Advanced Random Matrix Methods for Machine Learning

Malik Tiomoko

université
PARIS-SACLAY

Advanced Random Matrix Methods for Machine Learning

Méthodes avancées de la théorie des matrices aléatoires pour l'apprentissage automatique

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Traitement du signal et des images
Unité de recherche: Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue à Gif-sur-Yvette, le 7 Octobre 2021, par

Malik TIOMOKO

Composition du jury

Alexandre GRAMFORT Directeur de recherche, INRIA/ Université Paris-Saclay	Président
Alfred HERO Professeur, Université du Michigan	Rapporteur & Examinateur
Mylène MAIDA Professeure des universités, Université de Lille	Rapporteur & Examinatrice
Rémi BARDENET Chargé de recherche, CNRS / Université de Lille	Examinateur
Balazs KEGL Directeur de recherche, Noah's Ark Lab, Huawei Research	Examinateur

Direction de la thèse

Romain COUILLET Professeur, Université Grenoble-Alpes	Directeur de thèse
Frédéric PASCAL Professeur, CentraleSupelec	Co-directeur de thèse

Thèse de doctorat

NNT: 2021UPASG067

Contents

List of Acronyms

RMT Random Matrix Theory . . . 2

ML Machine Learning . . . 7

SVM Support Vector Machine . . . 8

AI Artificial Intelligence . . . 2

MTL Multi-Task Learning . . . 1

LS-SVM Least Square Support Vector Machine . . . 1

CLT Central Limit Theorem . . . 6

SSL Semi-Supervised Learning . . . 7

ESD Empirical Spectral Distribution . . . 15

LSD Limiting Spectral Distribution . . . 16

GAN Generative Adversarial Network . . . 29

SAR Synthetic Aperture Radar . . . 50

Acknowledgments

I would like to sincerely thank my thesis director Romain with whom I spent three wonderful years. He passed on to me all his passion for research and especially for the field of random matrix theory. I learned so much from him, whether it was in terms of work methodology, writing scientific articles or oral communication. He was very patient and always knew how to motivate me when almost all hope was lost. I won't forget those late nights or Sundays when Romain was correcting stuff on the article or those moments when a work session in his office could bring up lots of ideas for writing articles. His energy blew me away. He facilitated my professional insertion by introducing me to renowned researchers in machine learning and signal processing, and by inviting me to dinners with colleagues in random matrix theory.

I would also like to thank my co-supervisor Frédéric with whom I was able to work during my third year of thesis. He knew how to listen to me and adapt to my requests and my subjects of interest and I am grateful to him for that.

I would like to thank the reviewers of my thesis Alfred Hero and Mylène Maida. I was so honoured to have such experienced professors in the field of statistics as my reviewers and I was in no way disappointed as their remarks were beneficial for the improvement of the thesis manuscript. I particularly note all the suggestions made for the perspectives of this thesis which have inspired me a lot. I am very grateful to the members of the jury Alexandre Gramfort, Balazs Kegl and Remi Bardenet for their excellent questions and pertinent remarks on the thesis.

A special thanks to the Gipsa Lab where I spent years during my thesis and in particular to the CICS team which later became GAIA. I met some excellent researchers with whom I had book-clubs, interesting meetings and extraordinary collaborations. I would like to thank Florent Bouchard, Guillaume Ginholac, Steeve Zozor, Eric Moisan who have been exceptional collaborators.

To the doctoral students and friends with whom I spent these three years (Hafiz, Zhenyu, Xiaoyi, Mohamed, Lorenzo, Cosme, Charles, Cyprien, Tayeb, Minh Toan, Victor, Hugo, Roza, Yigit, Juliette, Julien, Fateme, Jeanne, Pierre, Imane, Ekkehard, Elaheh, Dawood, Cyril, Sara, Chika and so many others that I will not be able to mention due to lack of space), I would like to thank you all for the wonderful time you gave me. I have been incredibly lucky to be in such a close-knit group that brings joy to the work at all times.

I will finish with my family, who have been a great inspiration to me and on whom I have always been able to rely in moments of joy and pain. My father has always tried to show me the path of excellence, to the point of defending his thesis at the age of 60, to show us the example. My mother, what more can I say than cry, so much she shared my heart. She suffered like me in painful moments and felt the same joy as me in happy moments. She never failed to support me. Thank you Mum and Dad. To my brother

Hafiz who always showed me the right way in all areas (professional and social), I really say thank you. I thank my sister Chérifa, my second mother, always very close to me and always full of energy to make me feel the greatest joys of the world. To my brother Kemal, I thank you for the good moments of joy and laughter that we shared together, I hope that this thesis will inspire you to do even greater things. I would also like to thank all the aunts and uncles who have supported me during all these years. I dedicate this thesis especially to all this wonderful family.

Chapter 1

Introduction

Contents

1.1 From curse to blessing of dimensionality

The "*curse of dimensionality*". Most signal processing and machine learning methods (e.g. statistical tests, parameter estimators, classification, regression, etc) are based on non-trivial functionals of n observed random signal vectors. Under the assumption that the number of available data n is overwhelmingly larger than their dimension p, some theoretical results and insights can be derived because a deterministic behavior sometimes arises when $n \to \infty$, which simplifies the problem as exemplified by the celebrated law of large numbers and the central limit theorem. However, as already shown in the literature (Wigner, 1958; Marčenko & Pastur, 1967), some long-held beliefs supported by classical results (and intuitions) break down when n, p are comparably large, a problem often referred to as "*curse of dimensionality*". Under the current big data paradigm which sees loads of data being produced, exchanged and stored, we constantly face the situation where not only the size n but also the dimension of the data p are large.

On the other hand, the last decade has seen a vast increase both in the diversity of applications to which machine learning is applied (transfer learning, privacy, fairness, etc) and to the practical deployment of these applications. Therefore, machine learning is no longer just the engine behind ad placement and spam filters: it is now used to filter loan applicants (Fernández, 2019), deploy police officers (Rudin, 2013), make decision in criminal justice (Berk & Hyatt, 2015), etc making rapid headway into socio-technical systems. Algorithmic biases and misunderstood algorithms are one of the biggest risks of failure because it compromises the very purpose of machine learning since there has been heightened public concern about the impact of digital technology on society.

Random Matrix Theory as a solution. Understanding the resulting impact of popular statistical learning methods when n and p are both large and comparable is becoming a growing research concern in modern statistic. Despite the expected challenges

and difficulties in analyzing statistical methods in this high-dimensional setting, recent evidences, which we will further develop in the present manuscript, suggest that Random Matrix Theory (Random Matrix Theory (RMT)) provides the necessary tools to handle such a framework. The sample covariance matrix, being a dominant object of study in most of the multivariate analyses, is a suitable object to illustrate this "*curse of dimensionality*" paradigm. Let us consider a p-dimensional random vector x with zero-mean $\mathbb{E}[x] = 0$ and covariance $\Sigma = \mathbb{E}[xx^\mathsf{T}]$. Many topics in multivariate analysis (e.g., principal component analysis, factor analysis, multidimensional scaling, etc) deal with the study of functionals (trace, log determinant, etc) or spectral properties (eigenvalues, eigenvectors) of Σ.

Typically, Σ is unknown, and so has to be estimated using a sample of data. Given a sequence of independent random vectors $X = [x_1, \ldots, x_n]$ drawn from the same distribution as x, the usual estimate of the population covariance matrix Σ is the sample covariance matrix $\hat{\Sigma}$ defined as:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n} X X^\mathsf{T}.$$

For fixed p and $n \to \infty$, the sample covariance matrix is a consistent estimator for the population covariance matrix. If we denote $\lambda_1 \geq \ldots \geq \lambda_p$ the eigenvalues of Σ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ the eigenvalues of $\hat{\Sigma}$, Anderson et al. (1958, Theorem 13.5.1) proves more specifically, that if $x \sim \mathcal{N}(0, \Sigma)$, for $1 \leq j \leq p$, the sample eigenvalues $\hat{\lambda}_j$'s are asymptotically distributed according to:

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \to \mathcal{N}(0, 2\lambda_j^2) \text{ as } n \to \infty. \tag{1.1}$$

This result shows that when $n \to \infty$, p fixed, the j-th sample eigenvalue $\hat{\lambda}_j$ is a consistent estimator of the j-th population eigenvalue λ_j. However, this result doesn't apply when p, n are both large. A typical question would be to study the eigenvalue distribution of $\hat{\Sigma}$ in order to quantify its "deviation" from the eigenvalue distribution of the true covariance matrix Σ in the large dimensional setting of n, p both large.

The first result on the spectral behavior of sample covariance matrices is due to the seminal work of Marčenko and Pastur in 1967 (Marčenko & Pastur, 1967) where they obtained a self-consistent equation for the spectrum of $\hat{\Sigma}$ given $\Sigma = I_p$ as p, n go to infinity.

Theorem 1. *Suppose that X is a $p \times n$ matrix with i.i.d. real- or complex-valued entries with mean 0 and variance 1. Then, as $n, p \to \infty$ such that $p/n \equiv c_0 \to c_0^\infty$, the* empirical spectral measure $\mu_{\hat{\Sigma}} = \frac{1}{p}\sum_{i=1}^{p} \delta_{\hat{\lambda}_i}$ *of the eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ of $\frac{1}{n} X X^\mathsf{T}$, converges weakly, with probability one, towards a non-random distribution, known as the Marčenko–Pastur law and denoted by $\mu_{MP}^{c_0^\infty}$. If $c_0^\infty \in (0, 1)$, $\mu_{MP}^{c_0^\infty}$ has the probability density function:*

$$\mu_{MP}^{c_0^\infty}(dx) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi c_0^\infty x} dx$$

where $\lambda_\pm = (1 \pm \sqrt{c_0^\infty})^2$. *If* $c_0^\infty \in (1, \infty)$, *then* μ_{MP} *is a mixture of a point mass* 0 *and the pdf* $\mu_{MP}^{\frac{1}{c_0^\infty}}$ *with weights* $1 - \frac{1}{c_0^\infty}$ *and* $\frac{1}{c_0^\infty}$.

In particular, the influence of c_0^∞ appears precisely. Indeed, in the regime of p fixed and $n \to \infty$, the result of (Anderson et al., 1958) showing that the sample eigenvalues converge to the population eigenvalues, is recovered by the Marčenko-Pastur formula for $c_0^\infty \to 0$. However, when $c_0^\infty = \mathcal{O}(1)$, the same formula shows that all the sample eigenvalues become noisy estimators of the eigenvalues of the identity matrix no matter how large n is. More precisely, the distortion of the spectrum of $\hat{\Sigma}$ compared to the "true" one becomes more and more substantial as c_0^∞ becomes large (see Figure 1.1).

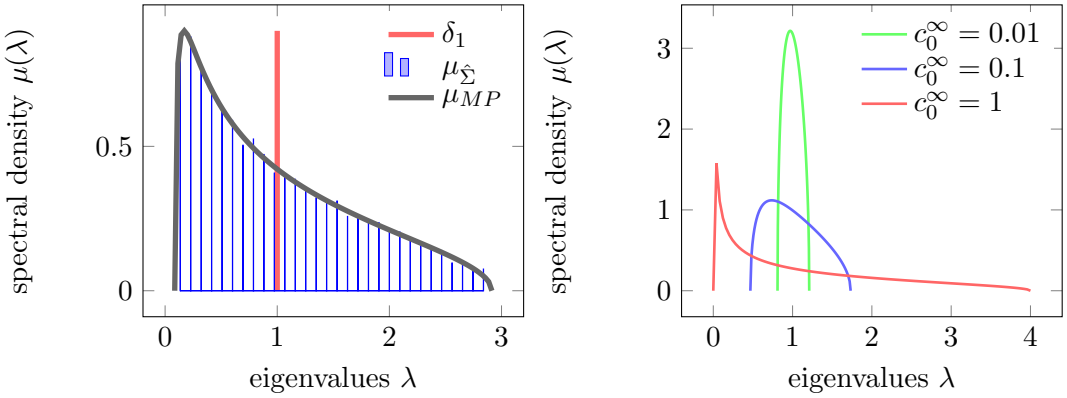

Figure 1.1: (Left) Spectral distribution of the empirical covariance matrix with $p = 1000$, $n = 2000$ versus the Marčenko-Pastur law with $c_0^\infty = 0.5$. (Right) Marčenko-Pastur law for different values of c_0^∞.

The heuristic behind this phenomenon is as follows. When the sample size n is very large, each individual coefficient of the population covariance matrix Σ can be estimated with negligible errors. But if p is also large and of the order of n, as is often the case in many situations, the sample estimator $\hat{\Sigma}$ becomes inconsistent. More specifically, the large number of simultaneous noisy coefficients of the sample covariance matrix creates important systematic errors in the computation of the eigenvalues of the matrix.

The Marčenko-Pastur result had a tremendous impact on the understanding of the "*curse of dimensionality*". This observation is the very essence of many applications in principal component analysis (Johnstone, 2001), sphericity test (Yuan et al., 2021), statistical inference (Mestre & Lagunas, 2008), etc.

To illustrate this "*curse of dimensionality*" paradigm, we present one of the important consequences of this deviation in the study of the sample generalized variance $\log(|\hat{\Sigma}|)$ appearing in many statistical methods (Quadratic Discriminant Analysis (Tharwat, 2016), hypothesis tests in multivariate statistic (Anderson et al., 1958), differential entropy in probability and information theory (Srivastava & Gupta, 2008), etc). These aspects are closely related to one of the topics covered in this thesis (Chapter 3).

Implications on a hypothesis test problem. An important statistic in multivariate analysis is the sample generalized variance:

$$\hat{\mathscr{L}}_n = \log(|\hat{\Sigma}|) = \sum_{j=1}^{p} \log\left(\hat{\lambda}_j\right).$$

When the dimension size p is fixed, for $\Sigma = I_p$, $\hat{\lambda}_j \to 1$ almost surely, as $n \to \infty$ and thus $\hat{\mathscr{L}}_n \to 0$. Further, by taking a Taylor expansion of $\log(1+x)$, one can show from (1.1) that $\sqrt{\frac{n}{p}}\hat{\mathscr{L}}_n \to \mathcal{N}(0,2)$. This suggests the possibility that $\hat{\mathscr{L}}_n$ remains asymptotically Gaussian for large p provided that $p = \mathcal{O}(n)$. However, when $p/n \equiv c_0 \to c_0^\infty \in (0,1)$ as $n \to \infty$, using results on the limiting spectral distribution of Theorem 1, we have (see the computation of the integral in (Bai & Silverstein, 2008)[Section 5]):

$$\frac{1}{p}\hat{\mathscr{L}}_n \to \int_{\lambda_-}^{\lambda_+} \frac{\log(x)}{2\pi\imath x c_0^\infty}\sqrt{(\lambda_+ - x)(\lambda_- - x)}dx \equiv d(c_0^\infty) = \frac{c_0^\infty - 1}{c_0^\infty}\log(1 - c_0^\infty) - 1 < 0$$

where $\lambda_- = (1 - \sqrt{c_0^\infty})^2$ and $\lambda_+ = (1 + \sqrt{c_0^\infty})^2$.

This shows that almost surely $\sqrt{\frac{n}{p}}\hat{\mathscr{L}}_n \sim d(c_0^\infty)\sqrt{pn} \to -\infty$. Thus the classical estimate $\sqrt{\frac{n}{p}}\hat{\mathscr{L}}_n$ is biased and will induce serious issues when implemented in hypothesis test problem.

The necessity of a new set of tools. Another problem of the classical regime (p fixed, $n \to \infty$) concerns the intricate nature of the random objects of interest in machine learning and signal processing algorithms which has sometimes limited the capability of learning theory in this regime to explain and predict the properties and the behavior of these algorithms. We illustrate this with the ridge regression problem for classification. Suppose a statistician observes n training examples $(x_i, y_i) \in \mathbb{R}^p \times \mathscr{Y}$, and wants to find a rule for predicting $\mathbf{y}$ on future unlabeled draws $\mathbf{x}$. In other words, the statistician seeks a function $h : \mathbb{R}^p \to \mathscr{Y}$, $h(x) = g(\omega^\mathsf{T} x)$ for which $\mathbb{E}[\ell(y, h(x))]$ is small where $\ell(.,.)$ is a loss function; in regression $\mathscr{Y} = \mathbb{R}$ and in binary classification $\mathscr{Y} \in \{-1, 1\}$. Such prediction problems lie at the heart of several scientific and industrial endeavors. The ridge regression problem consists in taking $\ell(y, h(x)) = \|y - h(x)\|^2$. Using a linear function $g(x) = x$ and normalizing the data matrix $X = [x_1, \ldots, x_n]$ by $\sqrt{n}$[1], the regression parameter vector ω is chosen in order to minimize the residual error in the training dataset X as:

$$\min_{\omega} \|y - \frac{X^\mathsf{T}}{\sqrt{n}}\omega\|^2 + \lambda\|\omega\|^2 \tag{1.2}$$

where $y = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^n$ and λ is the regularization parameter which controls the complexity of the model. Taking the derivative with respect to ω and setting it to zero,

[1]This normalization is carried out to prevent performance from diverging when n, p tends to infinity (See more details in (Louart & Couillet, 2018))

we get

$$\frac{1}{n}XX^\mathsf{T}\omega - \frac{1}{\sqrt{n}}Xy + \lambda\omega = 0. \tag{1.3}$$

Therefore the optimal regression parameter vector can be computed by

$$\omega = \left(\frac{1}{n}XX^\mathsf{T} + \lambda I_p\right)^{-1}\frac{X}{\sqrt{n}}y. \tag{1.4}$$

For a new test data $\mathbf{x}$, the decision is based on the sign of the test score :

$$g(\mathbf{x}) = \omega^\mathsf{T}\frac{\mathbf{x}}{\sqrt{n}} = \frac{1}{n}\mathbf{x}^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} + \lambda I_p\right)^{-1}Xy. \tag{1.5}$$

Even under a Gaussian mixture model for the data matrix X, to the best of our knowledge, there exist few theoretical works (Tsigler & Bartlett, 2020) in the classical regime that can study the statistical behavior of the random quantity $g(\mathbf{x})$ since it involves complex dependency in X in particular arising in the inverse matrix $\left(\frac{1}{n}XX^\mathsf{T} + \lambda I_p\right)^{-1}$. In Chapter 2, we will see that this object is natural and classical to study under the large dimensional setting of p and n both large. The quantity $\left(\frac{1}{n}XX^\mathsf{T} + \lambda I_p\right)^{-1}$ called the resolvent of the matrix $\frac{1}{n}XX^\mathsf{T}$ is at the heart of Random Matrix Theory (RMT) tools and will be discussed at length in this thesis.

Summarizing the "*curse of dimensionality*" paradigm and moving towards the "*blessing of dimensionality*" through RMT. This section discusses two problems induced by classical asymptotic in multivariate analysis:

1. Tools from Multivariate Analysis in classical statistic are not always sufficient to deal with the increasingly complex random objects appearing in machine learning and signal processing. This has been illustrated through the popular and simple ridge regression problem.

2. For the few methods that can be analyzed, classical asymptotic approximations based on the "fixed dimension, large sample size" regime are inadequate in capturing the effects of dimensionality when the data dimension p is not small compared to the sample size. The spectral behavior of the sample covariance matrix illustrates this phenomenon with consequences shown on the study of the sample generalized variance.

The "*curse of dimensionality*" initially introduced by Bellman (Bellman, 1966) in relation to complications occurring in dynamic programming has now become a common name for issues of theoretical nature arising in high dimensions. The recent advances in large dimensional random matrix theory have raised much interest for problems in statistic and signal processing under the assumption of large but similar population dimension p and sample size n. The "*curse of dimensionality*" can then be turned into the blessing of dimensionality using Random Matrix Theory as a consequence of the

measure concentration phenomena arising for high dimensional data. Therefore, *large dimensional setting should not be feared or avoided: it just has to be used properly.*

More specifically, we will show that by being a little too "stubborn" to assume $n \gg p$ even when it is not true, we can generate dramatic errors. Furthermore, one may wonder for which value of n, p, we start observing the curse of dimensionality. In fact, the $n, p \gg 1$ regime already occurs when n, p are rather small. By exploiting both randomness in p and n, the speed of convergence of many functionals discussed in this manuscript can be as high as $\mathcal{O}(1/\sqrt{pn})$ (unlike the classical Central Limit Theorem (CLT) which goes in $\mathcal{O}(1/\sqrt{n})$). For example, one could believe that $n = 100p$ with $p \gg 1$ is a "classical" regime, when in fact the RMT tools explain the phenomena better even in this case, and following the $n \gg p$ path to obtain intuitions when $n, p \gg 1$ is a bad strategy. The curse of dimensionality remains in fact true: the intuitions coming from $n \gg p$ are simply wrong; one has to wipe the slate clean, start on a radically different approach, which is what RMT offers. A major consequence is that one will get in this thesis new "intuitions" which, at first sight, are in fact rather very counter-intuitive.

In this thesis, we leverage the capability of random matrix theory to overcome the technical difficulties involved in recent machine learning algorithms and to deeply understand limitations and possible corrections of such "large p, large n" systems. Specifically, we will use the opportunity offered by random matrix tools to understand and improve two problems of great interest in machine learning and signal processing (distance between covariance matrices and Multi-Task Learning).

Before delving into the technical details, we next motivate the interest into these two problems starting with the estimation of distances between covariance matrices.

1.2 The blessing of dimensionality applied to the estimation of the distance between covariance matrices problem

Motivation of covariance matrix distance estimation problem. Similarities between covariance matrices are objects of interest for many engineering applications, among which machine learning problems (for instance, covariance-based data clustering regularly used in synthetic aperture radar, hyperspectral imaging (Chang, 2003), or EEG datasets (Richiardi et al., 2013)), dimensionality reduction (Carter, 2009), portfolio-optimization and asset clustering in finance (Tola et al., 2008), etc.

State-of-the-art estimation procedure. Depending on context and application, various metrics are available in the literature to compare semi-definite positive matrices (the Frobenius norm, the Fisher Information metric (Costa et al., 2015), the Bhattacharyya distance (Bhattacharyya, 1943), the Rényi or Kullback-Leibler divergence, the Wasserstein distance, etc.). If we define by Σ_1 and Σ_2 two covariance matrices of size $p \times p$, all these distances, which we will denote generically $D(\Sigma_1, \Sigma_2)$, can be expressed as functionals of the eigenvalues of the matrix $\Sigma_1^{-1}\Sigma_2$ (Fisher distance, Bhattacharyya distance, Kullback-

Leibler divergence, or Rényi divergence between centered Gaussians) or eigenvalues of the matrix $\Sigma_1\Sigma_2$ (Wasserstein distance between two centered Gaussians). Assuming that the number of samples n_1 and n_2 of data having Σ_1 and Σ_2 for covariance is very large compared to p, the law of large numbers guarantees that $D(\hat{\Sigma}_1, \hat{\Sigma}_2)$ is a consistent estimator for $D(\Sigma_1, \Sigma_2)$ with $\hat{\Sigma}_a = \frac{1}{n_a}\sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)T}$ for $a \in \{1, 2\}$ the empirical covariance matrix of the n_a centered samples $x_i^{(a)}$ with $x_i^{(a)} = \Sigma_a^{\frac{1}{2}} z_i^{(a)}$ and $z_i^{(a)}$ i.i.d. random vectors of zero mean and unit variance entries.

Inconsistency of the classical estimate. However, the classical estimator is strongly biased when $n_1, n_2 \sim p$ as shown in the middle column of Table 1.1. Using tools from

p	$D_{\mathrm{F}}(\Sigma_1, \Sigma_2)$	Classical	Proposed
2	0.0980	0.1002	0.0973
4	0.1456	0.1520	0.1461
8	0.1694	0.1820	0.1703
16	0.1812	0.2081	0.1845
32	0.1872	0.2363	0.1886
64	0.1901	**0.2892**	0.1920
128	0.1916	**0.3955**	0.1934
256	0.1924	**0.6338**	0.1942
512	0.1927	**1.2715**	0.1953
		(error > 50%)	**(error > 100%)** **(error > 500%)**

Table 1.1: Proposed versus classical estimator for the Fisher distance between Σ_1 and Σ_2 with $[\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$, $x_i^{(a)} \sim \mathcal{N}(0, \Sigma_a)$; $n_1 = 1024$ and $n_2 = 2048$ for different values of p. Averaged over 10 000 trials.

random matrix theory, this thesis proposes a general formula for a " universal " distance estimator $D(\Sigma_1, \Sigma_2)$ which is consistent within the limit where $p, n_1, n_2 \to \infty$ with $p/n_1 \to c_1 > 0$ and $p/n_2 \to c_2 > 0$, the estimated outputs of which are displayed in the rightmost column of Table 1.1. These aspects are investigated in Chapter 3.

Covariance matrices are at the heart of most statistical and machine learning methods and therefore knowing how to manage, estimate and understand covariance functions is central in Machine Learning (ML) since it allows for the efficient use of estimators when we have to manage several covariance matrices. These covariance matrix models, one way or another, are indirectly exploited in many machine learning algorithms. Large dimensional statistics have thus naturally provided new results to better understand, analyze and improve these algorithms, yet so far for "bottom of the shelf" algorithms (classical Support Vector Machine (SVM) (Liao & Couillet, 2019), spectral clustering (Couillet et al., 2016), graph-based Semi-Supervised Learning (SSL) (Mai & Couillet, 2018)). The richer and more complex methods such as multi-task learning, transfer learning, learning with fairness, privacy and safety in machine learning, etc which involve multiple biases that are difficult to trace (too many parameters, too much heterogeneity

in the data, etc.) are finally those that can gain the most from exploiting the RMT, which will identify these biases. We show that this is indeed the case, that there are multiple biases especially in the case of Multi-Task Learning, but that the RMT cleans them up one by one, while maintaining a remarkable algorithmic simplicity.

1.3 The blessing of dimensionality applied to Multi-Task Learning

From single to multiple task learning. The most advanced supervised machine learning algorithms generally require a large number of *labeled* training samples to achieve high levels of accuracy. Deep learning models are prototypical in their sometimes requiring millions of labeled samples for efficient training. In many (if not most) applications (say, for instance, medical imaging (Abdullah-Al-Zubaer Imran & Terzopoulos, 2019; Imran et al., 2020)), this is often too demanding, labeled samples being hard to collect. MTL (Caruana, 1997; Zhang & Yang, 2018, 2021), an offspring of which is better known as *transfer learning*, provides a potent workaround by appending the available small training dataset of interest with additional *somewhat similar* datasets on which similar (classification or regression) tasks can be performed; the additional data possibly being of a different nature, MTL effectively solves multiple tasks *in parallel* while exploiting task relatedness to enforce collaborative task learning.

How does MTL work? Precisely, multi-task learning simultaneously solves multiple related tasks and introduces shared hyperparameters or feature space, optimized to improve the performance of the individual tasks. The crux of the various multi-task learning algorithms lies in the means to both enforce and, most importantly, evaluate task relatedness: this is in general highly non-trivial as this implies to theoretically understand what common features of the parallel datasets can be adequately exploited by the MTL algorithm – the latter generally deriving from a classical single-task algorithm (such as a mere support vector machine). Several heuristics have been proposed which may be split into two groups: parameter-based versus feature-based MTL. In the parameter-based MTL approach, the tasks are assumed to share some common hyperparameters (Evgeniou & Pontil, 2004; Xu et al., 2013) (*e.g.*, the hyperplanes separating each class in a support vector machine flavor) or that these hyperparameters have a common prior distribution (Zhang & Yeung, 2012, 2014). Classical learning mechanisms (such as support vector machines (SVM), logistic regression, etc.) can then be appropriately adapted and turned into a multi-task version by enforcing these parameter relatedness assumptions. In this context, (Evgeniou & Pontil, 2004; Xu et al., 2013; Parameswaran & Weinberger, 2010) respectively adapt the SVM, least square-Support Vector Machine (SVM) (LS-SVM), and large margin nearest neighbor (LMNN) methods into a MTL paradigm. In the feature-based MTL approach, the tasks data are instead assumed to share a low-dimensional common representation. In this context, most of the works aim at determining a mapping of the ambient data space into a low-dimensional subspace (through sparse coding, deep neural network embeddings, principal component analysis, etc.) in which the tasks have

high similarity (Argyriou et al., 2008; Maurer et al., 2013; Zhang et al., 2016; Pan et al., 2010); other works simply use a feature selection method by merely extracting a subset of the original feature space (Obozinski et al., 2006; Wang & Ye, 2015; Gong et al., 2012).

The negative transfer plague. The strong underlying limitation of all these methods is their lack of theoretical tractability: as a result, many of the biases inherent to the base methods (SVM, LS-SVM, deep nets) are exacerbated in a multi-task setting. As a striking consequence, many of these heuristic MTL algorithms suffer from *negative transfer*, which corresponds to scenarios where the multi-task setup performs worse than a single-task approach (Rosenstein et al., 2005; Long et al., 2013); this is particularly the case when task relatedness is weaker than assumed so that the MTL method enforces fictitious similarities, thereby inducing strong biases.

A large dimensional analysis to redesign MTL. In this thesis, we focus on a very elementary (yet, as we shall see, already quite powerful) LS-SVM-based MTL approach and provide a thorough theoretical analysis from which we manage to automatically discard the negative transfer limitation. Specifically, placing ourselves under a large dimensional data setting, we exploit modern tools from large dimensional statistics (here random matrix theory) to theoretically investigate the key components that determine whether tasks interfere constructively or destructively under the MTL framework. This analysis provides powerful insight into the inner workings of the method and allows for a fundamental adaptation of the method which provably avoids all sorts of biases and most importantly discards the problem of negative transfer altogether. Methodologically, in its simplest approach, MTL algorithms can be obtained from a mere extension of support vector machines (SVM), accounting for more than one task. That is, instead of finding the hyperplane (through its normal vector ω) best separating the two classes of a unique dataset, (Evgeniou & Pontil, 2004) proposes to produce best separating hyperplanes (or normal vectors) $\omega_1, \ldots, \omega_k$ for each pair of data classes of k tasks, with the additional constraint that the normal vectors take the form $\omega_i = \omega_0 + v_i$ for some common vector w_0 and dedicated vectors v_i. The amplitude of the vectors v_i is controlled (through an additional hyperparameter) to enforce or relax task relatedness. We study this approach in chapter 4.

The conclusions drawn in this chapter thus allow for an optimal use of MTL LS-SVM with performance-maximizing hyperparameters and strong theoretical guarantees. As such, the work performed in Chapter 4 offers through MTL LS-SVM a viable fully-controlled (even better performing) alternative to state-of-the-art MTL.

1.4 Outline and contributions of the thesis

The classical asymptotic based on "$n \gg p$" hypothesis is not just "incorrect", it also induces terrible biases that completely destroy the functioning of the algorithms; the thesis proposes to shed light on this by showing the dramatic consequences that these

biases can induce. Furthermore, independently of its correctness or not, the $n \gg p$ hypothesis is often not enough to understand the behavior of the algorithms; the thesis shows that, paradoxically, the $n, p \to \infty$ hypothesis gives rise to analytically accessible expressions; the only difficulty is technical: it requires mastering the RMT tools and forgetting our reflexes and small dimensional biases. Particularly, the thesis is based on the fact that when $p, n \to \infty$, the statistics of the functionals of interest of the data (scores, performances, thresholds) depend only on first and second-order statistics (means and covariance matrices); this makes in particular the covariance matrices extremely rich objects that need to be understood.

This thesis first shows how classical estimates can destroy algorithms by introducing biases that are difficult to clean, whereas a consistent RMT estimation of the functionals of interest avoids biases. Furthermore, we illustrate how classical biases induced in "'bottom of the shelf" algorithms such as Support Vector Machine, Semi-Supervised Learning, etc are exacerbated in more involved algorithms like multi-task and transfer learning schemes. To that end:

- In Chapter 2, we introduce the necessary tools from Random Matrix Theory to grasp the technical ideas required for understanding the thesis. To that end, we consider two applications close to Chapters 3 and 4 of this thesis. In particular, we will look at how RMT works through simple examples that are closely related to the specific problems of this thesis. These examples will illustrate very simply the key problems evoked in the introductory chapter, in particular the curse of dimensionality and its cure through the RMT. These examples namely the spectral statistic estimation of the population covariance matrix and the asymptotic of ridge regression in the context of classification require to introduce some basic notions of RMT in particular the generalized Marčenko Pastur law which will begin the chapter.

- After having illustrated by the spectral statistical estimation of the population covariance matrix the limits of the classical asymptotic ($n \gg p$) and provided a consistent estimation by RMT, we will be able to introduce in the chapter 3, the problem of estimating the distance between covariance matrices while providing consistent estimators for these distances. An application to clustering of covariance-based dataset as well as an application to the covariance matrix estimation problem are provided. This chapter is the result of several contributions that started with the estimation of the most usual covariance matrices distances (Fisher distance, Battacharrya distance, Kullback-Leibler divergence for Gaussian). These distances can be expressed as eigenvalues functionals of the Fisher matrix $\Sigma_1^{-1}\Sigma_2$ for two population covariance matrices Σ_1 and Σ_2 to compare. This leads to the first contributions of this project:

  Journals:

  1. Romain Couillet, **Malik Tiomoko**, Steeve Zozor and Eric Moisan, "Random matrix-improved estimation of covariance matrix distances," Journal of

Multivariate Analysis, no. 174, pp. 104531, November 2019.

Conferences:

1. **Malik Tiomoko**, Romain Couillet, Steeve Zozor, and Eric Moisan, "Improved Estimation of the Distance between Covariance Matrices," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19), Brighton, UK, May 2019.

Although this family of metrics covered a plethora of distances used in signal processing, we realized that several distances of interest, notably the Wasserstein distance used in optimal transport, did not appear in this first class of distances. A study similar to the first one allowed us to cover this important case of distances that involve eigenvalues functionals of the matrix $\Sigma_1\Sigma_2$ among which are the Frobenius distance and the Wasserstein distance.

Conferences:

1. **Malik Tiomoko**, and Romain Couillet, "Random Matrix-Improved Estimation of the Wasserstein Distance between two Centered Gaussian Distributions," in European Signal Processing Conference (EUSIPCO'19), A Coruna, Spain, 2019, A Coruna, Spain, 2019, **Best Student Paper Award**.

Having a generic framework to consistently estimate all distances between covariance matrices, we develop a framework on the covariance matrix estimation itself based on an optimization problem involving distances between covariance matrices. Concretely, the estimation procedure consists in (i) writing the covariance matrix Σ as the solution to $\arg\min_{M\succ 0} D(M,\Sigma)$ for a wide range of metrics D (Fisher, Batthacharyya, Stein's loss, Wasserstein, etc.), (ii) based on Couillet et al. (2018), using the fact that $D(M,\Sigma) - \hat{D}(M,X) \to 0$ for some consistent estimator $\hat{D}$, valid for all deterministic M and samples $X = [x_1,\ldots,x_n] \in \mathbb{R}^{p\times n}$ having zero mean and covariance Σ, and (iii) proceeding to a gradient descent on $\hat{D}$ rather than on the unknown D itself. This forms the basis of the following contribution and can be seen as an application of the two previous contributions:

Conferences:

1. **Malik Tiomoko**, Florent Bouchard, Guillaume Ginholac, and Romain Couillet, "Random Matrix Improved Covariance Estimation for a Large Class of Metrics," in International Conference on Machine Learning (ICML'19), Long Beach, USA, 2019, Long Beach, USA, June 2019.

One of the major limitations of the three previous works was that they were only valid when the number of samples was larger than the feature size, which was a

real handicap for real applications. We therefore analyzed and understood this important problem and proposed solutions to solve it. This work has been done in the following contribution.

Conferences:

1. **Malik Tiomoko**, and Romain Couillet, "Estimation of Covariance Matrix Distances in the High Dimension Low Sample Size Regime," in IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'19), Guadeloupe, France, 2019.

- Having a generic framework for consistent estimates of distances between covariance matrices, we show in Chapter 4 how problems as complex as multi-task learning also depend on mean and covariance functionals that can be estimated consistently. But beyond that, the theoretical analysis highlights deep insights into the inner working of these algorithms and opens the way to improvements of the method. The project starts by deriving and analyzing a first simple algorithm based on the ridge regression problem. We provide in the following contribution a first asymptotic behavior of the latter. The main contributions of this work were technical to show that RMT can not only handle simple algorithms like spectral clustering, SVM, but also advanced machine learning methods such as MTL.

  Conferences:

  1. **Malik Tiomoko**, Cosme Louart and Romain Couillet, "Large Dimensional Asymptotics of Multi-Task Learning," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20), Barcelona, Spain, May 2020.

  This work led furthermore to a simplification and more interpretable formulas for the large dimensional analysis and most importantly to an improvement of the Multi-Task Learning schemes performed in the following contributions.

  Journals:

  1. **Malik Tiomoko**, Romain Couillet, and Hafiz Tiomoko, "Large Dimensional Analysis and Improvement of Multi-Task Learning,"Journal of Machine Learning Research, submitted September 2020.

  Conferences:

  1. **Malik Tiomoko**, Hafiz Tiomoko Ali and Romain Couillet, "Deciphering and Optimizing Multi-Task Learning: a Random Matrix approach ," (submitted) in International Conference on Learning Representations (ICLR'21), **Spotlight Article**.

- Chapter 5 lists several perspectives following the work performed in this thesis.

Chapter 2

Basics of Random Matrix Theory

Contents

The goal of this chapter is to introduce the necessary tools required to tackle the covariance matrix distance estimation problem and the Multi-Task Learning framework. In particular, these make respectively use of linear spectral statistic (discussed in Section 2.2) and of deterministic equivalent (discussed in Section 2.3) of random matrices. But they both rely on a fundamental result in Random Matrix Theory (the generalized Marčenko-Pastur law) introduced in Section 2.1 of this chapter. Two simple examples (estimation of the population generalized variance and ridge regression problem) are used to illustrate the utility of the linear spectral statistic and of the deterministic equivalent.

The interest of looking at two simple and classical problems of statistics is twofold: (i) it allows to illustrate on concrete examples the severe limitations of the classical asymptotic ($n \gg p$) highlighted in an abstract way in the introduction and to show how RMT allows to understand and correct them; (ii) these problems being simpler, they allow to highlight the main tools which will be used in this thesis. These examples have not been chosen randomly, especially as they are closely related to the problems of interest in Chapters 3 and 4.

2.1 Large dimensional spectral behavior of the sample covariance matrix

Setting the stage. Covariance matrices are ubiquitous in many machine learning and signal processing methods. They are used in several classical methods such as Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), etc. More specifically, they are at the heart of the objects that characterize the asymptotic performance of the algorithms studied in this thesis. The ability to understand the properties of the covariance matrix Σ that is typically unknown depends on the quality of the estimator. As mentioned in the introduction, the standard maximum likelihood estimator is the sample covariance matrix $\hat{\Sigma}$. A thorough understanding of the behavior and limitations of the methods developed in this thesis requires a good understanding of the spectral properties of $\hat{\Sigma}$.

There is a large body of work concerned with the limiting behavior of the eigenvalues of the sample covariance matrix $\hat{\Sigma}$ when p and n both go to ∞; it constitutes an important subset of Random Matrix Theory. The goal of this section is to introduce a fundamental result, the Marčenko-Pastur equation, that relates the asymptotic behavior of the eigenvalues of the sample covariance matrix to that of the population covariance matrix in the "large n, large p" asymptotic setting.

One of the practical uses of the asymptotic results in RMT is their universality with respect to the distribution of the matrix entries. This makes Marčenko-Pastur's result even more interesting for this thesis[1] than some results such as the Wishart distribution (Wishart, 1928) which characterizes the joint distribution of the eigenvalues of the sample covariance matrix $\hat{\Sigma}$ in the case of Gaussian data. Note that most of the methods of interest in this thesis are function of the global behavior of the spectrum of the covariance matrix, therefore at the difference of other results on the spectral behavior of the sample covariance matrix $\hat{\Sigma}$ (Wishart distribution (Wishart, 1928) for the joint distribution of the eigenvalues of $\hat{\Sigma}$, Tracy Widow law (Tracy & Widom, 1996) for the behavior of the largest eigenvalue of $\hat{\Sigma}$, etc), the Marčenko Pastur law (describing the global behavior of the limiting eigenvalue distribution) is sufficient for understanding the methods of interest in this thesis. Specifically, we are not interested by the joint distribution of the eigenvalues as well as their local behavior but rather by their behavior as a whole. This makes Marčenko-Pastur's law an indispensable and necessary tools for understanding the techniques in this thesis. However its formulation requires to introduce some concepts and notations.

Going from vectors to measures. One of the first problems to tackle is to find an efficient way to express the limit of a vector (the p eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$) whose size grows to ∞. A natural way to do so is to associate to any vector a probability measure. This leads to introducing the notion of Empirical Spectral Distribution which associates to any vector containing the eigenvalues of a target matrix a probability measure.

Empirical and Limiting Spectral Distribution. The *Empirical Spectral Distribution* (Empirical Spectral Distribution (ESD)) of a random matrix $M \in \mathbb{R}^{p \times p}$ is defined as

$$\mu_M(t) = \frac{1}{p} \sum_{j=1}^{p} \delta\left(t - \lambda_j(M)\right), \tag{2.1}$$

where δ is the Dirac delta function and $\lambda_j(M)$'s denote the p eigenvalues of M, including the multiplicity. The ESD is the normalized counting measure of the eigenvalues of M, i.e., the probability distribution that put mass $1/p$ at each one of the p eigenvalues of M. In general, the ESD is a probability measure on $\mathbb{C}$; it has support in $\mathbb{R}$ (resp. in $\mathbb{R}_+$) if M is Hermitian (resp. non-negative definite Hermitian). In this thesis, we are

[1] Since this thesis focuses on universal methods with respect to the data distribution and needs to be robust for real-world applications, a result on the spectral behavior of $\hat{\Sigma}$ independently of the data distribution is of particular interest.

mainly concerned by covariance matrices and since there are Hermitian and non-negative definite, the corresponding ESD's will have support on $\mathbb{R}_+$. In the rest of the thesis, the support of the probability measure μ_M will be denoted $\text{Supp}(\mu_M)$.

As $p, n \to \infty$ with constant ratio $p/n \equiv c_0 \to c_0^\infty$, the eigenvalues of the sequence of sample covariance matrices $\hat{\Sigma} \in \mathbb{R}^{p\times p}$ are random variables and the corresponding ESD's $\mu_{\hat{\Sigma}}$ are random probability measures on $\mathbb{R}_+$. A fundamental question in random matrix theory is about whether the sequence $\mu_{\hat{\Sigma}}$ (sometimes written μ_p when there are no ambiguities about the matrix under investigation) has a limit (in probability or almost surely). The *Limiting Spectral Distribution* (Limiting Spectral Distribution (LSD)) of M denoted μ_∞ is defined as the limit of (2.1) as $n, p \to \infty$ if it exists. In the rest of the thesis, the Empirical Spectral Distribution of the sample covariance matrix will be denoted by μ_p and its LSD will be denoted by μ. As for the sample covariance matrix, the ESD and LSD of the population covariance matrix Σ will be denoted respectively ν_p and ν. An important area of RMT is concerned with understanding the properties of μ as function of ν. To that end, we do not work directly with the LSD μ but with a tool that is similar in flavor to the characteristic function of a distribution: the "Stieltjes transform" of a measure.

Resolvent and Stieltjes transform. The resolvent of a random matrix $M \in \mathbb{R}^{p\times p}$ is defined $\forall z \in \mathbb{C} \setminus \text{Supp}(\mu_M)$ as

$$Q(z) = (M - zI_p)^{-1}.$$

This quantity displays several interesting properties, making it the relevant object to manipulate. First, it is a continuous function of z and it is easy to differentiate (compared to working directly on the ESD), providing a well-defined tool for mathematical analysis. Furthermore if M is symmetric, it contains the complete information about the eigenvalues λ_j's and the eigenvectors u_j's of the symmetric matrix M since it can be rewritten as:

$$Q(z) = \sum_{j=1}^{p} \frac{u_j u_j^\mathsf{T}}{\lambda_j - z}.$$

It is easy to see that the number of singularities of the resolvent is equal to the number of eigenvalues of M. While the statistics of the eigenvectors are an interesting and non-trivial subject in itself, we focus for now on the statistics of the eigenvalues through the ESD. For this aim, we define the normalized trace of the resolvent

$$m_{\mu_M}(z) = \frac{1}{p}\text{tr}Q(z).$$

We shall skip the index M and replace it by p as soon as there is no confusion about the matrix we are dealing with. As $p \to \infty$,

$$m_{\mu_p}(z) \xrightarrow{\text{a.s.}} m_\mu(z), \quad m_\mu(z) = \int \frac{\mu(dt)}{t - z} \tag{2.2}$$

which is known as the Stieltjes transform of μ.

The Stieltjes transform has a lot of interesting properties among which

1. $m_\mu(z)$ is holomorphic on $\mathbb{C} \setminus \text{Supp}(\mu)$
2. If $z \in \mathbb{C}^+$ then $m_\mu(z) \in \mathbb{C}^+$
3. If $\text{Supp}(\mu) \subset \mathbb{R}_+$ and $z \in \mathbb{C}_+$, then $zm_\mu(z) \in \mathbb{C}_+$,

where we recall that $\text{Supp}(\mu)$ is the support of the distribution μ and we note $\mathbb{C}_+ = \{z \in \mathbb{C}, \text{Im}(z) > 0\}$ with $\text{Im}(z)$ the imaginary part of the complex number z.

An appealing feature of the Stieltjes transform is its analyticity which implies that its local knowledge leads to its knowledge everywhere else. Furthermore if μ has bounded support, the Stieltjes transform can be equivalently be rewritten as

$$m_\mu(z) = -\sum_{n=0}^{\infty} \frac{1}{z^{n+1}} \int \lambda^n d\mu(\lambda).$$

Therefore the Stieltjes transform also known as the Cauchy transform is the generating function of the moments of the measure μ, (i.e., it is a power series in $1/z$ whose coefficients are the moments of μ). The spectral distribution can be recovered from the Stieltjes transform using the inversion formula:

$$\mu(\{x\}) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathbf{Im} \left[m_\mu \left(x + i\epsilon \right) \right]. \tag{2.3}$$

This important feature states that Stieltjes transform $m_\mu(z)$ and probability measure μ are one-to-one corresponding to each other.

Of particular interest, the Stieltjes transform can be connected to Cauchy's integral formula.

Theorem 2 (Cauchy's integral formula)**.** *For $\Gamma \subset \mathbb{C}$ a positively (i.e., counterclockwise) oriented closed curve and a complex function $f(z)$ analytic in a region containing Γ and its inside, then*

$$\begin{cases} \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z - z_0} = f(z_0) & \text{, if } \quad z_0 \in \mathbb{C} \quad \text{is enclosed by} \quad \Gamma \\ \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z - z_0} = 0 & \text{, otherwise} \end{cases}$$

Generalized Marčenko-Pastur distribution. In the study of covariance matrices, a fundamental result exists that describes the limiting behavior of the empirical spectral distribution μ, in terms of the limiting behavior of the population spectral distribution ν. The connection between these two measures is made through an equation that links the Stieltjes transform of the empirical spectral distribution to an integral against the population spectral distribution (Silverstein & Bai, 1995).

Theorem 3. *Suppose that the entries of the $p \times n$ matrix X are complex random variables that are independent identically distributed which satisfy $\mathbb{E}[X_{11}] = 0$, $\mathbb{E}[|X_{11}|^2] = 1$ and $\mathbb{E}[|X_{11}|^4] < \infty$. Also, assume that $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_j \in \mathbb{R}_+^\star$ and the distribution function of $\{\lambda_1, \ldots, \lambda_p\}$ converges almost surely to a probability distribution*

function ν as $p \to \infty$. Let $\tilde{\Sigma} = \frac{1}{n} X^\star \Sigma X$. Then as $n, p \to \infty$ such that $\frac{p}{n} \equiv c_0 \to c_0^\infty$, the ESD $\tilde{\mu}_p$ of $\tilde{\Sigma}$ converges to a non-random distribution $\tilde{\mu}$, where, for any $z \in \mathbb{C} \setminus \mathrm{Supp}(\tilde{\mu})$, its Stieltjes transform $\tilde{m} = m_{\tilde{\mu}}(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$z = -\frac{1}{\tilde{m}} + c_0^\infty \int \frac{t d\nu(t)}{1 + \tilde{m}t}, \tag{2.4}$$

which gives an explicit inverse for $m_{\tilde{\mu}}(z)$. Defining $m(z) = \frac{1}{c_0^\infty} \left(\tilde{m}(z) + \frac{1 - c_0^\infty}{z} \right)$ and noticing that the nonzero eigenvalues of $\frac{1}{n} \Sigma^{\frac{1}{2}} X X^\star \Sigma^{\frac{1}{2}}$ coincide with those of $\frac{1}{n} X^\star \Sigma X$, it can be easily deduced that the ESD μ_p of the sample covariance matrix $\hat{\Sigma}$ converges to a non-random distribution μ almost surely, where the Stieltjes transform $m = m_\mu(z)$ of μ is the unique solution in $\mathbb{C}_+$ of

$$m = \int \frac{d\nu(t)}{t(1 - c_0^\infty - c_0^\infty z m) - z}. \tag{2.5}$$

Equation (2.5) can be conveniently rewritten as:

$$m_\nu \left(-\frac{1}{m_{\tilde{\mu}}(z)} \right) = -z m_\mu(z) m_{\tilde{\mu}}(z). \tag{2.6}$$

We should note that the assumption of the bounded fourth moment entries of X in the theorem, i.e., $\mathbb{E}[|X_{11}|^4] < \infty$ ensures from (Silverstein & Bai, 1995) that the limiting distribution μ has bounded support.

Under the assumptions put forth in Theorem 3, the spectral distribution of the sample covariance matrix is asymptotically non-random. Furthermore, it is fully characterized by the true population spectral distribution, through the equation (2.5). A particular case of equation (2.5) is often of interest: the situation when all the population eigenvalues are equal to 1. In this case, $\nu = \delta_1$ and we recover the Marčenko-Pastur law introduced in Theorem 1. We should stress that Equation (2.6) is remarkable and extremely powerful. Indeed let's recall due to the inversion formula of the Stieltjes transform that the unique link between the measure μ and its Stieltjes transform $m_\mu(z)$ allows one to work with either. It turns out that the Stieltjes transform is more convenient and Equation (2.6) thus indirectly links μ and ν through an "explicit" link between m_ν and m_μ. This is an extremely powerful relationship which plays the role of $\mu \to \nu$ (i.e., $\hat{\lambda}_i \to \lambda_i$) in the case of the classical regime ($n \to \infty$, p fixed).

