}\mathbf{x} - c$ pour une constante de seuil c, où $\boldsymbol{\beta} = \mathbf{C}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Comme les paramètres statistiques $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ et $\mathbf{C}$ sont normalement inconnus en pratique, ils sont estimés à partir d'un ensemble d'échantillons de données $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ pour obtenir $\boldsymbol{\beta} = \hat{\mathbf{C}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, où $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$ est généralement l'estimation de la probabilité maximale (MLE) de $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{C})$ ou d'autres estimations. Bien que les performances de LDA soient garanties optimales dans la limite $n \gg p$ où $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}}) \to (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{C})$ pour tout estimateur consistant $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$, on ne peut pas en dire autant du régime où n, p sont proportionnellement grands. En effet, même avec le MLE, pour lequel nous avons des expressions assez simples de $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\mathbf{C}})$ (également appelés moyenne et covariance empiriques):

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i, \quad k = \{1, 2\},$$

où on note $i \in \mathcal{C}_k$ pour $i \in \{1, \ldots, n\}$ tels que $y_i = (-1)^k$; et

$$\hat{\mathbf{C}} = \frac{1}{n} \left[\sum_{i \in \mathcal{C}_1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^\mathsf{T} + \sum_{j \in \mathcal{C}_2} (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)^\mathsf{T} \right],$$

le comportement statistique de $\boldsymbol{\beta}$ est compliqué à caractériser pour des rapports n/p non triviaux, principalement en raison de l'existence de $\hat{\mathbf{C}}^{-1}$ dans l'expression de $\boldsymbol{\beta}$.

Comme expliqué précédemment, la grande dimensionnalité des données modernes induit le besoin d'études théoriques avancées sur les performances des algorithmes d'apprentissage, loin de la limite asymptotique conventionnelle $n/p \to \infty$, à laquelle les paramètres appris, tels que $\boldsymbol{\beta} = \hat{\mathbf{C}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$ dans l'exemple ci-dessus du LDA, deviennent des constantes déterministes. Néanmoins, cette grande dimensionalité offre en fait des avantages techniques. En effet, bien que le comportement statistique de $\hat{\mathbf{C}}^{-1}$, tel que la distribution de ses valeurs propres et des vecteurs propres associés, soit difficile d'accès pour des n, p modérés, un certain nombre de chercheurs pionniers se sont intéressés aux propriétés statistiques de matrices aléatoires telles que $\hat{\mathbf{C}}^{-1}$ dans la limite où p est large et n/p reste non trivial. En effet, en explorant les degrés de liberté supplémentaires fournis par la grande dimensionnalité des données, Marchenko et Pastur, dans [1], ont tout d'abord démontré que l'histogramme des valeurs propres de la covariance

empirique[1] issue de vecteurs de données $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ converge vers une certaine distribution continue déterministe, maintenant appelée distribution de Marchenko–Pastur. L'extension au cas où la covariance population est autorisée à être autre que la matrice d'identité peut être trouvée dans les travaux [2, 3] de Silverstein et Bai. Évidemment, connaître les propriétés spectrales de la matrice de covariance empirique $\hat{\mathbf{C}}$ équivaut à connaître celles de son inverse. En fait, de nombreuses investigations spectrales sur des matrices aléatoires telles que $\hat{\mathbf{C}}$ sont effectuées par le biais de manipulations techniques impliquant leur inverse. Sur la base des résultats de la théorie des matrices aléatoires (RMT), la performance de LDA a récemment été examinée dans [4], et sa variante plus élaborée QDA (analyse discriminante quadratique) dans [5].