In many applications, the population covariance Σ itself may not be the object of central interest. One is often rather interested in scalar functionals of Σ such as linear functionals of its eigenvalues. Based on the global knowledge of the LSD of the sample covariance matrix through the Marčenko-Pastur law, Random-matrix improved estimates of these functionals can be derived. To better introduce the random matrix improved estimate of the distance between covariance matrices in chapter 3, we present in the following section linear spectral statistic of random matrices in particular covariance matrices. An application to the estimation of the logarithm determinant of the population

covariance matrix (population generalized variance) already introduced in Chapter 2 will be taken as a preliminary example to tackle the more involved distance between covariance matrices in Chapter 3.

2.2 Linear spectral statistic

Setting the stage. We introduce previously the population covariance matrix Σ, the Fisher matrix $\Sigma_1^{-1}\Sigma_2$, and the matrix $\Sigma_1\Sigma_2$ appearing in the distance between two covariance matrices Σ_1 and Σ_2. Let M be one of these matrices. In one-sample and two-sample multivariate analysis, many statistics are functions of the eigenvalues $\{\lambda_j\}_{j=1}^p$ of the matrix M of the form

$$\mathscr{L} = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) = \int f(x) d\mu_M(x) \tag{2.7}$$

for any function f smooth and bounded over $\{z \in \mathbb{C}, \mathscr{R}[z] > 0\}$. Such statistic is called a linear spectral statistic of the matrix M.

The generalized variance: a linear spectral statistic example. For example, the so-called generalized variance introduced previously and discussed also later in this chapter, is

$$\mathscr{L} = \frac{1}{p}\log(|\Sigma|) = \frac{1}{p}\sum_{j=1}^{p}\log(\lambda_j). \tag{2.8}$$

So this particular $\mathscr{L}$ is a linear spectral statistic of the population covariance matrix Σ with function $f(x) = \log(x)$. In two-sample multivariate analysis with say covariance matrices Σ_1 and Σ_2, a lot of distances between Σ_1 and Σ_2 will still be of the previous form in (2.7), where however the eigenvalues λ_j will be those of the so-called Fisher matrix $\Sigma_1^{-1}\Sigma_2$ or $\Sigma_1\Sigma_2$ depending on the considered metric. Linear spectral statistic of these matrices is at the heart of the statistical tools developed in the first part of this thesis and more generally in modern statistic.

Therefore, understanding the asymptotic properties of eigenvalue statistics such as $\mathscr{L}$ above has paramount importance in data analysis when the dimension p is getting large with respect to the sample size n. In order to explain the methodology to deal with such quantity in random matrix theory, we will focus on the spectral statistic of the covariance matrix in this introductory part which has found several applications in signal processing and machine learning. In practice, one often estimates the covariance matrix using the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}XX^\mathsf{T}$ first and then uses it to compute the log-determinant. This estimate is known as the sample generalized variance.

Consistency of the sample generalized variance. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ the eigenvalues of the sample covariance $\hat{\Sigma}$ and $\lambda_1, \ldots, \lambda_p$ the eigenvalues of the population covariance

Σ. The sample generalized variance is defined as

$$\hat{\mathscr{L}}_n = \frac{1}{p}\log(|\hat{\Sigma}|) = \frac{1}{p}\sum_{j=1}^{p}\log\left(\hat{\lambda}_j\right).$$

We show in the introductory part that when the dimension size p is fixed, for $\Sigma = I_p$, $\hat{\lambda}_j \to 1$ almost surely as $n \to \infty$ and thus $\hat{\mathscr{L}}_n \to 0$. Therefore, $\hat{\mathscr{L}}_n$ is a good estimator for the logarithm determinant of the population covariance matrix. However, when $p, n \to \infty$ with $p/n \equiv c_0 \to c_0^\infty \in (0,1)$, since the limiting spectrum of $\hat{\Sigma}$ is almost surely bounded away from zero and upper-bounded, we have with probability 1 (see more details in (Bai & Silverstein, 2008)),

$$\hat{\mathscr{L}}_n \to \int_{\lambda_-}^{\lambda_+} \frac{\log(x)}{2\pi\imath x c_0^\infty}\sqrt{(\lambda_+ - x)(\lambda_- - x)}dx = \frac{c_0^\infty - 1}{c_0^\infty}\log(1 - c_0^\infty) - 1 < 0 \tag{2.9}$$

where $\lambda_- = (1 - \sqrt{c_0^\infty})^2$ and $\lambda_+ = (1 + \sqrt{c_0^\infty})^2$. This shows that the classical estimator is biased (except for $c_0^\infty = 0$).

This is due to the fact that since the eigenvalues of $\hat{\Sigma}$ follow asymptotically the Marčenko-Pastur Law discussed previously, the limiting spectrum is spread around 1. Therefore, taking the logarithm of such eigenvalues may lead to dramatic errors.

The solution of the real integral presented in Equation (2.9) has been derived using Poisson's integral formula (see more details in (Bai & Silverstein, 2008)[Section 5]). The methodology is not anymore applicable for more general spectral distributions and more complex functionals as will be the case in Chapter 3 of this thesis. Therefore, we present in the following a quite general methodology for recovering the n,p-consistent estimate of $\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j)$.

Design of n,p-consistent estimators. To that end, we recall the definition of the population and empirical eigenvalue distributions:

$$\mu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\hat{\Sigma})}, \quad \nu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\Sigma)}.$$

As $p, n \to \infty$ with $p/n \equiv c_0 \to c_0^\infty \in (0,\infty)$, $\mu_p \xrightarrow{\text{a.s.}} \mu$ and $\nu_p \xrightarrow{\text{a.s.}} \nu$ and we have the convergence of the corresponding Stieltjes transforms ($m_{\mu_p} \xrightarrow{\text{a.s.}} m_\mu$ and $m_{\nu_p} \xrightarrow{\text{a.s.}} m_\nu$).

Instead of using the classical asymptotic result $\mu_p \xrightarrow{\text{a.s.}} \nu_p$ (only valid when p is fixed and $n \to \infty$), the main idea of the estimation procedure consists in relating $\int f d\nu_p$, to the Stieltjes transform m_{ν_p} using the Cauchy's integral theorem. Using furthermore the relation between the Stieltjes transforms of the limiting empirical and population spectral distributions m_μ and m_ν from the generalized Marčenko-Pastur law, $\int f d\nu_p$ can be expressed as function of the Stieltjes transform of the limiting spectral distribution of the sample covariance matrix m_μ. It then remains to use the convergence $\nu_p \to \nu$ and $m_{\mu_p} \xrightarrow{\text{a.s.}} m_\mu$, along with the fact that the eigenvalues of $\hat{\Sigma}$ almost surely do not escape

the limiting support μ as $p \to \infty$. This is ensured from (Bai & Silverstein, 1998a), by assuming $\lim_{p\to\infty} \max(\|\Sigma\|, \|\Sigma^{-1}\|) < \infty$ with $\|.\|$ the operator norm.

The detailed steps are summarized as follows.

1. Express $\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j)$ as function of the Stieltjes transform m_{ν_p} of the empirical spectral distribution of the matrix Σ using Cauchy's integral formula:

$$\begin{aligned}\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) = \int f(t)d\nu_p(t) &= \frac{1}{2\pi\imath}\int \left[\oint_{\Gamma_\nu} \frac{f(\omega)}{\omega - t}d\omega\right] d\nu_p(t) \\ &= \frac{-1}{2\pi\imath}\oint_{\Gamma_\nu} f(\omega)m_{\nu_p}(\omega)d\omega\end{aligned}$$

   with Γ_ν a contour surrounding the limiting spectral distribution of Σ.

2. Relate the Stieltjes transform m_ν of the limiting spectrum of the population covariance matrix Σ to the Stieltjes transform m_μ of the limiting eigenvalue distribution of the sample covariance matrix $\hat{\Sigma}$ using the Marčenko-Pastur law given in equation (2.6) that we recall for convenience $\forall z \in \mathbb{C} \setminus \text{Supp}(\mu)$:

$$m_\nu\left(-\frac{1}{m_{\tilde{\mu}}(z)}\right) = -zm_\mu(z)m_{\tilde{\mu}}(z).$$

3. Deduce the expression of $\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j)$ as function of the Stieltjes transform m_{μ_p} of the empirical spectral distribution of the matrix $\hat{\Sigma}$. To that end, we need to proceed to the change of variable $\omega = \frac{-1}{m_{\tilde{\mu}}(z)}$ which ends up to

$$\begin{aligned}\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) &= -\frac{1}{2\pi\imath}\oint_{\Gamma_\mu} f\left(-\frac{1}{m_{\tilde{\mu}}(z)}\right) m_\nu\left(-\frac{1}{m_{\tilde{\mu}}(z)}\right)\frac{m_{\tilde{\mu}}^{'}(z)}{m_{\tilde{\mu}}(z)^2}dz \\ &= \frac{1}{2\pi\imath}\oint_{\Gamma_\mu} f\left(-\frac{1}{m_{\tilde{\mu}}(z)}\right) zm_{\tilde{\mu}}(z)m_\mu(z)\frac{m_{\tilde{\mu}}^{'}(z)}{m_{\tilde{\mu}}(z)^2}dz \\ &= \frac{1}{2\pi c_0^\infty\imath}\oint_{\Gamma_\mu} f\left(-\frac{1}{m_{\tilde{\mu}}(z)}\right) zm_{\tilde{\mu}}^{'}(z)dz.\end{aligned}$$

   Letting $g(z) = f(1/z)$ and $G(z)$ such that $G'(z) = g(z)$, integration by parts of the above expression along with $m_{\tilde{\mu}_p} \xrightarrow{\text{a.s.}} m_{\tilde{\mu}}$ further gives

$$\frac{1}{p}\sum_{j=1}^{p} f(\lambda_j) = \frac{1}{2\pi c_0\imath}\oint_{\Gamma_\mu} G\left(-m_{\tilde{\mu}_p}(z)\right)dz + o(1). \tag{2.10}$$

   We need furthermore to ensure that the change of variable performed brings any contour Γ_μ surrounding the support of μ to a contour Γ_ν surrounding only the support of ν (and not additional points such as singular point 0 in order to ensure that Cauchy's integral formula of step (1) still remains valid). We will show that this happens to be true only under the condition $c_0^\infty < 1$. In Chapter 3, these aspects are revisited and a solution is proposed to mitigate this important case.

4. Find if possible the solution of the complex integral in closed form using complex integration techniques.

Before going into the details of the integration contour determination, we handle step (4), by introducing some basics of integration in complex analysis.

Background on complex analysis for integration

Definition 1. *A function $f(z)$ is said to be analytic in a region R of the complex plane if $f(z)$ has a unique derivative at each point of R and if $f(z)$ is single valued.*

Note that in Definition 1, the definition of the differentiation of a complex function $f(z)$ at a point z_0 is defined similarly as in the real case as

$$f'(z_0) = \lim_{\delta z \to 0} \frac{f(z_0 + \delta z) - f(z_0)}{\delta z}.$$

To be differentiable, it is important that the limit be the same whichever direction we approach. Points at which a function $f(z)$ is not analytic are called singular points or singularities of $f(z)$. There are two different types of singular points:

- if there exists an integer n such that the product $(z-a)^n f(z)$ is analytic at $z = a$, then $f(z)$ has a pole of order n at $z = a$, if n is the smallest such integer. Example: $f(z) = 1/z^2$ has a pole of order 2 at $z = 0$.
- When $f(z)$ is a multivalued function, any point which is not in the region of definition of the single-valued branch of $f(z)$ is a singular branch point. Example: $f(z) = \sqrt{z-a}$, $f(z) = \log(z-a)$ have a branch point at $z = a$. The set of all branch points are called branch cuts.

Cauchy's residue theorem which is an extension of Cauchy's integral formula allows to compute a complex integral where the contour encloses several poles.

Theorem 4 (Cauchy's residue theorem)**.** *Let Γ be a closed path within and on which f is analytic except for m poles $\xi_1, \ldots, \xi_m$. Then*

$$\frac{1}{2\pi\imath} \oint_{\Gamma} f(z)dz = \sum_{j=1}^{m} Res_{\xi_j} f \tag{2.11}$$

where the residue $Res_{\xi_j} f$ of the pole ξ_j of order n is defined as $Res_{\xi_j} f = \lim_{z \to \xi_j} \frac{1}{(n-1)!} \frac{d^{n-1}}{dz^{n-1}} [(z-\xi_j)^n f(z)]$.

In the cases where the contour encloses not only poles but also branch points, it is mandatory to define an auxiliary contour to exclude them from the interior of the initial contour and try to relate the integral on the new contour to the integral on the initial one. The main guidelines to solve complex integrals can be summarized as follows.

- Determine the poles and the branch points of the integrand function.

- In case of branch points, deform the initial contour in order to exclude the branch points from the interior of the contour.
- Express the integral over the new contour as function of the integral over the initial contour and potentially other integrals which are evaluated by either applying residue Cauchy theorem (when surrounding poles) or by computing real integrals.

This procedure is applied to compute the complex integral in (2.10) for $f(z) = \log(z)$. For this choice of function, the corresponding function G is defined as $G(z) = z\,(\log(z) - 1)$. The branch points of $\log(z)$ are defined as $z \in \mathbb{R}$ such that $\mathscr{R}(z) < 0$ where $\mathscr{R}(z)$ denotes the real part of the complex value z. Therefore, the branch cuts of the integrand are defined as the z's such that $m_{\tilde{\mu}_p}(z) \geq 0$.

We recall from the definition of the Stieltjes transform that

$$m_{\tilde{\mu}_p}(z) = \frac{c_0}{p} \sum_{i=1}^{p} \frac{1}{\hat{\lambda}_i - z} + \frac{1 - c_0}{z}. \tag{2.12}$$

From equation (2.12), $m_{\tilde{\mu}_p}(z)$ is a rational function, therefore it can be written as

$$m_{\tilde{\mu}_p}(z) = \frac{\prod_{i=1}^{p} (z - \zeta_i)}{z \prod_{i=1}^{p} (\hat{\lambda}_i - z)}, \tag{2.13}$$

with $\zeta_1 < \ldots < \zeta_p$ the zeros of $m_{\tilde{\mu}_p}(z)$ and $\hat{\lambda}_1 < \ldots < \hat{\lambda}_p$ the eigenvalues of $\hat{\Sigma}$. Equations (2.12) and (2.13) will be important for future simplifications in the calculation of complex integrals. In particular, we have that:

$$\lim_{z \to \hat{\lambda}_j} (\hat{\lambda}_j - z) m_{\tilde{\mu}_p}(z) = \frac{c_0}{p}$$

$$\sum_{i=1}^{p} \log\left(\frac{\zeta_i}{\hat{\lambda}_i}\right) = \lim_{z \to 0} \log\left(-z m_{\tilde{\mu}_p}(z)\right) = \log(1 - c_0)$$

From equation (2.13), the branch cuts $\mathscr{B}$ are defined as $\mathscr{B} = [\zeta_1, \hat{\lambda}_1] \cup \ldots \cup [\zeta_p, \hat{\lambda}_p]$ as represented in Figure 2.1. These segments lie inside the integration contour Γ, which needs to be modified for proper integration; the new contour, denoted Γ_n is depicted in Figure 2.1. The complex integral defined in (2.10) over the contour Γ_n is the sum of several integrals, subdivided into four types:

- Integrals $\mathscr{I}_1$ over the circles surrounding $\{\zeta_j\}_{j=1}^{p}$ which, thanks to the variable change $z = \zeta_j + \epsilon\, \mathrm{e}^{\imath\theta}$, reduce to $\lim_{\epsilon \to 0} \int_{\epsilon}^{2\pi - \epsilon} \frac{G\left(-m_{\tilde{\mu}_p}(\zeta_j + \epsilon\, \mathrm{e}^{\imath\theta})\right)}{2\pi\imath c_0} \imath\epsilon\, \mathrm{e}^{\imath\theta}\, d\theta$ which leads for $G(z) = z(\log(z) - 1)$ to:

$$\frac{-1}{2\pi\imath c_0} \lim_{\epsilon \to 0} \int_{\epsilon}^{2\pi - \epsilon} m_{\tilde{\mu}_p}(\zeta_j + \epsilon\, \mathrm{e}^{\imath\theta}) \left(\log\left(-m_{\tilde{\mu}_p}(\zeta_j + \epsilon\, \mathrm{e}^{\imath\theta})\right) - 1\right) \imath\epsilon\, \mathrm{e}^{\imath\theta}\, d\theta = 0.$$

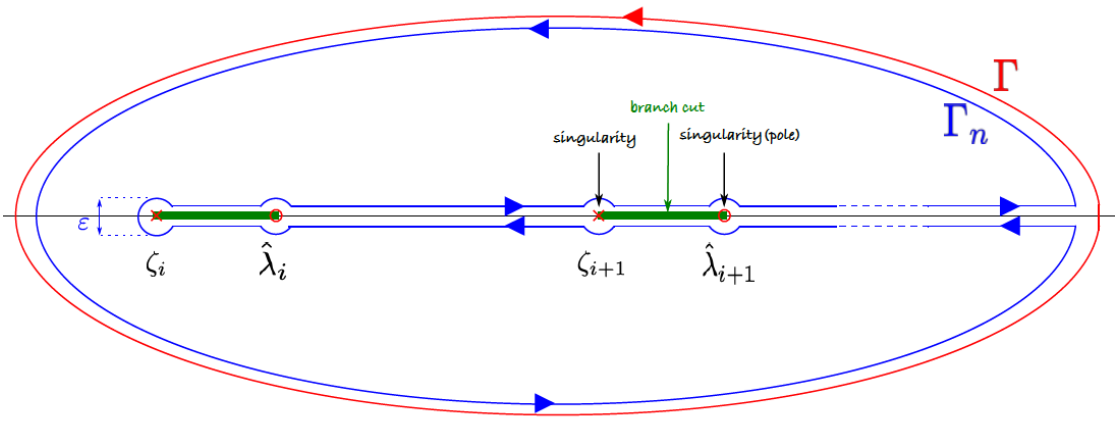

Figure 2.1: Contour deformation

- Integrals $\mathscr{I}_2$ over the circles surrounding the poles $\{\hat{\lambda}_j\}_{j=1}^p$ of order 1 which can be computed by using the residue theorem:

$$\begin{aligned}
\mathscr{I}_2 &= \frac{1}{c_0}\sum_{j=1}^{p}\lim_{z\to\hat{\lambda}_j} G\left(-m_{\tilde{\mu}_p}(z)\right)\left(z-\hat{\lambda}_j\right) \\
&= \frac{-1}{c_0}\sum_{j=1}^{p}\lim_{z\to\hat{\lambda}_j}(z-\hat{\lambda}_j)m_{\tilde{\mu}_p}(z)\left(\log\left(-m_{\tilde{\mu}_p}(z)\right)-1\right) \\
&= \frac{1}{p}\sum_{j=1}^{p}\lim_{z\to\hat{\lambda}_j}\left(\log\left(-m_{\tilde{\mu}_p}(z)\right)-1\right) \\
&= \frac{-1}{p}\sum_{j=1}^{p}\log(\hat{\lambda}_j)+\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p}\lim_{\epsilon\to 0}\log\left(\frac{\hat{\lambda}_j-\zeta_i+\epsilon}{\hat{\lambda}_i-\hat{\lambda}_j+\epsilon}\right)-1.
\end{aligned}$$

- Real integrals $\mathscr{I}_3$ over the segments $[\zeta_j,\hat{\lambda}_j]$ which can be computed by remarking that the log function has a discontinuity of $2\imath\pi$ at the branch cut.

$$\begin{aligned}
\mathscr{I}_3 &= \frac{-1}{2\imath\pi c_0}\sum_{j=1}^{p}\int_{\zeta_j+\epsilon}^{\hat{\lambda}_j-\epsilon} m_{\tilde{\mu}_p}(z)\left(\log\left(|m_{\tilde{\mu}_p}(z)|\right)+\imath\pi-\log\left(|m_{\tilde{\mu}_p}(z)|\right)+\imath\pi\right)dz \\
&= \frac{-1}{c_0}\sum_{j=1}^{p}\int_{\zeta_j+\epsilon}^{\hat{\lambda}_j-\epsilon} m_{\tilde{\mu}_p}(z)dz \\
&= \frac{-1}{c_0}\sum_{j=1}^{p}\int_{\zeta_j+\epsilon}^{\hat{\lambda}_j-\epsilon}\left(\frac{c_0-1}{z}+\frac{c_0}{p}\sum_{i=1}^{p}\frac{1}{\hat{\lambda}_i-z}\right)dz
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{p} \left[\frac{1-c_0}{c_0} \log(z) + \frac{1}{p} \sum_{i=1}^{p} \log(\hat{\lambda}_i - z) \right]_{\zeta_j+\epsilon}^{\hat{\lambda}_j-\epsilon} \\
&= \frac{1-c_0}{c_0} \sum_{j=1}^{p} \log(\frac{\hat{\lambda}_j}{\zeta_j}) + \lim_{\epsilon \to 0} \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} \log \left(\frac{\hat{\lambda}_j - \hat{\lambda}_i + \epsilon}{\hat{\lambda}_i - \zeta_j + \epsilon} \right) \\
&= \frac{c_0 - 1}{c_0} \log(1-c_0) - \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} \lim_{\epsilon \to 0} \log \left(\frac{\hat{\lambda}_i - \zeta_j + \epsilon}{\hat{\lambda}_j - \hat{\lambda}_i + \epsilon} \right)
\end{aligned}
$$

where in the last equality we used, among other algebraic simplifications, the fact that $\sum_{j=1}^{p} \log\left(\frac{\zeta_j}{\hat{\lambda}_j}\right) = \lim_{z \to 0} \log\left(-z m_{\tilde{\mu}_p}(z)\right) = \log(1-c_0)$.

- The sought for integral $\mathscr{I}_4$ over Γ for $\epsilon \to 0$.

Since the contour Γ_n doesn't contain any poles or branch points, the sum of the four above integrals reduces to zero. Combining these integrals then yields to the solution of the integral over the contour Γ denoted $\mathscr{I}_4$:

$$\mathscr{I}_4 = \frac{1}{p} \log(|\hat{\Sigma}|) + \frac{1-c_0}{c_0} \log(1-c_0) + 1,$$

where we retrieve the result of (Bai & Silverstein, 2008).

Integration contour determination. As already anticipated in Step (3) of the procedure, we need to ensure that the change of variable performed in Step (3) moves any complex contour closely encircling the support of μ onto a valid contour encircling the support of ν; we will in particular be careful that the resulting contour, in addition to encircling the support of ν, does not encircle additional values possibly bringing undesired residues (such as 0). We will proceed by showing that a contour encircling μ results on a contour encircling ν. These details rely heavily on the works of (Silverstein & Choi, 1995) and follow similar ideas as in e.g., (Couillet et al., 2011).

Let us consider a first contour Γ_μ closely around the support of μ (in particular not containing 0). We have to prove that any point z of this contour is mapped to a point ω of a contour Γ_ν closely around the support of ν.

The change of variable performed in Step (3) of the proposed methodology reads, for all $z \in \mathbb{C} \setminus \text{Supp}(\mu)$,

$$\omega \equiv \omega(z) = \frac{-1}{m_{\tilde{\mu}}(z)}.$$

It therefore remains to show that real z's (outside the support of μ) project onto properly located real ω's (i.e., on either side of the support of ν). This conclusion follows from the seminal work (Silverstein & Choi, 1995) on the spectral analysis of

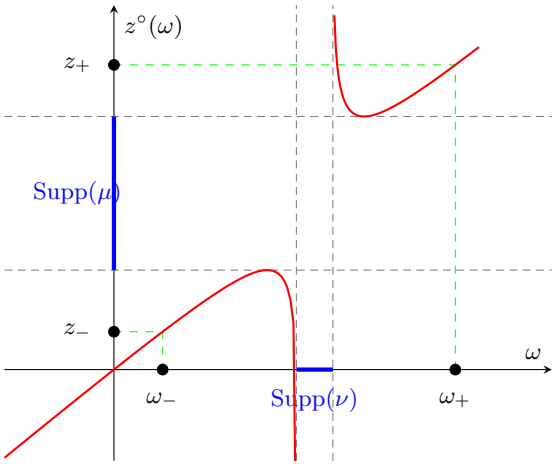

Figure 2.2: Variable change $\omega \mapsto z^\circ(\omega) = \omega + c_0^\infty \int \frac{t\nu(dt)}{1-t/\omega}$. Supp($\theta$) is the support of the probability measure θ.

sample covariance matrices. The essential idea is to note that, due to (2.4), the relation $\omega(z) = -1/m_{\tilde{\mu}}(z)$ can be inverted as

$$z \equiv z(\omega) = -\frac{1}{m_{\tilde{\mu}}} + c_0^\infty \int \frac{t\nu(dt)}{1 + tm_{\tilde{\mu}}} = \omega + c_0^\infty \int \frac{t\nu(dt)}{1 - \frac{t}{\omega}}.$$

In (Silverstein & Choi, 1995), it is proved that the image by $z(\cdot)$ of $\omega(\mathbb{R} \setminus \text{Supp}(\nu))$ coincides with the increasing sections of the function $z^\circ : \mathbb{R} \setminus \text{Supp}(\nu) \to \mathbb{R}$, $\omega \mapsto z(\omega)$. The latter being an explicit function, its functional analysis is simple and allows in particular to properly locate the real pairs (z, ω). Details of this analysis are provided in (Silverstein & Choi, 1995), which shall not be recalled here. The function z° is depicted in Figure 2.2 in the case of $c_0^\infty < 1$. In the case of $c_0^\infty > 1$, some undesired effects appear; these are discussed more in details in Chapter 3.

Summarizing and introducing Chapter 3. This section shows through the statistical estimation of $\log(|\Sigma|)$, how and why classical estimates fail and how Random Matrix Theory can be used to mitigate this problem. In chapter 3 of this thesis, we will go beyond this simple case. Specifically, we will be interested in computing the distance between two covariance matrices let's say Σ_1 and Σ_2 which can be expressed as a linear spectral statistic of the matrix $\Sigma_1^{-1}\Sigma_2$ or $\Sigma_1\Sigma_2$. Such distances will induce two technical issues compared to the previous case:

- The matrix of interest will not be a single matrix Σ but a more involved matrix $\Sigma_1^{-1}\Sigma_2$ known as F-matrix or $\Sigma_1\Sigma_2$ for which asymptotic spectral distribution should be derived.
- The complex integral obtained from Cauchy's integral will involve non-trivial real integrals.

These aspects will be discussed in detail in Chapter 3.

Beyond the eigenvalue functionals of the covariance matrix studied so far, some machine learning methods involve more complex functionals that involve the eigenvectors of the covariance matrix as will be the case in multi-task learning. Thus, it is important to introduce the tools necessary to characterize the general behavior of covariance matrices. Deterministic equivalents are tools that aim at this goal. This tool will be important to explain the behavior of algorithms such as ridge regression, multi-task learning, etc and is introduced next.

2.3 Deterministic equivalent of random matrices

Setting the stage. We previously raised in the introduction that most of the quantities involved in signal processing and machine learning applications involve complex dependency which limits the capability of classical learning theory to provide theoretical analysis. The example of ridge regression (close to the Least Square Support Vector Machine Multi-Task Learning tackled in Chapter 4) sounds particularly appealing. Indeed, the decision score for a new test data $\mathbf{x}$ is given from 1.4 by

$$g(\mathbf{x}) = \frac{1}{\sqrt{n}}\omega^{\mathsf{T}}\mathbf{x} = \frac{1}{n}\mathbf{x}^{\mathsf{T}}\left(\frac{1}{n}XX^{\mathsf{T}} + \lambda I_p\right)^{-1} Xy \tag{2.14}$$

which involves a bilinear form of $\left(\frac{1}{n}XX^{\mathsf{T}} + \lambda I_p\right)^{-1}$.

We should mention here that in the context of classification, the data matrix X will not be assumed anymore to be of zero-mean. We should therefore distinguish between the generalized population covariance matrix defined as $C = \mathbb{E}[\frac{1}{n}XX^{\mathsf{T}}]$ from the population covariance matrix $\Sigma = \mathbb{E}[\frac{1}{n}\left(X - \mathbb{E}[X]\right)\left(X - \mathbb{E}[X]\right)^{\mathsf{T}}]$ except for zero-mean data. We define similarly the generalized sample covariance matrix $\hat{C} = \frac{1}{n}XX^{\mathsf{T}}$ and define its resolvent as $Q(z) = \left(\hat{C} - zI_p\right)^{-1}$.

Beyond the eigenvalues of $\hat{C}$ (which are globally the same as those of $\hat{\Sigma}$) treated for now in the context of linear spectral statistic, our interest will also be on bilinear functionals of $Q(z) = \left(\hat{C} - zI_p\right)^{-1}$ therefore involving possibly the eigenvectors of $Q(z)$. As such, beyond studying the trace of the resolvent $Q(z)$, our interest is also on characterizing $Q(z)$ itself.

Precisely, we shall study so-called deterministic equivalents of $Q(z)$, the precise definition of which is given in Definition 2 and that can be described as deterministic matrices $\bar{Q}$ verifying $\mathrm{tr}A\left(Q(z) - \bar{Q}(z)\right) \xrightarrow{\text{a.s.}} 0$ when A has unit norm. Note that if we control $\mathrm{tr}\left(A(Q(z) - \bar{Q}(z))\right)$, we also control $u^{\mathsf{T}}Q(z)v = \mathrm{tr}\left(vu^{\mathsf{T}}Q(z)\right)$ for two deterministic vectors $u, v \in \mathbb{R}^p$ with unit norm.

Definition 2 (Deterministic equivalents)**.** *A deterministic equivalent, say $\bar{F} \in \mathbb{R}^{n\times p}$, of a given random matrix $F \in \mathbb{R}^{n\times p}$, denoted $F \leftrightarrow \bar{F}$, is defined by the fact that, for any deterministic linear functional $f : \mathbb{R}^{n\times p} \to \mathbb{R}$, $f(F - \bar{F}) \to 0$ almost surely (for instance,*

for u, v of unit norm, $u^\mathsf{T}(F - \bar{F})v \to 0$ and, for $A \in \mathbb{R}^{p\times n}$ deterministic of bounded operator norm, $\frac{1}{n}\mathrm{tr}A(F - \bar{F}) \to 0$).

In the following lines, for the sake of simplicity and in order to better grasp the main ideas of the derivation, we provide a high-level proof for the deterministic equivalent of the generalized sample covariance matrix $\hat{C}$.

Deterministic equivalent of $Q(z)$. First, let us note that a naive deterministic equivalent for $\left(\hat{C} - zI_p\right)^{-1}$ would be $(C - zI_p)^{-1}$. This turns out to be incorrect under $p, n \to \infty$. Instead, letting $\bar{Q}(z) = (\bar{C} - zI_p)^{-1}$ where $\bar{C}$ is some deterministic matrix to determine. To that end, we may first compute the difference using in particular the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for invertible matrix A and B:

$$\bar{Q} - \mathbb{E}[Q] = \mathbb{E}[Q(\frac{1}{n}XX^\mathsf{T} - \bar{C})\bar{Q}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Q(x_i x_i^\mathsf{T} - \bar{C})\bar{Q}]$$

where we recall that x_i is the i-th column of X.

Here, to go further, we need to make explicit the dependence between x_i and the matrix Q in order to evaluate the expectation of the product Qx_i. Let us denote $X_{-i} \in \mathbb{R}^{p\times(n-1)}$, the matrix X deprived of its i-th column, which leads us to define the matrices $\hat{C}_{-i} = \frac{1}{n}X_{-i}X_{-i}^\mathsf{T}$ and $Q_{-i} = (\hat{C}_{-i} - zI_p)^{-1}$.

To handle the dependence between x_i and Q, we will exploit the classical Schuur identities (Sherman-Morrison identity for example):

$$Q = Q_{-i} - \frac{1}{n}\frac{Q_{-i}x_i x_i^\mathsf{T} Q_{-i}}{1 + \frac{1}{n}x_i^\mathsf{T} Q_{-i}x_i}, \quad \text{and} \quad Qx_i = \frac{Q_{-i}x_i}{1 + \frac{1}{n}x_i^\mathsf{T} Q_{-i}x_i}.$$

The second inequality allows us to disentangle the relation between Q and x_i in the product Qx_i with a similar but easier to apprehend product $Q_{-i}x_i$ and a factor $\frac{1}{1+\frac{1}{n}x_i^\mathsf{T}Q_{-i}x_i}$ easily controllable term. This leads us to

$$\bar{Q} - \mathbb{E}[Q] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Q_{-i}\left(\frac{x_i x_i^\mathsf{T}}{1 + \frac{1}{n}x_i^\mathsf{T}Q_{-i}x_i} - \bar{C}\right)\bar{Q}\right] + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{Q_{-i}x_i x_i^\mathsf{T}Q_{-i}\bar{C}\bar{Q}}{1 + \frac{1}{n}x_i^\mathsf{T}Q_{-i}x_i}\right]. \tag{2.15}$$

Due to the supplementary factor $\frac{1}{n}$ and under appropriate assumption on the data matrix X, the norm of the rightmost random matrix is negligible (see more details in (Louart & Couillet, 2018)). Thus, if one assumes, say, that the random vectors $\{x_i\}_{i=1}^n$ follow the same law, one would choose naturally $\bar{C} = \frac{C}{1+\delta}$, with $\delta = \frac{1}{n}\mathbb{E}[x_i^\mathsf{T}Q_{-i}x_i] = \frac{1}{n}\mathrm{tr}\left(\mathbb{E}[Q_{-i}]\mathbb{E}[x_i x_i^\mathsf{T}]\right) = \frac{1}{n}\mathrm{tr}\left(C\bar{Q}\right)$. One may establish an implicit equation for δ not involving expectations over Q (or Q_{-i}). The deterministic equivalent is given by (Louart & Couillet, 2018).

Theorem 5. *Let a data matrix $X = [x_1, ..., x_n] \in \mathbb{R}^{p\times n}$ be distributed as a mixture of concentrated random vectors as per definition 3 in Appendix. Then the resolvent of*

the generalized sample covariance matrix defined as $Q(z) = \left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}$ *admits a deterministic equivalent* $\bar{Q}(z)$ *given by*

$$Q(z) \leftrightarrow \bar{Q}(z) = \left(\frac{C}{1+\delta(z)} - zI_p\right)^{-1}$$

where $\delta(z)$ *is the unique solution to the fixed point equation defined as*

$$\delta(z) = \frac{1}{n}\mathrm{tr}\left(C\left(\frac{C}{1+\delta(z)} - zI_p\right)^{-1}\right).$$

As introduced before, the interest of the deterministic equivalent is to allow to compute bilinear forms involving the random matrix $Q(z)$. In particular, $\frac{1}{p}\mathrm{tr}Q(z) - m_\mu(z) \to 0$ (where we retrieve the Generalized Marčenko-Pastur distribution) and $a^\mathsf{T}Q(z)b - a^\mathsf{T}\bar{Q}(z)b \to 0$ for deterministic $a, b \in \mathbb{R}^p$ of bounded Euclidean norm.

Universality. Theorem 5 has been proved in (Louart & Couillet, 2018) under a mixture of concentrated random vectors. The concentrated random vector assumption (introduced rigorously in definition 3 in Appendix) better models realistic datasets by imposing very little structure on the data. Exactly, it only constrains all *Lipschitz functionals* $\mathbb{R}^{p\times n} \to \mathbb{R}$ of X (i.e., its typical observations) to satisfy a concentration inequality; while this may seem demanding, the family of concentrated random vectors in fact contains all Lipschitz generative models (for instance neural networks) fed by Gaussian inputs (such as Generative Adversarial Network (GAN)s (Goodfellow et al., 2014)), as well as all further Lipschitz transformations of these vectors (for instance, features extracted by pretrained neural networks). It naturally comes that most derivations performed in Chapter 4 are *universal* in the sense of its being robust to a broad range of very realistic random data.

Revisiting the ridge regression. Let's consider data distributed in two classes $\mathscr{C}_1$ and $\mathscr{C}_2$ with opposite means and isotropic covariances. Specifically, we consider a data matrix $X = [X^{(1)}, X^{(2)}]$ with data of class ℓ, $X^{(\ell)} = [x_1^{(\ell)}, \ldots, x_{n_\ell}^{(\ell)}] \in \mathbb{R}^{p\times n_\ell}$, such that for $x_i^{(\ell)} \in \mathscr{C}_\ell$ for $\ell \in \{1, 2\}$,

$$x_i^{(\ell)} = (-1)^\ell \mu + \omega_i^\ell, \quad \omega_i^\ell \sim \mathscr{N}(0, I_p)$$

where $\mu \in \mathbb{R}^p$ with $\lim_{p\to\infty} \|\mu\| < \infty$. Let's suppose that the data are arranged in classes in the data matrix X and denoting $y = [\mathbb{1}_{n_1}, -\mathbb{1}_{n_2}]^\mathsf{T}$. Under this setting, the goal is to study the statistical behavior of the test score $g(\mathbf{x})$ for a new test data $\mathbf{x} \sim \mathscr{N}(\mu_x, I_p)$ independent from X:

$$g(\mathbf{x}) = \frac{1}{n}\mathbf{x}^\mathsf{T}\left(\frac{1}{n}XX^\mathsf{T} + \lambda I_p\right)^{-1} Xy = \frac{1}{n}\mathbf{x}^\mathsf{T}Q(-\lambda)Xy.$$

First, it can easily be proved by conditioning on the training data X that $g(\mathbf{x})$ is asymptotically Gaussian as the scalar product of a deterministic vector with a Gaussian isotropic vector. Moreover, the mean can be computed as follows given that $\mathbf{x} \in \mathscr{C}_\ell$:

$$\mathbb{E}[g(\mathbf{x})] = \mathbb{E}_X \left[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|X]\right] = \mathbb{E}_X \left[\sum_{i=1}^{n} \frac{(-1)^\ell}{n} \mu^\mathsf{T} Q(-\lambda) x_i y_i\right]$$

where we recall that $Q(-\lambda) = \left(\frac{1}{n} XX^\mathsf{T} + \lambda I_n\right)^{-1}$. We shall write sometimes for simplicity Q instead of $Q(z)$ when there are no ambiguities.

The above expectation requires to manage the statistical dependencies between Q and x_i, which can be handled using:

$$Qx_i = \frac{Q_{-i} x_i}{1 + \frac{1}{n} x_i^\mathsf{T} Q_{-i} x_i}. \tag{2.16}$$

As already seen in the derivation of the deterministic equivalent, the quadratic term $\frac{1}{n} x_i^\mathsf{T} Q_{-i} x_i \xrightarrow{\text{a.s.}} \delta(-\lambda)$. Therefore following the same line of reasoning as in (Seddik et al., 2020, Proposition A.3),

$$\begin{aligned}
\mathbb{E}[g(\mathbf{x})] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \frac{(-1)^\ell}{1+\delta} \mu^\mathsf{T} Q_{-i} x_i y_i\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \\
&= \frac{1}{n} \frac{(-1)^\ell}{1+\delta} \sum_{i=1}^{n} \mu^\mathsf{T} \bar{Q} \mathbb{E}[x_i] y_i + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)
\end{aligned} \tag{2.17}$$

where $\bar{Q}$ is the deterministic equivalent of Q.

As for the mean, the variance σ^2 can be computed as follows:

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\left(g(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})]\right)^2\right] \\
&= \frac{1}{n^2} \mathbb{E}\left[y^\mathsf{T} X^\mathsf{T} \left(\frac{1}{n} XX^\mathsf{T} + \lambda I_p\right)^{-2} Xy\right] \\
&= \frac{1}{n} \mathbb{E}\left[y^\mathsf{T} \left(\frac{1}{n} X^\mathsf{T} X + \lambda I_n\right)^{-2} \frac{X^\mathsf{T} X}{n} y\right] \\
&= \frac{1}{n} \mathbb{E}\left[y^\mathsf{T} \left(\frac{1}{n} X^\mathsf{T} X + \lambda I_n\right)^{-1} y - \lambda y^\mathsf{T} \left(\frac{1}{n} X^\mathsf{T} X + \lambda I_n\right)^{-2} y\right] \\
&= \frac{1}{n} \mathbb{E}\left[y^\mathsf{T} \tilde{Q} y - \lambda y^\mathsf{T} \tilde{Q}^2 y\right]
\end{aligned} \tag{2.18}$$

where $\tilde{Q} = \left(\frac{1}{n} X^\mathsf{T} X + \lambda I_n\right)^{-1}$.

We therefore need a deterministic equivalent of Q, $\tilde{Q}$ and $\tilde{Q}^2$ given by Corollary 1 as a consequence of Theorem 5.

Corollary 1 (Deterministic equivalents for $Q(z)$, $\tilde{Q}(z)$ and $\tilde{Q}(z)^2$)**.** *Under the setting described above, deterministic equivalents of $Q(z)$, $\tilde{Q}(z)$ and $\tilde{Q}^2(z)$ are given by:*

$$
\begin{aligned}
Q(z) &\leftrightarrow m(z)I_p - \frac{m(z)^2}{1+(c_0+\|\mu\|^2)m(z)}\mu\mu^\mathsf{T} \\
\tilde{Q}(z) &\leftrightarrow \tilde{m}(z)I_n - \frac{\tilde{m}(z)^2\|\mu\|^2}{1+(1+\|\mu\|^2)\tilde{m}(z)}yy^\mathsf{T} \\
\tilde{Q}(z)^2 &\leftrightarrow \tilde{m}'(z)I_n - \frac{2\tilde{m}'(z)\tilde{m}(z)\|\mu\|^2 + \tilde{m}'(z)\tilde{m}(z)^2\|\mu\|^2(1+\|\mu\|^2)}{\left(1+(1+\|\mu\|^2)\tilde{m}(z)\right)^2}yy^\mathsf{T}
\end{aligned}
$$

where

$$
\begin{aligned}
m(z) &= \frac{1-c_0-z+\sqrt{(1-c_0-z)^2-4zc_0}}{2zc_0} \\
\tilde{m}(z) &= \frac{-1+c_0-z+\sqrt{(1-c_0-z)^2-4zc_0}}{2z}
\end{aligned}
$$

and $\tilde{m}'(z)$ is the derivative of $m(z)$.

Corollary 1 is a special case of Theorem 5 for $C = I_p + \mu\mu^\mathsf{T}$ (see details in Appendix). Using the deterministic equivalents in Corollary 1, Theorem 6 unfolds directly.

Theorem 6 (Asymptotics of $g(\mathbf{x})$)**.** *Under the Gaussian mixture model introduced above, for $\mathbf{x} \sim \mathcal{N}(\mu_\mathbf{x}, I_p)$ with $\mu_\mathbf{x} = \mu_\ell$, as $p, n \to \infty$*

$$
g(\mathbf{x}) - G_\ell \to 0, \quad G_\ell \sim \mathcal{N}((-1)^\ell \mathfrak{m}, \sigma^2)
$$

in distribution where

$$
\begin{aligned}
\mathfrak{m} &= \frac{m(-\lambda)\|\mu\|^2}{(1+(c_0+\|\mu\|^2)m(-\lambda)))} \\
\sigma^2 &= \frac{\left(1+\tilde{m}(-\lambda)\right)^2\left(\tilde{m}(-\lambda)-\lambda\tilde{m}'(-\lambda)\right)+\tilde{m}^2(-\lambda)\|\mu\|^2\left(1+\tilde{m}(-\lambda)-\lambda\tilde{m}'(-\lambda)\right)}{\left(1+(1+\|\mu\|^2)\tilde{m}(-\lambda)\right)^2}.
\end{aligned}
$$

Since $g(\mathbf{x})$ has a Gaussian limit centered about $\pm\mathfrak{m}$, the (asymptotic) standard decision for $\mathbf{x}$ to be allocated to Class 1 ($\mathbf{x} \to \mathscr{C}_1$) or Class 2 ($\mathbf{x} \to \mathscr{C}_2$) is obtained by the "averaged-mean" test

$$
g(\mathbf{x}) \underset{\mathscr{C}_2}{\overset{\mathscr{C}_1}{\gtrless}} 0 \tag{2.19}
$$

the classification error rate $\epsilon \equiv \frac{1}{2}P(\mathbf{x} \to \mathscr{C}_2|\mathbf{x} \in \mathscr{C}_1) + \frac{1}{2}P(\mathbf{x} \to \mathscr{C}_1|\mathbf{x} \in \mathscr{C}_2)$ of which is then

$$
\epsilon \equiv P\left(g(\mathbf{x}) \underset{\mathscr{C}_2}{\overset{\mathscr{C}_1}{\gtrless}} 0\right) = \mathcal{Q}\left(\frac{\mathfrak{m}}{\sigma}\right) + o(1) \tag{2.20}
$$

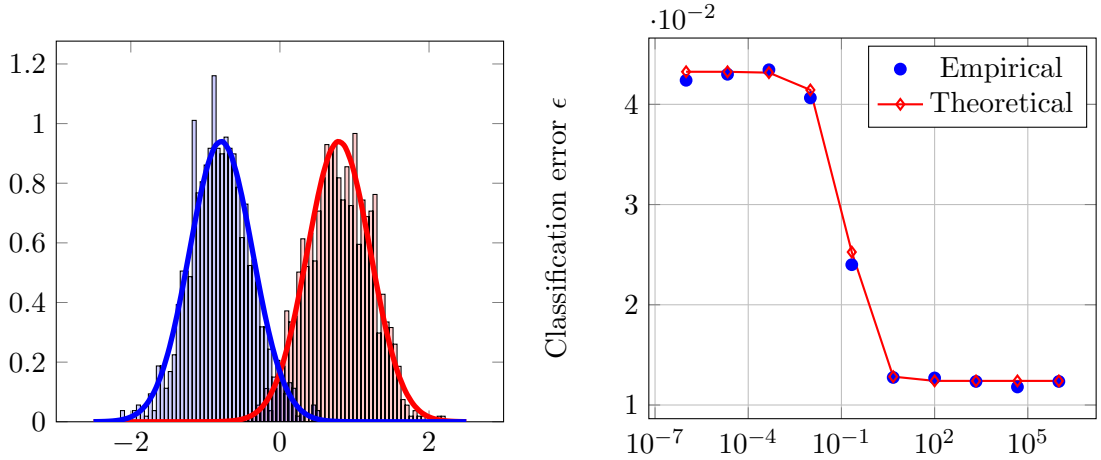

Figure 2.3: (Left) Score distribution [empirical histogram vs. theory in solid lines] for $\mathbf{x}$ of $\mathscr{C}_1$ (red) or Class $\mathscr{C}_2$ (blue) in a 2-class setting of isotropic balanced Gaussian mixtures. $p = 500$, $n = 600$, $\lambda = 1$ and $\mu \sim \frac{1}{10}\mathcal{N}(0, I_p)$ (Right) Theoretical and Empirical classification error averaged over 1000 test samples for the same setting as function of λ.

with $\mathscr{Q}(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2} dx$.

The close fit between the theoretical and the empirical prediction is illustrated in Figure 2.3.

This section illustrates that Random Matrix Theory provides the tools to predict theoretically the statistical behavior of $g(\mathbf{x})$ and therefore potentially to understand and optimize the ridge regression scheme. In chapter 4, the idea is to exploit these powerful tools to understand the functioning of more complex algorithms, closer to the reality of modern machine learning problems like multi-task learning with the following consequences:

- Multi-Task Learning is of huge interest in machine learning since it helps to leverage scarce labeled sample from similar tasks to help for the generalization performance. Since Transfer Learning phenomenon is prone to negative transfer, the proper understanding of their inner working mechanisms is of particular interest.

- From a technical point of view, in the multi-task learning framework, the data matrix X will be extended to a block diagonal matrix Z and therefore the Marčenko-Pastur law will not be applicable.

Chapter 3

Improved estimation of the distance between covariance matrices

Contents

3.1 Motivation and main findings

Motivation. Evaluating the distance between covariance matrices is at the core of many machine learning and signal processing applications. They are notably used for covariance features-based classification (for instance in brain signal or hyperspectral image classification), as well as for dimensionality reduction and representation of high dimensional points. Denote $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p\times p}$ two large dimensional covariance matrices for which we would like to compute the distance $D(\Sigma_1, \Sigma_2)$ based on few (p-dimensional) sample vectors. We assume that D can be written as a linear functional $\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i)$ of the eigenvalue distribution of either $\Sigma_1^{-1}\Sigma_2$ ($\lambda_i = \lambda_i(\Sigma_1^{-1}\Sigma_2)$) (as with the Fisher, Battacharrya, Kullback Leibler, Rényi divergences) or $\Sigma_1\Sigma_2$ ($\lambda_i = \lambda_i(\Sigma_1\Sigma_2)$) (for the Wasserstein distance, Frobenius distance).

Classical estimate. Based on a simple law-of-large-numbers argument, these metrics are commonly estimated from a simple replacement of the genuine $p \times p$-dimensional matrices Σ_1 and Σ_2 by their sample covariance estimates $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$. Such estimates, as shall be shown next, are however bound to sometimes extremely severe errors, particularly

when the respective numbers n_1 and n_2 of samples to estimate Σ_1 and Σ_2 are not large compared to p. This scenario is however frequently met in practice (short-time brain activity scans with high resolution EEG, large number of shortly-stationary assets in finance, high-resolution hyperspectral imaging, etc.) and therefore induces possibly weak data processing performances.

RMT-consistent estimates and main findings. To tackle these problems, in this work, we provide a random matrix improved consistent estimates for $D(\Sigma_1, \Sigma_2)$ for all aforementioned metrics. Technically speaking, our results rely on the following approach already introduced in chapter 2 in the simple case of the linear spectral statistic of the covariance matrix. We express a generic form of the metric under study under the form of a complex integral involving the *Stieltjes transform* of the (population) eigenvalue distribution of $\Sigma_1^{-1}\Sigma_2$ or $\Sigma_1\Sigma_2$. As the latter distribution is not accessible, we then link it to the (sample) eigenvalue distribution of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ (or $\hat{\Sigma}_1\hat{\Sigma}_2$), through a functional equation relating the Stieltjes transforms of population and empirical eigenvalue distribution similarly as in the generalized Marčenko-Pastur law. This results, through an appropriate change of variable, to a complex integral involving only the eigenvalues of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ (or $\hat{\Sigma}_1\hat{\Sigma}_2$), which may finally be evaluated using complex analysis techniques.

This approach already described in the linear spectral statistic of the covariance matrix in chapter 2 is notably inspired by the seminal work of Mestre (Mestre, 2008b) (see also (Couillet et al., 2011)) where functional estimates of the eigenvalue distribution of a single covariance matrix Σ is performed similarly from the corresponding eigenvalue distribution of the sample estimate $\hat{\Sigma}$. Aside from the more involved statistical models $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ and $\hat{\Sigma}_1\hat{\Sigma}_2$, the originality of the present work mostly lies in that the family of metrics involve non-smooth complex functionals (in particular logarithms and square roots) that result in more advanced technical considerations from real and complex analysis than in (Mestre, 2008b).

Moreover, although consistent for $n_1, n_2 \sim p$, these improved estimators still demand that $n_1, n_2 > p$ for all functions $f(z)$ having a singularity at $z = 0$ (e.g., $1/z$, $\log(z)$, $\log^2(z)$, $\sqrt{z}$). Based on a polynomial approximation of the functions of interest, we furthermore propose to retrieve consistent estimates for the challenging $n_2 < p$ scenario.

Chapter organization. The chapter is organized as follows. Section 3.2 introduces the main model and assumptions, Section 3.3 provides the general idea of the estimation and the problem induced by the challenging case $n_1 < p$ and $n_2 < p$ as well as a solution based on polynomial approximation. In Section 3.4, closed-form and numerically convenient expressions of the proposed estimators are derived for all functions of interest. Section 3.5 proposes a practical application to covariance matrix estimation and covariance-based feature classification.

Metrics	λ_i	$f(z)$
d_{F}^2	λ_i^-	$\log^2(z)$
d_{B}^2	λ_i^-	$-\frac{1}{4}\log(z)+\frac{1}{2}\log(1+z)-\frac{1}{2}\log(2)$
δ_{KL}	λ_i^-	$\frac{1}{2}z-\frac{1}{2}\log(z)-\frac{1}{2}$
$\delta_{\alpha\mathrm{R}}$	λ_i^-	$\frac{-1}{2(\alpha-1)}\log(\alpha+(1-\alpha)z)+\frac{1}{2}\log(z)$
d_{W}	λ_i^+	$-2\sqrt{z}$
d_{Fro}	λ_i^+	$-2z$

Table 3.1: Metrics and associated functions.

3.2 Models and Assumptions

For $a \in \{1,2\}$, let $X_a = [x_1^{(a)}, \ldots, x_{n_a}^{(a)}]$ be n_a independent and identically distributed random vectors with $x_i^{(a)} = \Sigma_a^{\frac{1}{2}} \tilde{x}_i^{(a)}$, where $\tilde{x}_i^{(a)} \in \mathbb{R}^p$ has zero-mean, unit variance and finite fourth-order moment entries. This holds in particular for $x_i^{(a)} \sim \mathcal{N}(0, \Sigma_a)$. In order to control the growth rates of n_1, n_2, p, we make the following assumption:

Assumption 1 (Growth Rates)**.** As $n_a \to \infty$, $p/n_a \equiv c_a \to c_a^\infty \in (0,\infty)$ and $\limsup_p \max\{\|\Sigma_a^{-1}\|, \|\Sigma_a\|\} < \infty$ for $\|\cdot\|$ the operator norm.

We should point out that the hypothesis $\limsup_p \max\{\|\Sigma_a^{-1}\|, \|\Sigma_a\|\} < \infty$ ensures that the limiting spectral distribution of the matrix of interest $\Sigma_1^{-1}\Sigma_2$ has bounded support away from zero. This will be particularly important to ensure that the limiting distribution of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ has bounded support which we will be used in particular to relate the non-asymptotic complex integral to their asymptotic counterpart. We define the sample covariance estimate $\hat{\Sigma}_a$ of Σ_a as

$$\hat{\Sigma}_a \equiv \frac{1}{n_a} X_a X_a^{\mathsf{T}} = \frac{1}{n_a} \sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\mathsf{T}}.$$

Our objective is to estimate the distance $D(\Sigma_1, \Sigma_2)$ between the covariance matrices Σ_1 and Σ_2 of the form:

$$D(\Sigma_1, \Sigma_2) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i)$$

where $\lambda_i = \lambda_i^-$ are the eigenvalues of $\Sigma_1^{-1}\Sigma_2$ or $\lambda_i = \lambda_i^+$ are the eigenvalues of $\Sigma_1\Sigma_2$. This form comprises, among others, the Fisher d_{F}^2 , Frobenius d_{Fro}, Wasserstein d_{W} and Battacharrya distances d_{B}^2, along with the Kullbach-Liebler δ_{KL} and Rényi divergences $\delta_{\alpha\mathrm{R}}$ as shown in details in table 3.1. The Wasserstein distance d_{W} between two zero-mean Gaussian distributions with covariances Σ_1 and Σ_2, respectively, assumes the form (Peyré & Cuturi, 2019, Remark 2.31): $d_{\mathrm{W}} = \mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\mathrm{tr}\left[(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\right]$. It is easily shown that, under Assumption 1, $\frac{1}{p}\mathrm{tr}\hat{\Sigma}_a - \frac{1}{p}\mathrm{tr}\Sigma_a \to 0$ almost surely. But estimating

$-2\mathrm{tr}\left[(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\right]$ is more involved and is the focus of this work. This explains why the corresponding function in the table is $-2\sqrt{z}$. The same holds for the Frobenius distance $d_{\mathrm{Fro}} = \mathrm{tr}\left(\Sigma_1^2 + \Sigma_2^2\right) - 2\mathrm{tr}\left(\Sigma_1\Sigma_2\right)$ for which the corresponding function of interest is $-2z$ since $\mathrm{tr}\left(\Sigma_i^2\right)$ can be estimated consistently (see more details in (Tiomoko & Couillet, 2019b)). The proposed estimate relies on random matrix theory tools developed in chapter 2 and particularly on the Stieltjes transform $m_\theta(z)$ of a probability distribution θ that we recall for convenience as:

$$m_\theta : \mathbb{C} \setminus \mathrm{supp}(\theta) \to \mathbb{C}, \quad z \mapsto \int (\lambda - z)^{-1} d\theta(\lambda).$$

The Stieltjes transform is here used to create a link between the population covariance eigenvalue distribution ν_p and the sample eigenvalue distribution μ_p (similarly as performed in the generalized Marčenko-Pastur law) defined by:

$$\nu_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}, \quad \mu_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\hat{\lambda}_i}$$

where $\hat{\lambda}_i = \hat{\lambda}_i^-$ are the eigenvalues of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ or $\hat{\lambda}_i = \hat{\lambda}_i^+$ are the eigenvalues of $\hat{\Sigma}_1\hat{\Sigma}_2$.

Similarly,

$$\nu_p^{\pm} = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i^{\pm}}, \quad \mu_p^{\pm} = \frac{1}{p}\sum_{i=1}^{p} \delta_{\hat{\lambda}_i^{\pm}}$$

and the corresponding Stieltjes transforms read in particular as

$$m_{\nu_p}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i - z}, \quad m_{\mu_p}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\hat{\lambda}_i - z}$$

for $\mu_p = \mu_p^{\pm}$ and $\nu_p = \nu_p^{\pm}$.

3.3 Improved estimate of the distance between covariance

Under the aforementioned setting, Theorem 7 provides a random matrix consistent estimate of these distances in the "easy" regime where $\lim p/n_a < 1$.

Theorem 7. *Let $f : \mathbb{C} \to \mathbb{C}$ be analytic on a contour $\Gamma_\mu \subset \{z \in \mathbb{C}, \mathscr{R}[z] > 0\}$ surrounding μ_p. Then under Assumption 1,*

$$\int f d\nu_p - \frac{1}{2\pi i}\oint_{\Gamma_\mu} f\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left[\frac{\psi_p'(z)}{\psi_p(z)} - \frac{\varphi_p'(z)}{\varphi_p(z)}\right]\frac{\psi_p(z)dz}{c_2} \xrightarrow{\text{a.s.}} 0$$

where

$$\varphi_p(z) = \begin{cases} z(1 + c_1 z m_{\mu_p}(z)), & \mu_p = \mu_p^- \\ \frac{z}{1 - c_1 - c_1 z m_{\mu_p}(z)}, & \mu_p = \mu_p^+ \end{cases}$$
$$\psi_p(z) = 1 - c_2 - c_2 z m_{\mu_p}(z).$$

The result of Theorem 7 has the strong advantage to be flexible to any smooth function f over $\{z \in \mathbb{C}, \mathscr{R}[z] > 0\}$, so in particular to $f(z) = \log^k(z)$ or $f(z) = \log^k(1 + \alpha z)$, which commonly appear in the distance between covariance matrices and divergences. The constraint $c_2 < 1$ is however mandatory and cannot be relaxed, unless f is analytic on all $\mathbb{C}$ (which fails for logarithm functions). Remark 2 also discusses these aspects (see also Section 6.2.2 for more details) and a solution is furthermore proposed in (Tiomoko & Couillet, 2019a) that we describe briefly in the end of the section.

Before getting to the proof, note that the formulation of Theorem 7 exhibits two important quantities, the functions φ_p and ψ_p, which both relate to the eigenvalue distribution of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ or $\hat{\Sigma}_1\hat{\Sigma}_2$ respectively through c_1 and c_2; each function therefore emphasizes the impact of the restricted number of data with respect to the dimension p.

We subsequently provide a sketch of proof of Theorem 7 (here for $\mu_p = \mu_p^-$). The detailed proofs are deferred into Section 6.2.1 in Appendix.

Sketch of Proof. Using the Cauchy integral formula similarly as in Step (1) of the procedure described in the linear spectral statistic of the population covariance matrix of Chapter 2, we have

$$\begin{aligned} D(\Sigma_1, \Sigma_2) &= \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i) = \int f(t)\nu_p(dt) \\ &= \frac{1}{2\pi\imath}\int \left[\oint_{\Gamma_\nu} \frac{f(z)}{z-t}\right] \nu_p(dt) = \frac{-1}{2\pi\imath}\oint_{\Gamma_\nu} f(z) m_{\nu_p}(z) dz. \end{aligned} \tag{3.1}$$

Thus, estimating $D(\Sigma_1, \Sigma_2)$ is equivalent to relating m_{ν_p} to m_{μ_p}. Since X_1 and X_2 are independent, we can condition first on X_2. By Theorem 3 given in chapter 2, the limiting eigenvalue distribution of $\Sigma_2\hat{\Sigma}_1^{-1}$, denoted ζ, can be written as a function of the limiting eigenvalue distribution of $\Sigma_2\Sigma_1^{-1}$, and similarly for the limiting eigenvalue distributions of $\hat{\Sigma}_2\hat{\Sigma}_1^{-1}$ and $\Sigma_2\hat{\Sigma}_1^{-1}$. This entails the two equations:

$$z m_{\mu_p}(z) = \varphi_p(z) m_{\zeta_p}(\varphi_p(z)) + o_p(1) \tag{3.2}$$

$$m_{\nu_p}(z/\psi_p(z)) = m_{\zeta_p}(z)\psi_p(z) + o_p(1). \tag{3.3}$$

where we follow the convention to use $o_p(1)$ for a sequence of random variables that convergences to zero in probability. Through the changes of variable $z \to \varphi_p(z)$ and $\omega \to \psi_p(\omega)$ applied in $\oint_{\Gamma_\nu} f(z) m_{\nu_p}(z) dz$, the result follows. □

Remark 1 (Known Σ_1)**.** *For Σ_1 known (which would mean that $\hat{\Sigma}_1 = \Sigma_1$ valid for p fixed and $n_1 \to \infty$), Theorem 7 is particularized by taking the limit $c_1 \to 0$, i.e.,*

$$\int f d\nu_p - \frac{1}{2\pi\imath}\oint_{\Gamma_\mu} f\left(\frac{z}{\psi_p(z)}\right)\left(-\frac{1}{z} + \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2} dz \xrightarrow{\text{a.s.}} 0$$

where now $m_{\mu_p}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i(\Sigma_1^{-1}\hat{\Sigma}_2) - z}$ and $m'_{\mu_p}(z) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{(\lambda_i(\Sigma_1^{-1}\hat{\Sigma}_2) - z)^2}$. Basic algebraic manipulations allow for further simplification, leading up to

$$\int f d\nu_p - \frac{-1}{2\pi\imath c_2}\oint_{\Gamma_\mu} f\left(\frac{-1}{m_{\tilde{\mu}_p}(z)}\right) m'_{\tilde{\mu}_p}(z) z dz \xrightarrow{\text{a.s.}} 0$$

where $\tilde{\mu}_p = c_2\mu_p + (1-c_2)\delta_0$ *is the eigenvalue distribution of* $\frac{1}{n_2}X_2^\intercal\Sigma_1^{-1}X_2$ *(and thus* $m_{\tilde{\mu}_p}(z) = c_2 m_{\mu_p}(z) - (1-c_2)/z$*). Letting* $g(z) = f(1/z)$ *and* $G(z)$ *such that* $G'(z) = g(z)$*, integration by parts of the above expression further gives*

$$\int f d\nu_p - \frac{1}{2\pi\imath c_2}\oint_{\Gamma_\mu} G\left(-m_{\tilde{\mu}_p}(z)\right) dz \xrightarrow{\text{a.s.}} 0.$$

For instance, for $f(z) = \log^2(z)$*,* $G(z) = z(\log^2(z) - 2\log(z) + 2)$*.*

We then retrieve the random matrix estimate of the population covariance matrix derived in chapter 2 for $\Sigma_1 = I_p$.

Remark 2 (On the need for $n_2 > p$ and $n_1 > p$ in Theorem 7)**.** *Since distances involving the eigenvalues of* $\Sigma_1^{-1}\Sigma_2$ *are estimated from the empirical matrix* $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$*, the constraint* $n_1 > p$ *is inevitable to ensure the existence of* $\hat{\Sigma}_1^{-1}$*. The requirement for* $n_2 > p$ *is less immediate. The two variable changes discussed in the proof of Theorem 7 are only licit if they realize a mapping from a contour* Γ_μ *enclosing the limiting support* $\mathrm{Supp}(\mu)$ *of* μ_p *to a valid contour* Γ_ν *enclosing the limiting support* $\mathrm{Supp}(\nu)$ *of* ν_p while *enclosing no additional singular points of the function* $f(z)$ *(otherwise the Cauchy formula used in* (3.1) *is incorrect). But for* $n_2 < p$*, it can be proved (see details in Section 6.2.2 of the appendix) that the pre-image of* Γ_μ *by the variable changes wraps around* $\mathrm{Supp}(\nu)$ and *around zero (with leftmost real crossing depending on the ratio* n_2/p*). This is a problem for all functions* $f(z)$ *singular at* $z = 0$*. In particular,* $1/z$*,* $\log(z)$*,* $\log^2(z)$ *and* $\sqrt{z}$ *are examples of such invalid functions which, for some, additionally have a branch-cut terminating at zero (that no valid contour may cross). This discussion is most conveniently illustrated in Figure 3.1. In this figure, we represent in blue the support of the distribution of* μ *(left) and* ν *(right). The branch cuts of the integrand are represented in red. For proper integration, the integration contour represented in green needs to encircle all the support* S_μ *while avoiding the branch cuts in red. For* $n_2 > p$*, this happens to be possible (top figure) while for* $n_2 < p$ *(bottom figure), this is impossible.*

Unfortunately, there seems to be no simple workaround to this situation. We (partially) solve in (Tiomoko & Couillet, 2019a) the problem by introducing entire *functions (thus analytical over* $\mathbb{C}$*) as substitutes for the locally non-analytic functions* $\log(z)$*,* $\log^2(z)$ *and* $\sqrt{z}$ *intervening in the distance* $D(\Sigma_1, \Sigma_2)$ *formulation.*

Our approach to extend the work to $p \geq n_2$ consists in approximating (arbitrarily closely) the analytic functions f under study that present singularities around zero by entire functions, and particularly degree-N polynomials $\tilde{f}_N(z)$ defined by $\tilde{f}_N(z) = \sum_{n=0}^N a_n z^n$.

Our central argument relies in the fact that, since $\|\Sigma_a\|$ and $\|\Sigma_a^{-1}\|$ are bounded (as per Assumption 1), the limiting support $\mathrm{Supp}(\nu)$ of ν_p is a compact set *strictly* away from zero. As such, one needs not approximate f on the whole $\mathbb{R}^+$ half-line (which would still pose problems in the vicinity of zero) but only on a subset $[a, b] \subset (0, \infty)$ over which polynomials are universal approximators.

This gives rise to the extension of Theorem 7 provided in (Tiomoko & Couillet, 2019a). For simplicity for the rest of this chapter, we will mostly focus on the case $p < n_2$ and

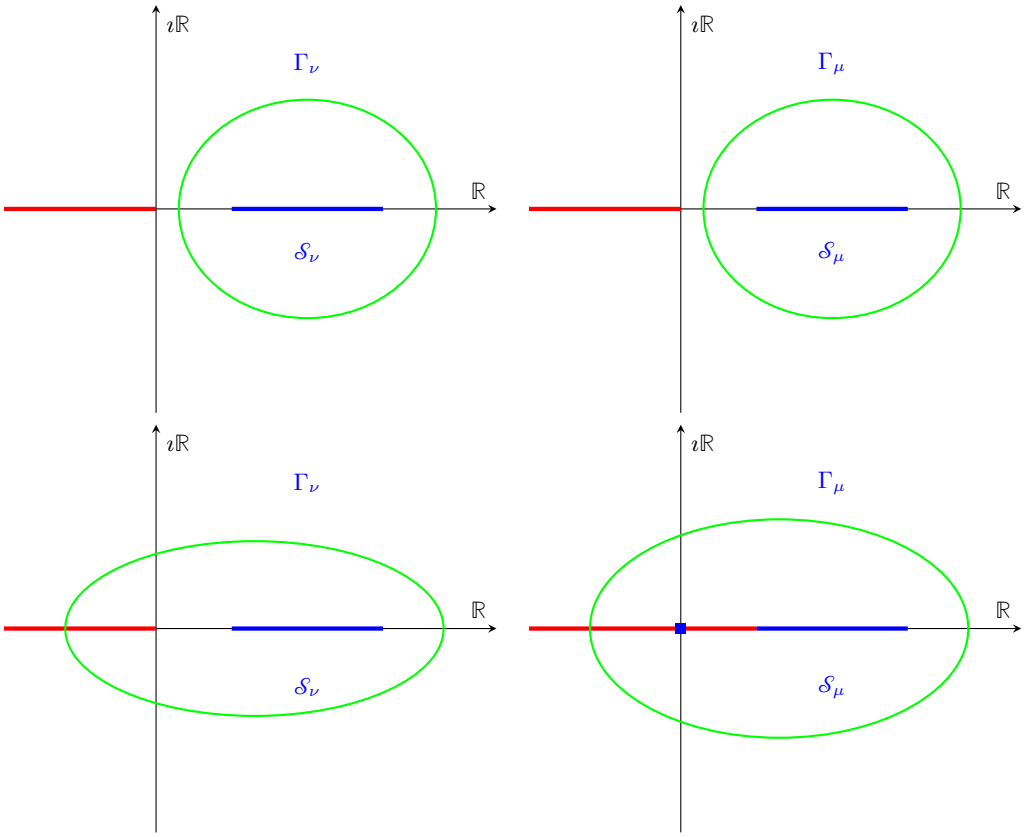

Figure 3.1: Illustration of the contours maps $\Gamma_\mu \mapsto \Gamma_\nu$ (from right to left) by the variable changes leading up to Theorem 7. (Top) $n_2 > p$. (Bottom) $n_2 < p$. For $n_2 < p$, the left real crossing of Γ_ν is necessarily negative (even if the mass at $\{0\}$ of $\mathrm{Supp}(\mu)$ were not included in Γ_μ). In case of singularities or branch-cuts (shown in red for the $\log(z)$ and $\sqrt{z}$ functions), the contours are invalid.

$p < n_1$. The interested reader can refer to (Tiomoko & Couillet, 2019a) for more details. We should however stress that (Tiomoko & Couillet, 2019a) only extends the result to $p > n_2$. The case $p > n_1$ is still an open research question.

3.4 Special cases

While Theorem 7 holds for all well-behaved f on Γ_μ, a numerical complex integral is required in practice to estimate $\int f d\nu_p$. It is convenient, when feasible, to assess the approximating complex integral in closed form, which is the objective of this section. When f is analytic in the inside of Γ_μ, the integral can be estimated merely through a residue calculus. This is the case notably of polynomials $f(t) = t^k$. If instead f exhibits singularities in the inside of Γ_μ, as for $f(t) = \log^k(t)$, more advanced contour integration arguments are required. These arguments (branch cuts in particular) were introduced in Chapter 2 and detailed for the generalized variance estimation problem. For the sake of readability, we briefly mention the technical difficulties involved for the current cases of interest and defer all the cumbersome algebraic simplifications in Section 6.2.3 in appendix.

Importantly, Theorem 7 is linear in f. Consequently, the contour integral calculus for elaborate functions f, such as met in most metrics of practical interest, can be reduced to the integral calculus of its linear components.

In the remainder of this section, we focus on the integral calculus for the atomic functions f listed in Table 3.1.

Corollary 2 (Case $f(t) = t$ and $\mu_p = \mu_p^-$)**.** *Under the conditions of Theorem 7,*

$$\int t d\nu_p(t) - (1 - c_1) \int t d\mu_p(t) \xrightarrow{\text{a.s.}} 0.$$

and in the case where $c_1 \to 0$, this is simply $\int t d\nu_p(t) - \int t d\mu_p(t) \xrightarrow{\text{a.s.}} 0$.

As such, the classical sample covariance matrix estimator $\int t d\mu_p(t)$ needs only be corrected by a product with $(1 - c_1)$. This result unfolds from Theorem 7 via a simple residue calculus.

Corollary 3 (Case $f(t) = \log(t)$ and $\mu_p = \mu_p^-$)**.** *Under the conditions of Theorem 7,*

$$\int \log(t) d\nu_p(t) - \left[\int \log(t) d\mu_p(t) - \frac{1 - c_1}{c_1} \log(1 - c_1) + \frac{1 - c_2}{c_2} \log(1 - c_2) \right] \xrightarrow{\text{a.s.}} 0.$$

When $c_1 \to 0$, $\int \log(t) d\nu_p(t) - \left[\int \log(t) d\mu_p(t) + \frac{1-c_2}{c_2} \log(1 - c_2) + 1 \right] \xrightarrow{\text{a.s.}} 0$.

Note interestingly that, for $f(t) = \log(t)$ and $c_1 = c_2$, the standard estimator is asymptotically p, n_1, n_2-consistent. This is no longer true though for $c_1 \neq c_2$ but only a fixed bias is induced. This result is less immediate as the complex extension of the logarithm function is multi-form, causing the emergence of branch-cuts inside the contour.

We evaluate the integral here, and in the subsequent corollaries, by means of a careful contour deformation subsequent to a thorough study of the function $\log(\varphi_p(z)/\psi_p(z))$ and to the identification of its branch-cut locations as performed in Chapter 2.

Corollary 4 (Case $f(t) = \log(1+st)$ and $\mu_p = \mu_p^-$)**.** *Under the conditions of Theorem 7, let $s > 0$ and denote κ_0 the unique* negative *solution to $1 + s\frac{\varphi_p(x)}{\psi_p(x)} = 0$. Then we have*

$$\int \log(1+st)d\nu_p(t) - \left[\frac{c_1+c_2-c_1c_2}{c_1c_2}\log\left(\frac{c_1+c_2-c_1c_2}{(1-c_1)(c_2-sc_1\kappa_0)}\right)\right.$$
$$\left.+\frac{1}{c_2}\log\left(-s\kappa_0(1-c_1)\right) + \int \log\left(1-\frac{t}{\kappa_0}\right)d\mu_p(t)\right] \xrightarrow{\text{a.s.}} 0.$$

In the case where $c_1 \to 0$, this is simply

$$\int \log(1+st)d\nu_p(t) - \left[\frac{1+s\kappa_0+\log(-s\kappa_0)}{c_2} + \int \log\left(1-\frac{t}{\kappa_0}\right)d\mu_p(t)\right] \xrightarrow{\text{a.s.}} 0.$$

The proof of Corollary 4 follows closely the proof of Corollary 3, yet with a fundamental variation on the branch-cut locations as the singularities of $\log(1+s\varphi_p(z)/\psi_p(z))$ differ from those of $\log(\varphi_p(z)/\psi_p(z))$.

As opposed to the previous scenarios, for the case $f(t) = \log^2(t)$, the exact form of the integral from Theorem 7 is non-trivial and involves dilogarithm functions (see its expression in (6.16) in the Appendix) which originate from numerous real integrals of the form $\int \log(x-a)/(x-b)dx$ appearing in the calculus. This involved expression can nonetheless be significantly simplified using a large-p approximation, resulting in an estimate only involving usual functions, as shown subsequently.

Corollary 5 (Case $f(t) = \log^2(t)$ and $\mu_p = \mu_p^-$)**.** *Let $0 < \eta_1 < \ldots < \eta_p$ be the eigenvalues of $\Lambda - \frac{\sqrt{\hat\lambda}\sqrt{\hat\lambda}^\mathsf{T}}{p-n_1}$ and $0 < \zeta_1 < \ldots < \zeta_p$ the eigenvalues of $\Lambda - \frac{\sqrt{\hat\lambda}\sqrt{\hat\lambda}^\mathsf{T}}{n_2}$, where $\hat\lambda = (\hat\lambda_1, \ldots, \hat\lambda_p)^\mathsf{T}$, $\Lambda = \operatorname{diag}(\hat\lambda)$, and $\sqrt{\hat\lambda}$ is understood entry-wise. Then, under the conditions of Theorem 7,*

$$\int \log^2(t)d\nu_p(t)$$
$$-\left[\frac{1}{p}\sum_{i=1}^p \log^2((1-c_1)\hat\lambda_i) + 2\frac{c_1+c_2-c_1c_2}{c_1c_2}\left\{\left(\Delta_\zeta^\eta\right)^\mathsf{T} M\left(\Delta_{\hat\lambda}^\eta\right) + (\Delta_{\hat\lambda}^\eta)^\mathsf{T} r\right\}\right.$$
$$\left.-\frac{2}{p}\left(\Delta_\zeta^\eta\right)^\mathsf{T} N 1_p - 2\frac{1-c_2}{c_2}\left\{\frac{1}{2}\log^2((1-c_1)(1-c_2)) + (\Delta_\zeta^\eta)^\mathsf{T} r\right\}\right] \xrightarrow{\text{a.s.}} 0$$

where we defined Δ_a^b the vector with $(\Delta_a^b)_i = b_i - a_i$ and, for $i, j \in \{1, \ldots, p\}$, $r_i = \frac{\log((1-c_1)\hat\lambda_i)}{\hat\lambda_i}$ and

$$M_{ij} = \begin{cases} \frac{\frac{\hat\lambda_i}{\hat\lambda_j} - 1 - \log\left(\frac{\hat\lambda_i}{\hat\lambda_j}\right)}{(\hat\lambda_i - \hat\lambda_j)^2} & , \ i \neq j \\ \frac{1}{2\hat\lambda_i^2} & , \ i = j \end{cases}, \quad N_{ij} = \begin{cases} \frac{\log\left(\frac{\hat\lambda_i}{\hat\lambda_j}\right)}{\hat\lambda_i - \hat\lambda_j} & , \ i \neq j \\ \frac{1}{\hat\lambda_i} & , \ i = j. \end{cases}$$

In the limit $c_1 \to 0$ (i.e., for Σ_1 known), this becomes

$$\int \log^2(t) d\nu_p(t) - \left[\frac{1}{p} \sum_{i=1}^{p} \log^2(\hat{\lambda}_i) + \frac{2}{p} \sum_{i=1}^{p} \log(\hat{\lambda}_i) - \frac{2}{p} \left(\Delta_\zeta^{\hat{\lambda}} \right)^\mathsf{T} Q 1_p \right.$$
$$\left. -2\frac{1-c_2}{c_2} \left\{ \frac{1}{2} \log^2(1-c_2) + \left(\Delta_\zeta^{\hat{\lambda}} \right)^\mathsf{T} q \right\} \right] \xrightarrow{\text{a.s.}} 0$$

with

$$Q_{ij} = \begin{cases} \frac{\hat{\lambda}_i \log\left(\frac{\hat{\lambda}_i}{\hat{\lambda}_j}\right) - (\hat{\lambda}_i - \hat{\lambda}_j)}{(\hat{\lambda}_i - \hat{\lambda}_j)^2} & , \ i \neq j \\ \frac{1}{2\hat{\lambda}_i} & , \ i = j \end{cases} \text{ , and } q_i = \frac{\log(\hat{\lambda}_i)}{\hat{\lambda}_i}.$$

Corollary 6 (Case $f(t) = \sqrt{(}t)$ and $\mu_p = \mu_p^+$)**.** *Let $\hat{\lambda}_1 \leq \ldots \leq \hat{\lambda}_p$, with $\hat{\lambda}_i \equiv \lambda_i(\hat{\Sigma}_1 \hat{\Sigma}_2)$, and define $\{\xi_i\}_{i=1}^p$ and $\{\eta_i\}_{i=1}^p$ the (increasing) eigenvalues of $\Lambda - \frac{1}{n_1}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T}$ and $\Lambda - \frac{1}{n_2}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T}$, respectively, where $\hat{\lambda} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_p\right)^\mathsf{T}$, $\Lambda = \mathrm{diag}(\hat{\lambda})$ and $\sqrt{\cdot}$ is understood entry-wise. Then, under Assumption 1,*

$$\int \sqrt{(}t) d\nu_p(t) - \hat{D}(X_1, X_2; \sqrt{\cdot}) \xrightarrow{\text{a.s.}} 0$$

where, if $n_1 \neq n_2$,

$$\hat{D}(X_1, X_2; \sqrt{\cdot}) = 2\sqrt{n_1 n_2} \frac{1}{p} \sum_{j=1}^{p} \sqrt{\hat{\lambda}_j}$$
$$+ \frac{2n_2}{\pi p} \sum_{j=1}^{p} \int_{\xi_j}^{\eta_j} \sqrt{-\frac{\varphi_p(x)}{\psi_p(x)}} \psi_p'(x) dx$$

with φ_p, ψ_p defined in Theorem 7 and, if $n_1 = n_2$,

$$\hat{D}(X_1, X_2; \sqrt{\cdot}) = \frac{2n_1}{p} \sum_{j=1}^{p} \left(\sqrt{\hat{\lambda}_j} - \sqrt{\xi_j} \right).$$

While still assuming an integral form (when $n_1 \neq n_2$), this formulation no longer requires the arbitrary choice of a contour Γ_μ and significantly reduces the computational time to estimate $D(\Sigma_1, \Sigma_2, \sqrt{\cdot})$. For $n_1 = n_2$, the expression is completely explicit and computationally only requires to evaluate the eigenvalues ξ_j of $\Lambda - \frac{1}{n_1}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T}$. The latter being a (negative definite) rank-1 perturbation of Λ, by Weyl's interlacing lemma (Franklin, 2012), the ξ_j's are interlaced with the $\hat{\lambda}_j$'s as

$$\xi_1 \leq \hat{\lambda}_1 \leq \xi_2 \leq \ldots \leq \xi_p \leq \hat{\lambda}_p.$$

As the $\hat{\lambda}_j$'s are of order $O(1)$ with respect to p, $|\hat{\lambda}_j - \xi_j| \leq |\hat{\lambda}_j - \hat{\lambda}_{j-1}| = O(p^{-1})$, therefore explaining why the expression of $\hat{D}(X_1, X_2; \sqrt{\cdot})$ is of order $O(1)$.

3.5 Applications

3.5.1 Confirmation of our results on synthetic data

In this section, we compare the Fisher distance estimate from Corollary 5 to the classical "plug-in" estimator $\frac{1}{p}\sum_{i=1}^{p}\log^2(\hat{\lambda}_i)$.

We first report in Table 3.2 the genuine versus estimated values of the Fisher distance on a synthetic setting (details in caption). A first surprising observation is that the plug-in estimator is extremely unfit to large values of p/n_1, p/n_2, bringing up to 500% error for $n_1 = 2p$; the proposed estimator is instead resilient to large p. Possibly more surprisingly, while Corollary 5 provably holds for asymptotically large p, n_1, n_2, our estimator already outperforms the standard approach for $p = 2$. This may be explained by the fact that the proposed approach essentially exploits randomness both from the size *and* the number of the dataset, with accuracies provably of order $\mathcal{O}(1/\sqrt{pn})$ thereby already reaching accurate values for not too large p (note that this in particular implies central limit theorems and thus convergence speed quadratically faster than in the large-n_1, n_2 alone setting).

p	$D_{\mathrm{F}}(\Sigma_1, \Sigma_2)$	Classical	Proposed
2	0.0980	0.1002	0.0973
4	0.1456	0.1520	0.1461
8	0.1694	0.1820	0.1703
16	0.1812	0.2081	0.1845
32	0.1872	0.2363	0.1886
64	0.1901	**0.2892**	0.1920
128	0.1916	**0.3955**	0.1934
256	0.1924	**0.6338**	0.1942
512	0.1927	**1.2715**	0.1953

(error > 50%) **(error > 100%)** **(error > 500%)**

Table 3.2: Proposed versus classical estimator for the Fisher distance between Σ_1 and Σ_2 with $[\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$, $x_i^{(a)} \sim \mathcal{N}(0, \Sigma_a)$; $n_1 = 1024$ and $n_2 = 2048$ for different values of p. Averaged over 10 000 trials.

3.5.2 Application to covariance features-based classification

In this section, we develop a practical application of our theoretical findings to the machine learning context of kernel spectral clustering (Von Luxburg, 2007). In several application domains, such as in brain signal processing (Rodrigues et al., 2018) or hyperspectral imaging (Chang, 2003), the relevant discriminative data "features" are the population covariance matrices of the data vectors (p-sensor brain activities, p-frequency spectra). Classification in these contexts is thus performed by comparing the distances between the population covariance matrices for each data-pair. Not being directly accessible, the

population covariances are classically substituted by their sample estimates. We will show here that this method has strong limitations that our proposed improved distance estimates overcome.

Specifically, we consider $m = 200$ data $X_1, \ldots, X_m$ to be clustered. Datum X_i is a $p \times n_i$ independent generation of n_i independent p-dimensional zero mean Gaussian samples $X_i = [x_1^{(i)}, \ldots, x_{n_i}^{(i)}] \in \mathbb{R}^{p \times n_i}$. Half of the samples have covariance $\mathrm{E}[x_j^{(i)} x_j^{(i)\mathsf{T}}] = \mathrm{E}[\frac{1}{n_i} X_i X_i^{\mathsf{T}}] = C_1$ (they will be said to belong to class 1) and half have covariance C_2 (they belong to class 2). For clarity, let us say that $\mathrm{E}[x_j^{(i)} x_j^{(i)\mathsf{T}}] = C_1$ for $i \leq m/2$ and $\mathrm{E}[x_j^{(i)} x_j^{(i)\mathsf{T}}] = C_2$ for $i > m/2$. We then define the kernel matrix $K \in \mathbb{R}^{m \times m}$ by

$$K_{ij} \equiv \exp(-\frac{1}{2} \hat{D}_{\mathrm{F}}(X_i, X_j)^2)$$

with $\hat{D}_{\mathrm{F}}$ either the classical (naive) estimator of D_{F} or our proposed random matrix-improved estimator. The purpose of spectral clustering is to retrieve the mapping between data indices and classes. It can be shown (see e.g., (Von Luxburg et al., 2008) but more fundamentally (Couillet et al., 2016) in this large dimensional context) that, for sufficiently distinct classes (so here covariance matrices), the eigenvectors v_1 and v_2 of K associated to its largest two eigenvalues are structured according to the classes. Thus, the classes can be read out of a two-dimensional display of v_2 versus v_1. This is our procedure in what follows.

A fundamental, but at first unsettling, outcome arises as we let $n_1 = \ldots = n_m$. In this case, for all tested values of p and m, the eigenvectors of K for either choice of $\hat{D}_{\mathrm{F}}$ are extremely similar. Consequently, spectral clustering performs the same, despite the obvious inaccuracies in the estimations of D_{F}. This result can be explained from the following fact: spectral clustering is of utmost interest when C_1 and C_2 are close matrices; in this case, the classical estimator for D_{F} is systematically biased by a constant, almost irrespective of the covariance (since both are essentially the same). This constant bias does affect K but not its dominant eigenvectors.

This observation collapses when the values of the n_i's differ. This is depicted in Figure 3.2. The top display provides a scatter plot of the dominant eigenvectors v_2 versus v_1 of K under the same conditions as above but now with n_i chosen uniformly at random in $[2p, 4p]$, with $p = 128$, $m = 200$, $C_1 = I_p$ and $[C_2]_{ij} = .05^{|i-j|}$. There, for the classical estimator, we observe a wide spreading of the eigenvector entries and a smaller inter-class spacing. This suggests poor clustering performance. On the opposite, the well-centered eigenvectors achieved by the proposed estimator imply good clustering performances. In a likely more realistic setting in practice, the bottom display considers the case where $n_1 = \ldots = n_{m-1} = 512$ and $n_m = 256$. This situation emulates a data retrieval failure for one observation (only half of the samples are seen). In this scenario, the classical estimator isolates one entry-pair in (v_1, v_2) (corresponding to their last entries). This is quite expected. However, more surprisingly, the presence of this outlier strongly alters the possibility to resolve the other data. This effect is further exacerbated when adding more outliers (not displayed here). This most likely follows from an adversarial effect

between the outliers and the genuine clusters, which all tend to "drive" the dominant eigenvectors.

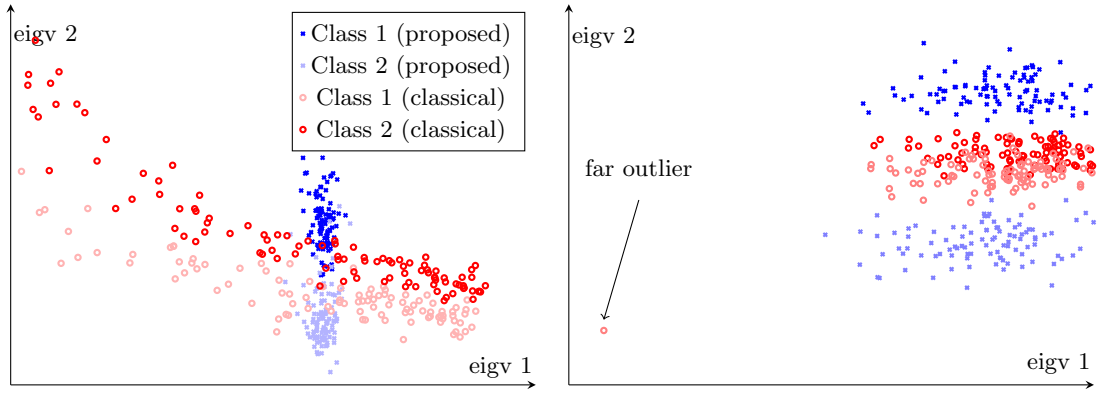

Figure 3.2: First and second eigenvectors of K for the traditional estimator (red circles) versus the proposed one (blue crosses); (top) random number of snapshots n_i; (bottom) $n_1 = \ldots = n_{m-1} = 512$ and $n_m = 256$.

3.5.3 Application to covariance matrix estimation

In this section we start presenting the covariance matrix estimation procedure for the family of metrics expressed as linear spectral statistic of $\Sigma_1^{-1}\Sigma_2$. We then furthermore explain the adaptations needed to handle the Wasserstein distance.

Estimation Method

In summary, our objective is to estimate Σ_2 as:

$$\check{\Sigma}_2 = \arg\min_M h(M) \tag{3.4}$$

$$h(M) = \hat{D}(M, X_2; f(\cdot))^2, \tag{3.5}$$

where $\hat{D}(M, X_2; f(\cdot))^2$ is the random-matrix estimate of the distance between any positive definite matrix M and the sought-for covariance matrix Σ_2 denoted $D(M, \Sigma_2) = \sum_{j=1}^{p} f(\lambda_j(M^{-1}\Sigma_2))$.

We solve (3.4) via a gradient descent algorithm on the Riemannian manifold S_p^{++} of positive definite $p \times p$ matrices.

The Riemannian gradient $\nabla h(M)$ of h at $M \in S_p^{++}$ is defined via the directional derivative $\mathrm{D}h(M)[\xi]$ of the functional $h : S_p^{++} \to \mathbb{R}^+$, at position $M \in S_p^{++}$ and in the direction of $\xi \in S_p$ (the vector space of symmetric $p \times p$ matrices), by (Absil et al., 2009)

$$\mathrm{D}h(M)[\xi] = \langle \nabla h(M), \xi \rangle_M^{S_p^{++}}$$

where $\langle \cdot, \cdot \rangle_{\cdot}^{S_p^{++}}$ is the Riemannian metric defined through

$$\langle \eta, \xi \rangle_M^{S_p^{++}} = \mathrm{tr}\left(M^{-1}\eta M^{-1}\xi\right).$$

$f(z)$	$G(z)$
$\log^2(z)$	$z\left(\log^2(z) - 2\log(z) + 2\right)$
$\log(z)$	$-z\log(z) + z$
$\log(1+sz)$	$s\log(s+z) + z\log\left(\frac{s+z}{z}\right)$
z	$\log(z)$

$f(z)$	$F(z)$
$\log^2(z)$	$z\left(\log^2(z) - 2\log(z) + 2\right)$
$\log(z)$	$z\log(z) - z$
$\log(1+sz)$	$\left(\frac{1}{s} + z\right)\log(1+sz) - z$
z	$\frac{1}{2}z^2$

Table 3.3: $G(z)$ and $F(z)$ for "atomic" $f(z)$ functions used in most distances and divergences under study; here $s > 0$ and $z \in \mathbb{C}$.

Differentiating $\hat{D}^2(M, X_2)$ at M in the direction ξ yields:

$$\begin{aligned}&\mathrm{D}h(M)[\xi]\\&= \frac{-\hat{D}(M,X_2)}{\pi i c_2}\oint_\Gamma g(-m_{\tilde{\mu}_p}(z,M))\mathrm{D}m_{\tilde{\mu}_p}(z,M)[\xi]dz.\end{aligned}$$

where $\hat{\Gamma}$ is a contour surrounding the support of the almost sure limiting eigenvalue distribution of $M^{-1}\hat{\Sigma}_2$. By using the fact that

$$\begin{aligned}&\mathrm{D}m_{\tilde{\mu}_p}(z,M)[\xi]\\&= \frac{c_2}{p}\mathrm{D}\mathrm{tr}\left(\left[M^{-1}\hat{\Sigma}_2 - zI_p\right]^{-1}\right)[\xi]\\&= \frac{c_2}{p}\mathrm{tr}\left(M^{-1}\hat{\Sigma}_2\left[M^{-1}\hat{\Sigma}_2 - zI_p\right]^{-2}M^{-1}\xi\right)\\&= \left\langle \frac{c_2}{p}\mathrm{sym}\left(\hat{\Sigma}_2\left[M^{-1}\hat{\Sigma}_2 - zI_p\right]^{-2}\right), \xi\right\rangle_M^{S_p^{++}}\end{aligned}$$

where $\mathrm{sym}(A) = \frac{1}{2}(A + A^\mathsf{T})$ is the symmetric part of $A \in \mathbb{R}^{p\times p}$, we retrieve the gradient of $h(M)$ as

$$\begin{aligned}&-\imath\pi p\frac{\nabla h(M)}{\hat{D}(M,X_2)}\\&= \oint_{\hat{\Gamma}} g\left(-m_{\tilde{\mu}_p}(z;M)\right)\mathrm{sym}\left(\hat{\Sigma}_2(M^{-1}\hat{\Sigma}_2 - zI_p)^{-2}\right)dz\end{aligned} \tag{3.6}$$

(recall that the right-hand side still depends on X_2 implicitly through $\tilde{\mu}_p$ and $\hat{\Sigma}_2$).

Once ∇h estimated, every gradient descent step in S_p^{++} corresponds to a small displacement on the geodesic starting at M and towards $-\nabla h(M)$, defined as the curve

$$\mathbb{R}_+ \to S_p^{++}$$

$$t \mapsto M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}} \nabla h(M) M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}$$

where, for $A = U\Lambda U^{\mathsf{T}} \in S_p^{++}$ in its spectral decomposition, $\exp(A) \equiv U \exp(\Lambda) U^{\mathsf{T}}$ (with exp understood here applied entry-wise on the diagonal elements of Λ).

That is, letting $M_0, M_1, \ldots$ and $t_0, t_1, \ldots$ be the successive iterates and step sizes of the gradient descent, we have, for some given initialization $M_0 \in S_p^{++}$,

$$M_{k+1} = M_k^{\frac{1}{2}} \exp\left(-t_k M_k^{-\frac{1}{2}} \nabla h(M_k) M_k^{-\frac{1}{2}}\right) M_k^{\frac{1}{2}}. \tag{3.7}$$

Our proposed method is summarized as Algorithm 1.

Algorithm 1 Proposed estimation algorithm.

Require $M_0 \in S_p^{++}$.

Repeat $M \leftarrow M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}} \nabla h(M) M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}$ with t either fixed or optimized by backtracking line search.

Until Convergence[1]

Return M.

Estimation of Σ_2^{-1}

In our framework, estimating Σ_2^{-1} rather than Σ_2 can be performed by minimizing $D(M, \Sigma_2^{-1})$ instead of $D(M, \Sigma_2)$. In this case, under Assumption 1,

$$D(M, \Sigma_2^{-1}) - \hat{D}^{\text{inv}}(M, X_2) \to 0$$

almost surely, for every deterministic M of bounded operator norm, where

$$\hat{D}^{\text{inv}}(M, X_2) \equiv \frac{1}{2\pi\imath c_2} \oint_{\hat{\Gamma}} F\left(-m_{\tilde{\mu}_p^{\text{inv}}}(z; M)\right) dz$$

for F such that $F'(z) \equiv f(z)$, where $\hat{\Gamma}$ is a contour surrounding the support of the almost sure limiting eigenvalue distribution of $M\hat{\Sigma}_2$ and $\tilde{\mu}_p^{\text{inv}} = \frac{p}{n_2}\mu_p^{\text{inv}} + (1 - \frac{p}{n_2})\delta_0$, where $\mu_p^{\text{inv}} \equiv \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(M\hat{\Sigma}_2)}$. The cost function to minimize under this setting is now given by $h^{\text{inv}}(M) \equiv (\hat{D}^{\text{inv}}(M, X_2))^2$ with gradient $\nabla h^{\text{inv}}(M)$ satisfying

$$\imath\pi p \frac{\nabla h^{\text{inv}}(M)}{\hat{D}^{\text{inv}}(M, X_2)} = \oint_{\hat{\Gamma}} f\left(-m_{\tilde{\mu}_p^{\text{inv}}}(z; M)\right) \text{sym}\left(M\hat{\Sigma}_2 (M\hat{\Sigma}_2 - zI_p)^{-2} M\right) dz.$$

With these amendments, Algorithm 1 can be adapted to the estimation of Σ_2^{-1}. Table 3.3 provides the values of F for the atomic functions f of interest.

[1] convergence reached whether one of the following conditions met: (i)Cost tolerance reached; (ii)Gradient norm tolerance reached; (iii)Maximum user time exceeded; (iv) Maximum iteration count reached

Application to Explicit Metrics

Algorithm 1 is very versatile as it merely consists in a gradient descent method for various metrics f through adaptable definitions of the function $h(M) = \hat{D}(M, X_2)^2$ and its resulting gradient. Yet, because of the integral form assumed by the gradient (Equation (3.6)), a possibly computationally involved complex integration needs to be numerically performed at each gradient descent step.

In this section, we specify closed-form expressions for the gradient for the atomic f functions of Table 3.3 (which is enough to cover the list of divergences in Table 3.1).

Let us denote

$$\nabla h(M) \equiv 2\hat{D}(M, X_2) \cdot \mathrm{sym}\left(\hat{\Sigma}_2 \cdot V \Lambda_\nabla V^{-1}\right)$$

where V are the eigenvectors of $M^{-1}\hat{\Sigma}_2$ and Λ_∇ is to be determined for each f.

For readability in the following, let us denote $\hat{\lambda}_i \equiv \lambda_i(M^{-1}\hat{\Sigma}_2)$, $i \in \{1, \dots, p\}$, the eigenvalues of the matrix $M^{-1}\hat{\Sigma}_2$ and $\xi_1, \dots, \xi_p$ the eigenvalues of $\Lambda - \frac{1}{n_2}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T}$ with $\Lambda = \mathrm{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_p)$ and $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_p)^\mathsf{T}$. Finally, for $s > 0$, let $\kappa_s \in (-1/(s(1-p/n_2)), 0)$ be the unique negative number t solution of the equation (see Section 6.2.3 of the appendix for details) $m_{\tilde{\mu}_p}(t) = -s$.

With these notations at hand, following the derivations in Section 6.2.3 of the appendix, we have the following determinations for Λ_∇.

Proposition 1 (Case $f(t) = t$)**.**

$$[\Lambda_\nabla]_{kk} = -\frac{1}{c_2} + \frac{1}{p}\sum_{i=1}^{p} \frac{1}{m'_{\tilde{\mu}_p}(\xi_i)\left(\hat{\lambda}_k - \xi_i\right)^2}$$

with $m'_{\tilde{\mu}_p}$ *the derivative of* $m_{\tilde{\mu}_p}$*.*

Proposition 2 (Case $f(t) = \log(t)$)**.** $[\Lambda_\nabla]_{kk} = \frac{-1}{p\hat{\lambda}_k}$.

Proposition 3 (Case $f(t) = \log(1+st)$)**.** *For* $s > 0$*,*

$$[\Lambda_\nabla]_{kk} = \frac{-1}{p(\hat{\lambda}_k - \kappa_s)}.$$

Proposition 4 (Case $f(t) = \log^2(t)$)**.** *For* $f(t) = \log^2(t)$*,*

$$[\Lambda_\nabla]_{kk} = \frac{2}{p}\log\left(\hat{\lambda}_k\right)\left[\sum_{i=1}^{p}\frac{1}{\hat{\lambda}_k - \xi_i} - \sum_{\substack{i=1\\ i\neq k}}^{p}\frac{1}{\hat{\lambda}_k - \hat{\lambda}_i} - \frac{1}{\hat{\lambda}_k}\right]$$
$$- \frac{2}{p}\sum_{i=1}^{p}\frac{\log(\xi_i)}{\hat{\lambda}_k - \xi_i} + \frac{2}{p}\sum_{\substack{i=1\\ i\neq k}}^{p}\frac{\log(\hat{\lambda}_i)}{\hat{\lambda}_k - \hat{\lambda}_i} - \frac{2 - 2\log(1-c_2)}{p\hat{\lambda}_k}.$$

These results unfold from residue calculus for entire functions f or advanced complex integration tools for logarithmic functions which is described in Section 6.2.4.