La solution de LDA est plutôt pratique pour effectuer des analyses théoriques en raison de sa forme explicite et du fait que sa seule non-linéarité est due à l'inverse de la matrice de covariance empirique $\hat{\mathbf{C}}^{-1}$, un objet largement étudié en RMT. La plupart des techniques d'apprentissage, telles que les méthodes à noyau, impliquent des non-linéarités plus complexes. Une autre complication dans l'analyse des systèmes d'apprentissage est qu'il peut ne pas exister d'expression explicite des résultats du système. L'absence de solution close est en réalité commune à de nombreuses méthodes d'apprentissage largement utilisées telles que la régression logistique, les machines à vecteurs de support (SVM) et les réseaux de neurones, pour lesquels les solutions sont définies comme un point de minimisation (local ou global) d'une certaine fonction de perte. Comme les résultats de RMT concernent généralement les propriétés statistiques de certains modèles de matrices aléatoires explicites spécifiques, ils ne sont pas adaptés pour caractériser des solutions implicites à des problèmes d'optimisation impliquant ces mêmes matrices aléatoires. D'autres approches sont donc nécessaires pour l'étude d'algorithmes d'apprentissage à solutions implicites.

À cet égard, la technique de perturbation "leave-one-out" s'est avérée efficace par une série de contributions. Le comportement statistique de la régression robuste avec M-estimateurs, qui ne suppose généralement pas l'existence d'une solution de forme close, est décrit dans [6, 7], en utilisant cette procédure de perturbation. Dans le même ordre d'idées, les travaux de [8, 9] sont axés sur la méthode de régression logistique pour la classification. L'idée principale de ces études est d'établir des équations statistiques des paramètres appris en capitalisant sur le fait que les résultats des algorithmes restent pratiquement inchangés après avoir exclu (i) un échantillon de l'ensemble de données d'apprentissage ou (ii) une caractéristique du vecteur de caractéristiques. Cette approche "double leave-one-out" est appliquée dans ces travaux en supposant que tous les échantillons de données sont centrés avec les entités i.i.d. gaussiennes (c'est-à-dire, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$), ce qui justifie notamment l'étape 'leave-one-feature-out' car toutes les caractéristiques sont statistiquement équivalentes et indépendantes. Contrairement aux modèles mixtes (comme celui considéré dans LDA), il n'y a pas de séparation de classe naturelle dans l'unique classe $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. Afin d'étudier les problèmes de classification dans ce contexte, les auteurs de [8, 9] ont imposé l'existence d'un signal de séparation de classe à l'intérieur du groupe $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. En tant que tels, les scénarios de classification courants avec des modèles de classe distincts (représentés par des composants différents dans les modèles de mélange) ne sont jusqu'à présent pas couverts par ce type d'analyse.

Contributions et organisation:

[1] Ceci fait référence à la mesure spectrale définie par $\frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i}(t)$ où λ_i sont les valeurs propres de la matrice de covariance empirique.

Le paradigme actuel du big data jette les bases du développement de nouveaux outils mathématiques pour l'analyse des algorithmes d'apprentissage dans le régime moderne où n, p sont de taille comparable. Contrairement aux analyses existantes dans ce régime, les approches techniques développées dans cette thèse exploitent à la fois des outils avancés de la théorie des matrices aléatoires et des arguments de type "leave-one-out". Grâce à la combinaison des avantages de la théorie des matrices aléatoires pour la manipulation de données structurées et de la puissance de la manipulation "leave-one-out" pour traiter des systèmes d'apprentissage complexes, nous sommes en mesure de mener des analyses plus complexes d'algorithmes d'apprentissage automatique dans des modèles de mélanges réalistes. Ces analyses entraînent des conséquences importantes dans l'application de méthodes d'apprentissage, dont certaines sont observées depuis longtemps sans avoir été bien comprises, d'autres inconnues des praticiens, voire contraires aux idées reçues. Nos analyses étant caractérisées de manière complète par les résultats de l'apprentissage, ces problèmes peuvent parfois être directement résolus par de simples mesures de correction, telles que la normalisation (ou le rééchelonnement) ou l'amélioration de la paramétrisation. Dans certains scénarios, des analyses en grandes dimensions peuvent même détecter des défauts fondamentaux dans la conception des algorithmes d'apprentissage et inspirer des approches supérieures, comme cela a été fait dans [10] et dans le travail qui a suivi [11], dans le cadre des contributions de cette thèse. De manière remarquable, les résultats théoriques dérivés au cours de la thèse prédisent de près les performances d'apprentissage sur des ensembles de données à la fois synthétiques et réels, ce qui suggère l'adéquation des modèles de données de mélange pour décrire le scénario d'apprentissage dans des applications réelles. Cette observation est notamment corroborée par les conclusions de [12] et de [13], où les auteurs démontrent que, sous certaines hypothèses (très légères) de "mélanges de données concentrées", une série d'objets aléatoires concernant la matrice de covariance empirique converge dans le régime des grands n, p vers une même limite, et ce quelle que soit la distribution des données.