The Wasserstein case We obtain the relation of the gradient for the Wasserstein distance ($\frac{1}{p}\mathrm{tr}(M+\hat{\Sigma}_2) - 2\hat{D}(M, X_2; \sqrt{\cdot})$) similarly as provided previously.

$$\pi\imath p \frac{\nabla h(M)}{2\sqrt{h(M)}} = \frac{1}{p}M^2 + \sum_{j=1}^{p} \int_{\xi_j}^{\hat{\lambda}_j} \sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}} \mathrm{sym}\left(M\hat{\Sigma}_2(M\hat{\Sigma}_2 - xI_p)^{-2}M\right) dx$$

where $\mathrm{sym}(A) = \frac{1}{2}(A + A^{\mathsf{T}})$ is the symmetric part of $A \in \mathbb{R}^{p\times p}$ and where the definition of ξ_j and $\hat{\lambda}_j$ is given in Corollary 6. We can write the latter as:

$$\nabla h(M) = 2\sqrt{h(M)}\left[\mathrm{sym}\left(V\Lambda_\nabla V^{-1}\right) + \frac{1}{p}M^2\right]$$

where V is the orthogonal matrix of the eigenvectors of $M\hat{\Sigma}_2$ and Λ_∇ is the diagonal matrix with

$$[\Lambda_\nabla]_{kk} = \frac{1}{\pi p}\sum_{j\neq k}\int_{\xi_j}^{\hat{\lambda}_j} \sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}}\frac{1}{(\hat{\lambda}_k - x)^2}dx + \frac{1}{\pi p}\sum_{j\neq k}\int_{\xi_k}^{\hat{\lambda}_k} \sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}}\frac{1}{(\hat{\lambda}_j - x)^2}dx.$$

Experiments on the Wasserstein distance

Figure 3.3 depicts the results of the algorithm. There is displayed the Wasserstein distance $D_W(\Sigma_2, \cdot)$ between a matrix Σ_2 having four distinct eigenvalues of equal multiplicity (precisely, $\nu_p = \frac{1}{4}(\delta_{.1} + \delta_3 + \delta_4 + \delta_5)$) and various estimators of Σ_2: the sample covariance matrix (SCM), the state-of-the-art "non-linear shrinkage" estimators QuEST1 (Ledoit & Wolf, 2015) (based on a Frobenius distance minimization) and QuEST2 (Ledoit et al., 2018) (based on a Stein loss minimization), and the result of the gradient descent approach proposed in this section. For fair comparison, the iterative QuEST1, QuEST2 and our proposed method are all initialized at M_0 the linear shrinkage estimator from (Ledoit & Wolf, 2004). Note that our proposed choice of Σ_2 is particularly suited to mimick an "optimal transport" problem of displacing the eigenvalues of M_0 to the discrete four positions of the eigenvalues of Σ_2.

In addition to the computational simplicity of our gradient-descent approach with respect to the QuEST estimators (see the numerical method details in (Ledoit & Wolf, 2017)), the figure demonstrates significant gains brought by our proposed approach for large values of p/n_2, where the SCM particularly fails.

3.6 Concluding Remarks

Conclusion. The present study has revealed a strong lack of consistency for the traditional "plug-in" covariance matrix-distance (and divergence) estimators, when the

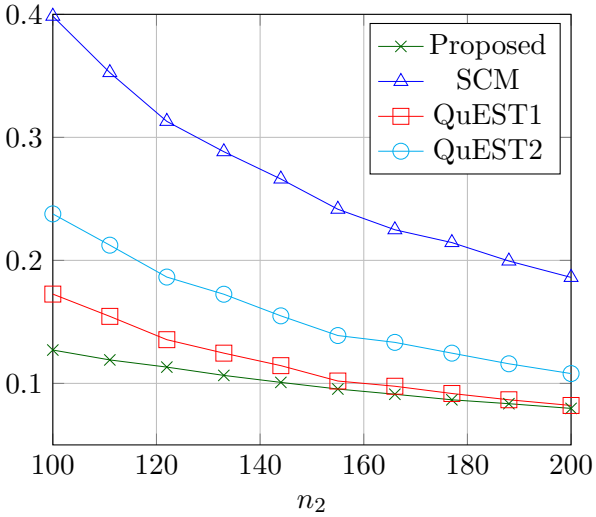

Figure 3.3: Wasserstein distance $D_W(\Sigma_2, \cdot)$ between Σ_2 with $\nu_p = \frac{1}{4}(\delta_{.1} + \delta_3 + \delta_4 + \delta_5)$ and (green) our proposed estimator, (blue) the sample covariance matrix, (red) and (light blue) the QuEST estimators proposed in (Ledoit et al., 2018; Ledoit & Wolf, 2015); for $p = 100$ and varying number of samples n_2 averaged over 10 realizations.

data dimension p is not small. This is particularly dramatic as p and the number of snapshots n are close. We provided a consistent solution to recover consistency, exploiting random matrix tools.

Importantly, by exploiting both randomness in p and n, our estimator converges as fast as $\mathcal{O}(1/\sqrt{pn})$, but a more precise central limit analysis is required to exactly assess confidence intervals, which is yet another avenue of research.

But the real strength and robustness of the proposed estimator will only be demonstrated when applied to real (non Gaussian) datasets and more exotic applications. Brain signal processing (or human-machine interaction) and radar imaging (Synthetic Aperture Radar (SAR) or hyperspectral) are both interesting application candidates that shall be investigated in the future.

Introducing the Multi-Task Learning analysis. Many analyses of machine learning algorithms have been conducted recently in high-dimensional statistics, but so far limited to"simple" algorithms (classical support vector machine (SVM) (Liao & Couillet, 2019), spectral clustering (Couillet et al., 2016), semi-supervised graph learning (SSL) (Mai & Couillet, 2018)). These studies have highlighted the prominent and unique role of means and covariance matrices in the understanding of these algorithms. This emphasizes once again the need for a thorough understanding of covariance functionals which has been the focus of the present chapter. However, richer and more complex methods such as multi-task learning, transfer learning, fairness learning, privacy and security in machine learning, etc. that involve multiple biases that are difficult to trace (too many parameters, too much heterogeneity in the data, etc.) are ultimately the ones that can gain the most from leveraging RMT, which will identify these biases. We show that this is indeed the

case, that there are multiple biases especially in the case of multi-task learning, but that the RMT cleans them one by one, while keeping a remarkable algorithmic simplicity. This is the focus and goal of the next chapter.

Chapter 4

Large dimensional analysis and improvements of Multi-Task Learning

Contents

4.1 Motivation and main findings

Multi-Task Learning motivation. The methodology for a long time considered in machine learning has consisted in tackling each given (classification, regression, estimation)

problem, hereafter referred to as a *task*, independently. This approach is in general counterproductive as it automatically discards a potentially rich source of data often available to perform more or less similar tasks. Multi-Task Learning (MTL) precisely aims to handle this deficiency by connecting datasets and tasks so to improve the generalization performance of one or several specific target tasks. This framework has recently gained renewed interest (Yang et al., 2020; Caruana, 1997; Collobert & Weston, 2008), given the availability of gigantic datasets (such as huge prelabeled image databases) and costly trained learning machines (such as deep neural nets), which must be useful to help solve learning tasks involving much fewer labeled data. Beyond this resurgence, numerous applications inherently benefit from an MTL approach, of which we may cite a few examples: prediction of student test results for a collection of schools (Aitkin & Longford, 1986), patient survival estimates in different clinics (Harutyunyan et al., 2017; Caruana et al., 1996), values of possibly related financial indicators (Allenby & Rossi, 1998), preference modeling of many individuals in a marketing context (Greene, 2000), etc.

Task relatedness modeling in MTL. Carefully modeling the relatedness between tasks has long been claimed to be the most critical determinant of the MTL algorithm performance. Several such models have been considered in the literature: task relatedness can be modeled by assuming that the parameters relating the tasks lie on a low dimensional manifold (Argyriou et al., 2007; Agarwal et al., 2010); these relating parameters may alternatively be assumed to be close in norm (Evgeniou & Pontil, 2004; Xu et al., 2013) or be distributed according to similar priors (Xue et al., 2007; Yu et al., 2005). However, for all these models, a failure in properly matching the task parameters is often likely to induce possibly severe cases of *negative learning*, that is occurrences where additional tasks play *against* rather than in favor of the target task objective. These cases of negative learning are difficult to anticipate as few theoretical works are amenable to prepare the experimenter to these scenarios. In the present work, we adopt a similar strategy as in (Evgeniou & Pontil, 2004), but with a strong theoretical background which will automatically eliminate the risks of negative learning.

A parameter-based modeling approach. In detail, the article (Evgeniou & Pontil, 2004), the spirit of which is followed here, is inspired by the natural extension of support vector machines (SVMs) (Vapnik, 2005) to a multiple, say k, task setting, by paralleling k SVMs but constraining their parameters (specifically, the k separating hyperplane normal vectors $\omega_1, \ldots, \omega_k$) to be "close" to each other. This is enforced by simply imposing that $\omega_i = \omega_0 + v_i$ for some common hyperplane normal vector ω_0 and dedicated hyperplane normal vectors v_i. The norm of the vectors v_i is controlled through an additional hyperparameter λ to strengthen or relax task relatedness. This is the approach followed in this chapter, to the noticeable exception that the fully explicit least-square SVM (LS-SVM) (Xu et al., 2013) rather than a margin-based SVM is considered. In addition to only marginally altering the overall behavior of the MTL algorithm of (Evgeniou & Pontil, 2004), the LS-SVM approach entails more explicit, more tractable, as well as more insightful results, let alone numerically cheaper implementations. As a matter of

fact, by a now well-established universality argument of large dimensional statistics, it has been shown in closely related works (Mai & Liao, 2019) that quadratic (least-square) cost functions are asymptotically optimal (as the data dimension and number increase) and uniformly outperform alternative costs (such as margin-based methods or logistic approaches) in terms of $0-1$ classification error, even in a classification setting; this argument further motivates to consider first and foremost the least square version of MTL-SVM.

Main findings and chapter contributions. This chapter develops a theoretical framework to exhaustively study the behavior and maximize the performance of a k-task m-class MTL LS-SVM framework, under the regime of numerous (n) and large (p) data, i.e., $n, p \to \infty$ with $n/p \to c_0 \in (0, \infty)$. The data are here modeled as a mixture of km *concentrated random vectors*, i.e., for $\mathbf{x}$ a data of class j ($j \in \{1, \ldots, m\}$) for Task i ($i \in \{1, \ldots, k\}$), $\mathbf{x} \sim \mathscr{L}_{ij}(\mu_{ij}, \Sigma_{ij})$, where $\mathscr{L}_{ij}(\mu, \Sigma)$ is the law of a Lipschitz-concentrated random vector (Ledoux, 2001) with statistical mean $\mu \in \mathbb{R}^p$ and covariance $\Sigma \in \mathbb{R}^{p \times p}$. For instance, $\mathbf{x} = \varphi_{ij}(\mathbf{z})$ for $\mathbf{z} \sim \mathcal{N}(0, I_q)$, $\varphi_{ij} : \mathbb{R}^q \to \mathbb{R}^p$ a 1-Lipschitz function and $\lim q/p \in (0, \infty)$. The statistical modeling for the data considered here has been briefly mentioned in chapter 2 as one of the big advantage of RMT tools in the sense that the theoretical insights and conclusions are robust to real data.

The main results and practical consequences of the chapter may be summarized as follows:

- we exhibit sufficient statistics, which concretely enable task comparison in the MTL LS-SVM algorithm; we show that, even when data are of large dimensions ($p \gg 1$), these statistics remain small dimensional (they only scale with the number k of tasks);
- while it is conventional to manually set labels associated to each dataset within $\{-1, 1\}$, we prove that this choice is largely suboptimal and may even cause MTL to severely fail (causing "negative transfer"); we instead provide the optimal values for the labels of each dataset, which depend on the sought-for objective: these optimal values are furthermore easily estimated from very few training data (i.e., no cross-validation is needed);
- for unknown new data $\mathbf{x}$, the MTL LS-SVM algorithm allocates a class based on the comparison of a score $g(\mathbf{x})$ to a threshold ζ, usually set to zero. We prove that, depending on the statistics and number of elements of the training dataset, a bias is naturally induced that makes $\zeta = 0$ a largely suboptimal choice in general. We provide a correction for this bias, which again can be estimated from the training data alone;
- we demonstrate on popular real datasets that our proposed optimized MTL LS-SVM is both resilient to real data and also manages, despite its not being a best-in-class MTL algorithm, to rival and sometimes largely outperform competing state-of-the-art algorithms.

In a nutshell, by exploiting random matrix theory tools explained in Section 2.3 (i,e,. deterministic equivalents of random matrices), the work provides a modern vision to multi-task and transfer learning. This vision is here turned into an elementary but cost-efficient algorithm, which relies on base principles, but which both largely outperforms competing (sometimes complex) methods and provides strong theoretical guarantees. As a side note, we must insist that our present objective is to study and improve "data-generic" multi-task learning mechanisms under no structural assumption on the data; this is quite unlike recent works exploiting convolutive techniques in deep neural nets to perform transfer or multi-task learning mostly for computer vision-oriented tasks, as in e.g., (Zhuang et al., 2020; Krishna & Kalluri, 2019).

Chapter organization. In order to best capture the main intuitions drawn from the large dimensional analysis, after a rigorous introduction of the multi-task learning framework in Section 4.2, a first highlight of our main contributions under the qualitatively more telling setting of binary tasks ($m = 2$) with data of equal identity covariance ($\Sigma_{ij} = I_p$) is proposed in Section 4.3. The technical details under the most generic data modeling setting as well as the most general technical result are then provided in Section 4.4. A broad series of applications is provided in Section 4.5. Extensive simulations are then proposed in Section 4.6, which corroborate our theoretical findings and show their resilience and compatibility to real data settings.

Notation. $e_m^{[n]} \in \mathbb{R}^n$ is the canonical vector of $\mathbb{R}^n$ with $[e_m^{[n]}]_i = \delta_{mi}$. Moreover, $e_{ij}^{[2k]} = e_{2(i-1)+j}^{[2k]}$. Similarly, $E_{ij}^{[n]} \in \mathbb{R}^{n\times n}$ is the matrix with $[E_{ij}^{[n]}]_{ab} = \delta_{ia}\delta_{jb}$. The notations $A \otimes B$ and $A \odot B$ for matrices or vectors A, B are respectively the Kronecker and Hadamard products. $\mathscr{D}_x$ is the diagonal matrix containing on its diagonal the elements of the vector x and $A_{i\cdot}$ is the i-th row of A. The notation $\mathring{A}$ is used when a centering operation is performed on the matrix or vector A. Uppercase calligraphic letters ($\mathscr{A}, \mathscr{K}, \Gamma, \mathscr{M}, \mathscr{V}$,...) are used for deterministic small dimensional matrices. Finally, $\mathbb{1}_m$ and I_m are respectively the vector of all one's of dimension m and the identity matrix of dimension $m \times m$. The index pair i, j generally refers to Class j in Task i.

4.2 The Multi-Task Learning Framework

4.2.1 The deterministic setting

Let $X \in \mathbb{R}^{p\times n}$ be a collection of n independent data vectors of dimension p. The data are divided into k subsets attached to individual "tasks", each task consisting of an m-class classification problem (m being the same for each task). Specifically, letting $X = [X_1, \ldots, X_k]$, Task i is a classification problem from the training samples $X_i = [X_i^{(1)}, \ldots, X_i^{(m)}] \in \mathbb{R}^{p\times n_i}$ with $X_i^{(j)} = [x_{i1}^{(j)}, \ldots, x_{in_{ij}}^{(j)}] \in \mathbb{R}^{p\times n_{ij}}$ the n_{ij} vectors of class $\mathscr{C}_j$, $j \in \{1, \ldots, m\}$, for Task i. In particular, $n = \sum_{i=1}^k n_i$ and $n_i = \sum_{j=1}^m n_{ij}$ for each $i \in \{1, \ldots, k\}$.

To each datum $x_{il}^{(j)} \in \mathbb{R}^p$ of the training set is attached a corresponding output vector (or score) $y_{il}^{(j)} \in \mathbb{R}^m$. Correspondingly to the notation X, X_i and $X_i^{(j)}$, let $Y = [Y_1^\mathsf{T}, \ldots, Y_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{n \times m}$ be the matrix of the m-dimensional outputs of all data, where $Y_i = [Y_i^{(1)\mathsf{T}}, \ldots, Y_i^{(m)\mathsf{T}}]^\mathsf{T} \in \mathbb{R}^{n_i \times m}$ and $Y_i^{(j)} = [y_{i1}^{(j)}, \ldots, y_{in_{ij}}^{(j)}]^\mathsf{T} \in \mathbb{R}^{n_{ij} \times m}$ the matrix of all outputs for Task i.

In the standard MTL learning approach (Evgeniou & Pontil, 2004; Xu et al., 2013), one would naturally set $y_{il}^{(j)} = e_j^{[m]}$, i.e., all data of class $\mathscr{C}_j$ are affected a hot-bit in position j. As claimed in the introduction and as we shall see, this hot-bit allocation approach is at the source of deleterious performances, such as negative transfer effects, and we thus voluntarily do not enforce any constraint on the vector $y_{il}^{(j)}$ at this point.

Before inserting the data-score pairs (X, Y) into the MTL LS-SVM framework, it is convenient to "center" the data X to eliminate additional sources of bias. This centering operation could be performed either on the whole dataset X, or task-wise on each X_i, or even class-wise on each $X_i^{(j)}$. In (Evgeniou & Pontil, 2004; Xu et al., 2013) this centering operation is not performed (which essentially boils down to centering X itself). We choose here to center the data task-wise, and this, for two reasons: (i) centering the whole dataset induces dependencies across tasks so that, even by enforcing the hyperplane controlling factor λ to decorrelate the tasks (i.e., $\lambda \to \infty$; see next), residual dependence must remain and negative transfer can still appear, (ii) class-wise centering has the double deleterious effect of canceling an important discrimination factor of the classes (i.e., their difference in statistical mean) and of necessitating a complex treatment to classify new (unlabeled) input data. Inappropriate centering choices would induce biases and undesired residual terms in our theoretical derivation, which further justifies our present task-wise centering choice (see e.g., Remark 4). Specifically, the MTL LS-SVM algorithm studied here is based, not on the data X_i but on their centered version

$$\mathring{X}_i = X_i \left(I_{n_i} - \frac{1}{n_i} \mathbb{1}_{n_i} \mathbb{1}_{n_i}^\mathsf{T} \right), \quad \forall i \in \{1, \ldots, k\},$$

and we will systematically consider the data-score pair $(\mathring{X}, Y)$, where $\mathring{X} = [\mathring{X}_1, \ldots, \mathring{X}_k]$ rather than (X, Y).

Having pre-treated the input data, we are in position to introduce the MTL LS-SVM framework. The MTL LS-SVM algorithm aims to predict, relative to each task i, an output score vector $\mathbf{y}_i \in \mathbb{R}^m$ for any new input vector $\mathbf{x} \in \mathbb{R}^p$. To this end, MTL LS-SVM determines k "hyperplane normal-vector" matrices $W = [W_1, W_2, \ldots, W_k] \in \mathbb{R}^{p \times km}$ which take the form $W_i = W_0 + V_i$ for some common W_0 and individual task-wise matrices $V = [V_1, \ldots, V_k]$ and biases $b = [b_1^\mathsf{T}, b_2^\mathsf{T}, \ldots, b_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{k \times m}$. These parameters are set to minimize the objective function

$$\min_{(W_0, V, b) \in \mathbb{R}^{p \times m} \times \mathbb{R}^{p \times km} \times \mathbb{R}^{k \times m}} \mathscr{J}(W_0, V, b) \tag{4.1}$$

where

$$\mathscr{J}(W_0, V, b) \equiv \frac{1}{2\lambda}\mathrm{tr}\left(W_0^\mathsf{T} W_0\right) + \frac{1}{2}\sum_{i=1}^{k}\frac{\mathrm{tr}\left(V_i^\mathsf{T} V_i\right)}{\gamma_i} + \frac{1}{2}\sum_{i=1}^{k}\mathrm{tr}\left(\xi_i^\mathsf{T}\xi_i\right)$$

$$\xi_i = Y_i - (\frac{\mathring{X}_i^\mathsf{T} W_i}{\sqrt{kp}} + \mathbb{1}_{n_i} b_i^\mathsf{T}), \quad \forall i \in \{1, \ldots, k\}.$$

This is a classical LS-SVM formulation in which the quadratic cost $\mathrm{tr}(\xi_i^\mathsf{T}\xi_i)$ replaces the boundary constraint of margin-based SVM and where the costs $\mathrm{tr}(W_0^\mathsf{T} W_0)$ and $\mathrm{tr}(V_i^\mathsf{T} V_i)$ are reminiscent of the hyperplane normal-vector norm minimization of classical SVM.

What is specific to the MTL approach is first the hyperparameter λ which enforces or relaxes the relatedness between tasks and the introduction of k extra parameters $\gamma_1, \ldots, \gamma_k$ which enforce a correct classification of the data in their respective classes. Similarly to (Evgeniou & Pontil, 2004), we place the hyperparameters γ_i as a prefactor of $\mathrm{tr}(V_i^\mathsf{T} V_i)$, rather than as a prefactor of $\mathrm{tr}(\xi_i^\mathsf{T}\xi_i)$; this differs from the normalization scheme proposed in (Xu et al., 2013). This choice is more flexible in the following sense: for a fixed value of λ, increasing all ratios $\frac{\lambda}{\gamma_i}$ "blurs" the difference between tasks and thus turns the optimization scheme into a single-task SVM (because the optimal V_i's need then be set to zero in the limit); for fixed values of the γ_i's instead, small ratios $\frac{\lambda}{\gamma_i}$ decorrelate the tasks (the optimal W_0 being close to zero). Note however that, unlike in (Evgeniou & Pontil, 2004), we choose to use here one hyperparameter γ_i per task instead of a common one. As will be seen next, this choice is more meaningful and of course offers more flexibility.

In passing, remark that the linear common-hyperplane condition $W_i = W_0 + V_i$, imposes by definition that all V_i's be of the same size $\mathbb{R}^{p\times m}$: this severely constrains (i) the data in each task to be of the same dimension p and (ii) the number of classes per task to be the same (m). Further linear or even non-linear relaxation schemes for W_i of the type $W_i = V_i + f_i(W_0)$ for some operator f_i could be envisioned to relax this constraint. This however goes beyond the scope of the chapter, which seeks to provide insights and optimality into a simplified (yet already non-trivial) form of MTL LS-SVM.

As for the choice of the hyperparameters λ, $\gamma_1, \ldots, \gamma_k$, as well as of the score matrix Y which we recall was left open, it is treated independently and is dictated, not by the present optimization scheme, but by an ultimate objective, such as minimizing the misclassification rate for a specific target class. These more applied considerations will be made in Section 4.5.

Remark 3 (LS-SVM classification versus regression)**.** *It may be disputed that the optimization framework* (4.1) *takes a regression rather than a classification form. It appears that, under a binary-class LS-SVM framework with scores $y_i \in \{\pm 1\}$, the classification constraint (of the form $y_i(W^T x_i + b_i) - 1 = \xi_i$) or the regression constraint (of the form $y_i - W^T x_i + b_i = \xi_i$) are associated to the same losses, thereby leading to the same classification solution and performance. Yet, as will become clear in the following, in addition for the solution of* (4.1) *to be explicit and theoretically tractable (which is not*

the case of alternative schemes such as margin-based SVM, logistic regression, Adaboost, etc.), the aforementioned flexibility in the score matrix Y largely outbalances the "failure" of treating a classification problem by means of a regression optimization scheme. Besides, under the large dimensional theoretical framework presently studied, recent works in related problems (Mai & Liao, 2019) forcefully suggest that the square loss is optimal to deal with large dimensional data as it uniformly outperforms all alternative cost functions in terms of $0-1$ classification error.

Being a quadratic cost optimization under linear constraints, (4.1) is easily solved using its dual formulation by introducing Lagrangian parameters $\alpha_i \in \mathbb{R}^{n_i \times m}$ for each task i (see details in Section 6.3.1). The solution is explicit and is as follows.

Proposition 5. *The solution to* (4.1) *is given by*

$$\begin{aligned} W_0 &= \left(\mathbb{1}_k^\mathsf{T} \otimes \lambda I_p\right) \frac{Z}{\sqrt{kp}}\alpha \\ W_i &= \left(e_i^{[k]\mathsf{T}} \otimes I_p\right) A \frac{Z}{\sqrt{kp}}\alpha \\ b &= (P^\mathsf{T} Q P)^{-1} P^\mathsf{T} Q Y \end{aligned}$$

where

$$\begin{aligned} Z &= \begin{pmatrix} \mathring{X}_1 & & \\ & \ddots & \\ & & \mathring{X}_k \end{pmatrix} \in \mathbb{R}^{kp \times n} \\ A &= \left(\mathscr{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k^\mathsf{T}\right) \otimes I_p \in \mathbb{R}^{kp \times kp} \\ \alpha &= Q(Y - Pb), \quad Q = \left(\frac{1}{kp} Z^\mathsf{T} A Z + I_n\right)^{-1} \in \mathbb{R}^{n \times n} \\ P &= \begin{pmatrix} \mathbb{1}_{n_1} & & \\ & \ddots & \\ & & \mathbb{1}_{n_k} \end{pmatrix} \in \mathbb{R}^{n \times k}. \end{aligned}$$

Despite the apparent intricate expression of W_i, it must be stressed that W_i "essentially" takes the form of the standard solution to a ridge regression considered in chapter 2 (or regularized least-square) problem as the term $AZQY$ (in which $Q = (\frac{1}{kp} Z^\mathsf{T} A Z + I_n)^{-1}$) appearing in the expended form of W_i confirms. From a technical standpoint, the large dimensional statistical behavior of the matrix Q, known as the *resolvent* of $\frac{1}{kp} Z^\mathsf{T} A Z$ in random matrix theory as defined in chapter 2, plays a central role in the analysis. More specific to the MTL framework, note the interesting isolation of the data subsets $\mathring{X}_i$ in the data matrix Z (it is not possible, to the best of our knowledge, to "linearly" express W_i as a function of $\mathring{X}$ itself); the elements $\mathring{X}_i$ are then "mixed" by the term $\lambda \mathbb{1}_k \mathbb{1}_k^\mathsf{T}$ appearing in matrix A, from which it naturally comes that, in the limit $\lambda \to 0$, MTL LS-SVM boils down to k independent LS-SVMs with $\mathscr{D}_\gamma$ imposing weights $\gamma_1, \ldots, \gamma_k$ on each data subset.

From Proposition 5, for any new data point $\mathbf{x} \in \mathbb{R}^p$, the classification score vector $g_i(\mathbf{x}) \in \mathbb{R}^m$ for Task i, is then defined by the linear model considered previously

$$g_i(\mathbf{x}) = \frac{1}{\sqrt{kp}} W_i^{\mathsf{T}} \mathring{\mathbf{x}} + b_i = \frac{1}{kp} \alpha^{\mathsf{T}} Z^{\mathsf{T}} A \left(e_i^{[k]} \otimes \mathring{\mathbf{x}} \right) + b_i \tag{4.2}$$

where $\mathring{\mathbf{x}} = \mathbf{x} - \frac{1}{n_i} X_i \mathbb{1}_{n_i}$ is a centered version of $\mathbf{x}$ with respect to the training dataset for Task i.

This formulation, along with the next remark, confirm again the relevance of a task-wise, rather than class-wise, centering of the data X, which allows for a well-defined expression of $\mathring{\mathbf{x}}$.

Remark 4 (Shift invariance of the scores)**.** *If the columns of $Y_i \in \mathbb{R}^{n_i \times m}$ are shifted by some constant vector $P\bar{\mathscr{Y}}$ for some (small dimensional) matrix $\bar{\mathscr{Y}} \in \mathbb{R}^{k \times m}$, i.e., if all data of the same task are affected by the same shift of their scores (or labels), then we find that the Lagrangian parameter α^{shift} after the shift is*

$$\alpha^{\text{shift}} = Q \left(I_n - P(P^{\mathsf{T}} Q P)^{-1} P^{\mathsf{T}} Q \right) (Y + P\bar{\mathscr{Y}}) = \alpha.$$

As such, the matrix $W_i = (e_i^{[k]\mathsf{T}} \otimes I_p) A \frac{Z}{\sqrt{kp}} \alpha$ and, consequently, the performance of MTL LS-SVM are insensitive to a simultaneous shift of all the scores of each task.

4.2.2 Statistical modeling and the large dimensional setting

In order to draw insights into the behavior of MTL LS-SVM and evaluate its performance, we propose to first model the dataset X as a mixture of concentrated random vectors and then to assume the dimensions p, n of X to be sufficiently large for deterministic (and predictable) concentration behavior to occur.

Assumption 2 (Distribution of X and $\mathbf{x}$)**.** There exist two constants $C, c > 0$ (independent of n, p) such that, for any 1-Lipschitz function $f : \mathbb{R}^{p \times n} \to \mathbb{R}$,

$$\forall t > 0, \ \mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq C e^{-(t/c)^2}$$

We further impose that the columns of X be independent and that the $x_{il}^{(j)}$, for $l \in \{1, \ldots, n_{ij}\}$, be distributed according to the same law $\mathscr{L}_{ij}$. These conditions guarantee the existence of a mean and covariance for the columns of X and we denote, for all $l \in \{1, \ldots, n_{ij}\}$,

$$\mu_{ij} \equiv \mathbb{E}[x_{il}^{(j)}]$$
$$\Sigma_{ij} \equiv \mathrm{Cov}[x_{il}^{(j)}].$$

Furthermore, the dummy variable $\mathbf{x} \in \mathbb{R}^p$ used for testing is independent of X, and distributed according to one of the laws $\mathscr{L}_{ij}$.

Assumption 2 notably encompasses the following scenarios: the $x_{il}^{(j)}$'s are (i) independent Gaussian random vectors $\mathcal{N}(\mu_{ij}, \Sigma_{ij})$, (ii) independent random vectors uniformly distributed on the $\mathbb{R}^p$ sphere of radius $\sqrt{p}$ and, most importantly, (iii) any 1-Lipschitz transformation $\varphi_{ij}(z_{il}^{(j)})$ with $z_{il}^{(j)}$ itself a concentrated random vector. Scenario (iii) is particularly relevant to model very realistic data by means of advanced non-linear generative models, as recently demonstrated in (Seddik et al., 2019) in the specific example of generative adversarial networks (GANs). As such, Assumption 2 offers the flexibility to assume either synthetic Gaussian mixture models, or very realistic and advanced generative data models. A core result of the chapter consists in showing that, for n, p large, either scenario leads to the same asymptotic performance for MTL LS-SVM (which thus only depends on the statistical means and covariances of the data).

Since all data $x_{il}^{(j)}$, $l \in \{1, \ldots, n_{ij}\}$, are identically distributed, we will further impose that their associated scores $y_{il}^{(j)} \in \mathbb{R}^m$ be identical. That is, $y_{i1}^{(j)} = \ldots = y_{in_{ij}}^{(j)} \equiv \mathscr{Y}_{ij}$ within every class j of each task i. The score matrix $Y \in \mathbb{R}^{n \times k}$ may then be reduced under the form

$$Y = \left[\mathscr{Y}_{11} \mathbb{1}_{n_{11}}^\mathsf{T}, \ldots, \mathscr{Y}_{km} \mathbb{1}_{n_{km}}^\mathsf{T} \right]^\mathsf{T} \in \mathbb{R}^{n \times m}$$

for $\mathscr{Y} = [\mathscr{Y}_{11}, \ldots, \mathscr{Y}_{km}]^\mathsf{T} \in \mathbb{R}^{km \times m}$. From Remark 4, it is also clear that, the performances of MTL LS-SVM being insensitive to a constant shift in the scores $\mathscr{Y}_{i1}, \ldots, \mathscr{Y}_{im}$ in every given task i, the centered version $\mathring{\mathscr{Y}} = [\mathring{\mathscr{Y}}_{11}, \ldots, \mathring{\mathscr{Y}}_{km}]^\mathsf{T}$ of $\mathscr{Y}$, where

$$\mathring{\mathscr{Y}}_{ij} \equiv \mathscr{Y}_{ij} - \sum_{j=1}^{m} \frac{n_{ij}}{n_i} \mathscr{Y}_{ij},$$

will naturally appear at the core of the upcoming results.

Although practical data will of course be considered to be of finite dimension p and number n, it will indeed be convenient, for technical reasons, to work under the following large dimensional random matrix assumption.

Assumption 3 (Growth Rate)**.** As $n \to \infty$, $n/p \to c_0 \in (0, \infty)$ and, for $1 \leq i \leq k$, $1 \leq j \leq m$, $n_{ij}/n \to c_{ij} \in (0, 1)$. We further denote $c_i = \sum_{j=m}^{k} c_{ij}$ and $c = [c_1, \ldots, c_k]^\mathsf{T} \in \mathbb{R}^k$.

With these notations and assumptions in place, we are in position to present the main results of the chapter. Yet, before entering the technical details of the large dimensional analysis of the performance of the MTL LS-SVM framework, the next section first provides a highlight of the main contributions and intuitions drawn by the analysis. To this end, it is convenient to temporarily restrict the setting to binary classes ($m = 2$) and to an isotropic mixture model for the data X, i.e., $\Sigma_{ij} = I_p$ for each measure $\mathscr{L}_{ij}$. The most general and slightly more technical setting ($m \geq 2$ and non-isotropic mixture data modeling) is considered in full in Section 4.4.

4.3 Highlights of the main results

To simplify the exposition of our main results, without impacting their core conclusions, in this section, Assumptions 2–3 are further restricted to the binary-classification setting ($m = 2$) and to measures $\mathscr{L}_{ij}$ of equal covariance $\Sigma_{ij} = I_p$, for all i, j.

The advantage of the isotropic ($\Sigma_{ij} = I_p$) condition is that all asymptotic results can be expressed under the form of low-dimensional matrix formulations (of size scaling with k but not with p, n). Adjoined to the $m = 2$ assumption, the isotropic model further guarantees a simplified form for (i) the (asymptotically) optimal labels Y, (ii) the optimal decision thresholds ζ_i, and (iii) the asymptotic performances of MTL LS-SVM, all of which can be estimated consistently as $p, n \to \infty$. Consequently, this simplified setting has the strong benefit to give rise to a first cost-efficient and robust multi-task classification algorithm (Algorithm 2) which, for practical data, makes the approximation that $\Sigma_{ij} \propto I_p$.

The binary setting does not a priori alter any of the previously introduced notations which stand with $m = 2$. Yet, it is particularly convenient in this setting to recast the score vectors $y_{il}^{(j)} \in \mathbb{R}^m$ into scalar scores $y_{il}^{\mathrm{bin}(j)} \in \mathbb{R}$. In a standard classification context, this would correspond to turning a two-dimensional hot-bit vector $e_j^{[2]}$ into a signed scalar ± 1; as we recall that $y_{il}^{(j)}$ is here considered as a *real* score (rather than a binary label) vector, to us this is equivalent to turning a score vector into a scalar score. Matrix $Y \in \mathbb{R}^{n \times m}$ similarly now becomes a score vector $y^{\mathrm{bin}} \in \mathbb{R}^n$, and in particular we define $\mathring{y}^{\mathrm{bin}} = [\mathring{y}_{11}^{\mathrm{bin}}, \ldots, \mathring{y}_{k2}^{\mathrm{bin}}]^\mathsf{T} \in \mathbb{R}^{2k}$ with

$$\mathring{y}_{ij}^{\mathrm{bin}} \equiv y_{ij}^{\mathrm{bin}} - \left(\frac{n_{i1}}{n_i} y_{i1}^{\mathrm{bin}} + \frac{n_{i2}}{n_i} y_{i2}^{\mathrm{bin}} \right) \in \mathbb{R}$$

where $y_{ij}^{\mathrm{bin}} \equiv y_{i1}^{\mathrm{bin}(j)} = \ldots = y_{in_{ij}}^{\mathrm{bin}\ (j)} \in \mathbb{R}$ is the common score assigned to the identically distributed data of class j for Task i. Correspondingly, the sought-for (W_i, b_i) collection of m hyperplanes of (4.1) becomes a single hyperplane $(w_i^{\mathrm{bin}}, b_i^{\mathrm{bin}})$ with $w_i^{\mathrm{bin}} \in \mathbb{R}^p$ and $b_i^{\mathrm{bin}} \in \mathbb{R}$. Yet, our present interest is only on the resulting score vector $g_i(\mathbf{x})$ which, replacing Y by y^{bin} in its expression (Equation 4.2), becomes the scalar test score

$$g_i^{\mathrm{bin}}(\mathbf{x}) \equiv \frac{1}{kp} (y^{\mathrm{bin}} - Pb)^\mathsf{T} Q Z^\mathsf{T} A \left(e_i^{[k]} \otimes \mathring{\mathbf{x}} \right) + [(P^\mathsf{T} Q P)^{-1} P^\mathsf{T} Q y^{\mathrm{bin}}]_i \in \mathbb{R}.$$

4.3.1 Theoretical analysis and large dimensional intuitions

Under the isotropic and binary-class setting, as $n, p \to \infty$ according to Assumption 3, the theoretical performance of MTL LS-SVM explicitly depends on two fundamental and isolated quantities: the data-related matrix $\mathscr{M} \in \mathbb{R}^{2k \times 2k}$ and the hyperparameter matrix $\mathscr{A} \in \mathbb{R}^{k \times k}$:

$$\mathscr{M} = \sum_{i,i'=1}^{k} \Delta\mu_i^\mathsf{T} \Delta\mu_{i'} \left(E_{ii'}^{[k]} \otimes \mathbb{c}_i \mathbb{c}_{i'}^\mathsf{T} \right)$$

$$\mathscr{A} = \left(I_k + \mathscr{D}_{\boldsymbol{\delta}^{[k]}}^{-\frac{1}{2}} \left(\mathscr{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k^\mathsf{T} \right)^{-1} \mathscr{D}_{\boldsymbol{\delta}^{[k]}}^{-\frac{1}{2}} \right)^{-1}$$

where we introduced the shortcut notations

$$\Delta\mu_i \equiv \mu_{i1} - \mu_{i2}, \quad \mathbb{c}_i \equiv \sqrt{c_{i1}/c_i}\sqrt{c_{i2}/c_i} \begin{bmatrix} \sqrt{c_{i2}/c_i} \\ -\sqrt{c_{i1}/c_i} \end{bmatrix}$$

and where $\boldsymbol{\delta}^{[k]} = [\boldsymbol{\delta}_1^{[k]}, \ldots, \boldsymbol{\delta}_k^{[k]}]^\mathsf{T}$ are the unique positive solutions to the implicit system of k equations

$$\boldsymbol{\delta}_i^{[k]} = \frac{c_i}{c_0} - \mathscr{A}_{ii}, \quad i \in \{1, \ldots, k\}. \tag{4.3}$$

In anticipation of future needs, it is convenient to further introduce the $2k$-dimensional variant $\boldsymbol{\delta}^{[2k]} = [\boldsymbol{\delta}_{11}^{[2k]}, \ldots, \boldsymbol{\delta}_{k2}^{[2k]}]^\mathsf{T} \in \mathbb{R}^{2k}$ where

$$\boldsymbol{\delta}_{ij}^{[2k]} = c_0 \frac{c_{ij}}{c_i} \boldsymbol{\delta}_i^{[k]}. \tag{4.4}$$

The asymptotic performances of MTL LS-SVM will be shown to solely depend on X through the matrices $\mathscr{M}$ and $\mathscr{A}$, which thus play the role of (asymptotically) *sufficient statistics*. It is particularly important to stress that, despite the quite generic concentration assumption on X (Assumption 2), when $\Sigma_{ij} = I_p$, only the k^2 inner products $\Delta\mu_i^\mathsf{T}\Delta\mu_{i'}$ and the $2k$ *class-wise* dimensionality ratios c_{ij}/c_i intervene in the expression of $\mathscr{M}$ – so in particular none of the higher order moments of X are accounted for, nor the absolute task-wise dimension ratios c_i. As for $\mathscr{A}$, it captures instead the information about the impact of the hyperparameters $\lambda, \gamma_1, \ldots, \gamma_k$ as well as the *task-wise* dimensionality ratios $c_1, \ldots, c_k$ and the data number-to-dimension ratio c_0. In the expression of the MTL LS-SVM performance, these two matrices combine into the core matrix $\in \mathbb{R}^{2k \times 2k}$

$$\Gamma = \left(I_{2k} + \left(\mathscr{A} \otimes \mathbb{1}_2 \mathbb{1}_2^\mathsf{T} \right) \odot \mathscr{M} \right)^{-1} \tag{4.5}$$

where we recall that '$\odot$' is the Hadamard (element-wise) matrix product.

Theorem 8 (Asymptotics of $g_i^{\text{bin}}(\mathbf{x})$)**.** *Under Assumptions 2–3, with $m = 2$ and $\Sigma_{ij} = I_p$, for a test data $\mathbf{x}$ with $\mathbb{E}[\mathbf{x}] = \mu_{ij}$ and $\text{Cov}[\mathbf{x}] = I_p$, as $p, n \to \infty$,*

$$g_i^{\text{bin}}(\mathbf{x}) - G_{ij} \xrightarrow{\text{a.s.}} 0, \quad G_{ij} \sim \mathcal{N}(m_{ij}, \sigma_i^2)$$

in distribution, where, letting $\boldsymbol{m} = [m_{11}, \ldots, m_{k2}]^\mathsf{T}$ and the normalized forms $\boldsymbol{y}^{\text{bin}} \equiv \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{\frac{1}{2}} y^{\text{bin}}$, $\mathring{\boldsymbol{y}}^{\text{bin}} = \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{\frac{1}{2}} \mathring{y}^{\text{bin}}$, $\boldsymbol{m} = \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{\frac{1}{2}} m$, and $\boldsymbol{\sigma}_i^2 = \boldsymbol{\delta}_i^{[k]} \sigma_i^2$,

$$\boldsymbol{m} = \boldsymbol{y}^{\text{bin}} - \Gamma \mathring{\boldsymbol{y}}^{\text{bin}}$$
$$\boldsymbol{\sigma}_i^2 = (\mathring{\boldsymbol{y}}^{\text{bin}})^\mathsf{T} \Gamma \mathscr{V}_i \Gamma \mathring{\boldsymbol{y}}^{\text{bin}}$$

with

$$\mathscr{V}_i = \mathscr{D}_{\mathscr{K}_{i\cdot}^\mathsf{T} \otimes \mathbb{1}_2} + \left(\mathscr{A} \mathscr{D}_{\mathscr{K}_{i\cdot}^\mathsf{T} + e_i^{[k]}} \mathscr{A} \otimes \mathbb{1}_2 \mathbb{1}_2^\mathsf{T} \right) \odot \mathscr{M}$$
$$\mathscr{K} = \frac{c_0}{k} [\mathscr{A} \odot \mathscr{A}] \left(\mathscr{D}_c - \frac{c_0}{k} [\mathscr{A} \odot \mathscr{A}] \right)^{-1}.$$

Theorem 8 interestingly indicates that the (asymptotic) statistics of the classification scores $g_i^{\text{bin}}(\mathbf{x})$, for $1 \leq i \leq k$, reduce to a mere functional of $2k$-*dimensional deterministic vectors and matrices*. In particular, $g_i^{\text{bin}}(\mathbf{x})$ depends on the data statistical means $\mu_{i'j'}$, $1 \leq i' \leq k$, $1 \leq j' \leq 2$, and on the hyperparameters λ and $\gamma_1, \ldots, \gamma_k$ mostly through the $2k$-dimensional matrix Γ (and more marginally through $\mathscr{V}_i$ and $\mathscr{K}$ for the variances).

Another non-trivial point to note is that, being in general non-diagonal, Γ acts on the centered scores (labels) $\mathscr{y}_{i'j'}^{\circ\text{bin}}$ of all classes j' and tasks i' which, therefore, all influence the performances. It can thus be anticipated that, for the decision on a particular Task i to be successful, not only the scores $\mathscr{y}_{i1}^{\text{bin}}$ and $\mathscr{y}_{i2}^{\text{bin}}$, but in fact all scores $\mathscr{y}_{i'j'}^{\text{bin}}$ across all classes and tasks, must be appropriately tuned.

Remark also that, in this isotropic ($\Sigma_{i'j'} = I_p$) setting, the variance σ_i^2 of the score $g_i^{\text{bin}}(\mathbf{x})$ with $\mathbb{E}[\mathbf{x}] = \mu_{ij}$ only depends on i, and not on j. This is particularly convenient, as shown next, to devise an optimal decision rule for classification into class 1 or 2 for Task i.

From a more technical standpoint, comparing the exact expression of $g_i(\mathbf{x})$ in (4.2) and that of $\mathscr{m}_{ij}$ (i.e., the large dimensional approximation of $\mathbb{E}[g_i(\mathbf{x})]$), we may interpret the matrix $\Gamma \in \mathbb{R}^{2k \times 2k}$ as a "condensed" form of $Q \in \mathbb{R}^{n \times n}$. From the expression $(I_{2k} + \mathscr{A} \otimes \mathbb{1}_2 \mathbb{1}_2^\mathsf{T})^{-1} \odot \mathscr{M}$, observe that: (i) if $\lambda \ll 1$, then $\mathscr{A}$ is diagonal dominant and thus "filters out" in the Hadamard product all off-diagonal entries of $\mathscr{M}$ – that is, all the cross-terms $\Delta\mu_i^\mathsf{T} \Delta\mu_j$ for $i \neq j$ –, therefore refusing to exploit the correlation between tasks; (ii) if instead $\lambda \sim 1$, then $\mathscr{A}$ may be developed (using the Sherman-Morrison matrix inverse formulas) as the sum of a diagonal matrix, which again filters out the $\Delta\mu_i^\mathsf{T} \Delta\mu_j$ for $i \neq j$, and of a rank-one matrix which instead performs a weighted sum (through the γ_i and the $\boldsymbol{\delta}_i^{[k]}$) of the entries of $\mathscr{M}$; specifically, letting $\gamma^{-1} = (\gamma_1^{-1}, \ldots, \gamma_k^{-1})^\mathsf{T}$, we have

$$\left(D_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k^\mathsf{T}\right)^{-1} = D_\gamma^{-1} - \frac{\lambda \gamma^{-1} (\gamma^{-1})^\mathsf{T}}{1 + \lambda \frac{1}{k} \sum_{i=1}^k \gamma_i^{-1}}.$$

As such, letting aside the regularization effect of the $\boldsymbol{\delta}_i^{[k]}$'s, the off-diagonal $\Delta\mu_i^\mathsf{T} \Delta\mu_j$ term intervening in the expression of $\mathscr{M}$ is weighted by a coefficient $(\gamma_i \gamma_j)^{-1}$: the impact of the $\gamma_{i'}$'s is thus strongly associated to the relevance of the correlation between tasks, and not only to the individual performances of the k isolated LS-SVM tasks.

Section 4.4 provides a more general version (Theorem 9) of Theorem 8 for $m \geq 2$ classes per task and generic Σ_{ij}. The technical derivation of these two results, of limited interest at this point, is also deferred to Section 4.4.

4.3.2 Decision threshold and label optimization

Since $g_i^{\text{bin}}(\mathbf{x})$ has a Gaussian limit centered about $\mathscr{m}_{ij}$ and with equal variance for $j = 1$ and $j = 2$, the (asymptotically) optimal decision for $\mathbf{x}$ to be allocated to class $\mathscr{C}_1$ or class $\mathscr{C}_2$ for Task i, i.e., the decision minimizing the averaged error probability under the

prior $\mathbb{P}(\mathbf{x} \in \mathscr{C}_1) = \mathbb{P}(\mathbf{x} \in \mathscr{C}_2)$, is obtained by the "averaged-mean" test

$$g_i^{\text{bin}}(\mathbf{x}) \underset{\mathscr{C}_2}{\overset{\mathscr{C}_1}{\gtrless}} \zeta_i \equiv \frac{1}{2}\left(m_{i1} + m_{i2}\right) \tag{4.6}$$

the associated misclassification rate being

$$\begin{aligned} \epsilon_{i1} &\equiv \mathbb{P}\left(g_i^{\text{bin}}(\mathbf{x}) \geq \frac{m_{i1} + m_{i2}}{2} \Big| \mathbf{x} \in \mathscr{C}_1\right) \\ &= \mathcal{Q}\left(\frac{m_{i1} - m_{i2}}{2\sigma_i}\right) + o(1) \end{aligned} \tag{4.7}$$

with m_{ij}, σ_i as in Theorem 8 and $\mathcal{Q}(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{u^2}{2}} du$.

It is of utmost interest at this point to recall that the asymptotics of $g_i^{\text{bin}}(\mathbf{x})$ from Theorem 8 (as from the more generic Theorem 9) depend in an elegant and simple manner on the training data scores $\mathscr{y}^{\text{bin}} = \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{-\frac{1}{2}} \boldsymbol{y}^{\text{bin}}$. Using again the independence of σ_i^2 on the genuine class of $\mathbf{x}$, the vector $\boldsymbol{y}^{\text{bin}\star}$ minimizing the misclassification rate for Task i simply reads:

$$\begin{aligned} \boldsymbol{y}^{\text{bin}\star} &= \underset{\boldsymbol{y}^{\text{bin}} \in \mathbb{R}^{2k}}{\arg\max} \frac{(m_{i1} - m_{i2})^2}{\sigma_i^2} \\ &= \underset{\boldsymbol{y}^{\text{bin}} \in \mathbb{R}^{2k}}{\arg\max} \frac{\|(\boldsymbol{y}^{\text{bin}})^\mathsf{T}(I_{2k} - \Gamma)\mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{-\frac{1}{2}}(e_{i1}^{[2k]} - e_{i2}^{[2k]})\|^2}{(\boldsymbol{y}^{\text{bin}})^\mathsf{T}\Gamma\mathscr{V}_i\Gamma\boldsymbol{y}^{\text{bin}}} \end{aligned}$$

for which the solution is explicitly defined, *up to an arbitrarily multiplicative constant (as it maximizes a ratio) and up to an arbitrary additive constant (as per Remark 4)*, by:

$$\boldsymbol{y}^{\text{bin}\star} = \Gamma^{-1}\mathscr{V}_i^{-1}[(\mathscr{A} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}) \odot \mathscr{M}]\mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{-\frac{1}{2}}(e_{i1}^{[2k]} - e_{i2}^{[2k]}). \tag{4.8}$$

and, for this choice of $\boldsymbol{y}^{\text{bin}\star}$, the corresponding (asymptotically) optimal classification error ϵ_{i1} defined in (4.7) is then

$$\epsilon_{i1}^\star = \mathcal{Q}\left(\frac{1}{2}\sqrt{(e_{i1}^{[2k]} - e_{i2}^{[2k]})^\mathsf{T}\mathscr{G}(e_{i1}^{[2k]} - e_{i2}^{[2k]})}\right) \tag{4.9}$$

for $\mathscr{G} = \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{\frac{1}{2}}[(\mathscr{A} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}) \odot \mathscr{M}]\mathscr{V}_i^{-1}[(\mathscr{A} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}) \odot \mathscr{M}]\mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{\frac{1}{2}}$. Of course, by symmetry, $\epsilon_{i2} \equiv P(g_i^{\text{bin}}(\mathbf{x}) \leq \frac{m_{i1}+m_{i2}}{2} | \mathbf{x} \in \mathscr{C}_2)$ has the same limiting optimal value $\epsilon_{i2}^\star = \epsilon_{i1}^\star$.

The only non-diagonal matrices in (4.8) are Γ and $\mathscr{V}_i$ in which $\mathscr{M}$ plays the role of a "variance profile" matrix. In particular, assume $\Delta\mu_i^\mathsf{T}\Delta\mu_{i'} = 0$ for all $i' \neq i$, i.e., the differences in statistical means of all tasks are orthogonal to those of Task i. Then the two rows and columns of $\mathscr{M}$ associated to Task i are all zero but on the 2×2 diagonal block. Therefore, $\mathscr{y}^{\text{bin}\star}$ will have all zero entries but on its Task i two elements. All

other choices for the null entries of $\mathcal{y}^{\text{bin}\star}$ (such as the usual $\mathcal{y}^{\text{bin}} = [1, -1, \ldots, 1, -1]^\mathsf{T}$) would be suboptimal and (possibly severely) detrimental to the classification performance of Task i, not by altering the means m_{i1}, m_{i2} but *by increasing the variance* σ_i^2. This extreme example strongly suggests that, in order to maximize the MTL performance on a targeted Task i, one must impose low absolute scores $\mathcal{y}^{\text{bin}}_{i'j}$ to all Tasks i' strongly different from Task i.

The choice $\mathcal{y}^{\text{bin}} = [1, -1, \ldots, 1, -1]^\mathsf{T}$ can also be very detrimental when $\Delta\mu_i^\mathsf{T}\Delta\mu_{i'} < 0$ for some pair i, i': that is, when the mapping of the two classes within each task is reversed (e.g., if class $\mathscr{C}_1$ in Task 1 is closer to class $\mathscr{C}_2$ than class $\mathscr{C}_1$ in Task 2). In this setting, it is easily seen that $\mathcal{y}^{\text{bin}} = [1, -1, \ldots, 1, -1]^\mathsf{T}$ works against the classification and performs much worse than a single-task LS-SVM.

Another interesting conclusion arises from the simplified setting of equal number of samples per task and per class, i.e., $n_{11} = \ldots = n_{k2}$. In this case, $\boldsymbol{\delta}_{11}^{[k2]} = \ldots = \boldsymbol{\delta}_{k2}^{[k2]}$ and, since $\mathcal{y}^{\text{bin}\star}$ is defined up to a multiplicative constant, we have

$$\mathcal{y}^{\text{bin}\star} = \Gamma^{-1}\mathscr{V}_i^{-1}\left((\mathscr{A} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}) \odot \mathscr{M}\right)(e_{i1}^{[2k]} - e_{i2}^{[2k]})$$

in which all matrices are organized in 2×2 blocks of equal entries. This immediately implies that $\mathcal{y}^{\text{bin}\star}_{i'1} = -\mathcal{y}^{\text{bin}\star}_{i'2}$ for all i'. So in particular, the detection threshold $\frac{1}{2}(m_{i1} + m_{i2})$ of the averaged-mean test (4.6) is zero (as conventionally assumed). In all other settings for the $n_{i'j}$'s, it is very unlikely that $\mathcal{y}^{\text{bin}\star}_{i1} = -\mathcal{y}^{\text{bin}\star}_{i2}$ and the optimal decision threshold *must* also be estimated. As a matter of fact, following up on Remark 4, the aforementioned optimal value $\mathcal{y}^{\text{bin}\star}$ for $\mathcal{y}^{\text{bin}}$ is not unique and could be shifted by any constant vector. This extra degree of freedom will be of much relevance in the application Section 4.5, as commented in the following remark.

Remark 5 (Setting the decision threshold to zero)**.** *As per Remark 4, the addition of a constant term to* $\mathcal{y}^{\text{bin}}$ *does not affect the ultimate performance of MTL LS-SVM. Yet, it affects the value of the limiting means* m_{ij} *of* $g_i^{\text{bin}}(\mathbf{x})$*, so in particular the value of the limiting optimal threshold* $\frac{1}{2}(m_{i1} + m_{i2})$*. Specifically, one may shift all entries of* $\mathcal{y}^{\text{bin}}$ *in such a way that* $\frac{1}{2}(m_{i1} + m_{i2}) = 0$ *and thus recenter the decision threshold to zero. For* $\bar{y} \in \mathbb{R}$ *this constant shift, this boils down to solving in the variable* $\bar{y}$ *the equation*

$$0 = \frac{1}{2}(m_{i1} + m_{i2}) = \frac{1}{2}(\mathcal{y}^{\text{bin}} + \bar{y}e_i^{[k]} \otimes \mathbb{1}_2)^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\boldsymbol{\delta}^{[2k]}}(I_{2k} - \mathscr{Z}_e\Gamma)(e^{[2k]}_{2(i-1)+1} + e^{[2k]}_{2i})$$

where $\mathscr{Z}_e = I_{2k} - \sum_{i'=1}^k E^{[k]}_{i'i'} \otimes \mathscr{c}_{i'}$ *and* $\mathscr{c}_{i'} = \mathbb{1}_2\left[\begin{smallmatrix} \frac{n_{i'1}}{n_{i'}} & \frac{n_{i'2}}{n_{i'}} \end{smallmatrix}\right]$*. Similarly, one may instead impose that* $m_{i1} = 0$*: this will appear to be fundamental to* align classifiers *in the multi-class "one-versus-all" extension of the present binary classification scheme (see details in Section 4.5.2).*

Remark 6 (Tuning the hyperparameters)**.** *The previous section provided a high-level interpretation for the impact of the vector parameter* $\gamma \in \mathbb{R}^k$ *and the scalar parameter* $\lambda \in \mathbb{R}$*, the effect of which is to respectively regularize LS-SVM learning and to set the*

throttle between individual versus collective learning. These hyperparameters intervene deeply inside our theoretical formulas (so far in Theorem 8 but later in Theorem 9) and are not amenable to simple optimization. Yet, as will be confirmed by experiments (see in particular Figure 4.3), the proposed optimization of the input scores $\mathcal{Y}^{\mathrm{bin}}$ partly compensates for suboptimal choices in γ, λ. As such, an "informed guess", based on our previous discussion of the effects of these parameters, is in general sufficient for highly performing MTL LS-SVM. A further gradient descent operation (or local grid search) on the theoretical performance approximation, initialized at the informed guess values, can further improve the overall learning performance.

4.3.3 Practical implementation of improved MTL LS-SVM

As already pointed out, a fundamental aspect of Theorem 8 lies in the performances of the *large dimensional* $(n, p \gg 1)$ classification problem at hand boiling down to $2k$-dimensional statistics. More importantly from a practical perspective, these $2k$-dimensional "sufficient statistics" are easily amenable to fast and efficient estimation: it indeed only requires a few training data samples to estimate all quantities involved in the theorem (which, as a corollary, lets one envision the possibility of efficient transfer learning methods based on very scarce data samples).

Remark 7 (On the estimation of m_{ij} and σ_i)**.** *All quantities defined in Theorem 8 are a priori known, apart from the quantities $\mathscr{M} \equiv \sum_{i,i'} \Delta\mu_i^\mathsf{T} \Delta\mu_{i'} \left(E_{ii'}^{[k]} \otimes \mathbb{c}_i \mathbb{c}_{i'}^\mathsf{T} \right)$ and most specifically the inner products $\Delta\mu_i^\mathsf{T} \Delta\mu_{i'}$. For these, define, for $j = 1, 2$, two sets $\mathcal{S}_{ij}, \mathcal{S}'_{ij} \subset \{1, \ldots, n_{ij}\}$ and the corresponding indicator vectors $\mathbb{j}_{ij}, \mathbb{j}'_{ij} \in \mathbb{R}^{n_i}$ with $[\mathbb{j}_{ij}]_a = \delta_{a \in \mathcal{S}_{ij}}$ and $[\mathbb{j}'_{ij}]_a = \delta_{a \in \mathcal{S}'_{ij}}$. We further impose that $\mathcal{S}'_{ij} \cap \mathcal{S}_{ij} = \emptyset$. Then, for $i \neq i'$, the following estimates hold:*

$$
\begin{aligned}
&\Delta\mu_i^\mathsf{T} \Delta\mu_{i'} - \left(\frac{\mathbb{j}_{i1}}{|\mathcal{S}_{i1}|} - \frac{\mathbb{j}_{i2}}{|\mathcal{S}_{i2}|} \right)^\mathsf{T} \mathring{X}_i^\mathsf{T} \mathring{X}_{i'} \left(\frac{\mathbb{j}_{i'1}}{|\mathcal{S}_{i'1}|} - \frac{\mathbb{j}_{i'2}}{|\mathcal{S}_{i'2}|} \right) \\
&= O\left((p \min_{l \in \{1,2\}} \{|\mathcal{S}_{il}|, |\mathcal{S}_{i'l}|\})^{-\frac{1}{2}} \right) \\
&\Delta\mu_i^\mathsf{T} \Delta\mu_i - \left(\frac{\mathbb{j}_{i1}}{|\mathcal{S}_{i1}|} - \frac{\mathbb{j}_{i2}}{|\mathcal{S}_{i2}|} \right)^\mathsf{T} \mathring{X}_i^\mathsf{T} \mathring{X}_i \left(\frac{\mathbb{j}'_{i1}}{|\mathcal{S}'_{i1}|} - \frac{\mathbb{j}'_{i2}}{|\mathcal{S}'_{i2}|} \right) \\
&= O\left((p \min_{l \in \{1,2\}} \{|\mathcal{S}_{il}|, |\mathcal{S}'_{il}|\})^{-\frac{1}{2}} \right).
\end{aligned}
$$

Observe in particular that a single sample (two when $i = i'$) per task and per class ($|\mathcal{S}_{il}| = 1$) is sufficient to obtain a consistent estimate for all quantities, so long that p is large. In a transfer learning setting where some tasks may contain few labeled data, it is thus still possible to optimize the MTL algorithm. Of course, when more data are available, under our assumption that $p \sim n$, taking all samples in the averaging, the convergence speed is of order $O(1/\sqrt{np}) = O(1/n)$, which is a quadratic increase in the speed of the usual central-limit theorem.

Estimating m_{ij} and σ_i not only allows one to anticipate theoretical performances but also enables the actual estimation of the decision threshold $\frac{1}{2}(m_{i1} + m_{i2})$ of the test (4.6) and, as shown previously, opens the possibility to largely optimize MTL LS-SVM through an (asymptotically) optimal choice of the training scores y^{bin}.

The series of theoretical and practical results of this section may be synthetized under the form of Algorithm 2.

Algorithm 2 Proposed binary Multi-Task Learning algorithm.

Input: Training samples $X = [X_1, \ldots, X_k]$ with $X_{i'} = [X_{i'}^{(1)}, X_{i'}^{(2)}]$ and test data $\mathbf{x}$.
Output: Estimated class $\hat{j} \in \{1, 2\}$ of $\mathbf{x}$ for target Task i.
Center and normalize data per task: for all $i' \in \{1, \ldots, k\}$,

- $\mathring{X}_{i'} \leftarrow X_{i'} \left(I_{n_{i'}} - \frac{1}{n_{i'}} \mathbb{1}_{n_{i'}} \mathbb{1}_{n_{i'}}^\mathsf{T} \right)$
- $\mathring{X}_{i'} \leftarrow \mathring{X}_{i'} / \frac{1}{n_{i'} p} \text{tr}(\mathring{X}_{i'} \mathring{X}_{i'}^\mathsf{T})$

Estimate: Matrix M from Remark 7 and $\boldsymbol{\delta}^{[k]}$ by solving (4.3).
Create scores $y^{\text{bin}} = y^{\text{bin}\star}$ according to (4.8).
Compute the threshold ζ_i from (4.6), with m_{ij} defined in Theorem 8 for $y^{\text{bin}} = y^{\text{bin}\star}$.

(Optional) Estimate the theoretical classification error $\epsilon_{i1} = \epsilon_{i1}(\lambda, \gamma)$ from (4.7) and minimize over (λ, γ).[1]
Compute classification score $g_i(\mathbf{x})$ according to (4.2).
Output: $\hat{j}$ such that $g_i(\mathbf{x}) \underset{\hat{j}=2}{\overset{\hat{j}=1}{\gtrless}} \zeta_i$.

4.3.4 Empirical evidence

This section shortly illustrates the ideas and intuitions developed so far (such as the relevance of an optimal choice of the data labels and decision threshold) through the performances of Algorithm 2 on a transfer learning benchmark application. Sections 4.5–4.6 will cover a much larger spectrum of applications and experiments, under the most general data setting discussed in the subsequent sections.

For optimal comparison, we consider here the standard Office+Caltech256 real image classification benchmark (Saenko et al., 2010; Griffin et al., 2007), consisting of four tasks and $m = 10$ categories shared by all tasks. The dataset X consists here of the VGG features of size $p = 4096$ extracted from these images. We place ourselves under a $k = 2$ transfer learning setting where Task 1 is the source task and Task 2 is the target task (the performance of which we aim to optimize), taken from two of the four

[1] As per Remark 6, this operation involves reevaluating $\boldsymbol{\delta}^{[k]}$ and thus $y^\star$, and thus m for each (λ, γ). It can be performed either on a static grid or by gradient descent until a local minimum is reached.

tasks of the dataset (Caltech, Webcam, Amason, dslr). For testing, the samples of the target task are randomly selected from the test dataset of Office+Caltech256 and the classification accuracy is averaged over 20 trials. Table 4.1 reports the accuracy for all possible pairs ($4 \times 3 = 12$ of them) of source and transfer tasks, obtained by Algorithm 2 (Ours) versus the non-optimized LS-SVM of (Xu et al., 2013) (LS-SVM) and versus other state-of-the-art transfer learning algorithms: the max-margin domain transform of (Hoffman et al., 2013) (MMDT) which seeks a linear transform to match the source data to the target data and then applies an SVM on the resulting target domain; the cross-domain landmark selection (CDLS) of (Hubert Tsai et al., 2016), which learns a feature subspace which matches the cross-domain data distribution and eliminates the domain differences; and the invariant latent space (ILS) of (Herath et al., 2017), which, similar to MMDT, learns an invariant latent space in which the discrepancy between source and target is minimized. As already pointed out in introduction, since we aim to propose an improved classification algorithm independent of the feature representation, it is fair to compare it to methods which use the same data features. The algorithms compared in the table all systematically use VGG features. It would be unfair to compare these against "end to end" MTL learning methods including a (explicit or implicit) step of feature learning like recent deep neural networks methods(Zhuang et al., 2020; Krishna & Kalluri, 2019).

Since $m = 10$ here, Algorithm 2 cannot rigorously be used as it stands. We apply instead a *naive* "one-versus-all" extension consisting in running in parallel $m = 10$ times Algorithm 2 by considering, for each class $\mathscr{C}_j$ of Task i, $1 \leq j \leq m$, a binary setting where the fictitious "Class $\tilde{\mathscr{C}}_1$" coincides with $\mathscr{C}_j$ and the second fictitious "Class $\tilde{\mathscr{C}}_2$" is the union of all $\mathscr{C}_{j'}$ for $j' \neq j$. Following up on Remark 5, each classifier $\ell \in \{1, \ldots, m\}$ can be set in such a way that $\mathbb{E}[g_i^{\text{bin}}(\mathbf{x}; \ell)] = 0$ when $\mathbf{x} \in \mathscr{C}_\ell$. For a new datum $\mathbf{x}$, of all m classifiers $g_i^{\text{bin}}(\mathbf{x}; 1), \ldots, g_i^{\text{bin}}(\mathbf{x}; m)$, the one reaching the greatest score is the selected allocation class for $\mathbf{x}$.

Table 4.1 demonstrates that our proposed improved MTL LS-SVM, despite its simplicity and unlike the competing methods used for comparison, has stable performances and is extremely competitive. It either outperforms all other methods or is second-to-best. But, most importantly, the method comes along with performance predictions and guarantees, which none of the competing works are able to provide.[2]

These preliminary results are already very conclusive and reveal the strength of our proposed methodology. Yet, the assumptions in place so far are restricted to random concentrated data with identity covariance and to a binary classification setting (which, as already observed, needs be adapted to account for more than two classes per task). The next sections elaborate on the more generic setting of $m \geq 2$ classes per task with more realistic data models. The theoretical results no longer reduce to compact expressions as

[2]In the present context of the *naive* "one-versus-all", this claim should be taken with care: the performance can indeed be predicted provided the binary class model $\mathcal{N}(\mu_{i1}, I_p)$ versus $\mathcal{N}(\mu_{i2}, I_p)$ correctly matches the actual data distribution; this is likely not the case here as the collected fictitious "$\tilde{\mathscr{C}}_2$" is rather a mixture of Gaussian rather than a unique Gaussian. In Section 4.5.2, a more elaborate, and theoretically better supported, version of the one-versus-all approach will be discussed.

Table 4.1: Classification accuracy over Office+Caltech256 database. c(Caltech), w(Webcam), a(Amazon), d(dslr), for different "Source to target" task pairs ($S \to T$) based on VGG features. Best score in boldface, second-to-best in italic.

S/T	c→ w	w→ c	c→ a	a→ c	w→ a	a→ d	d→ a	w→ d	c→ d	d→ c	a→ w	d→ w	Mean score
LS-SVM	96.69	**89.90**	92.90	*90.00*	93.80	78.70	93.50	95.00	85.00	**90.20**	94.70	**100**	91.70
MMDT	93.90	87.05	90.83	84.40	*94.17*	86.25	**94.58**	97.50	86.25	87.23	92.05	*97.35*	90.96
ILS	77.89	73.55	86.85	76.22	86.22	71.34	74.53	82.80	68.15	63.49	78.98	92.88	77.74
CDLS	*97.60*	88.30	*93.54*	88.30	93.54	*92.50*	93.54	93.75	**93.75**	88.30	*97.35*	96.70	*93.10*
Ours	**98.68**	**89.90**	**94.40**	**90.60**	**94.40**	**93.80**	*94.20*	**100**	*92.50*	*89.90*	**98.70**	*99.30*	**94.70**

in the previous sections but are easily understood having already delineated the main take-away messages and ideas.

4.4 The General Framework

The results from the previous section are extended here to the more realistic setting where the data arise from a mixture of $m \geq 2$ concentrated random vectors with generic covariance Σ_{ij}. New insights, and most importantly, more general and application-driven algorithms will be introduced. In addition, the results are presented here with a sketched development of their main technical arguments, the full proofs being deferred to the appendix.

4.4.1 Main ideas

Taking for the moment for granted the Gaussian limit for $g_i(\mathbf{x}) \in \mathbb{R}^m$ as $p, n \to \infty$ (for $1 \leq i \leq k$), the main technical task to obtain our main result (Theorem 9, which generalizes the already introduced Theorem 8) is to evaluate the large dimensional behavior $\boldsymbol{m}_{ij}$ and $\boldsymbol{\sigma}_{ij}$ of the statistical mean $\mathbb{E}[g_i(\mathbf{x})] = \boldsymbol{m}_{ij} + o(1)$ and covariance matrix $\mathrm{Cov}[g_i(\mathbf{x})] = \boldsymbol{\sigma}_{ij} + o(1)$ of the classification score $g_i(\mathbf{x})$ in (4.2) for data vectors $\mathbf{x}$ in class $\mathcal{C}_j$ (i.e., such that $\mathbb{E}[\mathbf{x}] = \mu_{ij}$ and $\mathrm{Cov}[\mathbf{x}] = \Sigma_{ij}$), respectively given from (4.2) by:

$$\boldsymbol{m}_{ij} = \mathbb{E}\left[\frac{1}{kp}(Y - Pb)^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} \left(e_i^{[k]} \otimes \mu_{ij}\right) + b_i\right] \tag{4.10}$$

$$\boldsymbol{\sigma}_{ij} = \mathbb{E}\left[\frac{1}{(kp)^2}(Y - Pb)^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z (Y - Pb)\right] \tag{4.11}$$

with $S_{ij} = e_i^{[k]} {e_i^{[k]}}^\mathsf{T} \otimes \Sigma_{ij}$ and $\tilde{Q} = \left(\frac{A^{\frac{1}{2}} Z Z^\mathsf{T} A^{\frac{1}{2}}}{kp} + I_{kp}\right)^{-1}$.

Our technical approach to evaluate these terms, in the large dimensional regime of Assumption 3 and for data distributed as per Assumption 2, consists in determining *deterministic equivalents* for the matrices $\tilde{Q}$, $\tilde{Q} A^{\frac{1}{2}} Z$, $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z$ which are at the core of the formulation of $\boldsymbol{m}_{ij}$ and $\boldsymbol{\sigma}_{ij}$.

Lemma 5, deferred to Section 6.3.2 of the appendix (as the result in itself does not bring any deep insight worth discussing here), provides the necessary deterministic equivalents for these matrices. It is interesting to point out though that, from a technical standpoint, the block structure followed by the core data matrix Z introduced in Proposition 5 makes the large dimensional random matrix analysis more challenging and the result less straightforward than in similar previous works (Mai & Liao, 2019; Liao & Couillet, 2019) and in Section 2.3 for the ridge regression problem. Even in the simplest setting where the $x_{il}^{(j)}$ would be vectors of i.i.d. $\mathcal{N}(0,1)$ entries, the matrix Z is *not* a matrix of i.i.d. entries (due to precisely located blocks of zeros) and the singular values of Z do not asymptotically follow the popular Marčenko-Pastur distribution, as was the case in Section 2.3.

The main information to be extracted from Lemma 5 (again, see its complete form in the appendix) is the central role played by the deterministic matrices

$$M = \left(e_1^{[k]} \otimes [\mu_{11}, \ldots, \mu_{1m}], \ldots, e_k^{[k]} \otimes [\mu_{k1}, \ldots, \mu_{km}] \right)$$
$$C_{ij} = A^{\frac{1}{2}} \left(e_i^{[k]} e_i^{[k]\mathsf{T}} \otimes (\Sigma_{ij} + \mu_{ij}\mu_{ij}^{\mathsf{T}}) \right) A^{\frac{1}{2}}$$

which generalize the matrices $\mathscr{M}$ and $\mathscr{A}$ discussed at length in Section 4.3 when $\Sigma_{ij} = I_p$. While gaining in genericity, unlike $\mathscr{M}$, the matrices M and C_{ij} preserve their large dimensions: this is the main price paid by the generalization to $\Sigma_{ij} \neq I_p$. Yet, the central small dimensional matrix Γ defined in (4.5) remains small and now becomes

$$\Gamma = \left(I_{mk} + \mathbb{M}^{\mathsf{T}} \bar{\tilde{Q}}_0 \mathbb{M} \right)^{-1}$$
$$\bar{\tilde{Q}}_0 = \left[\sum_{i=1}^{k} \sum_{j=1}^{m} (\mathscr{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k)^{\frac{1}{2}} e_i^{[k]} e_i^{[k]\mathsf{T}} (\mathscr{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k)^{\frac{1}{2}} \otimes \boldsymbol{\delta}_{ij}^{[mk]} \Sigma_{ij} + I_{kp} \right]^{-1}$$
$$\mathbb{M} = A^{\frac{1}{2}} M \mathscr{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}$$

and the mk scalars $\boldsymbol{\delta}_{ij}^{[mk]}$ are the unique positive solutions of the fixed point equations

$$\boldsymbol{\delta}_{ij}^{[mk]} = \frac{c_{ij}}{c_0 \left(1 + \frac{1}{kp} \mathrm{tr}(C_{ij} \bar{\tilde{Q}}) \right)}$$
$$\bar{\tilde{Q}} = \left(\sum_{i=1}^{k} \sum_{j=1}^{m} \boldsymbol{\delta}_{ij}^{[mk]} C_{ij} + I_{kp} \right)^{-1}.$$

Here, $\bar{\tilde{Q}}$ is a deterministic equivalent of $\tilde{Q}$. Finally, the matrix $\mathscr{K}$ appearing in the variance term of Theorem 8 now becomes

$$\mathscr{K} = c_0 \bar{T} \left(\mathscr{D}_c - c_0 \mathscr{T} \right)^{-1}$$

$$\bar{T}_{ij,i'j'} = \frac{\boldsymbol{\delta}_{ij}^{[mk]}\boldsymbol{\delta}_{i'j'}^{[mk]}}{kp}\mathrm{tr}\left(C_{i'j'}\bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}}\right)$$
$$\mathscr{T}_{ij,i'j'} = \frac{\boldsymbol{\delta}_{ij}^{[mk]}\boldsymbol{\delta}_{i'j'}^{[mk]}}{kp}\mathrm{tr}(C_{ij}\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}})$$

where $\bar{T}_{ij,i'j'}$ is the element at row $m(i-1)+j$ and column $m(i'-1)+j'$ of $\bar{T}$ (and similarly for $\mathscr{T}$) and $\kappa_{ij,\cdot} \in \mathbb{R}^{mk}$ represents the $m(i-1)+j$ row of matrix $\kappa \in \mathbb{R}^{mk\times mk}$.

With these technical elements at hand, we are in position to enunciate the main result of the chapter.

4.4.2 Classification score asymptotics

Theorem 9. *Under Assumptions 2 and 3, for a test data* $\mathbf{x}$ *with* $\mathbb{E}[\mathbf{x}] = \mu_{ij}$ *and* $\mathrm{Cov}[\mathbf{x}] = \Sigma_{ij}$*, as* $p, n \to \infty$*,*

$$g_i(\mathbf{x}) - G_{ij} \to 0, \quad G_{ij} \sim \mathscr{N}(\boldsymbol{m}_{ij}, \boldsymbol{\sigma}_{ij})$$

in law where, letting $\boldsymbol{m} = [\boldsymbol{m}_{11}, \ldots, \boldsymbol{m}_{km}]^\mathsf{T} \in \mathbb{R}^{km\times m}$ *and the normalized forms* $\boldsymbol{\mathcal{Y}} \equiv \mathscr{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\mathcal{Y}$, $\mathring{\boldsymbol{\mathcal{Y}}} = \mathscr{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\mathring{\mathcal{Y}}$, $\boldsymbol{m} = \mathscr{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\boldsymbol{m}$,

$$\boldsymbol{m} = \boldsymbol{\mathcal{Y}} - \Gamma\mathring{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^m$$
$$\boldsymbol{\sigma}_{ij} = \mathring{\boldsymbol{\mathcal{Y}}}^\mathsf{T}\Gamma\mathscr{V}_{ij}\Gamma\mathring{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^{m\times m}$$

with

$$\mathscr{V}_{ij} = \mathscr{D}_{\kappa_{ij,\cdot}} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathbb{M}$$
$$\mathbb{V}_{ij} = A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}} + \sum_{i'=1}^{k}\sum_{j'=1}^{m}\boldsymbol{\delta}_{i'j'}^{[mk]}\kappa_{ij,i'j'}A^{\frac{1}{2}}S_{i'j'}A^{\frac{1}{2}}$$
$$\kappa_{ij,i'j'} = \frac{\mathscr{K}_{ij,i'j'}}{\boldsymbol{\delta}_{ij}^{[mk]}}.$$

Proof. See Section 6.3.4 of the appendix. □

In the particular case of $\Sigma_{ij} = I_p$ and $m = 2$, Theorem 9 reduces to Theorem 8 (see details in Section 6.3.4 of the appendix) by remarking that $\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M} = \left(\mathscr{A}\otimes\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)\odot\mathscr{M}$ and $\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0V_{ij}\bar{\tilde{Q}}_0\mathbb{M} = \left(\mathscr{A}\mathscr{D}_{\kappa_i+e_i^{[k]}}\mathscr{A}\otimes\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)\odot\mathscr{M}$ which, as already pointed out, have the advantage to be defined as the product of exclusively small dimensional matrices. Still, although more technical, Theorem 9 follows the same structure as Theorem 8.

Before concretely applying the result of Theorem 9 to practical learning problems (multi-task, transfer learning, hypothesis testing), a few comments and immediate corollaries are in order.

Remark 8 (Optimization of $\mathscr{y}^{\text{bin}}$ for $m = 2$ and generic Σ_{ij})**.** *As suggested in Section 4.3, for binary classification ($m = 2$), it is particularly convenient to recast the score vectors $y_{il}^{(j)} \in \mathbb{R}^m$ into scalar scores $y_{il}^{\text{bin}(j)} \in \mathbb{R}$ (this being irrespective of the nature of Σ_{ij}). Inspired by Section 4.3, one can trivially extend Theorem 9 to this binary setting. In this case, $g_i(\mathbf{x}) \in \mathbb{R}^m$ is now turned into a scalar $g_i^{\text{bin}}(\mathbf{x}) \in \mathbb{R}$ well approximated by $\mathcal{N}(\boldsymbol{m}_{ij}, \boldsymbol{\sigma}_{ij})$ where now $\boldsymbol{m}_{ij}$ and $\boldsymbol{\sigma}_{ij}$ are scalar, obtained by simply replacing $y_{il}^{(j)} \in \mathbb{R}^m$ by $y_{il}^{\text{bin}(j)} \in \mathbb{R}$ in their respective expressions.*

With these notations, setting the decision threshold of $g_i^{\text{bin}}(\mathbf{x})$ to $\zeta \in \mathbb{R}$ and assuming equal prior probability for the genuine class of $\mathbf{x}$, the classification error rate for a target task i is

$$E = \frac{1}{2}\mathscr{Q}\left(\frac{\zeta - \boldsymbol{m}_{i1}}{\sqrt{\boldsymbol{\sigma}_{i1}}}\right) + \frac{1}{2}\mathscr{Q}\left(\frac{\zeta - \boldsymbol{m}_{i2}}{\sqrt{\boldsymbol{\sigma}_{i2}}}\right).$$

As in Section 4.3, if $\Sigma_{i1} = \Sigma_{i2}$, then $\boldsymbol{\sigma}_{i1} = \boldsymbol{\sigma}_{i2} \equiv \boldsymbol{\sigma}_i$ and the decision threshold ζ minimizing the classification error is

$$\zeta^\star = \frac{\boldsymbol{m}_{i1} + \boldsymbol{m}_{i2}}{2}$$

from which the optimal vector $\mathscr{y}^{\text{bin}}$ for Task i is computed as

$$\begin{aligned}\mathscr{y}^{\text{bin}\star} &= \underset{\mathscr{y}^{\text{bin}} \in \mathbb{R}^{2k}}{\arg\max} \frac{(\boldsymbol{m}_{i1} - \boldsymbol{m}_{i2})^2}{\boldsymbol{\sigma}_i} \\ &= \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{-\frac{1}{2}} \Gamma^{-1} \mathscr{V}_{ij}^{-1} \mathbb{M}^\mathsf{T} \bar{\tilde{Q}}_0 \mathbb{M} \mathscr{D}_{\boldsymbol{\delta}^{[2k]}}^{-\frac{1}{2}} (e_{i1}^{[2k]} - e_{i2}^{[2k]}). \end{aligned} \tag{4.12}$$

It is important to recall here that, while $\mathscr{y}^{\text{bin}\star}$ expresses here solely as a function of terms involving the index i, all other statistics of the tasks $i' \neq i$ are in fact "embedded" inside these terms and are thus, of course, accounted for in the optimization.

When $\boldsymbol{\sigma}_{i1} \neq \boldsymbol{\sigma}_{i2}$ (which is the case in general), one may minimize E by resorting to numerical optimization techniques. We suggest to use a gradient descent method initialized to the expression obtained in (4.12)*. So long that Σ_{i1} and Σ_{i2} are not drastically different, this approach shows good performances (see our results in Section 4.6).*

This said, the specific setting of binary classification may in practice be one of hypotheses testing. Under this scenario, one may not demand that the average error E be minimized (i.e., that data from either class is equally well identified) but rather that the probability of misclassification of a given class (say, a Type-I error) be bounded to some $\eta > 0$ while minimizing the error rate for the other class (Type-II error). In this context, if, say, one fixes $\mathscr{Q}(\frac{\zeta - \boldsymbol{m}_{i1}}{\sqrt{\boldsymbol{\sigma}_{i1}}}) \equiv \eta$, then the classification error for the second class $\mathscr{Q}(\frac{\zeta - \boldsymbol{m}_{i2}}{\sqrt{\boldsymbol{\sigma}_{i2}}})$ is minimized by now choosing

$$\mathscr{y}^{\text{bin}\star} = \underset{\mathscr{y}^{\text{bin}} \in \mathbb{R}^{2k}}{\arg\max} \frac{(\sqrt{\boldsymbol{\sigma}_{i1}}\mathscr{Q}^{-1}(\eta) + \boldsymbol{m}_{i1} - \boldsymbol{m}_{i2})^2}{\boldsymbol{\sigma}_{i2}}$$

where $\mathscr{Q}^{-1}$ is the inverse function of the $\mathscr{Q}$ function. This again can be solved using numerical convex optimization initialized at the value of (4.12)*. More details on this hypothesis testing setting, along with concrete experiments, are developed in Section 4.6.*

Remark 9 (Estimation of $\boldsymbol{m}_{ij}$ and $\boldsymbol{\sigma}_{ij}$)**.** *In order to anticipate the performances and best set the decision thresholds for classification, one needs to access all quantities arising in Theorem 9. Yet, as opposed to Remark 7, where the low dimensional quantities of interest (mainly the inner products between statistical means) are easily estimated, the low dimensional quantities involved in Theorem 9 are less convenient to estimate, this being due to the presence of the a priori unknown covariance matrices Σ_{ij}. We propose here two strategies:*

1. *either make the assumption that $\Sigma_{ij} \simeq \alpha_{ij} I_p$ with α_{ij} estimated by $\frac{1}{pn_{ij}} \mathrm{tr} \mathring{X}_i^{(j)} \mathring{X}_i^{(j)\mathsf{T}}$; then normalize the data as $\mathring{X}_i^{(j)} \leftarrow \mathring{X}_i^{(j)} / \frac{1}{pn_{ij}} \mathrm{tr} \mathring{X}_i^{(j)} \mathring{X}_i^{(j)\mathsf{T}}$ in the spirit of Algorithm 2. This places the experimenter under an isotropic data setting for all classes and tasks, from which the considerations of Section 4.3 (possibly generalized to $m > 2$) apply.*

2. *either estimate Σ_{ij} by means of the sample covariance matrix $\frac{1}{n_{ij}} \mathring{X}_i^{(j)} \mathring{X}_i^{(j)\mathsf{T}}$; this procedure is known to be biased, particularly so if n_{ij} is not large compared to p; yet, as demonstrated in our experiments in Section 4.6, this only marginally (if not at all) alters the performance of our proposed algorithms.*[3]

The choice of strategy mainly depends on the belief from the experimenter that the genuine covariance matrices are "well-conditioned" (i.e., their eigenvalues do not spread much) in which case Option 1 would be favored or "ill-conditioned" (typically when the space spanned by the data is much lower than p) in which case Option 2 would be more appropriate.

Remark 10 (On universality)**.** *As pointed out in the introduction, the input data X follows a very generic* concentrated random vector *assumption (Assumption 2). This choice provides both a technical, but most importantly, a fundamentally practical, advantage:*

1. *from a technical standpoint, the concentration of measure phenomenon provides efficient and fast mathematical tools (Louart & Couillet, 2018; Ledoux, 2001) to analyze the random quantities appearing in the classification test scores $g_i(\mathbf{x})$ of MTL LS-SVM (which, in essence, is a mere functional $\mathbb{R}^{p \times n} \to \mathbb{R}$ of the random input data X). More specifically, alternative random matrix tools based on Gaussian (Pastur & Shcherbina, 2011) or independent entries assumptions (Bai & Silverstein, 2009) of X would both be less general (at least for our machine learning purpose) and more computationally intense;*

2. *on the practical side, as underlined in Section 4.2.2, the concentrated random vector assumption better models realistic datasets by imposing very little structure on the*

[3] It must be pointed out that similar random matrix-based studies propose consistent estimates for low dimensional quantities such as those met in Theorem 9; however, these would assume cumbersome forms which, we believe, go against our present request for simple, intuitive but well parameterized algorithms for multi-task and transfer learning.

data. Exactly, it only constrains all Lipschitz functionals $\mathbb{R}^{p\times n} \to \mathbb{R}$ *of* X *(i.e., its typical observations) to satisfy a concentration inequality; while this may seem demanding, the family of concentrated random vectors in fact contains all Lipschitz generative models (for instance neural networks) fed by Gaussian inputs (such as GANs (Goodfellow et al., 2014)), as well as all further Lipschitz transformations of these vectors (for instance, features extracted by pretrained neural networks). As such, provided that the assumption of a common statistical mean and covariance per class and per task is reasonable, Theorem 9 ensures for instance that the performance of MTL LS-SVM applied to classes of the popular VGG or ResNet representations of GAN images is predictable. From this remark, it naturally comes that the proposed method is* universal *in the sense of its being robust to a broad range of very realistic random data, and it is not daring to claim that it is equally valid on genuinely real data. This is confirmed by our numerical results of Section 4.6.*

With these elements in place, we are in position to apply our findings to a host of applications in statistical learning and to test the resulting algorithms against state-of-the-art alternatives.

4.5 Applications

This section provides various applications and optimizations of the proposed MTL LS-SVM framework based on the findings of the previous sections in the context of multi-class classification.

Having access to the large dimensional behavior of the classification test score in Theorem 9 (i.e., for $m \geq 2$ classes per task) is more fundamental than one may think. It indeed allows for a fine-tuning of the hyperparameters to be set to extend the usually considered binary MTL framework of (Evgeniou & Pontil, 2004; Xu et al., 2013) to a multiclass-per-task MTL.

4.5.1 Multi-class classification preliminary

The literature (Bishop, 2006; Rocha & Goldenstein, 2013) describes broad groups of approaches for dealing with $m > 2$ classes. We focus here on the most common methods, namely one-versus-all, one-versus-one, and one hot encoding. Being so far theoretically intractable, these methods inherently suffer from sometimes severe limitations; these are partly tackled by adapting the theoretical results discussed in Section 4.4:

1. **one-versus-all**: in this method, focusing on Task i, m individual binary classifiers $g_i^{\text{bin}}(\ell)$ for $\ell = 1, \ldots, m$ are trained, each of them separating Class $\mathscr{C}_\ell$ from the other $m-1$ classes $\mathscr{C}_{\ell'}$, $\ell' \neq \ell$. Each test sample is then allocated to the class with the highest score among the m classifiers. Although quite used in practice, the approach first suffers a severe data unbalancing effect when using binary (± 1) labels as the set of negative labels in each binary classification is on average $m-1$ times larger than the set of positive labels, and also suffers a centering-scale issue when ultimately

comparing the outputs of the m decision functions $g_i^{\text{bin}}(\mathbf{x};\ell)$, $\ell = 1,\ldots,m$, whose average locations and ranges may greatly differ; these issues lead to undesirable effects, as reported in (Bishop, 2006, section 7.1.3)).

In Section 4.5.2, these problems are simultaneously addressed: specifically, having access to the theoretical statistics of the classification scores allows us to appropriately center and scale the scores. Moreover, each binary classifier is optimized by appropriately choosing the class labels (no longer binary) so to minimize the resulting classification error (see Figure 4.1 for an illustration of the improvement induced by the proposed approach).

2. **one-versus-one**: here, $\frac{1}{2}m(m-1)$ binary classifiers are trained (one for each pair j, j' of classes, solving a binary classification). For each test sample, each binary classifier decides on – or "votes" for – the more relevant class. The test sample is then attributed to the class having the majority of votes. Although the number of binary classifiers is greater than in the one-versus-all approach, the training process for each classifier is faster since the training database is much smaller for each binary classifier. Besides, the method is more robust to class imbalances (since only pairwise comparisons are made) but suffers from an undecidability limitation in the case of equal numbers of majority votes for two or more classes.

   In Section 4.5.3, each binary classifier will be optimized according to Algorithm 2 by choosing appropriate labels and appropriate decision thresholds, thereby largely improving over the basal classifier performance.

3. **one-hot encoding approach**, also known as one-per-class coding: in this approach, each class is encoded using the m-dimensional canonical vector of the class (the code vector for class j has a 1 at position j and 0's elsewhere). When testing an unknown sample $\mathbf{x}$, the index of the encoding output vector $g_i(\mathbf{x}) \in \mathbb{R}^m$ with maximum value is selected as the class of $\mathbf{x}$.

   Exploiting the asymptotic performance of this approach from Theorem 9 allows us to derive a different label (or score) encoding for each class which theoretically minimizes the classification error. This is developed in detail in Section 4.5.4.

In the remainder of the section, each of the three classifiers is studied, optimized and their asymptotic performances are analyzed according to our previous results except for one-versus-one classification which involves difficult combinatorial aspects. While this does not provide a definite and general answer as to which of the three classifiers is best, it however provides an accurate assessment of their asymptotic performances; most importantly, these performances may be evaluated *before* running the classifier, thereby helping practitioners to anticipate and optimize the method best suited for the application at hand, without resorting to any cross-validation procedure.

Let us finally insist that, for the two multi-class extensions based on binary classifiers (one-versus-one, one-versus-all), each binary classifier will be optimized independently following Remark 8: i.e., by recasting the score vectors $y_{il}^{(j)} \in \mathbb{R}^m$ into scalar scores

$y_{il}^{\mathrm{bin}(j)} \in \mathbb{R}$. As such, from now on, for each binary classifier ℓ, $g_i(\mathbf{x};\ell) \in \mathbb{R}^m$ will be systematically turned into a scalar $g_i^{\mathrm{bin}}(\mathbf{x},\ell) \in \mathbb{R}$ well approximated by $\mathscr{N}(\boldsymbol{m}_{ij}, \boldsymbol{\sigma}_{ij})$ where now $\boldsymbol{m}_{ij}$ and $\boldsymbol{\sigma}_{ij}$ are scalar, obtained by simply replacing $y_{il}^{(j)}(\ell) \in \mathbb{R}^m$ by $y_{il}^{\mathrm{bin}(j)}(\ell) \in \mathbb{R}$ in their respective expressions.

4.5.2 One-versus-all multi-class classification

For every Task i, the one-versus-all approach solves m binary MTL LS-SVM algorithms with target class $\mathscr{C}_\ell$, for each $\ell \in \{1,\ldots,m\}$, versus all other classes $\mathscr{C}_{\ell'}$, $\ell' \neq \ell$. Calling $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ the output of the classifier ℓ for a new datum $\mathbf{x}$, the class allocation decision is traditionally based on the largest among all scores $g_i^{\mathrm{bin}}(\mathbf{x};1),\ldots,g_i^{\mathrm{bin}}(\mathbf{x};m)$. This approach generalizes the naive, yet simpler, method proposed in Algorithm 2 which, despite its good performances (recall Table 4.1), is fundamentally "incorrect" in its assuming that, for each ℓ, all classes $\mathscr{C}_{\ell'}$ ($\ell' \neq \ell$) have the same statistics.

However, this presumes that the distribution of the scores $g_i^{\mathrm{bin}}(\mathbf{x};1)$ when $\mathbf{x} \in \mathscr{C}_1$, $g_i^{\mathrm{bin}}(\mathbf{x};2)$ when $\mathbf{x} \in \mathscr{C}_2$, etc., have more or less the same mean and variance. This is not the case in general, as depicted in the first column of Figure 4.1, where data from class $\mathscr{C}_1$ are more likely to be allocated to class $\mathscr{C}_3$ (compare the red curves).

By providing an accurate estimate of the distribution of the scores $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ for all ℓ and all genuine classes of $\mathbf{x}$, Theorem 9 allows us to predict the various positions of the Gaussian curves in Figure 4.1. In particular, exploiting the theorem along with Remark 5, it is possible, for binary classifier ℓ to shift the corresponding input scores $\boldsymbol{y}^{\mathrm{bin}}(\ell)$ by a constant term $\bar{\boldsymbol{y}}(\ell) \in \mathbb{R}$ in such a way that $\mathbb{E}_{\mathbf{x}\in\mathscr{C}_\ell}[g_i^{\mathrm{bin}}(\mathbf{x};\ell)] \simeq \boldsymbol{m}_{i\ell}(\ell) = 0$ and $\mathrm{Var}_{\mathbf{x}\in\mathscr{C}_\ell}[g_i^{\mathrm{bin}}(\mathbf{x};\ell)] \simeq C_{i\ell}(\ell) = 1$. This operation prevents the centering and scale problems depicted in the first column of Figure 4.1, the result being visible in the second column of Figure 4.1.

This first improvement step simplifies the algorithm which still boils down to determining the largest $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$, $\ell \in \{1,\ldots,m\}$, output but now limiting the risks induced by the inherent centering and scale issues previously discussed.

This being said, our theoretical analysis further allows to adapt the input scores $\boldsymbol{y}^{\mathrm{bin}}(\ell)$ in such a way to optimize the expected output. Ideally, assuming $\mathbf{x}$ genuinely belongs to class ℓ, one may aim at increasing the distance between the output score $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ and the other output scores $g_i^{\mathrm{bin}}(\mathbf{x};\ell')$ for $\ell' \neq \ell$. This however demands to simultaneously adapt all input scores $\boldsymbol{y}^{\mathrm{bin}}(1),\ldots,\boldsymbol{y}^{\mathrm{bin}}(m)$. Instead, we resort to maximizing the distance between the output score $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ for $\mathbf{x} \in \mathscr{C}_\ell$ and the scores $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ for $\mathbf{x} \notin \mathscr{C}_\ell$. By "mechanically" pushing away all wrong decisions, this ensures that, when $\mathbf{x} \in \mathscr{C}_\ell$, $g_i^{\mathrm{bin}}(\mathbf{x};\ell)$ is greater than $g_i^{\mathrm{bin}}(\mathbf{x};\ell')$ for $\ell' \neq \ell$. This is visually seen in the third column of Figure 4.1, where the distances between the rightmost Gaussians and the other two is increased when compared to the second column, and we retrieve the desired behavior.

Specifically, our proposed score optimization consists in solving, for each $i \in \{1,\ldots,k\}$ and each $\ell \in \{1,\ldots,m\}$ the optimization problems:

$$\boldsymbol{y}^{\mathrm{bin}\star}(\ell) = \underset{\boldsymbol{y}^{\mathrm{bin}}(\ell)\in\mathbb{R}^{km}}{\arg\min} \max_{j\neq\ell} \mathcal{Q}\left(\frac{\boldsymbol{m}_{i\ell}(\ell) - \boldsymbol{m}_{ij}(\ell)}{\sqrt{\boldsymbol{\sigma}_{tj}}}\right)$$

$$= \underset{\mathcal{Y}^{\text{bin}}(\ell)\in\mathbb{R}^{km}}{\arg\min} \max_{j\neq\ell} \mathcal{Q}\left(\frac{\mathcal{Y}^{\text{bin}}(\ell)^{\mathsf{T}}\left(I_{mk} - \mathcal{D}_{\boldsymbol{\delta}^{[mk]}}^{-\frac{1}{2}}\Gamma\mathcal{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\right)(e^{[mk]}_{m(i-1)+\ell} - e^{[mk]}_{m(i-1)+j})}{\sqrt{\mathcal{Y}^{\text{bin}}(\ell)^{\mathsf{T}}\mathcal{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\Gamma\mathcal{V}_{ij}\Gamma\mathcal{D}_{\boldsymbol{\delta}^{[mk]}}^{\frac{1}{2}}\mathcal{Y}^{\text{bin}}(\ell)}} \right) \tag{4.13}$$

with $\mathcal{Q}$ the Gaussian q-function.

Being a non-convex and non-differentiable (due to the max) optimization, Equation 4.13 cannot be solved straightforwardly. An approximated solution consists in relaxing the max operator $\max(x_1,\ldots,x_n)$ into the differentiable operator $\frac{1}{\gamma n}\log(\sum_{j=1}^{n}\exp(\gamma x_j))$ for some $\gamma > 0$, and use a standard gradient descent optimization scheme here initialized at $\mathcal{Y}^{\text{bin}}(\ell) \in \mathbb{R}^{mk}$ filled with 1's at every $m(i'-1)+\ell$, for $i' \in \{1,\ldots,m\}$, and with -1's everywhere else.

In effect, the optimized vector $\mathcal{Y}^{\text{bin}\star}(\ell)$ is evaluated first *before* the constant shift scalar $\bar{\mathcal{Y}}$ (ensuring that $\mathbb{E}_{\mathbf{x}\in\mathscr{C}_\ell}[g_i^{\text{bin}}(\mathbf{x};\ell)]$ is close to zero) is added to $\mathcal{Y}^{\text{bin}\star}(\ell)$. This order of treatment is mandatory as $\mathbb{E}_{\mathbf{x}\in\mathscr{C}_\ell}[g_i^{\text{bin}}(\mathbf{x};\ell)]$ depends explicitly on the value of the input score vector $\mathcal{Y}^{\text{bin}}$. This global procedure is described in Algorithm 3 below.

Algorithm 3 Proposed one-versus-all multi-task learning algorithm.

Input: Training samples $X = [X_1,\ldots,X_k]$ with $X_i = [X_i^{(1)},\ldots,X_i^{(m)}]$, $X_i^{(j)} \in \mathbb{R}^{p\times n_{ij}}$ and test data $\mathbf{x}$.
Output: Estimated class $\hat{\ell} \in \{1,\ldots,m\}$ of $\mathbf{x}$ for Task i.
for $\ell = 1$ **to** m **do**
Center and normalize data per task: for all $i' \in \{1,\ldots,k\}$,

- $\mathring{X}_{i'} \leftarrow X_{i'}\left(I_{n_{i'}} - \frac{1}{n_{i'}}\mathbb{1}_{n_{i'}}\mathbb{1}_{n_{i'}}^{\mathsf{T}}\right)$
- $\mathring{X}_{i'} \leftarrow \mathring{X}_{i'}/\frac{1}{n_{i'}p}\text{tr}(\mathring{X}_{i'}\mathring{X}_{i'}^{\mathsf{T}})$.

Estimate: $\mathbb{M}^{\mathsf{T}}\bar{\tilde{Q}}_0\mathbb{M}$ and $\mathcal{V}_{i\ell}$ according to Remark 9.
Create scores $\mathcal{Y}^{\text{bin}\star}(\ell)$ by numerically solving (4.13) (see discussion following (4.13)).