Voici les contributions principales de cette thèse organisées en chapitres:

- Sur le plan technique, l'apport de cette thèse réside dans le développement d'une approche combinant les techniques de RMT et de la procédure de "leave-one-out", adaptable à l'analyse d'une série de problèmes d'apprentissage importants, comme en témoignent nos études présentées dans cette thèse. Les outils basiques de RMT et l'idée fondamentale de la manipulation "leave-one-out" sont présentés dans Chapitre 2, avant une démonstration sur comment les combiner pour des analyses plus complexes à l'aide d'un exemple illustratif.

- Passant aux contributions principales, la première partie concerne l'apprentissage semi-supervisé sur des graphes, constitués de Chapter 3 et Chapter 4:

  - Dans Chapter 3, nous présentons l'analyse de grande dimension sur une famille d'algorithmes d'apprentissage semi-supervisés basés sur des graphes, souvent appelés méthodes de régularisations laplaciennes. Notre analyse explique pourquoi la plupart de ces algorithmes semi-supervisés couramment utilisés échouent en grande dimension, à l'exception de celui avec la matrice laplacienne de la marche aléatoire (également appelé algorithme PageRank). L'étude révèle également plusieurs conséquences importantes induites par la grande dimensionnalité des données. Des mesures de correction et des conseils pratiques sont données en vue de ces résultats. Une conclusion très importante de cette analyse est que les performances de tous ces algorithmes

manifestent une croissance négligeable de performance avec l'ajout des données non-étiquetées. Ceci suggère l'existence d'un défaut fondamental dans l'approche de régularisation laplacienne, qui la rend inadéquate pour réaliser un apprentissage semi-supervisé efficace sur des données de grande dimension. Les résultats de ce chapitre sont rappelés de

> X. Mai, R. Couillet,"A Random Matrix Analysis and Improvement of Semi-Supervised Learning for Large Dimensional Data", Journal of Machine Learning Research, vol. 19, no. 79, pp. 1-27, 2018.
>
> X. Mai, R. Couillet,"The Counterintuitive Mechanism of Graph-based Semi-Supervised Learning in the Big Data Regime", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

– Suite à la dernière remarque de l'analyse des régularisations laplaciennes, nous proposons dans Chapitre 4 un nouvel algorithme de régularisation capable d'apprendre efficacement à la fois des données étiquetées et non-étiquetées de grande dimension, dans un sens que l'exactitude de la classification augmente d'une manière non-négligeable lorsque un des ratios $n_{[l]}/p, n_{[u]}/p$ de taille pour les données étiquetées ($[l]$) et les données non-étiquetées ($[u]$) est plus grand. L'algorithme proposé présente un avantage indiscutable sur les méthodes laplaciennes, car les performances de ces dernières ne dépendent que de $n_{[l]}/p$, la taille relative à la dimension des données étiquetées. Notre nouvelle approche implique une opération de centrage sur les similitudes. Une analyse approfondie des performances est également effectuée. La méthode proposée et son analyse de performance présentée dans ce chapitre sont basées sur les contributions suivantes

> X. Mai, R. Couillet, "Revisiting and Improving Semi-Supervised Learning: A Large Dimensional Approach", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019.
>
> X. Mai, R. Couillet, "Consistent Semi-Supervised Graph Regularization for High Dimensional Data", submitted to Journal of Machine Learning Research, 2019.

- La deuxième partie est consacrée à l'étude des algorithmes sans solution explicite, Chapitre 5 étant dédié à la méthode de SVMs et Chapitre 6 à la régression logistique.

  – La méthode des machines à vecteurs de support doit son nom au fait que le paramètre appris $\boldsymbol{\beta} \in \mathbb{R}^p$ est déterminé par un sous-ensemble d'échantillons d'apprentissage, appelés vecteurs de support. En fait, puisque nous avons $\boldsymbol{\beta} = \sum_{i=1}^n c_i \mathbf{x}_i$ où $c_i \geq 0$, un vecteur $\mathbf{x}_i$ de données d'entraînement associé à un non-nul c_i est un vecteur de support. Nous caractérisons dans Chapitre 5 le comportement des vecteurs de support en grandes dimensions via la distribution statistique de c_i. Ensuite, nous montrons comment la distribution statistique de $\boldsymbol{\beta}$ est liée à celle de c_i, ce qui nous permet de tirer des remarques importantes sur l'impact de l'hyperparamètre dans la méthode SVM. Cette analyse est présentée dans l'article

  > X. Mai, R. Couillet, "Statistical Behavior and Performance of Support Vector Machines for Large Dimensional Data", in preparation, 2019.

– La régression logistique est l'un des algorithmes définis par le principe de la minimisation du risque empirique, avec une perte de vraisemblance logarithmique négative. Comme la régression logistique donne une estimation du maximum de vraisemblance pour les paramètres $\boldsymbol{\beta}, \beta_0$, option par défaut et généralement considérée comme optimale lorsque l'hypothèse de distribution des données est satisfaite, nous proposons de vérifier l'optimalité de la régression logistique par une analyse conjointe des algorithmes de la minimisation du risque empirique avec fonctions lisses de perte. De manière remarquable, nos résultats prouvent que, contrairement à la conviction générale, la régression logistique basée sur la maximization de vraisemblance ne produit pas la meilleure performance de classification. Nous élaborons également des stratégies d'amélioration de ces algorithmes à partir de nos résultats théoriques avant d'en discuter les limites. Le chapitre est basé sur les contributions suivantes

  X. Mai, Z. Liao, R. Couillet, "A Large Scale Analysis of Logistic Regression: Asymptotic Performance and New Insights", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, 2019.

  X. Mai, Z. Liao, "High Dimensional Classification via Empirical Risk Minimization: Statistical Analysis and Optimality", in preparation, 2019.

Bibliography

[1] V. Marchenko and L. Pastur, "The eigenvalue distribution in some ensembles of random matrices," *Math. USSR Sbornik*, vol. 1, pp. 457–483, 1967.

[2] J. W. Silverstein and Z. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate analysis*, vol. 54, no. 2, pp. 175–192, 1995.

[3] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. Springer, 2010, vol. 20.

[4] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE transactions on signal processing*, vol. 63, no. 11, pp. 2804–2814, 2015.

[5] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "Asymptotic performance of regularized quadratic discriminant analysis based classifiers," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

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

[7] D. Donoho and A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing," *Probability Theory and Related Fields*, vol. 166, no. 3-4, pp. 935–969, 2016.

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

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

[10] X. Mai and R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[11] ——, "Consistent semi-supervised graph regularization for high dimensional data," 2019.

[12] C. Louart and R. Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," 2019.

[13] M. E. A. Seddik, M. Tamaazousti, and R. Couillet, "Kernel random matrices of large concentrated data: the example of gan-generated images," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7480–7484.

[14] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and remote sensing*, vol. 32, no. 5, pp. 1087–1095, 1994.

[15] F. G. Cozman, I. Cohen, and M. Cirelo, "Unlabeled data can degrade classification performance of generative classifiers." in *Flairs conference*, 2002, pp. 327–331.

[16] S. Ben-David, T. Lu, and D. Pál, "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning." in *COLT*, 2008, pp. 33–44.

[17] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[20] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[21] R. Couillet, F. Benaych-Georges *et al.*, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[22] H. Huang, "Asymptotic behavior of support vector machine for spiked population model," *Journal of Machine Learning Research*, vol. 18, no. 45, pp. 1–21, 2017.

[23] N. E. Karoui, "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results," *arXiv preprint arXiv:1311.2445*, 2013.

[24] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

[25] P. Billingsley, *Probability and Measure*, 3rd ed. Hoboken, NJ: John Wiley and Sons, Inc., 1995.