Shift scores $\mathcal{Y}^{\text{bin}\star}(\ell)$ according to Remark 5.
Estimate $\boldsymbol{\sigma}_{i\ell}(\ell)$ from Theorem 9 and Remark 9.
Compute classification scores $g_i^{\text{bin}}(\mathbf{x};\ell)$ according to (4.2).
end for
Output: $\hat{\ell} = \arg\max_{\ell\in\{1,\ldots,m\}}\left\{\frac{g_i^{\text{bin}}(\mathbf{x};\ell)}{\sqrt{\boldsymbol{\sigma}_{i\ell}(\ell)}}\right\}$.

As an immediate corollary of Theorem 9, for large dimensional data, the classification accuracy of Algorithm 3 can be precisely estimated, as follows.

Proposition 6. *Under the notations of Theorem 9, the probability of correct classification*

$P_i^{(j)}(\mathbf{x})$ *for Task* i *of a test data* $\mathbf{x} \in \mathscr{C}_j$ *is given by*

$$P_i^{(j)}(\mathbf{x}) = \underbrace{\int \cdots \int_0^\infty}_{m-1} \frac{1}{\sqrt{(2\pi)^{m-1}|\boldsymbol{\sigma}(j)|}} \exp\left(-\frac{1}{2}(x-\mu(j))^\mathsf{T}\boldsymbol{\sigma}(j)^{-1}(x-\mu(j))\right) dx + o(1)$$

where $\mu(j) = \mathscr{Y}_{-j}(I_{mk} - \mathscr{D}_{\boldsymbol{\delta}}^{-\frac{1}{2}}\Gamma\mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}})e_{m(i-1)+j}^{[mk]} \in \mathbb{R}^{m-1}$ *and* $\boldsymbol{\sigma}(j) = \mathscr{Y}_{-j}\mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}}\Gamma\mathscr{V}_{ij}\Gamma\mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}}\mathscr{Y}_{-j}^\mathsf{T} \in \mathbb{R}^{(m-1)\times(m-1)}$, *with* $\mathscr{Y}_{-j} = \{\mathscr{y}^{\text{bin}}(j)^\mathsf{T} - \mathscr{y}^{\text{bin}}(j')^\mathsf{T}\}_{j'\neq j} \in \mathbb{R}^{(m-1)\times km}$.

Figure 4.1, succinctly introduced above, illustrates the successive improvements of the proposed algorithms. Specifically, it shows the gains of the centering-scaling operation on the input and output scores (2nd column) and of the optimization of the input scores (3rd column) when compared with the standard approach (1st column). Here synthetic data arising from a Gaussian mixture model are considered in a two-task ($k = 2$) and three-class ($m = 3$) setting in which $x_{1l}^{(j)} \sim \mathcal{N}(\mu_{1j}, I_p)$ and $x_{2l}^{(j)} \sim \mathcal{N}(\mu_{2j}, I_p)$, where $\mu_{2j} = \beta\mu_{1j} + \sqrt{1-\beta^2}\mu_{1j}^\perp$, with $\mu_{1j} = 2e_j^{[p]}$ and $\mu_{1j}^\perp = e_{p-j}^{[p]}$. Here $p = 200$, $\beta = 0.2$, $[n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}] = [393, 309, 394, 20, 180, 480]$ and the optimization framework used for input score (label) $\mathscr{y}^{\text{bin}}$ optimization is a standard interior point method (Boyd & Vandenberghe, 2004).[4]

4.5.3 One-versus-one multi-class classification

For a given Task i, the one-versus-one multi-class method trains $\frac{1}{2}m(m-1)$ binary classifiers for each pair $j \neq j' \in \{1, \ldots, m\}$. As intensively discussed in the previous section, as well as in Section 4.3 and Remark 8, each resulting binary classifier $g_i^{\text{bin}}(\mathbf{x}; j, j')$ can be optimized by choosing optimal input labels $\mathscr{y}^{\text{bin}}(j, j')$. This leads to Algorithm 4 described below.

In order to derive the asymptotic correct classification of class ℓ based on Algorithm 4, it is necessary to enumerate all scenarios which lead to the prediction of the class ℓ. Although this could be done in theory, the combinatorics, already for three classes, are cumbersome and not worth developing here. For the specific one-versus-one setting, we therefore do not provide a theoretical performance analysis.

4.5.4 One-hot encoding approach

For a given Task i in a one-hot encoding approach, using the canonical vector encoding for each class (i.e., $\mathscr{Y}_{ij} = e_j^{[m]}$ encodes all training input data $x_{il}^{(j)}$ of class $\mathscr{C}_j$), the class allocated to an unknown test sample $\mathbf{x}$ is the index of the output vector $g_i(\mathbf{x}) \in \mathbb{R}^m$ with maximum value.

We disrupt here from this approach by explicitly not imposing a one-hot encoding for $\mathscr{Y}_{ij}$. Instead we consider a generic encoding $\mathscr{Y} \in \mathbb{R}^{km\times m}$, which will be optimized in such a way to maximize the classification accuracy.

[4] Here we use the `fmincon` function implemented in Matlab.

[5] The mode of a set of indices is defined as the most frequent value. When multiple indices occur equally frequently, the smallest of those indices is considered by convention.

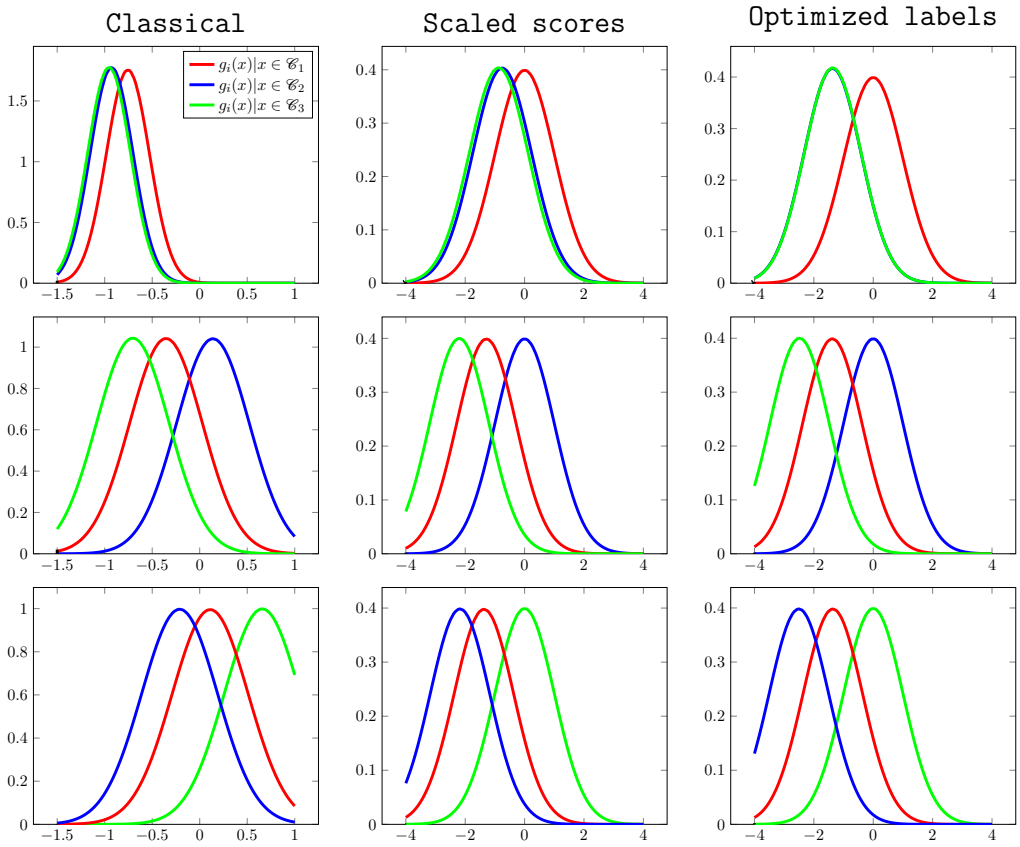

Figure 4.1: Test score distribution in a 2-task and 3 classes-per-task setting, using a one-versus-all multi-class classification. Every graph in row ℓ depicts the limiting distributions of $g_i(\mathbf{x}; \ell)$ for $\mathbf{x}$ in different classes. Column 1 (Classical) is the standard implementation of the one-versus-all approach. Column 2 (Scaled scores) is the output for centered and scaled $g_i(\mathbf{x}; \ell)$ for $\mathbf{x} \in \mathscr{C}_\ell$. Column 3 (Optimized labels) is the same as Column 2 but with optimized input scores (labels) $y^{\text{bin}\star}(\ell)$. Under the "classical" approach, data from $\mathscr{C}_1$ (red curves) will often be misclassified as $\mathscr{C}_2$. With "optimized labels", the discrimination of scores for $\mathbf{x}$ in either class $\mathscr{C}_2$ or $\mathscr{C}_3$ is improved (blue curve in 2nd row further away from blue curve in 1st row; and similarly for green curve in 3rd versus 1st row).

Algorithm 4 Proposed one-versus-one multi-task learning algorithm.

Input: Training samples $X = [X_1, \ldots, X_k]$ with $X_i = [X_i^{(1)}, \ldots, X_i^{(m)}]$, $X_i^{(j)} \in \mathbb{R}^{p \times n_{ij}}$ and test data $\mathbf{x}$.
Output: Estimated class $\hat{\ell} \in \{1, \ldots, m\}$ of $\mathbf{x}$ for Task i.
Center and normalize data per task: for all $i' \in \{1, \ldots, k\}$,

- $\mathring{X}_{i'} \leftarrow X_{i'} \left(I_{n_{i'}} - \frac{1}{n_{i'}} \mathbb{1}_{n_{i'}} \mathbb{1}_{n_{i'}}^\mathsf{T} \right)$
- $\mathring{X}_{i'} \leftarrow \mathring{X}_{i'} / \frac{1}{n_{i'} p} \mathrm{tr}(\mathring{X}_{i'} \mathring{X}_{i'}^\mathsf{T})$.

for $j = 1$ **to** m **do**
  for $j' \in \{1, \ldots, m\} \setminus \{j'\}$ **do**
    Estimate: $\mathbb{M}^\mathsf{T} \tilde{\bar{Q}}_0 \mathbb{M}$ and $\mathscr{V}_{ij}$ according to Remark 9.
    Create optimal scores $\mathscr{y}^{\mathrm{bin}\star}(j', j)$ according to Remark 8.
    Compute classification scores according to (4.2) and deduce the predicted class $c(j, j') = j$ or $c(j, j') = j'$ based on the decision rule in (4.6).
  end for
end for
Output: $\hat{j} = \underset{j', j \in \{1, \ldots, m\}}{\mathrm{mode}} \{c(j, j')\}$.[5]

Proposition 7. *Under a "one-hot encoding" scheme with generic $\mathscr{Y}$, the probability of correct classification $P_i^{(j)}(\mathbf{x})$ for Task i of a test data $\mathbf{x} \in \mathscr{C}_j$ is given by*

$$P_i^{(j)}(\mathbf{x}) = \underbrace{\int \cdots \int_0^\infty}_{m-1} \frac{1}{\sqrt{(2\pi)^{m-1} |\boldsymbol{\sigma}(j)|}} \exp\left(-\frac{1}{2} (x - \mu(j))^\mathsf{T} \boldsymbol{\sigma}(j)^{-1} (x - \mu(j)) \right) dx,$$

where $\mu(j) = \mathscr{C}_j \mathscr{Y}^T \left(I_{mk} - \mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}} \Gamma \mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}} \right) e_{(m(i-1)+j)}^{[km]} \in \mathbb{R}^{m-1}$ *and* $\boldsymbol{\sigma}(j) = \mathscr{C}_j \mathscr{Y}^\mathsf{T} \mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}} \Gamma \mathscr{V}_{ij} \Gamma \mathscr{D}_{\boldsymbol{\delta}}^{\frac{1}{2}} \mathscr{Y} \mathscr{C}_j^\mathsf{T} \in \mathbb{R}^{(m-1)\times(m-1)}$ *with* $\mathscr{C}_j = \{(e_j^{(m)} - e_{j'}^{(m)})^\mathsf{T}\}_{j \neq j'} \in \mathbb{R}^{(m-1)\times m}$.

A natural objective is to set $\mathscr{Y}$ so to maximize the average correct classification accuracy $\frac{1}{m} \sum_{j=1}^m P_i^{(j)}(\mathbf{x})$ (under assumed uniform prior on $\mathbf{x}$). This form again is not convex in $\mathscr{Y}$ but may be approximated by gradient descent starting from the one-hot encoding solution, as described in Algorithm 5.

4.6 Experiments

This section has a double objective. The first part (Section 4.6.1) devises numerical experiments on binary classification settings to corroborate the theoretical analyses and conclusions drawn in this previous section. Here, the target is threefold: (i) empirically illustrate the effects of the bias in the threshold decision and in the label optimization

Algorithm 5 Proposed "one-hot encoding" multi-task learning algorithm.

Input: Training samples $X = [X_1, \ldots, X_k]$ with $X_i = [X_i^{(1)}, \ldots, X_i^{(m)}]$, $X_i^j \in \mathbb{R}^{p \times n_{ij}}$ and test data $\mathbf{x}$.
Output: Estimated class $\hat{\ell} \in \{1, \ldots, m\}$ of $\mathbf{x}$ for target Task i.
Center and normalize data per task: for all $i' \in \{1, \ldots, k\}$,

- $\mathring{X}_{i'} \leftarrow X_{i'} \left(I_{n_{i'}} - \frac{1}{n_{i'}} \mathbb{1}_{n_{i'}} \mathbb{1}_{n_{i'}}^\mathsf{T} \right)$
- $\mathring{X}_{i'} \leftarrow \mathring{X}_{i'} / \frac{1}{n_{i'} p} \mathrm{tr}(\mathring{X}_{i'} \mathring{X}_{i'}^\mathsf{T})$.

Estimate Matrix $\mathbb{M}^\mathsf{T} \tilde{Q}_0 \mathbb{M}$ and $\mathscr{V}_{ij}$ according to Remark 9.
Compute the theoretical score $\mu(j)$ and covariance $\boldsymbol{\sigma}(j)$ and derive the asymptotic classification accuracy $P_i(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^{m} P_i^{(j)}(\mathbf{x})$.
Create optimal scores $\mathscr{Y}^\star = \arg\max_{\mathscr{Y}} P_i(\mathbf{x})$.
Compute classification scores $g_i(\mathbf{x})$ according to (4.2).
Output: $\hat{\ell} = \arg\max_{\ell \in \{1,\ldots,m\}} g_i(\mathbf{x}; \ell)$.

scheme discussed in Section 4.3, (ii) discuss the impact of numerous tasks ($k > 2$) in the binary class setting, thereby emphasizing the effects of negative transfer and its correction through input score (label) optimization, and (iii) exemplify the relevance of our theoretical findings to a specific application to hypothesis testing in a multi-task setting.

In a second part (Section 4.6.2), experiments on both synthetic and real data for multi-class classification are realized, which demonstrate, even for real data: (i) the extreme accuracy of the theoretical performance predictions of Propositions 6–7 against empirical data and (ii) the large performance gains induced by the various improvements introduced at length in Section 4.5.

4.6.1 Experiments on binary classification

Effect of input score (label) choice

In the present experiment, the effects of the bias in the decision threshold (in general not centered on zero) and of the input score (label) optimization are demonstrated on both synthetic data and real data.

Specifically, MTL LS-SVM is first applied to the following two-task ($k = 2$) and two-class ($m = 2$) setting: for Task 1, $x_{1l}^{(j)} \sim \mathcal{N}((-1)^j \mu_1, I_p)$ (evenly distributed in both classes) and for Task 2, $x_{2l}^{(j)} \sim \mathcal{N}((-1)^j \mu_2, I_p)$ (evenly distributed in both classes), where $\mu_2 = \beta \mu_1 + \sqrt{1 - \beta^2} \mu_1^\perp$ and $\mu_1^\perp$ is any vector orthogonal to μ_1 and $\beta \in [0, 1]$. This setting allows us to tune, through β, the similarity between tasks. For four different values of β, Figure 4.2 depicts the distribution of the binary output scores $g_i^{\mathrm{bin}}(\mathbf{x})$ both for the classical MTL LS-SVM (top displays) and for our proposed random matrix improved

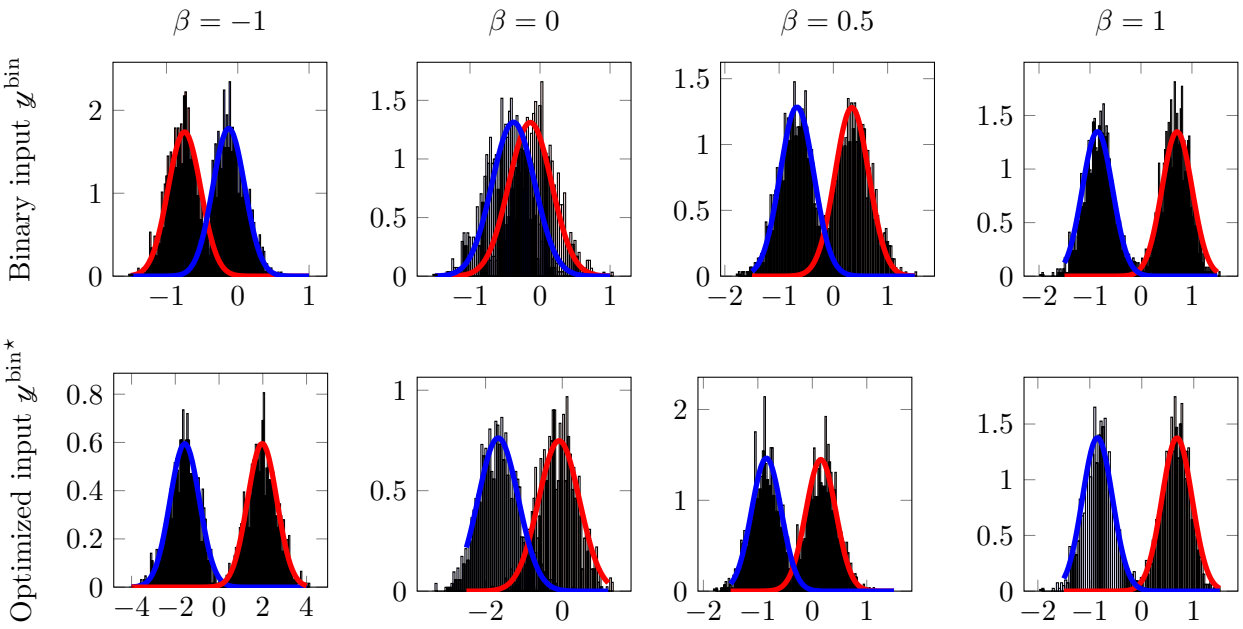

Figure 4.2: Score distribution for new datum $\mathbf{x}$ of Class $\mathscr{C}_1$ (red) and Class $\mathscr{C}_2$ (blue) for Task 2 in a 2-task ($k = 2$) and 2 class-per-task ($m = 2$) setting of isotropic Gaussian mixtures for: **(top)** classical MTL LS-SVM with no optimization and a threshold assumed at $\zeta = 0$; **(bottom)** proposed optimized MTL LS-SVM with estimated threshold ζ; decision thresholds ζ represented in dashed vertical lines; differently related tasks ($\beta = 0$ for orthogonal means, $\beta > 0$ for positively correlated means and $\beta < 0$ for negatively correlated means), $p = 100$, $[c_{11}, c_{12}, c_{21}, c_{22}] = [0.3, 0.4, 0.1, 0.2]$, $\gamma = \mathbb{1}_2$, $\lambda = 10$. Histograms drawn from 1 000 test samples of each class. The figure clearly depicts the deviation from 0 of the decision threshold in unbalanced classes and the deleterious effect of "negative transfer" when β is small; these problems are well handled by the proposed optimized scheme.

scheme, with optimized input labels (bottom display).

As a first remark, note that both theoretical prediction and empirical outputs closely fit for all values of β, thereby corroborating our theoretical findings. In practical terms, the figure supports (i) the importance to estimate the threshold decision which is non-trivial (not always close to zero) and (ii) the relevance of an appropriate choice of the input labels to improve the discrimination performance between both classes, especially when the two tasks are not quite related. In effect, the entries of $y^{\text{bin}\star}$ naturally drop to zero for all unrelated tasks and classes, thereby discarding the undesired use of the latter; the classical binary input labels instead inappropriately exploit these data and induce a negative learning effect, sometimes to such an extent to completely switch the final decision (as here when $\beta = -1$).

For experiments on real data, the MNIST dataset (Deng, 2012) is considered. Specif-

ically, the setting is that of a binary classification for two tasks, mimicking a transfer learning setting: there, the "target" Task 2 aims to discriminate Class $\mathscr{C}_1$ and Class $\mathscr{C}_2$ respectively composed of images of digit 1 and digit 4. The "source" Task 1 is here used as a support for classification in the target task, and consists of the classification of other pairs of digits: either $(5, 9)$, $(9, 5)$, $(6, 2)$ or $(8, 3)$ (we recall that the order of the set of digits (X, Y) is important for the non-optimized MTL LS-SVM since the source and target tasks labels are "paired"; thus $(5, 9)$ or $(9, 5)$ digits for the source task will bring different results). We compare here again the non-optimized MTL LS-SVM with labels $\mathscr{y}^{\text{bin}} = [-1, 1, -1, 1]^\mathsf{T}$ to our proposed optimized scheme (as detailed in Remark 8). For both methods, the optimal theoretical threshold decision ζ is used (rather than $\zeta = 0$ for the non-optimized setup) in order to emphasize the influence of input score (label) optimization.

Figure 4.3 depicts the performance for both methods as a function of the hyper-parameter λ. We recall that, as $\lambda \to 0$, the multi-task scheme becomes equivalent to independent single-task classifiers, while as $\lambda \to \infty$, both source and target tasks are considered together as one task. Figure 4.3 raises the stability of optimal input labeling with respect to λ: this is explained by the fact that $\mathscr{y}^{\text{bin}\star}$ is a *function of* λ and thus adapts to each value of λ, even if suboptimal. Besides, for appropriate values of λ, the proposed improved labeling can largely outperform the non-optimized setting, even here on real data.

Table 4.2 complements the figure by effectively displaying the optimal vectors $\mathscr{y}^{\text{bin}\star}$ at the optimal value for λ. The table demonstrates the appropriate adjustment of the labels to the data correlation $\frac{\Delta\mu_1^\mathsf{T}\Delta\mu_2}{\|\Delta\mu_2\|^2}$. Specifically, for a negative correlation between the classes of both tasks, the method naturally "switches" the labels (the input data scores) by opposing the signs of $\mathscr{y}^{\text{bin}\star}$ in entries $1, 3$ (Class $\mathscr{C}_1$ in each task) and $2, 4$ (Class $\mathscr{C}_2$ in each task). For rather orthogonal tasks (here typically $(8, 3)$), the entries of $\mathscr{y}^{\text{bin}\star}$ corresponding to the source task (entries 1 and 2) are almost zero, thereby discarding the source data and avoiding negative transfer. It is also interesting to note that, for moderately correlated tasks (here for the source digits $(5, 9)$), despite the fact that the source task offers ten times more data ($n_{1j} = 100$, $n_{2j} = 10$) and is thus deemed trustworthy for classification, the corresponding entries 1 and 2 in $\mathscr{y}^{\text{bin}\star}$ are much smaller than the entries 3, 4 corresponding to the target task: the algorithm thus judges the few target data more relevant to target classification than the many related source tasks.

Analysis of increasing number of tasks

This next experiment illustrates the effect of adding more tasks for the transfer learning setting on synthetic and MNIST datasets. For synthetic data, Gaussian classes with mean $\mu_{ij} = \beta\mu_{i1} + \sqrt{1-\beta^2}\mu_{i1}^\perp$ and various values of β are successively added. For the MNIST dataset, different classifications of digits are added progressively to help classify the specific pair of digits $(1, 4)$. Figure 4.4 depicts the classification error after each new task addition, both for a classical binary input label choice and for the proposed optimized input labels. The figure forcefully illustrates that our proposed framework

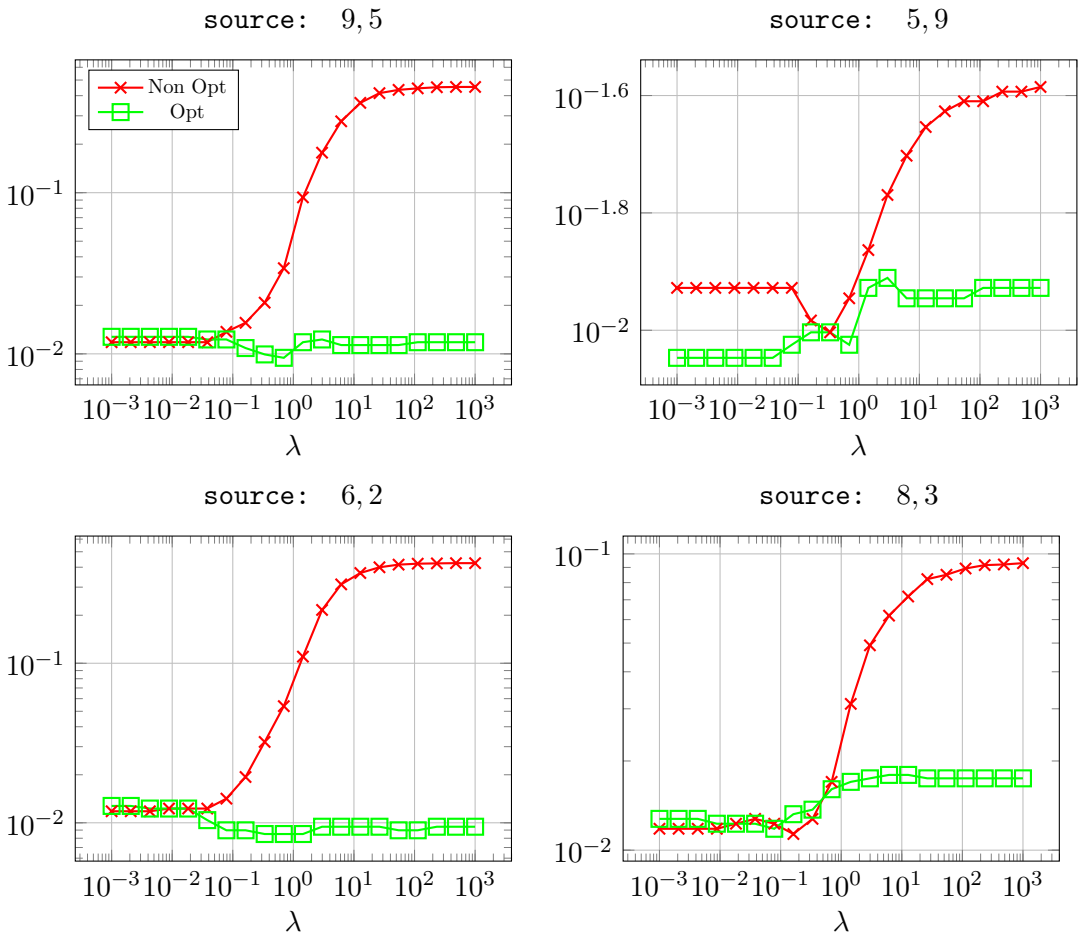

Figure 4.3: Classification error of digit pair $(1, 4)$ with different source training pairs for classical LS-SVM and optimized LS-SVM. $n_{11} = n_{12} = 100$, $n_{21} = n_{22} = 10$ and $\gamma = \mathbb{1}_2$. A PCA preprocessing is performed on each image to extract their $p = 100$ principal components; the accuracy is performed over $n_{\text{test}} = 1\,135$ test samples. The proposed method shows a low sensitivity to λ.

[Source]	(9,5)	(5,9)	(6,2)	(8,3)
$\frac{\Delta\mu_1^\intercal \Delta\mu_2}{\|\Delta\mu_2\|^2}$	-0.2450	0.2450	-0.1670	-0.0818
$y^{\text{bin}\star} = \begin{bmatrix} y_{11}^{\text{bin}\star} \\ y_{12}^{\text{bin}\star} \\ y_{21}^{\text{bin}\star} \\ y_{22}^{\text{bin}\star} \end{bmatrix}$	$\begin{bmatrix} -0.2808 \\ 0.2808 \\ 0.6489 \\ -0.6489 \end{bmatrix}$	$\begin{bmatrix} 0.2808 \\ -0.2808 \\ 0.6489 \\ -0.6489 \end{bmatrix}$	$\begin{bmatrix} -0.2879 \\ 0.2879 \\ 0.6459 \\ -0.6459 \end{bmatrix}$	$\begin{bmatrix} -0.0400 \\ 0.0400 \\ 0.7060 \\ -0.7060 \end{bmatrix}$

Table 4.2: Optimal input label $y^{\text{bin}\star}$ as a function of the source data pair in the (λ-optimal) configuration of Figure 4.3.

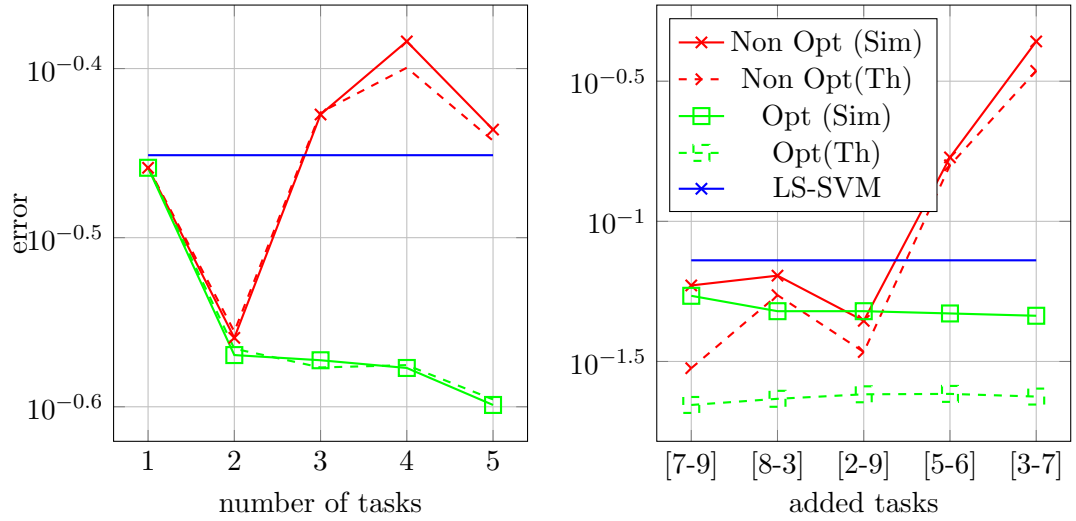

Figure 4.4: Classification accuracy for increasing number of tasks. **(Left)** Synthetic data with task correlations $\beta = 1, .9, .5, .2, .8$ in this order, $p = 100$ and $c = [.07, .11, .10, .10, .06, .08, .09, .12, .10, .11, .03, .03]^\mathsf{T}$; accuracy evaluated out of 10 000 test samples. **(Right)** MNIST dataset with digits $(1, 4)$ as target task, each added task being shown in x-axis; 100 training samples are used for each class of the source tasks and 10 training samples for each class of the target class; HOG features with $p = 144$ for each image digit; accuracy evaluated out of $n_{\text{test}} = 1\,135$ test samples. For both setting, $\gamma = \mathbb{1}_k$ and $\lambda = 10$. The optimized scheme avoids negative transfer by systematically benefiting from additional tasks.

avoids negative transfer, as the classification error of MTL never increases as the number of tasks grows. This is quite unlike the non-optimized scheme which severely suffers from negative transfer.

Hypothesis testing

The next experiments, on synthetic data, apply the results of MTL LS-SVM to a hypothesis test on a *target* Task t based on training samples both from a source Task s and the target Task t. For data $\mathbf{x}$ in the target task, the test

$$g_t^{\text{bin}}(\mathbf{x}) \underset{\mathscr{H}_0}{\overset{\mathscr{H}_1}{\gtrless}} \zeta$$

is performed, where $\mathscr{H}_0$ is the null hypothesis (say, Class 2) and $\mathscr{H}_1$ the alternative (say, Class 1) and $\zeta = \zeta(\eta)$ is a decision threshold here selected in such a way to enforce the false alarm rate constraint $P(g_t(\mathbf{x}) \geq \zeta(\eta) \mid \mathbf{x} \in \mathscr{H}_0) \leq \eta$, for a given $\eta \in (0, 1)$. The objective is then to maximize over the input scores $\mathscr{y}^{\text{bin}}$ the correct detection rate $P(g_t^{\text{bin}}(\mathbf{x}) \geq \zeta(\eta) \mid \mathbf{x} \in \mathscr{H}_1)$: this induces a different value for the optimal scores $\mathscr{y}^{\text{bin}\star}$ than proposed in (4.12), which can be constructed following Remark 8.

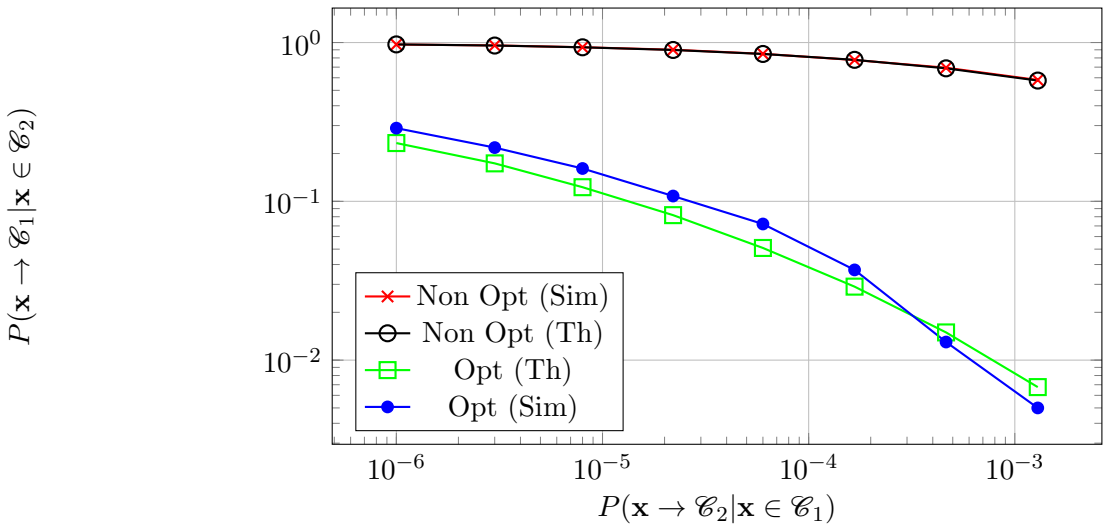

Figure 4.5: ROC curve for proposed optimized versus standard MTL LS-SVM. Synthetic data with $p = 128$, $n_{11} = 384$, $n_{12} = 256$, $n_{21} = 64$, $n_{22} = 40$, $\mu_{11} = -\mu_{12} = [1, 0, \ldots, 0]^\mathsf{T}$, $\mu_{21} = -\mu_{22} = [.87, .5, 0, \ldots, 0]^\mathsf{T}$. The accuracy of the theoretical anticipation is remarkable and allows for a precise setting of the decision threshold ensuring a desired false alarm rate.

The experimental synthetic data is here a two-task ($k = 2$) setting in which $x_{1j} \sim \mathcal{N}(\pm\mu_{11}, I_p)$ (i.e., $\mu_{12} = -\mu_{11}$) and $x_{2j} \sim \mathcal{N}(\pm\mu_{21}, I_p)$, where $\mu_{21} = \beta\mu_{11} + \sqrt{1-\beta^2}\mu_{11}^\perp$, μ_{11} is a unit-norm vector and $\mu_{11}^\perp$ any unit-norm vector orthogonal to μ_{11}. We take here $\beta = 0.5$, so that both tasks are "slightly" correlated.

Figure 4.5 depicts the algorithm performance through a receiver-operating curve (ROC) for false alarm rates η on synthetic data. Both theoretical (Th) asymptotics (used to set the decision threshold ζ) and actual performances (Sim) are displayed, for the optimal (Opt) choice of $\mathscr{y}^{\text{bin}}$ (Opt) and for $\mathscr{y}^{\text{bin}} = [-1, 1, -1, 1]^\mathsf{T}$ (Non-Opt).

Figure 4.5 confirm, here under the hypothesis testing problem, the large superiority of our proposed optimized MTL LS-SVM over the standard non-optimized alternative. Besides, the theoretical classification error prediction is an accurate fit to the actual empirical performance, even for not so large values of p and the n_{ij}'s, and even for small error values.[6] This remark is here all the more fundamental that, in practice, η can be set a priori, using Theorem 9 with no need for heavy, unreliable, and data-consuming cross-validation procedures.

[6]Since our main result (Theorem 9) is a central limit theorem, it is not expected to be particularly accurate in the "tails" of the distribution of the output scores $g_i^{\text{bin}}(\mathbf{x})$; as such, the observed high accuracy for small error values is remarkable.

4.6.2 Experiments on multi-class classification

We here consider the complete setting of a $k \geq 2$, $m > 2$ multi-class learning scenario, first on synthetic and then on real image datasets.

Experiments on synthetic dataset

In the synthetic data experiment, the scenario is a two-task ($k = 2$) setting in which $x_{1l}^{(j)} \sim \mathcal{N}(\mu_{1j}, I_p)$ and $x_{2l}^{(j)} \sim \mathcal{N}(\mu_{2j}, I_p)$, where $\mu_{2j} = \beta\mu_{1j} + \sqrt{1-\beta^2}\mu_{1j}^{\perp}$, with $\mu_{1j} = 2e_j^{[p]}$ and $\mu_{1j}^{\perp} = e_{p-j}^{[p]}$, and β varies from 0.1 to 0.8.

Table 4.3 provides the empirical classification accuracy achieved by one-versus-all (Algorithm 3), one-versus-one (Algorithm 4) and one-hot (Algorithm 5) learning versus their standard (non-optimized) algorithm equivalent on 10 000 test samples. The table also reports the theoretical classification accuracies predicted by the empirical estimation of the quantities involved in Propositions 6–7 (therefore without any cross-validation) for the one-versus-all and one-hot methods.

Table 4.3: Classification accuracy for synthetic data $x_{1l}^{(j)} \sim \mathcal{N}(\mu_{1j}, I_p)$ and $x_{2l}^{(j)} \sim \mathcal{N}(\mu_{2j}, I_p)$, $\mu_{2j} = \beta\mu_{1j} + \sqrt{1-\beta^2}\mu_{1j}^{\perp}$, for different values of the data-correlation $\beta > 0$ and various multi-class learning algorithms. Theoretical performance predictions are provided in parentheses. Here $m = 5$, $p = 100$, $c_{1j} = .16$, $c_{2j} = .04$, for $j \in \{1, \ldots, 5\}$, $\lambda = 1$ and $\gamma = \mathbb{1}_k$. The performance gains of the proposed optimal scheme is particularly clear in tasks with low correlation.

β	Method	one-vs-all	one-vs-one	one-hot
$\beta = 0.1$	Classical	61.43 (59.87)	65.31	65.61 (64.35)
	Optimized	67.63 (67.57)	74.98	67.63 (67.55)
$\beta = 0.5$	Classical	65.47 (66.00)	71.30	67.41 (67.90)
	Optimized	68.00 (68.52)	76.31	68.03 (68.48)
$\beta = 0.8$	Classical	71.16 (70.63)	78.20	70.97 (70.58)
	Optimized	71.19 (70.76)	78.55	71.14 (70.67)

The output performance scores naturally show an improvement using the proposed MTL LS-SVM framework and confirm again the extremely accurate prediction of performance by the theoretical formulas. Most importantly, the table reveals that the gap between the non-optimized and optimized schemes is all the more important that the correlation between task (through the parameter β) is small; this indicates that the optimized MTL LS-SVM learning framework better exploits the (even little) correlation arising between tasks or, alternatively, that the non-optimized scheme suffers from negative learning when "over-emphasizing" the weight of data from the other task (through the binary input labels $\mathcal{Y}$).

As for the comparison of the three classification methods (one-versus-all, one-versus-one and one-hot), it shows here an overall superiority of the one-versus-one approach. This result should nonetheless be interpreted with extreme care as no optimization over

the hyperparameters γ, λ is conducted in any scenario.

Image classification

Similarly as in Section 4.3, we now turn to the popular Office+Caltech256 multi-task image classification benchmark (Saenko et al., 2010; Griffin et al., 2007) often exploited for transfer learning. The overall database consists of 10 categories shared by both Office and Caltech256 datasets. As in Table 4.1, we consider in sequence the transfer learning of one out of four possible source tasks, each of which consisting in classifying data from one sub-database (images issued from the Caltech set (c), Webcam images (w), Amazon pictures (a) or dslr images (d)), towards another task; this boils down to $4 \times (4-1) = 12$ source-target comparison pairs.)

The results in Table 4.1 using VGG features for the image representations are extremely close to 100%, already for the "naive" approach consisting in a simplified one-versus-all extension of Algorithm 2. Little would be gained (at least not in computational efforts) by running the more involved Algorithm 3 on the same database. For this reason, for the present experiment, we compare the more challenging (since less discriminative) $p = 800$ SURF-BoW features of the Office+Caltech256 images instead of their VGG features.

Half of the samples of the target task are randomly selected as test data and the accuracy is evaluated over 20 independent trials. For complexity reasons, as in Section 4.3, for each experiment, the naive version of the one-versus-all algorithm is run 10 times, considering a fictitious two-class $\tilde{\mathscr{C}}_1$-versus-$\tilde{\mathscr{C}}_2$ setting where, for the classifier focusing on class $\mathscr{C}_\ell$, class $\tilde{\mathscr{C}}_1 = \mathscr{C}_\ell$ while class $\tilde{\mathscr{C}}_2$ is the union of all other classes $\mathscr{C}_{\ell'}$, $\ell' \neq \ell$.

Table 4.4 reports the accuracy obtained by the algorithm (Proposed) versus the non-optimized MTL LS-SVM from (Xu et al., 2013) (LS-SVM) and state-of-the-art transfer learning algorithms already introduced in Section 4.3. Table 4.4 again demonstrates that our proposed improved MTL LS-SVM, despite its simplicity and unlike the competing methods used for comparison, has stable performances and is highly competitive.

Table 4.4: Classification accuracy for transfer learning on the Office+Caltech256 database, against state-of-the-art alternatives. Here with c(Caltech), w(Webcam), a(Amazon), d(dslr) based on SURF-BoW features. Our proposed approach is systematically best or second to best and best on average.

S/T	c→ w	w→ c	c→ a	a→ c	w→ a	a→ d	d→ a	w→ d	c→ d	d→ c	a→ w	d→ w	Mean score
LS-SVM	79.47	47.70	68.10	49.65	68.13	57.50	70.00	73.75	67.50	46.45	74.83	84.11	65.60
MMDT	69.47	42.55	68.95	39.70	65.24	59.50	62.16	**86.06**	56.94	27.92	68.54	**87.88**	61.24
ILS	24.5	20.92	25.21	21.10	22.92	26.25	27.08	43.75	30.00	26.95	15.23	57.62	28.46
CDLS	*82.28*	**54.21**	*73.75*	**54.49**	**71.52**	*68.56*	*70.54*	69.44	*69.44*	**53.86**	**81.59**	82.78	*69.37*
Ours	**86.09**	*49.65*	**75.00**	*50.35*	*68.83*	**73.75**	**71.25**	*72.50*	**77.50**	*48.05*	*80.13*	*85.43*	**69.88**

4.7 Concluding remarks

Through the example of multi-task learning, as well as its particularization to transfer learning, we demonstrate the ability of random matrix theory to predict the performance of advanced machine learning schemes (here based on an extension of LS-SVM) and most importantly to propose improved learning mechanisms, which are competitive with, if not largely outperforming, elaborate state-of-the-art alternatives.

Interestingly, as already reported in recent works (Mai & Couillet, 2018; Mai et al., 2019), the proposed random-matrix-optimized framework is largely counter-intuitive and comes along with novel insights on the overall learning mechanisms of large dimensional data classification. Here specifically, the proposed input score (label) optimization is at odds with the conventional binary input label insights of most machine learning schemes, but is key to optimize the exploitation of other tasks and to discard altogether the long standing problem of negative transfer.

The random-matrix framework also draws a significant advantage in its being *universal* to data distributions. As shown here, our main results (Theorem 9) are valid for data modeled as mixtures of concentrated random vectors which go quite beyond the usually assumed Gaussian mixtures, as they encompass extremely realistic synthetic data models (such as GAN images). This universality phenomenon, possible surprising at first, in fact holds for a wide range of large dimensional "dense" (as opposed to sparse) data representation vectors, encompassing not only images but also likely other forms of data representations, such as word embeddings in natural language processing, vectors of moments of graphons in statistical graph analysis, etc.

To conclude, we importantly emphasize a fundamental underlying take-away message of the present work: recalling that LS-SVM is nothing but an explicit and computationally-cheap linear regression method, the fact that it competes or even outperforms elaborate MTL methods testifies of the possibility, when dealing with large dimensional data, to design highly performing elementary and cost-efficient random-matrix-based learning schemes. This remark is in line with the recent parallel analysis of information-theoretic bounds on the performances of machine learning problems, such as in (Lelarge & Miolane, 2019) for semi-supervised learning (SSL); similar to the present work, in (Mai & Couillet, 2018), the authors propose a random-matrix-based optimization of standard graph SSL learning which they demonstrate to tightly reach the information-theoretic upper bound of (Lelarge & Miolane, 2019). This strongly suggests the practical relevance of "reinvesting" research efforts in simple, cost-efficient, theoretically tractable, controllable, and usually more stable machine learning schemes, rather than in complex and theoretically intractable techniques.

Chapter 5

Perspectives and future work

Contents

5.1 Short-term perspectives related to chapters 3 and 4

The analyses carried out throughout this thesis provide a theoretical understanding and corrections of learning schemes involving covariance matrices as well as theoretical and practical insights into Multi-Task and Transfer Learning algorithms. However, some research directions which are discussed next will enable the work to have a significant societal and broader impact.

Application to real data. The random matrix improved estimation of the distance between covariance matrices has been applied to a covariance-based feature classification but the real strength and robustness of the proposed estimator will only be demonstrated when applied to real (non-Gaussian) datasets and more exotic applications (brain signal classification, hyperspectral image classification, etc). Moreover, the random matrix improved estimation scheme proposed can be used in the context of testing the equality of population covariance matrix widely used in signal processing (Krzanowski, 1979; Boik, 1988; Schott, 1991). This would however need to access the fluctuations of the random matrix estimator proposed in theorem 7. More concretely, it would be convenient to obtain a result similar to (Yao et al., 2012) in the present context, that is a central limit theorem for the fluctuations of the estimator of Theorem 7. This would allow in addition to the applications mentioned previously to access both a consistent estimator for their sought-for matrix distance as well as a confidence margin. This investigation demands even more profound calculi (as can be seen from the detailed derivations of (Yao et al., 2012)).

On the other hand, for the metrics we are dealing with, it is crucial that the smallest eigenvalue of the covariance matrices Σ_1 and Σ_2 does not tend to 0 as $p \to \infty$ which leads to singularities of the covariance matrices in the Positive Semi-Definite Riemannian manifold. This happens quite often in over-parametrized settings as in Electroencephalography (EEG) datasets (Rodrigues et al., 2017). The randomness involved in the measurements exacerbates the singularities of the sample covariance matrices $\hat{\Sigma}_1$, $\hat{\Sigma}_2$ and makes even more challenging the estimation procedure. A first step to tackle this problem may

consist to consider instead the regularized population covariance matrices $\Sigma_1 + \lambda_1 I_p$ and $\Sigma_2 + \lambda_2 I_p$ for some constants $\lambda_1, \lambda_2 > 0$. Several questions, therefore, need to be properly addressed. What is the loss incurred by the regularization scheme as function of the strength of λ_1 and λ_2? Is there an optimal value of λ_1 and λ_2 as function of the spectrum of Σ_1 and Σ_2?

The High Dimension Low Sample Size Regime. We point out the difficulty to handle the high dimension low sample size regime (i.e the regime $n_i \leq p$). It is important to note that this regime is very common in real-life applications where data collection is often difficult and the dimensions are relatively large. Managing this critically important setting will help to get closer to real-life problems. We propose one solution with polynomial approximation in (Tiomoko & Couillet, 2019a) which only handles partially the case $n_2 < p$. Random projections and regularization methods can be alternative methods to tackle this scenario, however possibly to the detriment of the estimator consistency. More concretely, relying on the random projection approach one would wonder for a random matrix $W \in \mathbb{R}^{p \times q}$, what would be the loss induced by projecting the covariance matrices Σ_1 and Σ_2 in the q-dimensional space generated by the columns of W? In other words, this consists to study the statistical behavior of the random quantity $D(W^\mathsf{T}\Sigma_1 W, W^\mathsf{T}\Sigma_2 W) - D(\Sigma_1, \Sigma_2)$ as $p, n \to \infty$. This would be adequate in the case of $p > n_1, n_2$ to choose $q < n_1, n_2$ and then apply the estimation framework derived in this thesis and one would furthermore quantify the loss incurred as the ratio q/p decreases. From a technical point of view, this requires choosing an appropriate model for the random matrix W (Haar random matrix, Gaussian matrix,...) and to deal with the asymptotic eigenvalue distribution of $(W^\mathsf{T}\Sigma_1 W)^{-1} W^\mathsf{T}\Sigma_2 W$ which can be quite involved due to the inverse matrix. Free probability theory (Mingo & Speicher, 2017) can be an adequate tool to handle such asymptotic eigenvalue distribution since it allows for computing the asymptotic eigenvalue distribution of rational functions of random matrices. But a first step would rather consider the family metric involving the eigenvalue distribution of $W^\mathsf{T}\Sigma_1 W W^\mathsf{T}\Sigma_2 W$.

Towards an understanding of covariance matrix estimation scheme. We proposed a generic framework to estimate the covariance matrix Σ under different metrics. For the subsequent discussion, we briefly recall the estimation procedure which is based on remarking that $\Sigma \equiv \arg\min_{M \succ 0} D(M, \Sigma)$, for some distance between any deterministic matrix M and the sought-for covariance matrix Σ. Relying on the proposed improved estimate $\hat{D}(M, X)$ for $D(M, \Sigma)$ based on samples $X = [x_1, \ldots, x_n]$ (of zero mean and covariance matrix Σ), the optimization problem introduced above can be approximated as $\check{\Sigma} \equiv \arg\min_{M \succ 0} h_X(M)$ with $h_X(M) = \hat{D}(M, X)^2$ and solved using a gradient descent algorithm.

Since each distance (Fisher distance, Wasserstein distance, ...) leads to a different estimate, the natural question would be to compare the different estimators in order to give which suits for a specific application. This however requires to understand theoretically the gradient descent step performed which is quite involved and sub-optimal.