[26] R. Couillet, M. Debbah, and J. Silverstein, "A deterministic equivalent for the analysis of correlated mimo multiple access channels," *IEEE Transactions on Information Theory*, vol. 6, no. 57, pp. 3493–3514, 2011.

[27] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[28] M. S. T. Jaakkola and M. Szummer, "Partially labeled classification with markov random walks," *International Conference in Neural Information Processing Systems*, vol. 14, pp. 945–952, 2002.

[29] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Citeseer, Tech. Rep., 2002.

[30] K. Avrachenkov, P. Goncalves, A. Mishenin, and M. Sokol, "Generalized optimization framework for graph-based semi-supervised learning," *arXiv preprint arXiv:1110.4278*, 2011.

[31] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on Machine Learning*, vol. 3, 2003, pp. 912–919.

[32] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *International Conference on Computational Learning Theory (COLT)*. Springer, 2004, pp. 624–638.

[33] T. Joachims *et al.*, "Transductive learning via spectral graph partitioning," in *International Conference on Machine Learning*, vol. 3, 2003, pp. 290–297.

[34] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," vol. 16, 2004, pp. 321–328.

[35] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[36] A. B. Goldberg, X. Zhu, A. Singh, Z. Xu, and R. Nowak, "Multi-manifold semi-supervised learning," 2009.

[37] A. Moscovich, A. Jaffe, and B. Nadler, "Minimax-optimal semi-supervised regression on unknown manifolds," *arXiv preprint arXiv:1611.02221*, 2016.

[38] L. Wasserman and J. D. Lafferty, "Statistical analysis of semi-supervised regression," in *International Conference on Neural Information Processing Systems*, 2008, pp. 801–808.

[39] P. J. Bickel, B. Li *et al.*, "Local polynomial regression on unknown manifolds," in *Complex Datasets and Inverse Problems*. Institute of Mathematical Statistics, 2007, pp. 177–186.

[40] A. Globerson, R. Livni, and S. Shalev-Shwartz, "Effective semi-supervised learning on manifolds," in *International Conference on Learning Theory (COLT)*, 2017, pp. 978–1003.

[41] S. K. Narang, A. Gadde, and A. Ortega, "Signal processing techniques for interpolation in graph structured data," in *IEEE International Conference Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 5445–5449.

[42] S. K. Narang, A. Gadde, E. Sanou, and A. Ortega, "Localized iterative methods for interpolation in graph structured data," in *Global Conference on Signal and Information Processing*. IEEE, 2013, pp. 491–494.

[43] A. Gadde, A. Anis, and A. Ortega, "Active semi-supervised learning using sampling theory for graph signals," in *International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 492–501.

[44] A. Anis, A. El Gamal, S. Avestimehr, and A. Ortega, "Asymptotic justification of bandlimited interpolation of graph signals for semi-supervised learning," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5461–5465.

[45] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *arXiv preprint arXiv:1510.03547*, 2015.

[46] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International conference on database theory*. Springer, 1999, pp. 217–235.

[47] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*. Springer, 2001, pp. 420–434.

[48] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?" in *26th Internat. Conference on Very Large Databases*, 2000, pp. 506–515.

[49] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.

[50] F. Angiulli, "On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness," *Journal of Machine Learning Research*, vol. 18, no. 170, pp. 1–60, 2018. [Online]. Available: http://jmlr.org/papers/v18/17-151.html

[51] B. Nadler, N. Srebro, and X. Zhou, "Semi-supervised learning with the graph laplacian: the limit of infinite unlabelled data," in *International Conference on Neural Information Processing Systems*, 2009, pp. 1330–1338.

[52] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.

[53] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT press, 2006.

[54] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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

[56] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 90–94.

[57] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[58] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[59] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.

[60] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[61] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[63] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

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

[65] S. Ben-David, N. Eiron, and P. M. Long, "On the difficulty of approximately maximizing agreements," *Journal of Computer and System Sciences*, vol. 66, no. 3, pp. 496–514, 2003.

[66] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?" *Neural Computation*, vol. 16, no. 5, pp. 1063–1076, 2004.

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

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

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

[70] S. Portnoy *et al.*, "Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. i. consistency," *The Annals of Statistics*, vol. 12, no. 4, pp. 1298–1309, 1984.

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

[72] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

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

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

[75] G. Fumera, R. Fabio, and S. Alessandra, "A theoretical analysis of bagging as a linear combination of classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1293–1299, 2008.

[76] H. T. Ali, A. Kammoun, and R. Couillet, "Random matrix asymptotics of inner product kernel spectral clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2441–2445.

[77] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[78] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[79] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[80] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To transfer or not to transfer," in *NIPS 2005 workshop on transfer learning*, vol. 898, 2005, p. 3.

[81] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[82] M. A. Woodbury, "Inverting modified matrices," *Memorandum report*, vol. 42, no. 106, p. 336, 1950.

[83] R. Walter, "Real and complex analysis," 1987.

[84] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.

[85] Z.-D. Bai, J. W. Silverstein *et al.*, "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices," *The Annals of Probability*, vol. 26, no. 1, pp. 316–345, 1998.

[86] F. Benaych-Georges and R. Couillet, "Spectral analysis of the gram matrix of mixture models," *ESAIM: Probability and Statistics*, vol. 20, pp. 217–237, 2016.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Méthodes des matrices aléatoires pour l'apprentissage en grandes dimensions.

Mots clés : Apprentissage en grandes dimensions, théorie des matrices aléatoires, apprentissage semi-supervisé, machines à vecteurs de support, régression logistique.

Résumé : Le défi du BigData entraîne un besoin pour les algorithmes d'apprentissage automatisé de s'adapter aux données de grande dimension et de devenir plus efficace. Récemment, une nouvelle direction de recherche est apparue qui consiste à analyser les méthodes d'apprentissage dans le régime moderne où le nombre n et la dimension p des données sont grands et du même ordre. Par rapport au régime conventionnel où $n \gg p$, le régime avec n, p sont grands et comparables est particulièrement intéressant, car les performances d'apprentissage dans ce régime restent sensibles à l'ajustement des hyperparamètres, ouvrant ainsi une voie à la compréhension et à l'amélioration des techniques d'apprentissage pour ces données de grande dimension.

L'approche technique de cette thèse s'appuie sur des outils avancés de statistiques de grande dimension, nous permettant de mener des analyses allant au-delà de l'état de l'art. La première partie de la thèse est consacrée à l'étude de l'apprentissage semi-supervisé sur des grandes données. Motivés par nos résultats théoriques, nous proposons une alternative supérieure à la méthode semi-supervisée de régularisation laplacienne. Les méthodes avec solutions implicites, comme les SVMs et la régression logistique, sont ensuite étudiées sous des modèles de mélanges réalistes, fournissant des détails exhaustifs sur le mécanisme d'apprentissage. Plusieurs conséquences importantes sont ainsi révélées, dont certaines sont même en contradiction avec la croyance commune.

Title: Methods of random matrices for large dimensional statistical learning.

Keywords: Large dimensional learning, random matrix theory, semi-supervised learning, support vector machines, logistic regression.

Abstract: The BigData challenge induces a need for machine learning algorithms to evolve towards large dimensional and more efficient learning engines. Recently, a new direction of research has emerged that consists in analyzing learning methods in the modern regime where the number n and the dimension p of data samples are commensurately large. Compared to the conventional regime where $n \gg p$, the regime with large and comparable n, p is particularly interesting as the learning performance in this regime remains sensitive to the tuning of hyperparameters, thus opening a path into the understanding and improvement of learning techniques for large dimensional datasets.

The technical approach employed in this thesis draws on several advanced tools of high dimensional statistics, allowing us to conduct more elaborate analyses beyond the state of the art. The first part of this dissertation is devoted to the study of semi-supervised learning on high dimensional data. Motivated by our theoretical findings, we propose a superior alternative to the standard semi-supervised method of Laplacian regularization. The methods involving implicit optimizations, such as SVMs and logistic regression, are next investigated under realistic mixture models, providing exhaustive details on the learning mechanism. Several important consequences are thus revealed, some of which are even in contradiction with common belief.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France