Indeed, the considered approach suffers a profound limitation: $\hat{D}(M, X)$ only estimates $D(M, \Sigma)$ for M independent of X. This poses a formal problem when implemented in a gradient descent. More concretely, in chapter 3, it is precisely shown that, for every deterministic sequence of matrices $\{M^{(p)},\ p = 1, 2, \ldots\}$ and $\{\Sigma^{(p)},\ p = 1, 2, \ldots\}$, with $M^{(p)}, \Sigma^{(p)} \in \mathbb{R}^{p \times p}$ and $\max(\|\Sigma^{(p)}\|, \|M^{(p)}\|) < K$ for some constant K independent of p, we have that, for $X^{(p)} = [x_1^{(p)}, \ldots, x_n^{(p)}]$ with $x_i^{(p)} = \Sigma^{(p)\frac{1}{2}} z_i^{(p)}$ and $z_i^{(p)}$ i.i.d. vectors of i.i.d. zero mean and unit variance entries,

$$D(M^{(p)}, \Sigma^{(p)}) - \hat{D}(M^{(p)}, X^{(p)}) \to 0$$

almost surely as $n, p \to \infty$ and $p/n \to c_0 \in (0, 1)$. If we denote M_k the matrix obtained at the iteration k of the gradient descent step, the proposed methodology supposes that $\hat{D}(M_k, X)$ is a good approximation for the sought-for $D(M_k, \Sigma)$. This, however, only holds true so long that M_k *is independent of* X which clearly does not stand when proceeding to successive gradient descent steps in the direction of $\nabla h_X(M)$ which depends explicitly on X. As such, while initializations with, say, $M_0 = I_p$, allow for a close approximation of $D(M_k, \Sigma)$ in the very first steps of the descent, for larger values of k, the descent is likely to drive the optimization in less accurate directions. This needs be tackled: (i) either by estimating the introduced bias so to infer the loss incurred or, better, (ii) by accounting for the dependence to provide a further estimator $\hat{D}(M(X), X)$ of $D(M(X), \Sigma)$ for all X-*dependent* matrices $M(X)$ following a specific form.

Furthermore, if we denote by U and $\hat{U}$ respectively the eigenvectors of the population covariance matrix Σ and the sample covariance matrix $\hat{\Sigma}$, we can prove that the set of covariance matrix estimators $\mathscr{H} = \{\hat{U} D \hat{U}^\mathsf{T} \mid D \text{ diagonal}\}$ is a stable set of Algorithm 1 in the sense that $M_k \in \mathscr{H} \Rightarrow M_{k+1} \in \mathscr{H}$. One may thus wonder if $\mathscr{H}$ is also a global attractor: i.e., does every trajectory $\{M_1, M_2, \ldots\}$ necessarily converge to $\mathscr{H}$? Extensive simulations initialized randomly (say with M_0 a random Wishart matrix) indeed suggest that, after a few iterates, the eigenvectors of M_k do converge to those of $\hat{\Sigma}$. This is, however, not everywhere true. Indeed, in the extreme scenario where $M_0 = \Sigma$, $0 = D(\Sigma, \Sigma) \simeq \hat{D}(\Sigma, X)$ which consistently estimates zero for large n, p (and irrespective of n/p), the gradient descent does not progress much from M_0 and is thus unlikely converging within $\mathscr{H}$ (which would mean the existence of a D such that $|\hat{D}(UDU^\mathsf{T}, X)| < |\hat{D}(\Sigma, X)|$). These aspects need to be clarified and well-understood. Indeed, in the absence of any a priori knowledge about the structure of Σ, such as sparseness or a factor model, several authors (Ledoit & Wolf, 2020; Karoui, 2008; Mestre, 2008a) argue that it is natural to only consider estimators of Σ that are rotation-equivariant (that belong to $\mathscr{H}$). Therefore, one could ask if the estimation scheme proposed does at least better than the rotation-equivariant estimators? If not we should gain more to restrict the estimation procedure to a non-linear shrinkage of the eigenvalues of the sample covariance matrix, that is, to first consistently estimate $D(\hat{U} D \hat{U}, \Sigma)$ for deterministic diagonal matrices D and perform a gradient descent step on D.

Unifying feature-based and parameter-based MTL approaches. MTL approaches are usually divided into parameter-based versus feature-based learning schemes. In the

parameter-based MTL approach, the tasks are assumed to share some parameters (e.g., the hyperplanes best separating each class) as discussed in chapter 4 of the present work. In the feature-based MTL approach, the tasks data are instead assumed to share a low-dimensional common representation. In this context, most of the works aim to determine a mapping of the ambient data space into a low-dimensional subspace (through sparse coding, deep neural networks, principal component analysis, etc.) in which the tasks have high similarity (Argyriou et al., 2007; Maurer et al., 2013; Zhang et al., 2016; Pan et al., 2010).

We should note that every year the literature on transfer learning evolves at an incredible speed with new methods that complement the previous long list of existing algorithms. In this context, the practitioner is often confused as to which method to consider. Due to the lack of theoretical understanding, most of these schemes are over-parametrized and difficult to tune for practitioners and the optimality of the algorithms is difficult to prove. The work carried out in Chapter 4 is part of this deeper understanding of the methods developed so far in order to identify their similarities, give the user-specific instructions on how to use them since the underlying idea is generally pertinent and should not necessarily be discarded.

Using Random Matrix theory, it will be interesting to completely address MTL learning scheme by studying feature-based methods in order to understand from a statistical point of view the difference between the two approaches, to assess the conditions where one approach is better than the other in order to provide theoretical guidelines for practitioners. In this context, the Principal Component Analysis (PCA), the Transfer Component Analysis (TCA)(Pan et al., 2010) and the sparse coding may be theoretically investigated using Random Matrix Theory. The ultimate goal is to fully control the dimensionality reduction-based transfer learning problem and to improve them by providing theoretical guidelines for the choice of some hyperparameters (number of the components to extract,...). The idea is to understand the interplay between the generally huge amount of hyperparameters and the sufficient statistics so that to make the method optimal with respect to the theoretic information bounds and at the same time decrease the computational cost by discarding redundant hyperparameters. More concretely, any practitioner would be able to identify which algorithm of transfer learning fits its application based on its optimality or not with respect to a theoretical bound, its computational advantage and be able to use efficiently the chosen algorithm by a proper understanding of its inner working statistics and the guidelines for the hyperparameter tuning.

Task relatedness assumption. Furthermore, as pointed in the introduction, the relatedness assumption is central in the multi-task learning schemes. In chapter 4, we use a relatedness assumption on the separating hyperplane of the SVM. One interesting question could be to try to derive other relatedness assumptions mostly regarding the data. An idea could try to find a projection matrix that should be optimized so that the projected data for source tasks and target tasks have a low discrepancy in the projected space. Concretely, the approach should search for two matrices A and B such that

$f(A, B) = D(AX_1, BX_2)$ is as low as possible where $D(AX_1, BX_2)$ refers to a distance function between the data AX_1 and BX_2 with X_1 and X_2 the samples of task 1 and 2. Subsequent learning schemes can then be applied on the optimal low dimensional space. In this setting, one would rather be interested in the theoretical analysis of traditional learning schemes (SVM, Empirical risk minimizers, ...) in the aforementioned optimal subspace and what would be the impact of the dimension of the subspace. To that end, it is of particular interest for tractability to choose a convex function $f(A, B)$. Further constraints on the structure of the matrix A and B can be added to the optimization problem depending on the context.

Towards optimal real-world multi-class algorithms. The improvements performed in Chapter 4 were mainly concerned with the binary classification and the optimality of the algorithm can only be proved for binary classification. This is due to the fact that the support vector machine algorithm was introduced to solve binary classification problems. The treatment of multi-class classification is always tricky and potentially sub-optimal. Generally, we rely on splitting the multi-class classification into several binary classifiers or relying on a one-hot encoding approach. The results obtained by the different extensions as provided in chapter 4 are generally different. Therefore, several questions naturally arise. Are we making optimal use of the improvement achieved in the binary case when building multi-classifiers? To answer this question, one would have to derive theoretical bounds on multi-class classification, compare different multi-class learning algorithms, and choose/design the optimal algorithm or at least quantify the sub-optimality of each multi-class extension. Some attempts to compare existing schemes (one-versus-all, one-versus-one, one-shot coding) have been proposed in the literature, but they are mainly based on experimental studies (see for example (Rifkin & Klautau, 2004)). This discussion stems from the overriding importance of multi-class classification in real-world applications compared to binary classification and especially because multi-class classification is by far more difficult than the binary case.

On the other hand, in some real-world problems, data are usually missing and contain outliers that need to be handled properly. More concretely, suppose we have missing data in the dataset inputs and we want to evaluate the capability of the imputation strategies to recover the classes with the least amount of defects.

A possible elementary model to study is the kernel matrix

$$\frac{1}{p}(X \odot S + A)^{\mathsf{T}}(X \odot S + A)$$

with $X = [x_1, \ldots, x_n]$, S the missing values matrix, and A the imputation matrix (with non-zero values a priori on the same locations as those of S.

An interesting study could try to access the performance of multi-task learning schemes under this model and to choose potentially the optimal imputation strategy. More concretely assuming that S has a simple statistical model (e.g. with Bernoulli i.i.d. inputs), what are the performances (in terms of classification error) obtained by specific simple imputation models for A (e.g., $A = \alpha(\mathbb{1}\mathbb{1}^{\mathsf{T}} - S)$ is constant over the set of non-zero

inputs of S, or A_{ij} is chosen as an average of the non-zero inputs of $X_{i,\dots}$, etc). A similar study has been performed in (Seddik et al., 2020) in the single task to characterize the performance in terms of phase transition and can be inspiring for the proposed study.

We conclude this part by pointing out that class imbalance problems have drawn growing interest recently because of their classification difficulty. Class imbalance learning refers to a type of classification problem, where some classes are highly underrepresented compared to other classes. The skewed distribution makes many conventional machine learning algorithms less effective, especially in predicting minority class examples. The learning objective can be generally described as obtaining a classifier that will provide high accuracy for the minority class without severely impacting the accuracy of the majority class. In this case, the least-square loss is generally known to perform poorly. Instead, several adaptations have been proposed ((Tang et al., 2008; Zheng et al., 2015)) using classical asymptotic heuristics. Random Matrix Theory analysis can be performed to take this important case into account.

The thesis and the perspectives proposed below are part of an urgent need to adapt to the new challenges of the modern world. We propose in the following section a modern vision of machine learning that we believe is possible and of which this thesis could be one of the founding ideas.

5.2 Towards an efficient, low-cost, controllable and Green AI

The traditional vision in machine learning. The classical methodology in machine learning and signal processing is to consider that the number of samples is very large compared to the dimension of the samples. As we have shown in this thesis, this induces many biases that lead to dramatic and unexpected behavior in the algorithms. The traditional explanation for such a failure is often and systematically attributed to the small number of samples. Therefore, the natural solution that follows from this analysis is to increase the number of samples and the model parameters (generally making the model more complex and "black-boxed"), as is generally the case in methods such as neural networks, ensemble learning, etc. The consequences of such approaches can be dramatic in the sense that the price to reach a given performance is excessively high (lots of computational units, lots of data with the carbon footprint, and the dramatic environmental consequences that this induces).

Towards a modern vision of machine learning. For the sake of efficiency, it would make more sense to (i) determine the best performance that would be induced by the a priori knowledge of the problem and (ii) determine algorithms whose performance is predictable and that are demonstrably optimal with respect to this bound. This approach to the problem ensures that the resources are used sparingly, efficiently, and more interestingly that we deeply understand the model parameter behind the problem so as to connect the intrinsic physical model of the problem with the model learned by the algorithm. More concretely, this would allow us to determine the minimum resources

needed to achieve a given performance and thus build low-cost algorithms that achieve this bound or at least quantify how far the proposed algorithm is from the optimal one. Indeed, the knowledge of the theoretical bound allows to know in addition the gap between the proposed algorithm and the optimal one, allowing to evaluate the sub-optimality of the designed algorithm versus the gain induced by the reduced computational cost. This reflection is in line with the increasingly urgent need to use data/resources smartly in our daily life.

What can Random Matrix Theory do? Random matrix Theory tools are part of this logic since they allow to analyze and control machine learning algorithms. Although the performance is determined asymptotically, the analyses remain true even for settings that were traditionally considered to be in the classical regime and for finite dimensions. This thesis proposes through random matrix theory a deeper understanding of the functionals of covariance matrices by exploiting randomness both in the sample size and in the dimension size. The improved scheme has important environmental consequences. To illustrate that, let's assume we want to estimate the Fisher distance between two covariance matrices Σ_1 and Σ_2 denoted $D_F(\Sigma_1, \Sigma_2)$ with $[\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$ based on n_a samples $x_i^{(a)} \sim \mathcal{N}(0, \Sigma_a)$ for $p = 64$ and $a \in \{1, 2\}$. Using $n_1 = n_2 = 128$ samples, a relative error of $\sim 5\%$ could be achieved using the proposed scheme while with $100\times$ samples more ($n_1 = n_2 = 12,800$), only $\sim 6\%$ relative error can be achieved using the classical asymptotic. This example illustrates that to achieve the same relative error, one needs to collect more than $100\times$ more samples with the carbon footprint that this induces. Even though experimental, this example illustrates how Random matrix Theory can be part of this change in the field of artificial intelligence to meet the new challenges of the world.

Regarding the Multi-Task Learning Least Square Support Vector Machine analyzed in chapter 4, the classical training cost involved by setting the hyperparameters λ and γ_i's which traditionally relies on cross-validation is time and data consuming since it involves making a grid search over ~ 1000 hyperparameter candidates. Moreover, to evaluate the pertinence of the different candidates, one needs furthermore to dedicate a specific number of training samples for the validation. This induces training at a small number of samples and therefore achieves low performance. The proposed learning approach instead relies on an explicit expression of the optimal labels which, when optimized, lead to a low sensitivity of the hyperparameters as illustrated in chapter 4. Due to the aforementioned reasons, the carbon footprint of the proposed method is likely lower than the one of the classical method. Furthermore, as will be discussed next, the proposed algorithm has been proven in a preliminary work to be optimal with respect to information-theoretic bound.

The new challenges. Based on this strategy of designing theoretical bounds given a problem, it would be interesting to use Bayesian and Information theory tools to derive theoretical bounds that any algorithm can achieve given the a priori knowledge of the problem (sparsity, feature-based covariance model, etc). This study, already conducted

for learning domains such as semi-supervised learning and specific model for the data (Gaussian mixture model) (Lelarge & Miolane, 2019) which paradoxically did not bring much attention, could be applied to transfer learning and compared to the algorithm proposed in this thesis. This work has already been preliminarily tackled (on-going research) and has shown promising results in the sense that it is possible to prove that the optimized method proposed in Chapter 4 of this thesis is close to optimal with respect to this bound. In the same vein of idea, regarding the covariance feature-based algorithm, an optimal bound on the classification algorithm could be designed and compared with the proposed algorithm in Chapter 3. On the other hand, Semi-supervised multi-task learning which has the advantage of using not only unlabeled data (which are easier to obtain than labeled data) but also data belonging to other tasks is another promising research direction. Similar to the information-theoretic bounds provided in (Lelarge & Miolane, 2019) for semi-supervised single-task learning, theoretical bounds of multi-task semi-supervised learning can be derived using statistical physics tools. Furthermore, using the preliminary results of (Mai & Liao, 2019), a graph-based algorithm can be derived which reaches this bound.

However, the information-theoretic bounds using information theory and the algorithms derived using the random matrix theory do not take for now into account the structure of the data (images, text, time series, sparse data, etc). A very interesting approach would be to incorporate the a priori (sparse data, images, text) in the knowledge for deriving the theoretical bounds and design data-dependent algorithms. This will prevent the use of features generally extracted by hand (SIFT features, neural networks, etc). Practically, it would be possible to design a theoretical bound given that the means of the data are sparse (as for MNIST data) or that the covariance matrix presents a specific structure (sparse inverse as in the case of brain applications(Cai et al., 2018), etc). The introduction of such a priori will also allow reaching higher performances at lower costs. This study will however require relevant modeling of the data. Several models can be used as starting point as (Peyré, 2009) for a manifold model for images, (Gerber et al., 2010) for a manifold of brain population analysis, (Cross & Jain, 1983) for a random Markov field model for images, (Dong et al., 2009) for a hidden Markov model for time series, etc. But the main technical difficulty will be the theoretical tractability of the introduction of such models into the information-theoretic analysis.

On the other hand, in this learning perspective at lower cost, an idea would be, given a ML question, to determine the minimal information to keep in order to reach a certain level of performance and to derive the optimal feature selection to remove the redundant information through "sparsification", multiplication by random matrices, etc as preliminary proposed in recent works (Couillet et al., 2021; Dall'Amico et al., 2021, 2020). These ideas can lead to a modern vision for machine learning integrating the new economical and environmental challenges.

Chapter 6

Appendix

Contents

6.1 Appendix for Chapter 2

Definition 3 (Mixture of Concentrated random vector(Louart & Couillet, 2018)). *Let $X = [x_1, \ldots, x_n] \in \mathscr{M}_{p,n}$ be a data matrix which is constituted of n random vectors distributed on k different classes $\mathscr{C}_1, \ldots, \mathscr{C}_k$ such that the data classes are characterized by the moments, for $x_i \in \mathscr{C}_\ell$*

$$\mathbb{E}[x_i] = \mu_\ell, \quad \mathbb{E}[x_i x_i^\mathsf{T}] = \Sigma_\ell + \mu_\ell \mu_\ell^\mathsf{T}$$

In particular the data matrix X satisfy a concentration assumption in the sense that for any 1-Lipschitz function $f : \mathscr{M}_{p,n} \to \mathbb{R}$ with $\mathscr{M}_{p,n}$ enrolled by the Frobenius norm $\|\|_F$, for $q > 0$, there exists $C, \sigma > 0$ independent of p and n such that

$$\forall t > 0, \quad \mathbb{P}\left(\|f(X) - \mathbb{E}[f(X)]\| \geq t\right) \leq Ce^{-(t/\sigma)^q}$$

6.1.1 Proof of corollary 1

In this section, we provide the proof of Corollary 1 which finds the deterministic equivalents of $Q(z)$, $\tilde{Q}(z)$ and $\tilde{Q}(z)^2$ as a consequence of Theorem 5 in the particular case of $C = I_p + \mu\mu^\mathsf{T}$.

We recall from Theorem 5 that the resolvent of the generalized sample covariance matrix defined as $Q(z) = \left(\frac{1}{n}XX^\mathsf{T} - zI_p\right)^{-1}$ admits a deterministic equivalent $\bar{Q}(z)$ given by

$$Q(z) \leftrightarrow \bar{Q}(z) = \left(\frac{C}{1+\delta(z)} - zI_p\right)^{-1} \tag{6.1}$$

where $\delta(z)$ is the unique solution to the fixed point equation defined as

$$\delta(z) = \frac{1}{n}\mathrm{tr}\left(C\left(\frac{C}{1+\delta(z)} - zI_p\right)^{-1}\right). \tag{6.2}$$

Plugging in $C = \Sigma + \mu\mu^\mathsf{T}$ in Equation (6.1) and applying Sherman Morrison identity matrix i.e., $(A+uu^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}uu^\mathsf{T}A^{-1}}{1+u^\mathsf{T}A^{-1}u}$ for any invertible matrix A and any vector u, we obtain

$$\bar{Q}(z) = m(z)I_p - \frac{m(z)^2}{1+\delta(z)+m(z)\|\mu\|^2}$$

where $m(z) \equiv \left(\frac{1}{1+\delta(z)} - z\right)^{-1}$. Further, in order to find an explicit expression of $\delta(z)$, we rely on the following rank-1 perturbation lemma for the resolvent of a matrix M.

Lemma 1 (Perturbation lemma (Silverstein & Bai, 1995))**.** *Let $A, M \in \mathbb{R}^{p\times p}$ some symmetric and non negative definite matrices, $u \in \mathbb{R}^p$, $\lambda > 0$ and $z < 0$, then,*

$$|\mathrm{tr}A\left(M + \lambda uu^\mathsf{T} - zI_p\right)^{-1} - \mathrm{tr}A\left(M - zI_p\right)^{-1}| \leq \frac{\|A\|}{|z|}.$$

Note that the bound in Lemma 1 does not depend on $\|u\|$. In particular denoting

$$\delta(z) = \frac{1}{n}\mathrm{tr}\left((I_p + \mu\mu^\mathsf{T})\left(\frac{I_p+\mu\mu^\mathsf{T}}{1+\delta(z)} - zI_p\right)^{-1}\right), \quad \delta'(z) = \frac{1}{n}\mathrm{tr}\left(\frac{1}{1+\delta(z)}I_p - zI_p\right)^{-1},$$

we obtain $\forall z < 0$,

$$\delta(z) = \delta'(z) + \mathcal{O}(n^{-1}).$$

Therefore as $p, n \to \infty$ with $p/n \equiv c_0 \to c_0^\infty$, particularizing to $C = I_p + \mu\mu^\mathsf{T}$ (or equivalently to $C = I_p$ following Lemma 1), Equation (6.2) has an explicit expression uniquely defined (using the definition of the Stieltjes transform) as:

$$\delta(z) = \frac{\sqrt{(c_0+z-1)^2 - 4c_0 z} - (c_0+z-1)}{2z}$$

The expression of $m(z)$ is retrieved by using the relation $m(z) = \left(\frac{1}{1+\delta(z)} - z\right)^{-1}$. We conclude by remarking that $\delta(z) = c_0 m(z)$.

Similarly as for $Q(z)$, the deterministic equivalent of $\tilde{Q}(z)$ is retrieved from Theorem 5 for $C = I_n + \|\mu\|^2 yy^\mathsf{T}$. Finally, the deterministic equivalent of $\tilde{Q}(z)^2$ follows by remarking that $\tilde{Q}(z)^2 = \frac{\partial}{\partial z}\tilde{Q}(z)$.

6.2 Appendix for Chapter 3

We provide here the technical developments for the proof of Theorem 7 as well as all subsequent corollaries (Corollaries 2–5) for the family of metrics expressed as linear spectral statistic of $\Sigma_1^{-1}\Sigma_2$ for simplicity.

The appendix is structured as follows: Appendix 6.2.1 provides the proof of Theorem 7 following the same approach as in (Couillet et al., 2011), relying mostly on the results from (Silverstein & Bai, 1995; Silverstein & Choi, 1995). Appendix 6.2.2 discusses in detail the question of the position of the complex contours when affected by change of variables. Appendix 6.2.3 then provides the technical details of the calculi behind Corollaries 2–5; this is undertaken through a first thorough characterization of the singular points of φ_p and ψ_p and functionals of these (these singular points are hereafter denoted $\hat{\lambda}_i$, η_i, ζ_i and κ_i), allowing for a proper selection of the integration contour, and subsequently through a detailed calculus for all functions $f(t)$ under study. Appendix 6.2.4 provides the details framework for the covariance matrix estimation as well as with the technical arguments.

6.2.1 Integral Form

Relating m_ν to m_μ

We start by noticing that we may equivalently assume the following setting:

- $x_1^{(1)}, \ldots, x_{n_1}^{(1)} \in \mathbb{R}^p$ vectors of i.i.d. zero-mean and unit variance entries
- $x_1^{(2)}, \ldots, x_{n_2}^{(2)} \in \mathbb{R}^p$ of the form $x_i^{(2)} = \Sigma^{\frac{1}{2}}\tilde{x}_i^{(2)}$ with $\tilde{x}_i^{(2)} \in \mathbb{R}^p$ a vector of i.i.d. zero-mean and unit variance entries

where $\Sigma \equiv \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$.

Indeed, with our first notations, $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2 = \frac{1}{n_1}\Sigma_1^{-\frac{1}{2}}\tilde{X}_1\tilde{X}_1^\mathsf{T}\Sigma_1^{-\frac{1}{2}}\frac{1}{n_2}\Sigma_2^{\frac{1}{2}}\tilde{X}_2\tilde{X}_2^\mathsf{T}\Sigma_2^{\frac{1}{2}}$ (here $\tilde{X}_a = [\tilde{x}_1^{(a)}, \ldots, \tilde{x}_{n_a}^{(a)}]$), the eigenvalue distribution of which is the same as that of the matrix $(\frac{1}{n_1}\tilde{X}_1\tilde{X}_1^\mathsf{T})(\frac{1}{n_2}\Sigma_1^{-\frac{1}{2}}\Sigma_2^{\frac{1}{2}}\tilde{X}_2\tilde{X}_2^\mathsf{T}\Sigma_2^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}})$ and we may then consider that the $x_i^{(1)}$'s actually have covariance I_p, while the $x_i^{(2)}$'s have covariance $\Sigma = \Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$, without altering the spectra under study. With these new definitions, we first condition with respect to the $x_i^{(2)}$'s, and study the spectrum of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$, which is the same as that of $\hat{\Sigma}_2^{\frac{1}{2}}\hat{\Sigma}_1^{-1}\hat{\Sigma}_2^{\frac{1}{2}}$. A useful remark is the fact that $\hat{\Sigma}_2^{\frac{1}{2}}\hat{\Sigma}_1^{-1}\hat{\Sigma}_2^{\frac{1}{2}}$ is the "inverse spectrum" of

$\hat{\Sigma}_2^{-\frac{1}{2}}\hat{\Sigma}_1\hat{\Sigma}_2^{-\frac{1}{2}}$, which is itself the same spectrum as that of $\frac{1}{n_1}X_1^\mathsf{T}\hat{\Sigma}_2^{-1}X_1$ except for $n_1 - p$ additional zero eigenvalues.

Denoting $\tilde{\mu}_p^{-1}$ the eigenvalue distribution of $\frac{1}{n_1}\tilde{X}_1^\mathsf{T}\hat{\Sigma}_2^{-1}\tilde{X}_1$, we first know from (Silverstein & Bai, 1995) that, under Assumption 1, as $p \to \infty$, $\tilde{\mu}_p^{-1} \xrightarrow{\text{a.s.}} \tilde{\mu}^{-1}$, where $\tilde{\mu}^{-1}$ is the probability measure with Stieltjes transform $m_{\tilde{\mu}^{-1}}$ defined as the unique (analytical function) solution to

$$m_{\tilde{\mu}^{-1}}(z) = \left(-z + c_1^\infty \int \frac{td\xi_2^{-1}(t)}{1 + tm_{\tilde{\mu}^{-1}}(z)}\right)^{-1}$$

with ξ_2 the almost sure limiting spectrum distribution of $\hat{\Sigma}_2$ and m_{ξ_2} its associated Stieltjes transform (note importantly that, from (Silverstein & Bai, 1995) and Assumption 1, ξ_2 has bounded support and is away from zero). Recognizing a Stieltjes transform from the right-hand side integral, this can be equivalently written

$$m_{\tilde{\mu}^{-1}}(z) = \left(-z + \frac{c_1^\infty}{m_{\tilde{\mu}^{-1}}(z)} - \frac{c_1^\infty}{m_{\tilde{\mu}^{-1}}(z)^2} m_{\xi_2^{-1}}\left(-\frac{1}{m_{\tilde{\mu}^{-1}}(z)}\right)\right)^{-1}. \tag{6.3}$$

Accounting for the aforementioned additional zero eigenvalues, $\tilde{\mu}^{-1}$ is related to μ^{-1}, the almost sure limiting spectrum distribution of $\hat{\Sigma}_2^{-1/2}\hat{\Sigma}_1\hat{\Sigma}_2^{-1/2}$, through the relation $\tilde{\mu}^{-1} = c_1^\infty\mu^{-1} + (1 - c_1^\infty)\delta_0$ with δ_x the Dirac measure at x and we have

$$m_{\tilde{\mu}^{-1}}(z) = c_1^\infty m_{\mu^{-1}}(z) - (1 - c_1^\infty)\frac{1}{z}.$$

Plugging this last relation in (6.3) leads then to

$$m_{\xi_2^{-1}}\left(\frac{z}{1 - c_1^\infty - c_1^\infty z m_{\mu^{-1}}(z)}\right) = m_{\mu^{-1}}(z)\left(1 - c_1^\infty - c_1^\infty z m_{\mu^{-1}}(z)\right). \tag{6.4}$$

Now, with the convention that, for a probability measure θ, θ^{-1} is the measure defined through $\theta^{-1}([a,b]) = \theta\left([\frac{1}{a}, \frac{1}{b}]\right)$, we have the Stieltjes transform relation

$$m_{\theta^{-1}}(z) = -\frac{1}{z} - \frac{1}{z^2} m_\theta\left(\frac{1}{z}\right).$$

Using this relation in (6.3), we then deduce

$$\begin{aligned} zm_\mu(z) &= (z + c_1^\infty z^2 m_\mu(z)) m_{\xi_2}(z + c_1^\infty z^2 m_\mu(z)) \\ &= \varphi(z) m_{\xi_2}(\varphi(z)) \end{aligned} \tag{6.5}$$

where we recall that $\varphi(z) = z(1 + c_1^\infty z m_\mu(z))$. It will come in handy in the following to differentiate this expression along z to obtain

$$m'_{\xi_2}(\varphi(z)) = \frac{1}{\varphi(z)}\left(\frac{m_\mu(z) + zm'_\mu(z)}{\varphi'(z)} - m_{\xi_2}(\varphi(z))\right)$$

which might be conveniently rewritten as

$$m'_{\xi_2}(\varphi(z)) = \frac{1}{\varphi(z)} \left(-\frac{\psi'(z)}{c_2^\infty \varphi'(z)} - m_{\xi_2}(\varphi(z)) \right). \tag{6.6}$$

We next determine m_{ξ_2} as a function of ν. Since $\hat{\Sigma}_2$ is itself a sample covariance matrix, we may apply again the results from (Silverstein & Bai, 1995). Denoting $\tilde{\xi}_2$ the almost sure limiting spectrum distribution of $\frac{1}{n_2}\tilde{X}_2^\mathsf{T}\Sigma\tilde{X}_2$, we first have

$$m_{\tilde{\xi}_2}(z) = \left(-z + c_2^\infty \int \frac{t d\nu(t)}{1 + t m_{\tilde{\xi}_2}(z)} \right)^{-1}. \tag{6.7}$$

Similar to previously, we have the Stieltjes transform relation $m_{\tilde{\xi}_2}(z) = c_2^\infty m_{\xi_2}(z) - \frac{(1-c_2^\infty)}{z}$ which yields, when plugged in (6.7)

$$m_\nu \left(-\frac{z}{c_2^\infty z m_{\xi_2}(z) - (1-c_2^\infty)} \right) = -m_{\xi_2}(z) \left(c_2^\infty z m_{\xi_2}(z) - (1-c_2^\infty) \right). \tag{6.8}$$

The two relations (6.5) and (6.7) will be instrumental to relating $\int f d\nu$ to the observation measure μ_p, as described in the next section.

Remark 11 (The case $c_2^\infty > 1$)**.** *The aforementioned reasoning carries over to the case $c_2^\infty > 1$. Indeed, since the equation* (6.3) *is now meaningless (as the support of ξ_2 contains the atom $\{0\}$), consider the model $\hat{\Sigma}_1^{-1}(\hat{\Sigma}_2 + \varepsilon I_p) = \hat{\Sigma}_1^{-1}\hat{\Sigma}_2 + \varepsilon\hat{\Sigma}_1^{-1}$ for some small $\varepsilon > 0$. Then* (6.5) *holds with now ξ_2 the limiting empirical spectral distribution of $\hat{\Sigma}_2 + \varepsilon I_p$. Due to ε, Equation* (6.7) *now holds with $m_{\tilde{\xi}_2}(z)$ replaced by $m_{\tilde{\xi}_2}(z+\varepsilon)$. By continuity in the small ε limit, we then have that* (6.5) *and* (6.8) *still hold in the small ε limit. Now, since $\hat{\Sigma}_1^{-1}(\hat{\Sigma}_2 + \varepsilon I_p) - \hat{\Sigma}_1^{-1}\hat{\Sigma}_2 = \varepsilon\hat{\Sigma}_1^{-1}$, the operator norm of which almost surely vanishes as $\varepsilon \to 0$ (as per the almost sure boundedness of $\limsup_p \|\hat{\Sigma}_1^{-1}\|$), we deduce that $\mu_p \to \mu$ defined through* (6.5) *and* (6.8)*, almost surely, also for $c_2^\infty > 1$.*

Integral formulation over m_ν

With the formulas above, we are now in position to derive the proposed estimator. We start by using Cauchy's integral formula to obtain

$$\int f d\nu = -\frac{1}{2\pi\imath} \oint_{\Gamma_\nu} f(z) m_\nu(z) dz$$

for Γ_ν a complex contour surrounding the support of ν but containing no singularity of f in its inside. This contour is carefully chosen as the image of the mapping $\omega \mapsto z = -\omega/(c_2^\infty \omega m_{\xi_2}(\omega) - (1-c_2^\infty))$ of another contour Γ_{ξ_2} surrounding the limiting support of ξ_2; the details of this (non-trivial) contour change are provided in Appendix 6.2.2 (where it is seen that the assumption $c_2^\infty < 1$ is crucially exploited). We shall admit here that this change of variable is licit.

Operating the aforementioned change of variable gives

$$\int f d\nu = \frac{1}{2\pi\imath} \oint_{\Gamma_{\xi_2}} \frac{f\left(\frac{-\omega}{c_2^\infty \omega m_{\xi_2}(\omega) - (1-c_2^\infty)}\right) m_{\xi_2}(\omega) \left(c_2^\infty \omega^2 m'_{\xi_2}(\omega) + (1-c_2^\infty)\right)}{c_2^\infty \omega m_{\xi_2}(\omega) - (1-c_2^\infty)} d\omega \quad (6.9)$$

where we used (6.8) to eliminate m_ν.

To now eliminate m_{ξ_2} and obtain an integral form only as a function of m_μ, we next proceed to the variable change $u \mapsto \omega = \varphi(u) = u + c_1^\infty u^2 m_\mu(u)$. Again, this involves a change of contour, which is valid as long as Γ_{ξ_2} is the image by φ of a contour Γ_μ surrounding the support of μ, which is only possible if $c_1^\infty < 1$ (see Appendix 6.2.2 for further details). With this variable change, we can now exploit the relations (6.5) and (6.6) to obtain, after basic algebraic calculus (using in particular the relation $um_\mu(u) = (-\psi(u) + 1 - c_2^\infty)/c_2^\infty$)

$$\int f d\nu = \frac{1}{2\pi\imath} \oint_{\Gamma_\mu} f\left(\frac{\varphi(u)}{\psi(u)}\right) \frac{\psi(u)}{c_2^\infty} \left[\frac{\varphi'(u)}{\varphi(u)} - \frac{\psi'(u)}{\psi(u)}\right] du - \frac{1-c_2^\infty}{c_2^\infty} \frac{1}{2\pi\imath} \oint_{\Gamma_\mu} f\left(\frac{\varphi(u)}{\psi(u)}\right) \left[\frac{\varphi'(u)}{\varphi(u)} - \frac{\psi'(u)}{\psi(u)}\right] du.$$

Performing the variable change backwards ($u \mapsto z = \frac{\varphi(u)}{\psi(u)}$), the rightmost term is

$$\frac{1}{2\pi\imath} \oint_{\Gamma_\mu} f\left(\frac{\varphi(u)}{\psi(u)}\right) \left[\frac{\varphi'(u)}{\varphi(u)} - \frac{\psi'(u)}{\psi(u)}\right] du = \frac{1}{2\pi\imath} \oint_{\Gamma_\mu} f\left(\frac{\varphi(u)}{\psi(u)}\right) \frac{\psi(u)}{\varphi(u)} \left(\frac{\varphi(u)}{\psi(u)}\right)' du = \frac{1}{2\pi\imath} \oint_{\Gamma_\nu} \frac{f(z)}{z} dz = 0,$$

since $\Gamma_\nu \subset \{z \in \mathbb{C}, \mathscr{R}[z] > 0\}$. Note that if Γ_ν were to contain 0 (which occurs when $c_2^\infty > 1$), then an additional residue equal to $-\frac{1-c_2^\infty}{c_2^\infty} f(0)$ would have to be accounted for. We conclude that

$$\int f d\nu = \frac{1}{2\pi\imath} \oint_{\Gamma_\mu} f\left(\frac{\varphi(u)}{\psi(u)}\right) \frac{\psi(u)}{c_2^\infty} \left[\frac{\varphi'(u)}{\varphi(u)} - \frac{\psi'(u)}{\psi(u)}\right] du.$$

To ensure that m_μ can be replaced by m_{μ_p} in the the above expression, one however needs to ensure that dominated convergence on the compact set Γ_μ holds. For this, two ingredients are needed: (i) First we need to guarantee that the convergence $m_{\mu_p} \xrightarrow{\text{a.s.}} m_\mu$ is uniform on Γ_μ, which easily follows from the analytic nature of Stieltjes transforms, and most importantly (ii) prove that $f\left(\frac{\varphi_p(u)}{\psi_p(u)}\right) \frac{\psi_p(u)}{c_2} \left[\frac{\varphi_p'(u)}{\varphi_p(u)} - \frac{\psi_p'(u)}{\psi_p(u)}\right]$ is uniformly bounded on Γ_μ. This second item follows from (Bai & Silverstein, 1998b) which prove that the eigenvalues of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$ are asymptotically assembled in contiguous bulks and almost surely do not escape the limiting support μ as $p \to \infty$ under the condition that the fourth moment of the entries of the matrices $\tilde{X}_1$ and $\tilde{X}_2$ is finite and that the limiting spectrum ν of $\Sigma_1^{-1}\Sigma_2$ has bounded support away from zero (due to the assumption $\limsup_p \max\{\|\Sigma_a^{-1}\|, \|\Sigma_a\|\} < \infty$) along with the analyticity of the involved functions inside the contour. This allows to retrieve Theorem 7 by uniform convergence on the compact contour (see also (Couillet et al., 2011) for a similar detailed derivation).

Remark 12 (Case Σ_1 known)**.** *The case where Σ_1 is known is equivalent to setting $c_1 \to 0$ above, leading in particular to $m_\mu = m_{\xi_2}$ and to the unique functional equation*

$$m_\nu\left(\frac{z}{1-c_2^\infty - c_2^\infty z m_\mu(z)}\right) = m_\nu(z)\left(1 - c_2^\infty - c_2^\infty z m_\nu(z)\right).$$

In particular, if $\Sigma_1 = \Sigma_2$, this reduces to

$$1 = -m_\mu(z)(z - \psi(z))$$

with $\psi(z) = 1 - c_2^\infty - c_2^\infty z m_\mu(z)$, which is the functional Stieltjes-tranform equation of the popular Marčenko–Pastur law (Marčenko & Pastur, 1967).

6.2.2 Integration contour determination

This section details the complex integration steps sketched in Appendix 6.2.1. These details rely heavily on the works of (Silverstein & Choi, 1995) and follow similar ideas as in e.g., (Couillet et al., 2011).

Our objective is to ensure that the successive changes of variables involved in Appendix 6.2.1 move any complex contour closely encircling the support of μ onto a valid contour encircling the support of ν; we will in particular be careful that the resulting contour, in addition to encircling the support of ν, does not encircle additional values possibly bringing undesired residues (such as 0). We will proceed in two steps, first showing that a contour encircling μ results on a contour encircling ξ_2 and a contour encircling ξ_2 results on a contour encircling ν.

Let us consider a first contour Γ_{ξ_2} closely around the support of ξ_2 (in particular not containing 0). We have to prove that any point ω of this contour is mapped to a point of a contour Γ_ν closely around the support of ν.

The change of variable performed in (6.8) reads, for all $\omega \in \mathbb{C} \setminus \mathrm{Supp}(\xi_2)$,

$$z \equiv z(\omega) = \frac{-\omega}{-(1-c_2^\infty) + c_2^\infty \omega m_{\xi_2}(\omega)} = \frac{-1}{m_{\tilde{\xi}_2}(\omega)}$$

where we recall that $\tilde{\xi}_2 = c_2^\infty \xi_2 + (1-c_2^\infty)\delta_0$. Since $\Im[\omega]\Im[m_{\tilde{\xi}_2}(\omega)] > 0$ for $\Im[\omega] \neq 0$, we already have that $\Im[z]\Im[\omega] > 0$ for all non-real ω.

It therefore remains to show that real ω's (outside the support of ξ_2) project onto properly located real z's (i.e., on either side of the support of ν). This conclusion follows from the seminal work (Silverstein & Choi, 1995) on the spectral analysis of sample covariance matrices. The essential idea is to note that, due to (6.7), the relation $z(\omega) = -1/m_{\tilde{\xi}_2}(\omega)$ can be inverted as

$$\omega \equiv \omega(z) = -\frac{1}{m_{\tilde{\xi}_2}} + c_2^\infty \int \frac{t d\nu(t)}{1 + t m_{\tilde{\xi}_2}} = z + c_2^\infty \int \frac{t d\nu(t)}{1 - \frac{t}{z}}.$$

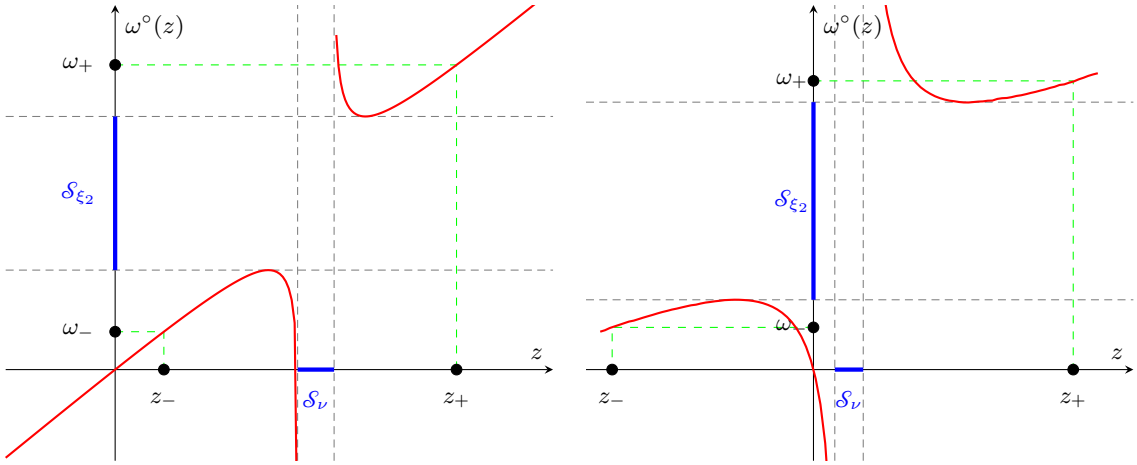

Figure 6.1: Variable change $z \mapsto \omega^\circ(z) = z + c_2^\infty \int \frac{z d\nu(t)}{z-t}$ for $c_2^\infty < 1$ (left) and $c_2^\infty > 1$ (right). $\mathcal{S}_\theta$ is the support of the probability measure θ. For $0 < \omega_- = \omega(z_-) < \inf \mathrm{Supp}(\nu)$, the pre-image z_- is necessarily negative for $c_2^\infty > 1$.

In (Silverstein & Choi, 1995), it is proved that the image by $\omega(\cdot)$ of $z(\mathbb{R} \setminus \mathrm{Supp}(\xi_2))$ coincides with the increasing sections of the function $\omega^\circ : \mathbb{R} \setminus \mathrm{Supp}(\nu) \to \mathbb{R}$, $z \mapsto \omega(z)$. The latter being an explicit function, its functional analysis is simple and allows in particular to properly locate the real pairs (ω, z). Details of this analysis are provided in (Silverstein & Choi, 1995) as well as in (Couillet et al., 2011), which shall not be recalled here. The function ω° is depicted in Figure 6.1; we observe and easily prove that, for $c_2^\infty < 1$, any two values $z_- < \inf(\mathrm{Supp}(\nu)) \leq \sup(\mathrm{Supp}(\nu)) < z_+$ have respectively images ω_- and ω_+ satisfying $w_- < \inf(\mathrm{Supp}(\xi_2)) \leq \sup(\mathrm{Supp}(\xi_2)) < w_+$ as desired. This is however not the case for $c_2^\infty > 1$ where $\{z_-, z_+\}$ enclose not only $\mathrm{Supp}(\nu)$ but also 0 and therefore do not bring a valid contour. This essentially follows from the fact that $(\varphi_p/\psi_p)'(0)$ is positive for $c_2^\infty < 1$ and negative for $c_2^\infty > 1$. Even though in Figure 6.1 the support μ is considered compact of one component, the analysis and conclusions remains true even for several disjoint supports.

The same reasoning now holds for the second variable change. Indeed, note that here

$$\omega = u(1 + c_1^\infty u m_\mu(u)) = u\left(1 - c_1^\infty - \frac{c_1^\infty}{u} m_{\mu^{-1}}\left(\frac{1}{u}\right)\right) = -m_{\tilde{\mu}^{-1}}\left(\frac{1}{u}\right).$$

Exploiting (6.3) provides, as above, a functional inverse given here by

$$u \equiv u(\omega) = \left(\frac{1}{\omega} + c_1^\infty \int \frac{d\xi_2(t)}{t - \omega}\right)^{-1}$$

the analysis of which follows the same arguments as above (see display in Figure 6.2 of the extension to $u^\circ(\omega) = u(\omega)$ for all $\omega \in \mathbb{R} \setminus \mathrm{Supp}(\xi_2)$).

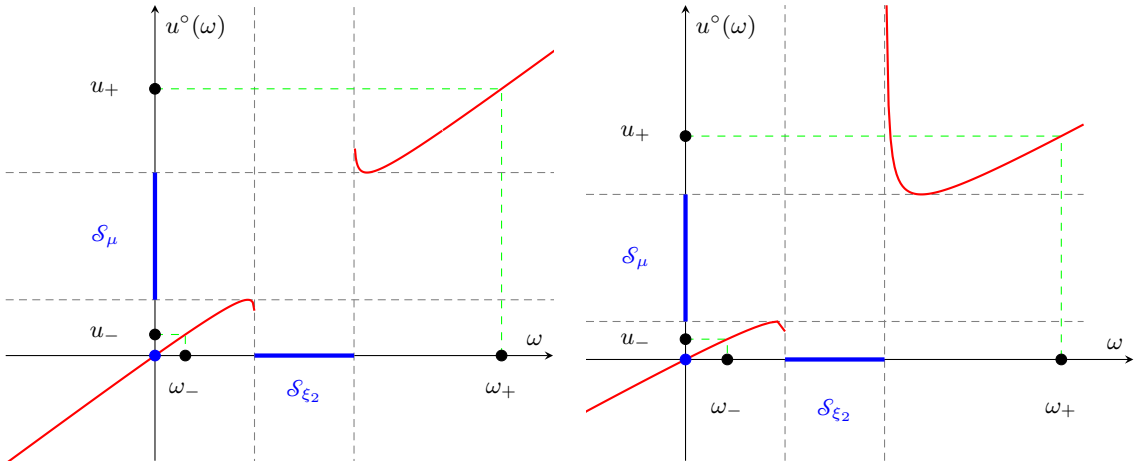

Figure 6.2: Variable change $u^\circ(\omega) = (\frac{1}{\omega} + c_1^\infty \int \frac{1}{t-\omega} d\xi_2(t))^{-1}$ for $c_2^\infty < 1$ (left) and $c_2^\infty > 1$ (right). $\mathcal{S}_\theta$ is the support of the probability measure θ.

6.2.3 Integral Calculus

To compute the complex integral in theorem 7, note first that, depending on f, several types of singularities in the integral may arise. Of utmost interest (but not always exhaustively, as we shall see for $f(t) = \log(1+st)$) are: (i) the eigenvalues $\hat{\lambda}_i$ of $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2$, (ii) the non-null values η_i such that $\varphi_p(\eta_i) = 0$, (iii) the values ζ_i such that $\psi_p(\zeta_i) = 0$.

In the following, we first introduce a sequence of intermediary results of interest for most of the integral calculi.

Rational expansion

At the core of the subsequent analysis is the function $\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}$. As this is a rational function, we first obtain the following important partial fraction decomposition, that will be repeatedly used in the sequel:

$$\begin{aligned}&\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2} \\ &= \left(\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\sum_{j=1}^{p}\frac{1}{z-\hat{\lambda}_j} + \frac{1-c_2}{c_2}\frac{1}{z} + \frac{c_1+c_2-c_1c_2}{c_1c_2}\sum_{j=1}^{p}\frac{1}{z-\eta_j}.\end{aligned} \tag{6.10}$$

This form is obtained by first observing that the $\hat{\lambda}_j$'s, η_j's and 0 are the poles of the left-hand side expression. Then, pre-multiplying the left-hand side by $(z-\hat{\lambda}_j)$, z, or $(z-\eta_j)$ and taking the limit when these terms vanish, we recover the right-hand side, using in particular the following estimates (which easily entail from the definitions of φ_p

and ψ_p):

$$\varphi_p(z) = \frac{c_1}{p}\frac{\hat{\lambda}_i^2}{\hat{\lambda}_i - z} - 2c_1\frac{\hat{\lambda}_i}{p} + \hat{\lambda}_i + \frac{c_1}{p}\sum_{j\neq i}\frac{\hat{\lambda}_i^2}{\hat{\lambda}_j - \hat{\lambda}_i} + O(\hat{\lambda}_i - z)$$

$$\varphi_p'(z) = \frac{c_1}{p}\frac{\hat{\lambda}_i^2}{(\hat{\lambda}_i - z)^2} + O(1)$$

$$\psi_p(z) = -\frac{c_2}{p}\frac{\hat{\lambda}_i}{\hat{\lambda}_i - z} + \frac{c_2}{p} + 1 - c_2 - \frac{c_2}{p}\sum_{j\neq i}\frac{\hat{\lambda}_i}{\hat{\lambda}_j - \hat{\lambda}_i} + O(\hat{\lambda}_i - z)$$

$$\psi_p'(z) = -\frac{c_2}{p}\frac{\hat{\lambda}_i}{(\hat{\lambda}_i - z)^2} + O(1)$$

in the vicinity of $\hat{\lambda}_i$, along with $\psi_p(\eta_i) = \frac{c_1+c_2-c_1c_2}{c_1}$ and $\psi_p(0) = 1 - c_2$.

To retrieve the constant term of the partial fraction decomposition, we further compute $\lim_{z\to\infty}\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}$ which is shown to be zero by using the expressions of $\varphi_p(z)$ and $\psi_p(z)$.

From this expression, we have the following immediate corollary.

Remark 13 (Residue for f analytic at $\hat{\lambda}_i$)**.** *If $f \circ (\varphi_p/\psi_p)$ is analytic in a neighborhood of $\hat{\lambda}_i$, i.e., if f is analytic in a neighborhood of $-(c_1/c_2)\hat{\lambda}_i$, then $\hat{\lambda}_i$ is a first-order pole for the integrand, leading to the residue*

$$\mathrm{Res}(\hat{\lambda}_i) = -f\left(-\frac{c_1}{c_2}\hat{\lambda}_i\right)\left[\frac{c_1 + c_2 - c_1c_2}{c_1c_2} - \frac{1}{p}\right].$$

Characterization of η_i and ζ_i, and φ_p/ψ_p

First note that the η_i (the zeros of $\varphi_p(z)$) and ζ_i (the zeros of $\psi_p(z)$) are all real as one can verify that, for $\Im[z] \neq 0$, $\Im[\varphi_p(z)]\Im[z] > 0$ and $\Im[\psi_p(z)]\Im[z] < 0$.

Before establishing the properties of φ_p and ψ_p in the vicinity of η_i and ζ_i, let us first locate these values. A study of the function $M_p : \mathbb{R} \to \mathbb{R}$, $x \mapsto xm_{\mu_p}(x)$ (see Figure 6.3) reveals that M_p is increasing (since $x/(\hat{\lambda}_i - x) = -1 + 1/(\hat{\lambda}_i - x)$) and has asymptotes at each $\hat{\lambda}_i$ with $\lim_{x\uparrow\hat{\lambda}_i} M_p(x) = \infty$ and $\lim_{x\downarrow\hat{\lambda}_i} M_p(x) = -\infty$. As a consequence, since $\varphi_p(x) = 0 \Leftrightarrow M_p(x) = -\frac{1}{c_1} < -1$, there exists exactly one solution to $\varphi_p(x) = 0$ in the set $(\hat{\lambda}_i, \hat{\lambda}_{i+1})$. This solution will be subsequently called η_i. Since $M_p(x) \to -1$ as $x \to \infty$, there exists a last solution to $\varphi_p(x) = 0$ in $(\hat{\lambda}_p, \infty)$, hereafter referred to as η_p. Similarly, $\psi_p(x) = 0 \Leftrightarrow M_p(x) = (1 - c_2)/c_2 > 0$ and thus there exists exactly one solution, called ζ_i in $(\hat{\lambda}_{i-1}, \hat{\lambda}_i)$. When $x \to 0$, $M_p(x) \to 0$ so that a further solution is found in $(0, \hat{\lambda}_1)$, called ζ_1. Besides, due to the asymptotes at every $\hat{\lambda}_i$, we have that $\zeta_1 < \hat{\lambda}_i < \eta_1 < \zeta_2 < \ldots < \eta_p$.

As such, the set Γ defined in Theorem 7 exactly encloses all η_i, $\hat{\lambda}_i$, and ζ_i, for $i = 1, \ldots, p$, possibly to the exception of the leftmost ζ_1 and the rightmost η_p (as

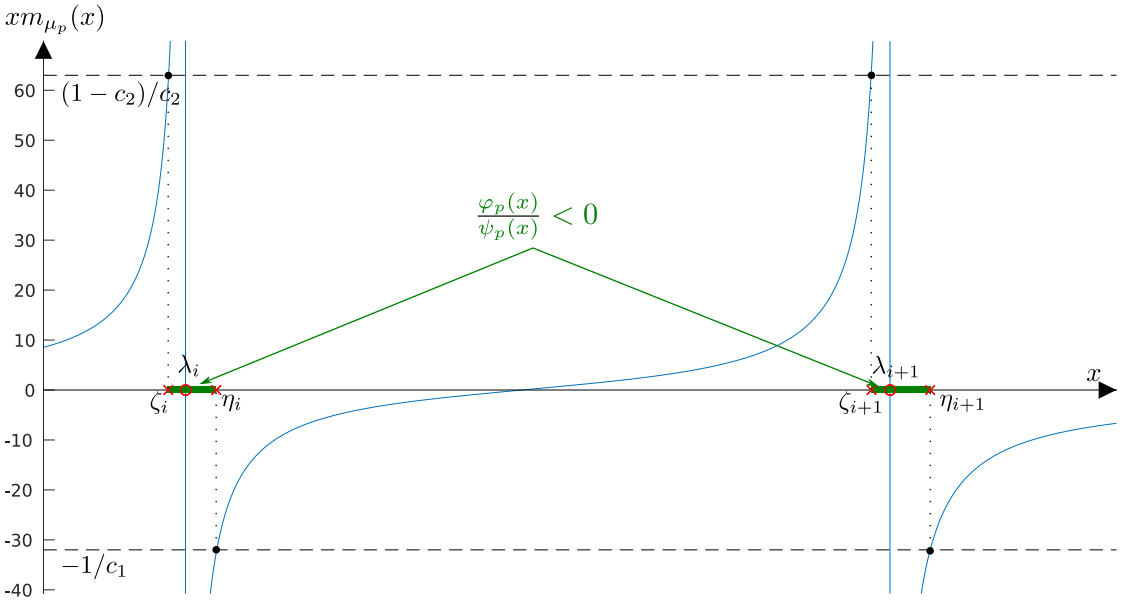

Figure 6.3: Visual representation of $x \mapsto M_p(x) = xm_{\mu_p}(x)$; here for $p = 4$, $n_1 = 8$, $n_2 = 16$. Solutions to $M_p(x) = -1/c_1$ (i.e., η_i's) and to $M_p(x) = (1-c_2)/c_2$ (i.e., ζ_i's) indicated in red crosses. Green solid lines indicate sets of negative φ_p/ψ_p.

those are not comprised in a set of the form $[\hat{\lambda}_{i+1}, \hat{\lambda}_i]$). To ensure that the latter do asymptotically fall within the interior of Γ, one approach is to exploit Theorem 7 for the elementary function $f(t) = 1$. There we find that

$$\frac{1}{2\pi\imath}\oint_{\Gamma_\nu} m_\nu(z)dz - \frac{1}{2\pi\imath}\oint_{\Gamma}\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz \xrightarrow{\text{a.s.}} 0.$$

The left integral is easily evaluated by residue calculus and equals -1 (each $\lambda_i(\Sigma_1^{-1}\Sigma_2)$, $1 \le i \le p$, is a pole with associated residue $-1/p$), while the right integral can be computed from (6.10) again by residue calculus and equals $-1 + \frac{c_1+c_2-c_1c_2}{c_1c_2}(p - \#\{\eta_i \in \Gamma^\circ\})$ with Γ° the "interior" of Γ. As such, since both integrals are (almost surely) arbitrarily close in the large p limit, we deduce that $\#\{\eta_i \in \Gamma^\circ\} = p$ for all large p and thus, in particular, η_p is found in the interior of Γ. To obtain the same result for ζ_1, note that, from the relation $\psi_p(z) = \frac{c_1+c_2-c_1c_2}{c_1} - \frac{c_2}{c_1}\frac{\varphi_p(z)}{z}$ along with the fact that $\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}$ is an exact derivative (of $\log(\varphi_p/\psi_p)$), the aforementioned convergence can be equivalently written

$$\frac{1}{2\pi\imath}\oint_{\Gamma_\nu} m_\nu(z)dz - \frac{-1}{2\pi\imath}\oint_{\Gamma}\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\varphi_p(z)}{zc_1}dz \xrightarrow{\text{a.s.}} 0.$$

Reproducing the same line of argument (with an expansion of $(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)})\frac{\varphi_p(z)}{zc_1}$ equivalent to (6.10)), the same conclusion arises and we then proved that both ζ_1

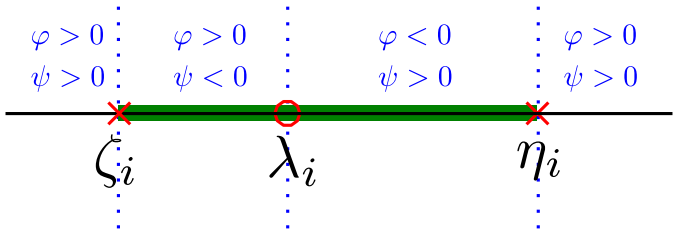

Figure 6.4: Visual representation of the signs of φ_p and ψ_p around singularities.

and η_p (along with all other ζ_i's and η_i's) are asymptotically found within the interior of Γ.

One can also establish that, on its restriction to $\mathbb{R}^+$, φ_p is everywhere positive but on the set $\cup_{i=1}^p(\hat{\lambda}_i, \eta_i)$. Similarly, ψ_p is everywhere positive but on the set $\cup_{i=1}^p(\zeta_i, \hat{\lambda}_i)$. As a consequence, the ratio φ_p/ψ_p is everywhere positive on $\mathbb{R}^+$ but on the set $\cup_{i=1}^p(\zeta_i, \eta_i)$.

These observations are synthesized in Figure 6.4.

In terms of monotonicity on their restrictions to the real axis, since $\psi_p(x) = 1 - c_2 \int \frac{t}{t-x} d\mu_p(t)$, ψ_p is decreasing. As for φ_p, note that

$$\begin{aligned}\varphi_p'(x) &= 1 + 2c_1 \int \frac{x}{t-x} d\mu_p(t) + c_1 \int \frac{x^2}{(t-x)^2} d\mu_p(t) \\ &= \int \frac{t^2 - 2(1-c_1)xt + (1-c_1)x^2}{(t-x)^2} d\mu_p(t).\end{aligned}$$

Since $c_1 < 1$, we have $1 - c_1 > (1 - c_1)^2$, and therefore

$$\varphi_p'(x) > \int \frac{(t - (1-c_1)x)^2}{(t-x)^2} d\mu_p(t) > 0$$

ensuring that φ_p is increasing on its restriction to $\mathbb{R}$.

Showing that $x \mapsto \varphi_p(x)/\psi_p(x)$ is increasing is important for the study of the case $f(t) = \log(1 + st)$ but is far less immediate. This unfolds from the following remark, also of key importance in the following.

Remark 14 (Alternative form of φ_p and ψ_p)**.** *It is interesting to note that, in addition to the zero found at $z = 0$ for φ_p, we have enumerated all zeros and poles of the rational functions φ_p and ψ_p (this can be ensured from their definition as rational functions) and it thus comes that*

$$\varphi_p(z) = (1 - c_1) z \frac{\prod_{j=1}^p (z - \eta_j)}{\prod_{j=1}^p (z - \hat{\lambda}_j)} \tag{6.11}$$

$$\psi_p(z) = \frac{\prod_{j=1}^p (z - \zeta_j)}{\prod_{j=1}^p (z - \hat{\lambda}_j)} \tag{6.12}$$

where the constants $1 - c_1$ *and* 1 *are found by observing that, as* $z = x \in \mathbb{R} \to \infty$, $\varphi_p(x)/x \to 1 - c_1$ *and* $\psi_p(x) \to 1$. *In particular*

$$\frac{\varphi_p(z)}{\psi_p(z)} = (1 - c_1) z \frac{\prod_{j=1}^p (z - \eta_j)}{\prod_{j=1}^p (z - \zeta_j)}. \tag{6.13}$$

A further useful observation is that the η_i's are the eigenvalues of

$$\Lambda - \frac{1}{p - n_1} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T}$$

where $\Lambda = \operatorname{diag}(\{\hat{\lambda}_i\}_{i=1}^p)$ and $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)^\mathsf{T}$. Indeed, these eigenvalues are found by solving

$$\begin{aligned}
0 &= \det \left(\Lambda - \frac{\sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T}}{p - n_1} - x I_p \right) \\
&= \det(\Lambda - x I_p) \det \left(I_p - (\Lambda - x I_p)^{-1} \frac{\sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T}}{p - n_1} \right) \\
&= \det(\Lambda - x I_p) \left(1 - \frac{1}{p - n_1} \sqrt{\hat{\lambda}}^\mathsf{T} (\Lambda - x I_p)^{-1} \sqrt{\hat{\lambda}} \right) \\
&= \det(\Lambda - x I_p) \left(1 - \frac{1}{p - n_1} \sum_{i=1}^p \frac{\hat{\lambda}_i}{\hat{\lambda}_i - x} \right)
\end{aligned}$$

which, for x away from the $\hat{\lambda}_i$ (not a solution to $\varphi_p(x) = 0$), reduces to $\frac{1}{p} \sum_{i=1}^p \frac{\hat{\lambda}_i}{\hat{\lambda}_i - x} = 1 - \frac{1}{c_1}$, which is exactly equivalent to $m_{\mu_p}(x) = -\frac{1}{c_1 x}$, i.e., $\varphi_p(x) = 0$.

Similarly, the ζ_i's are the eigenvalues of the matrix

$$\Lambda - \frac{1}{n_2} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T}.$$

These observations allow for the following useful characterization of φ_p/ψ_p:

$$\begin{aligned}
\frac{\varphi_p(z)}{\psi_p(z)} &= (1 - c_1) z \frac{\det \left(z I_p - \Lambda - \frac{1}{n_1 - p} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T} \right)}{\det \left(z I_p - \Lambda + \frac{1}{n_2} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T} \right)} \\
&= (1 - c_1) z \left(1 - \frac{n_1 + n_2 - p}{n_2 (n_1 - p)} \sqrt{\hat{\lambda}}^\mathsf{T} \left(z I_p - \Lambda + \frac{1}{n_2} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T} \right)^{-1} \sqrt{\hat{\lambda}} \right)
\end{aligned}$$

(after factoring out the matrix in denominator from the determinant in the numerator) the derivative of which is, after simplification,

$$\left(\frac{\varphi_p(z)}{\psi_p(z)} \right)' = (1 - c_1) \left(1 + \frac{n_1 + n_2 - p}{n_2 (n_1 - p)} \sqrt{\hat{\lambda}}^\mathsf{T} Q \left(\Lambda - \frac{1}{n_2} \sqrt{\hat{\lambda}} \sqrt{\hat{\lambda}}^\mathsf{T} \right) Q \sqrt{\hat{\lambda}} \right).$$

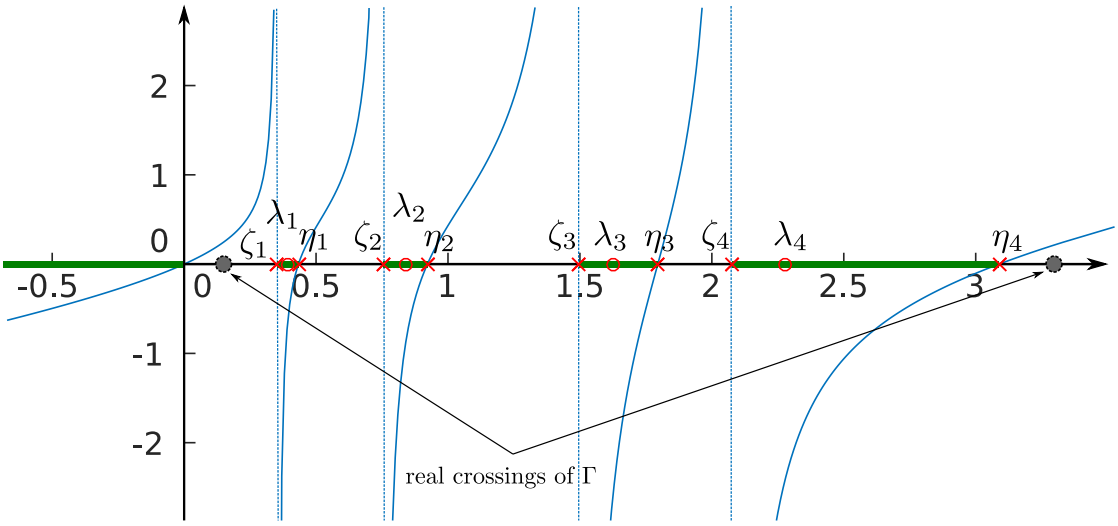

Figure 6.5: Example of visual representation of $\varphi_p/\psi_p : \mathbb{R} \to \mathbb{R}$, $x \mapsto \varphi_p(x)/\psi_p(x)$; here for $p = 4$, $n_1 = 8$, $n_2 = 16$. In green solid lines are stressed the sets over which $\varphi_p(x)/\psi_p(x) < 0$ (which correspond to branch cuts in the study of $f(z) = \log^k(z)$). Possible real crossings of the contour Γ are indicated, notably showing that no branch cut is passed through when $f(z) = \log^k(z)$.

for $Q = (zI_p - \Lambda + \frac{1}{n_2}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T})^{-1}$. Since $\Lambda - \frac{1}{n_2}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^\mathsf{T}$ is positive definite (its eigenvalues being the ζ_i's), on the real axis the derivative is greater than $1 - c_1 > 0$ and the function $x \mapsto \varphi_p(x)/\psi_p(x)$ is therefore increasing.

Figure 6.5 displays the behavior of φ_p/ψ_p when restricted to the real axis.

Since we now know that the contour Γ from Theorem 7 encloses exactly all η_i's and ζ_i's, it is sensible to evaluate the residues for these values when $f(z)$ is analytic in their neighborhood.

Remark 15 (Residue for f analytic at η_i and ζ_i)**.** *If f is analytic with no singularity at zero, then the integral has a residue at η_i easily found to be*

$$\mathrm{Res}(\eta_i) = f(0)\frac{c_1 + c_2 - c_1 c_2}{c_2}.$$

Similarly, if $f(\omega)$ has a well defined limit as $|\omega| \to \infty$, then no residue is found at ζ_i.

As a consequence of Remarks 13 and 15, we have the following immediate corollary.

Remark 16 (The case $f(t) = t$)**.** *In the case where $f(t) = t$, a singularity appears at ζ_i, which is nonetheless easily treated by noticing that the integrand then reduces to*

$$f\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2} = \frac{\varphi_p'(z)}{c_2} - \frac{\psi_p'(z)\varphi_p(z)}{c_2\psi_p(z)}$$

and thus, with $\psi_p(z) = (z-\zeta_i)\psi_p'(\zeta_i) + O((z-\zeta_i)^2)$*, we easily find the residue*

$$\mathrm{Res}_{\{f(t)=t\}}(\zeta_i) = -\frac{\varphi_p(\zeta_i)}{c_2} = -\zeta_i \frac{c_1+c_2-c_1c_2}{c_2^2}.$$

Together with Remarks 13 and 15, along with the fact that Γ *encloses all* η_i *and* $\hat{\lambda}_i$*, for* $i = 1, \ldots, p$*, we then find that*

$$\int t d\nu_p(t) - \left[\frac{c_1+c_2-c_1c_2}{c_2^2} \sum_{i=1}^{p} (\hat{\lambda}_i - \zeta_i) - \frac{c_1}{c_2} \frac{1}{p} \sum_{i=1}^{p} \hat{\lambda}_i \right] \xrightarrow{\text{a.s.}} 0.$$

By then noticing that $\sum_i \zeta_i = \mathrm{tr}(\Lambda - \frac{1}{n_2}\sqrt{\hat{\lambda}}\sqrt{\hat{\lambda}}^{\mathsf{T}}) = (1-c_2/p)\sum_i \hat{\lambda}_i$*, we retrieve Corollary 2.*

Development for $f(t) = \log(t)$

The case $f(t) = \log(t)$ leads to an immediate simplification since, then, $\log\det(\hat{\Sigma}_1^{-1}\hat{\Sigma}_2) = \log\det(\hat{\Sigma}_2) - \log\det(\hat{\Sigma}_1)$; one may then use previously established results from the random matrix literature (e.g., the G-estimators in (Girko, 1987) or more recently (Kammoun et al., 2013)) to obtain the sought-for estimate. Nonetheless, the full explicit derivation of the contour integral in this case is quite instructive and, being simpler than the subsequent cases where $f(t) = \log^2(t)$ or $f(t) = \log(1+st)$ that rely on the same key ingredients, we shall here conduct a thorough complex integral calculus.

For $z \in \mathbb{C}$, define first $f(z) = \log(z)$ where $\log(z) = \log(|z|)e^{i\arg(z)}$, with $\arg(z) \in (-\pi, \pi]$. For this definition of the complex argument, since $\varphi_p(x)/\psi_p(x)$ is everywhere positive but on $\cup_{i=1}^p (\zeta_i, \eta_i)$, we conclude that $\arg(\varphi_p(z)/\psi_p(z))$ abruptly moves from π to $-\pi$ as z moves from $x + 0^+\imath$ to $x + 0^-\imath$ for all $x \in \cup_{i=1}^p (\zeta_i, \eta_i)$. This creates a set of p branch cuts $[\zeta_i, \eta_i]$, $i = 1, \ldots, p$ as displayed in Figure 6.5. This naturally leads to computing the complex integral estimate of $\int f d\nu$ based on the contour displayed in Figure 6.6, which avoids the branch cuts.

This contour encloses no singularity of the integrand and therefore has a null integral. With the notations of Figure 6.6, the sought-for integral (over Γ) therefore satisfies

$$0 = \oint_\Gamma + \sum_{i=1}^{p} \left(\int_{I_i^A} + \int_{I_i^B} + \int_{I_i^C} + \int_{I_i^D} + \int_{I_i^E} \right).$$

We start by the evaluation of the integrals over I_i^B and I_i^D, which can be similarly handled. To this end, note that, since $\arg(\frac{\varphi_p}{\psi_p})$ moves from π to $-\pi$ across the branch cut, we have

$$\begin{aligned} \frac{1}{2\pi\imath} \int_{I_i^B} &= \frac{1}{2\pi\imath} \int_{\zeta_i+\varepsilon}^{\hat{\lambda}_i-\varepsilon} \left[\log\left(-\frac{\varphi_p(x)}{\psi_p(x)}\right) + \imath\pi - \log\left(-\frac{\varphi_p(x)}{\psi_p(x)}\right) + \imath\pi \right] \\ &\quad \times \left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)} \right) \frac{\psi_p(x)}{c_2} dx \end{aligned}$$

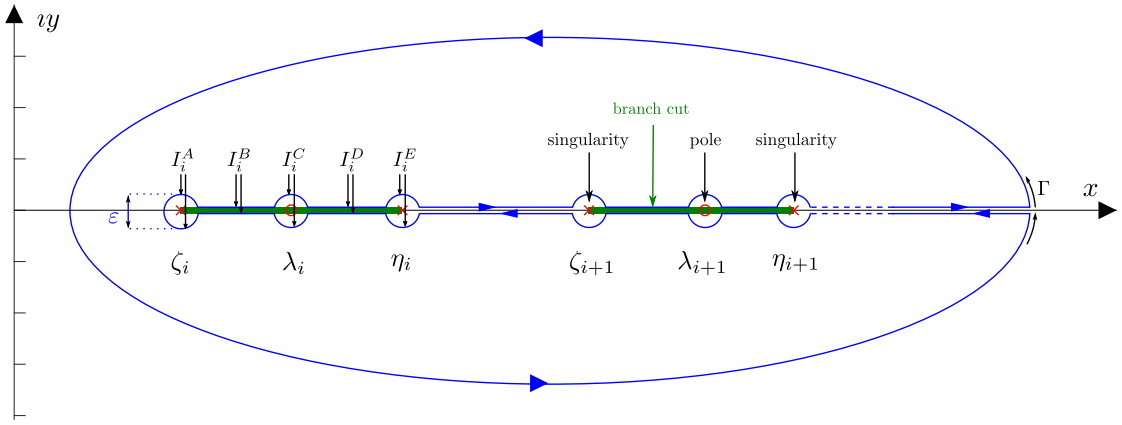

Figure 6.6: Chosen integration contour. The set I_i^B is the disjoint union of the segments $[\zeta_i + \varepsilon + 0^+\imath, \hat{\lambda}_i - \varepsilon + 0^+\imath]$ and $[\zeta_i + \varepsilon + 0^-\imath, \hat{\lambda}_i - \varepsilon + 0^-\imath]$. Similarly the set I_i^D is the disjoint union of the segments $[\hat{\lambda}_i + \varepsilon + 0^+\imath, \eta_i - \varepsilon + 0^+\imath]$ and $[\hat{\lambda}_i + \varepsilon + 0^-\imath, \eta_i - \varepsilon + 0^-\imath]$. The sets I_i^A, I_i^C and I_i^E are the disjoint unions of semi-circles (in the upper- or lower-half complex plane) of diameters ε surrounding ζ_i, $\hat{\lambda}_i$ and η_i respectively.

$$= \int_{\zeta_i+\varepsilon}^{\hat{\lambda}_i-\varepsilon} \left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)} \right) \frac{\psi_p(x)}{c_2} dx.$$

We first exploit the rational form expansion (6.10) of $(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)})\frac{\psi_p(x)}{c_2}$ to obtain the integral over I_i^B

$$\frac{1}{2\pi\imath}\int_{I_i^B} = \left(\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\left(\sum_{j\neq i}\log\left|\frac{\hat{\lambda}_i-\hat{\lambda}_j}{\zeta_i-\hat{\lambda}_j}\right| + \log\left|\frac{\varepsilon}{\zeta_i-\hat{\lambda}_i}\right|\right)$$
$$+ \frac{1-c_2}{c_2}\log\frac{\hat{\lambda}_i}{\zeta_i} + \frac{c_1+c_2-c_1c_2}{c_1c_2}\sum_{j=1}^{p}\log\left|\frac{\hat{\lambda}_i-\eta_j}{\zeta_i-\eta_j}\right| + o(\varepsilon).$$

The treatment is similar for the integral over I_i^D which results, after summation of both integrals, to

$$\frac{1}{2\pi\imath}\int_{I_i^B\cup I_i^D} = \left(\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\sum_{j=1}^{p}\log\left|\frac{\eta_i-\hat{\lambda}_j}{\zeta_i-\hat{\lambda}_j}\right| + \frac{1-c_2}{c_2}\log\frac{\eta_i}{\zeta_i}$$
$$+ \frac{c_1+c_2-c_1c_2}{c_1c_2}\left(\sum_{j\neq i}\log\left|\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right| + \log\left|\frac{\varepsilon}{\zeta_i-\eta_i}\right|\right) + o(\varepsilon).$$

Note here the asymmetry in the behavior of the integrand in the neighborhood of ζ_i $(+\varepsilon)$ and η_i $(-\varepsilon)$; in the former edge, the integral is well defined while in the latter it diverges as $\log\varepsilon$ which must then be maintained.

Summing now over $i \in \{1, \dots, p\}$, we recognize a series of identities. In particular, note that from the product form (6.13),

$$\begin{aligned}
\sum_{j=1}^{p}\sum_{i=1}^{p} \log\left|\frac{\eta_i - \hat{\lambda}_j}{\zeta_i - \hat{\lambda}_j}\right| &= \sum_{j=1}^{p} \log\left|\frac{\frac{\psi_p}{\varphi_p}(\hat{\lambda}_j)}{(1-c_1)\hat{\lambda}_j}\right| \\
&= \log\left(\frac{c_1}{c_2(1-c_1)}\right) \\
\sum_{j=1}^{p} \log\frac{\eta_i}{\zeta_i} &= \lim_{z\to 0} \log\left(\frac{\frac{\psi_p}{\varphi_p}(z)}{(1-c_1)z}\right) \\
&= -\log\left((1-c_1)(1-c_2)\right) \\
\sum_{i=1}^{p}\sum_{j\neq i} \log\left|\frac{\eta_i - \eta_j}{\zeta_i - \eta_j}\right| + \sum_{i=1}^{p}\log\left|\frac{1}{\zeta_i - \eta_i}\right| &= \lim_{z\to\eta_i} \sum_{j=1}^{p} \log\left|\frac{\frac{\psi_p}{\varphi_p}(z)}{(1-c_1)z(z-\eta_j)}\right| \\
&= \sum_{j=1}^{p} \log\left|\frac{\left(\frac{\psi_p}{\varphi_p}\right)'(\eta_j)}{(1-c_1)\eta_j}\right|.
\end{aligned}$$

As such, we now find that

$$\begin{aligned}
\frac{1}{2\pi\imath}\sum_{i=1}^{p}\int_{I_i^B \cup I_i^D} &= \log\left(\frac{c_1}{c_2(1-c_1)}\right) - \frac{1-c_2}{c_2}\log\left((1-c_1)(1-c_2)\right) \\
&\quad - \frac{c_1+c_2-c_1c_2}{c_1c_2}\left(p\log\left(\frac{c_1}{c_2(1-c_1)}\right) - \sum_{j=1}^{p}\log\left|\frac{\left(\frac{\psi_p}{\varphi_p}\right)'(\eta_j)}{(1-c_1)\eta_j}\right|\right) \\
&\quad + \frac{c_1+c_2-c_1c_2}{c_1c_2} p\log\varepsilon + o(\varepsilon).
\end{aligned}$$

The diverging term in $\log\varepsilon$ is compensated by the integral over I_i^E. Indeed, letting $z = \eta_i + \varepsilon e^{i\theta}$, we may write

$$\begin{aligned}
&\frac{1}{2\pi\imath}\int_{I_i^E} \log\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi(z)}{c_2}dz \\
&= \frac{\varepsilon}{2\pi c_2}\left[\int_{\pi}^{0^+} + \int_{0^-}^{-\pi}\right]\log\left((1-c_1)(\eta_i+\varepsilon e^{i\theta})\frac{\prod_{j=1}^{p}(\eta_i - \eta_j + \varepsilon e^{i\theta})}{\prod_{j=1}^{p}(\eta_i - \zeta_j + \varepsilon e^{i\theta})}\right) \\
&\times \left(\sum_{j=1}^{p}\frac{\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}}{\eta_i + \varepsilon e^{i\theta} - \hat{\lambda}_j} + \frac{\frac{1-c_2}{c_2}}{\eta_i + \varepsilon e^{i\theta}} + \sum_{j=1}^{p}\frac{\frac{c_1+c_2-c_1c_2}{c_1c_2}}{\eta_i + \varepsilon e^{i\theta} - \eta_j}\right) e^{i\theta}d\theta.
\end{aligned}$$

To evaluate the small ε limit of this term, first remark importantly that, for small ε, the

term in the logarithm equals

$$(1-c_1)\frac{\eta_i}{\eta_i-\zeta_i}\frac{\prod_{j\neq i}(\eta_i-\eta_j)}{\prod_{j\neq i}(\eta_i-\zeta_j)}\varepsilon e^{i\theta}+o(\varepsilon)$$

the argument of which equals that of θ. As such, on the integral over $(\pi,0)$, the log term reads $\log|\cdot|+\imath\theta+o(\varepsilon)$, while on $(0,-\pi)$, it reads $\log|\cdot|-\imath\theta+o(\varepsilon)$. With this in mind, keeping only the non-vanishing terms in the small ε limit (that is: the term in $\log\varepsilon$ and the term in $\frac{1}{\varepsilon}$) leads to

$$\begin{aligned}&\frac{1}{2\pi\imath}\int_{I_i^E}\log\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi(z)}{c_2}dz\\&=\frac{c_1+c_2-c_1c_2}{c_1c_2}\log\varepsilon+\frac{c_1+c_2-c_1c_2}{c_1c_2}\log\left|\left(\frac{\varphi_p}{\psi_p}\right)'(\eta_i)\right|+o(\varepsilon)\end{aligned}$$

where we used the fact that $\lim_{\varepsilon\to 0}\frac{1}{\varepsilon e^{\imath\theta}}\left(\frac{\varphi_p}{\psi_p}\right)(\eta_i+\varepsilon e^{\imath\theta})=\left(\frac{\varphi_p}{\psi_p}\right)'(\eta_i)$.

We proceed similarly to handle the integral over I_i^C

$$\begin{aligned}&\frac{1}{2\pi\imath}\int_{I_i^C}\log\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi(z)}{c_2}dz\\&=\frac{\varepsilon}{2\pi c_2}\left[\int_{\pi}^{0^+}+\int_{0^-}^{-\pi}\right]\log\left((1-c_1)(\hat\lambda_i+\varepsilon e^{i\theta})\frac{\prod_{j=1}^p(\hat\lambda_i-\eta_j+\varepsilon e^{i\theta})}{\prod_{j=1}^p(\hat\lambda_i-\zeta_j+\varepsilon e^{i\theta})}\right)\\&\times\left(\sum_{j=1}^p\frac{\frac{1}{p}-\frac{c_1+c_2-c_1c_2}{c_1c_2}}{\hat\lambda_i+\varepsilon e^{\imath\theta}-\hat\lambda_j}+\frac{\frac{1-c_2}{c_2}}{\hat\lambda_i}+\varepsilon e^{\imath\theta}+\sum_{j=1}^p\frac{\frac{c_1+c_2-c_1c_2}{c_1c_2}}{\hat\lambda_i+\varepsilon e^{\imath\theta}-\eta_j}\right)e^{i\theta}d\theta.\end{aligned}$$

Here, for small ε, the angle of the term in the argument of the logarithm is that of

$$\begin{aligned}&\frac{\varphi_p}{\psi_p}(\hat\lambda_i)+\left(\frac{\varphi_p}{\psi_p}\right)'(\hat\lambda_i)\varepsilon e^{\imath\theta}+o(\varepsilon)\\&=-\frac{c_1}{c_2}\hat\lambda_i+\varepsilon e^{\imath\theta}\frac{c_1}{c_2}\left(p\frac{c_1+c_2-c_1c_2}{c_1c_2}-1\right)+o(\varepsilon).\end{aligned}$$

That is, for all large p, the argument equals $\pi+o(\varepsilon)<\pi$ uniformly on $\theta\in(0,\pi)$ and $-\pi+o(\varepsilon)>-\pi$ uniformly on $\theta\in(-\pi,0)$; thus the complex logarithm reads $\log|\cdot|+\imath\theta+o(\varepsilon)$ on $(\pi,0)$, while on $(0,-\pi)$, it reads $\log|\cdot|-\imath\theta+o(\varepsilon)$. Proceeding as previously for the integral over I_i^E, we then find after calculus that

$$\begin{aligned}&\frac{1}{2\pi\imath}\int_{I_i^C}\log\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi(z)}{c_2}dz\\&=\left(\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right)\log\left(\frac{c_1}{c_2}\hat\lambda_i\right).\end{aligned}$$

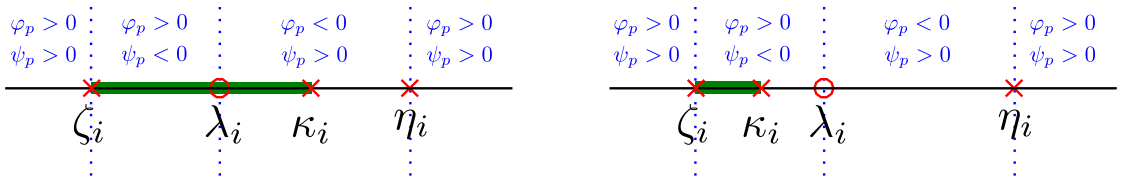

Figure 6.7: Visual representation of the signs of φ_p and ψ_p around singularities for the function $f(t) = \log(1 + st)$. Left: case where $\kappa_i > \hat{\lambda}_i$. Right: case where $\kappa_i > \hat{\lambda}_i$.

Note that this expression is reminiscent of a "residue" at $\hat{\lambda}_i$ (with negatively oriented contour), according to Remark 13, however for the function $\log|\cdot|$ and not for the function $\log(\cdot)$, due to the branch cut passing through $\hat{\lambda}_i$.

The final integral over I_i^A is performed similarly. However, here, it is easily observed that the integral is of order $O(\varepsilon \log \varepsilon)$ in the small ε limit, and thus vanishes.

Finally, summing up all contributions, we have

$$\begin{aligned}
\frac{1}{2\pi\imath}\oint_\Gamma &= -\sum_{i=1}^{p} \frac{1}{2\pi\imath}\left(\int_{I_i^A} + \int_{I_i^B} + \int_{I_i^C} + \int_{I_i^D} + \int_{I_i^E}\right) \\
&= -\log\left(\frac{c_1(1-c_2)}{c_2}\right) + \frac{1}{c_2}\log((1-c_1)(1-c_2)) + \frac{1}{p}\sum_{i=1}^{p}\log\left(\frac{c_1}{c_2}\hat{\lambda}_i\right) \\
&\quad - \frac{c_1+c_2-c_1c_2}{c_1c_2}\left(p\log(1-c_1) + \sum_{i=1}^{p}\log\hat{\lambda}_i - \sum_{i=1}^{p}\log((1-c_1)\eta_i)\right) \\
&= -\log(1-c_2) + \frac{1}{c_2}\log((1-c_1)(1-c_2)) + \frac{1}{p}\sum_{i=1}^{p}\log\hat{\lambda}_i \\
&\quad + \frac{c_1+c_2-c_1c_2}{c_1c_2}\sum_{i=1}^{p}\log\left(\frac{\eta_i}{\hat{\lambda}_i}\right) \\
&= \frac{1}{p}\sum_{i=1}^{p}\log\hat{\lambda}_i + \frac{1-c_2}{c_2}\log(1-c_2) - \frac{1-c_1}{c_1}\log(1-c_1)
\end{aligned}$$

where in the last equality we used, among other algebraic simplifications, the fact that $\sum_{i=1}^{p}\log(\frac{\eta_i}{\hat{\lambda}_i}) = \lim_{x\to 0}\log(\frac{\psi_p(x)}{(1-c_1)x}) = -\log(1-c_1)$. This is the sought-for result.

Development for $f(t) = \log(1 + st)$

The development for $f(t) = \log(1 + st)$ is quite similar to that of $f(t) = \log(t)$, with some noticeable exceptions with respect to the position of singularity points.

A few important remarks are in order to start with this scenario. First note from Figure 6.4 and the previous discussions that the function $z \mapsto \log(1 + s\varphi_p(z)/\psi_p(z))$ has a singularity at $z = \kappa_i$, $i = 1, \ldots, p$, for some $\kappa_i \in (\zeta_i, \eta_i)$ solution to $1 + s\varphi_p(x)/\psi_p(x) = 0$ (indeed, $\varphi_p(x)/\psi_p(x)$ is increasing on (ζ_i, η_i) with opposite asymptotes and thus κ_i exists

and is uniquely defined). In addition, $\log(1 + s\varphi_p(z)/\psi_p(z))$ has a further singularity satisfying $1 + s\varphi_p(x)/\psi_p(x) = 0$ in the interval $(-\infty, 0)$ which we shall denote κ_0.

A few identities regarding κ_i are useful. Using the relation between φ_p and ψ_p, we find in particular that

$$\begin{aligned}\varphi_p(\kappa_i) &= -\frac{1}{s}\frac{c_1 + c_2 - c_1c_2}{c_2}\frac{\kappa_i}{-\frac{1}{s} + \frac{c_1}{c_2}\kappa_i}\\ \psi_p(\kappa_i) &= \frac{c_1 + c_2 - c_1c_2}{c_2}\frac{\kappa_i}{-\frac{1}{s} + \frac{c_1}{c_2}\kappa_i}\\ (\psi_p + s\varphi_p)\left(\frac{c_2}{c_1 s}\right) &= \frac{c_1 + c_2 - c_1c_2}{c_1}.\end{aligned}$$

With the discussions above, we also find that

$$1 + s\frac{\varphi_p(z)}{\psi_p(z)} = (1 - c_1)s(z - \kappa_0)\frac{\prod_{i=1}^p (z - \kappa_i)}{\prod_{i=1}^p (z - \zeta_i)} \tag{6.14}$$

$$\psi_p(z) + s\varphi_p(z) = s(1 - c_1)(z - \kappa_0)\frac{\prod_{i=1}^p (z - \kappa_i)}{\prod_{i=1}^p (z - \hat{\lambda}_i)}. \tag{6.15}$$

Note now importantly that $\hat{\lambda}_i > \frac{c_1}{c_2 s}$ is equivalent to $-\frac{c_2}{c_1}\hat{\lambda}_i < -\frac{1}{s}$ which is also $\varphi_p(\hat{\lambda}_i)/\psi_p(\hat{\lambda}_i) < \varphi_p(\kappa_i)/\psi_p(\kappa_i)$; then, as φ_p/ψ_p is increasing, $\hat{\lambda}_i > \frac{c_1}{c_2 s}$ is equivalent to $\hat{\lambda}_i < \kappa_i$. On the opposite, for $\hat{\lambda}_i < \frac{c_1}{c_2 s}$, we find $\hat{\lambda}_i > \kappa_i$. As such, to evaluate the contour integral in this setting, one must isolate two sets of singularities (see Figure 6.7): (i) those for which $\kappa_i > \hat{\lambda}_i$ (which are all the largest indices i for which $\hat{\lambda}_i > \frac{c_1}{c_2 s}$) and (ii) those for which $\kappa_i < \hat{\lambda}_i$. This affects the relative position of the branch cut with respect to $\hat{\lambda}_i$ and therefore demands different treatments. In particular, the integrals over I_i^B and I_i^D may be restricted to integrals over shorter (possibly empty) segments. Nonetheless, the calculus ultimately reveals that, since the branch cut does not affect the local behavior of the integral around $\hat{\lambda}_i$, both cases entail the same result. In particular, in case (i) where $\hat{\lambda}_i > \kappa_i$, recalling (6.10), one only has to evaluate

$$\begin{aligned}&\int_{\zeta_i+\varepsilon}^{\kappa_i-\varepsilon}\left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)}\right)\frac{\psi_p(x)}{c_2}dx\\ &= \int_{\zeta_i+\varepsilon}^{\kappa_i-\varepsilon}\left(\frac{1}{p} - \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\right)\sum_{j=1}^p \frac{1}{x - \hat{\lambda}_j} + \frac{1 - c_2}{c_2}\frac{1}{x}\\ &+ \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\sum_{j=1}^p \frac{1}{x - \eta_j}dx\\ &= \frac{1}{p}\sum_{j=1}^p \log\left|\frac{\kappa_i - \hat{\lambda}_j}{\zeta_i - \hat{\lambda}_j}\right| + \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\sum_{j=1}^p\left(\log\left|\frac{\kappa_i - \eta_j}{\kappa_i - \hat{\lambda}_j}\right| - \log\left|\frac{\zeta_i - \eta_j}{\zeta_i - \hat{\lambda}_j}\right|\right)\\ &+ \frac{1 - c_2}{c_2}\log\left|\frac{\kappa_i}{\zeta_i}\right| + o(\varepsilon).\end{aligned}$$

In case (ii), subdividing the integral as $\int_{\zeta_i+\varepsilon}^{\hat\lambda_i-\varepsilon}+\int_{\hat\lambda_i+\varepsilon}^{\kappa_i-\varepsilon}$ brings immediate simplification of the additional terms in $\hat\lambda_i$ and thus the result remains the same.

The integral over I_i^C is slightly more delicate to handle. In case (i), in the limit of small ε,

$$1+s\frac{\varphi_p}{\psi_p}(\hat\lambda_i+\varepsilon e^{i\theta})=1-s\frac{c_1}{c_2}\hat\lambda_i+\varepsilon s\frac{c_1}{c_2}\left(p\frac{c_1+c_2-c_1c_2}{c_1c_2}-1\right)e^{i\theta}+o(\varepsilon)$$

the angle of which is $0+o(\varepsilon)$ uniformly on $\theta\in(-\pi,\pi]$ (since $1-s\frac{c_1}{c_2}\hat\lambda_i>0$). As such, for all small ε, the sum of the integrals over $(-\pi,0)$ and $(0,\pi]$ reduces to the integral over $(-\pi,\pi]$, leading up to a mere residue calculus, and

$$\begin{aligned}&\frac{1}{2\pi\imath}\oint_{I_i^C}\log\left(1+s\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz\\&=\log\left(1-s\frac{c_1}{c_2}\hat\lambda_i\right)\left(\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right)+o(\varepsilon).\end{aligned}$$

In case (ii), $1-s\frac{c_1}{c_2}\hat\lambda_i<0$ and thus the angle of $1+s\frac{\varphi_p}{\psi_p}(\hat\lambda_i+\varepsilon e^{i\theta})$ is close to π; for $\theta\in(0,\pi)$, this leads to an argument equal to $\pi+o(\varepsilon)<\pi$ and for $\theta\in(-\pi,0)$ to an argument equal to $-\pi+o(\varepsilon)>-\pi$. All calculus made, we then find that in either case (i) or (ii)

$$\begin{aligned}&\frac{1}{2\pi\imath}\oint_{I_i^C}\log\left(1+s\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz\\&=\log\left|1-s\frac{c_1}{c_2}\hat\lambda_i\right|\left(\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right)+o(\varepsilon).\end{aligned}$$

As in the case of $f(t)=\log(t)$, the integral over I_i^A is of order $o(\varepsilon)$ and vanishes. As a consequence, summing over $i\in\{1,\dots,p\}$, we find that

$$\begin{aligned}\frac{1}{2\pi\imath}\oint_\Gamma&=-\frac{1}{p}\sum_{i,j=1}^p\log\left|\frac{\kappa_i-\hat\lambda_j}{\zeta_i-\hat\lambda_j}\right|-\frac{1-c_2}{c_2}\sum_{i=1}^p\log\frac{\kappa_i}{\zeta_i}\\&\quad+\frac{c_1+c_2-c_1c_2}{c_1c_2}\sum_{i,j=1}^p\left(\log\left|\frac{\zeta_i-\eta_j}{\zeta_i-\hat\lambda_j}\right|-\log\left|\frac{\kappa_i-\eta_j}{\kappa_i-\hat\lambda_j}\right|\right)\\&\quad-\left(\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right)\sum_{i=1}^p\log\left|1-s\frac{c_1}{c_2}\hat\lambda_j\right|+o(\varepsilon).\end{aligned}$$

Before reaching the final result, note that, from (6.14),

$$\sum_{i=1}^p\frac{1}{p}\sum_{j=1}^p\log\frac{|\kappa_i-\hat\lambda_j|}{|\zeta_i-\hat\lambda_j|}$$

$$= \frac{1}{p}\sum_{j=1}^{p} \log\left|\left(1 + s\frac{\varphi_p(\hat{\lambda}_j)}{\psi_p(\hat{\lambda}_j)}\right)\frac{1}{\hat{\lambda}_j - \kappa_0}\frac{1}{(1-c_1)s}\right|$$
$$= \frac{1}{p}\sum_{j=1}^{p} \log\left|1 - \frac{c_1}{c_2}s\hat{\lambda}_i\right| - \frac{1}{p}\sum_{j=1}^{p}\log(\hat{\lambda}_j - \kappa_0) - \log((1-c_1)s)$$

and similarly

$$\sum_{i,j=1}^{p} \log\left|\frac{\zeta_i - \eta_j}{\zeta_i - \hat{\lambda}_j}\right| = \sum_{i=1}^{p}\log\left|\frac{\varphi_p(\zeta_i)}{(1-c_1)\zeta_i}\right| = p\log\left|\frac{c_1 + c_2 - c_1c_2}{c_2(1-c_1)}\right|$$
$$\sum_{i,j=1}^{p} \log\left|\frac{\kappa_i - \eta_j}{\kappa_i - \hat{\lambda}_j}\right| = \sum_{i=1}^{p}\log\left|\frac{\varphi_p(\kappa_i)}{(1-c_1)\kappa_i}\right| = \sum_{i=1}^{p}\log\left|\frac{c_1 + c_2 - c_1c_2}{c_2(1-c_1)}\frac{1}{1 - \frac{c_1}{c_2 s}\kappa_i}\right|$$
$$\sum_{i=1}^{p}\log\frac{\kappa_i}{\zeta_i} = \log\left(\frac{1 + s\frac{\varphi_p}{\psi_p}(0)}{-(1-c_1)s\kappa_0}\right) = -\log\left(-(1-c_1)s\kappa_0\right).$$

Using now (6.15), we find that

$$\sum_{i=1}^{p}\log\left(\frac{1 - s\frac{c_1}{c_2}\hat{\lambda}_i}{1 - s\frac{c_1}{c_2}\kappa_i}\right) = \sum_{i=1}^{p}\log\left(\frac{\frac{c_2}{c_1 s} - \hat{\lambda}_i}{\frac{c_2}{c_1 s} - \kappa_i}\right) = \log\left(\frac{\psi_p\left(\frac{c_2}{c_1 s}\right) + s\varphi_p\left(\frac{c_2}{c_1 s}\right)}{s(1-c_1)\left(\frac{c_2}{c_1 s} - \kappa_0\right)}\right)$$
$$= \log\left(\frac{c_1 + c_2 - c_1c_2}{sc_1(1-c_1)\left(\frac{c_2}{c_1 s} - \kappa_0\right)}\right).$$

Combining the previous results and remarks then leads to

$$\frac{1}{2\pi\imath}\oint_{\Gamma} = \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\log\left(\frac{c_1 + c_2 - c_1c_2}{(1-c_1)(c_2 - sc_1\kappa_0)}\right)$$
$$+ \frac{1-c_2}{c_2}\log\left(-s\kappa_0(1-c_1)\right) + \log((1-c_1)s) + \frac{1}{p}\sum_{i=1}^{p}\log(\hat{\lambda}_i - \kappa_0)$$
$$= \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\log\left(\frac{c_1 + c_2 - c_1c_2}{(1-c_1)(c_2 - sc_1\kappa_0)}\right) + \frac{1}{c_2}\log\left(-s\kappa_0(1-c_1)\right)$$
$$+ \frac{1}{p}\sum_{i=1}^{p}\log\left(1 - \frac{\hat{\lambda}_i}{\kappa_0}\right).$$

This concludes the proof for the case $c_1 > 0$. In the limit where $c_1 \to 0$, it suffices to use the Taylor expansion of the leftmost logarithm in the small c_1 limit (i.e., $\log(c_2(1-c_1) + c_1) \sim \log(c_2(1-c_1)) + c_1/(c_2(1-c_1))$ and $\log(c_2(1-c_1) - sc_1\kappa_0(1-c_1)) \sim \log(c_2(1-c_1)) - sc_1\kappa_0/c_2$).

Development for $f(t) = \log^2(t)$

The function $f(t) = \log^2(t)$ is at the core of the Fisher distance and is thus of prime importance in many applications. The evaluation of the complex integral in Theorem 7 for this case is however quite technical and calls for the important introduction of the dilogarithm function. We proceed with this introduction first and foremost.

The dilogarithm function The (real) dilogarithm is defined as the function

$$\mathrm{Li}_2(x) = -\int_0^x \frac{\log(1-u)}{u} du.$$

for $x \in (-\infty, 1]$.

The dilogarithm function will intervene in many instances of the evaluation of the contour integral of Theorem 7, through the subsequently defined function $F(X, Y; a)$. This function assumes different formulations depending on the relative position of X, Y, a on the real axis.

Lemma 2 (Dilogarithm integrals)**.** *We have the following results and definition*

$$\begin{aligned}
(X, Y \geq a > 0) \quad & \int_Y^X \frac{\log(x-a)}{x} dx \equiv F(X, Y; a) \\
& = \mathrm{Li}_2\left(\frac{a}{X}\right) - \mathrm{Li}_2\left(\frac{a}{Y}\right) + \frac{1}{2}\left[\log^2(X) - \log^2(Y)\right] \\
(X, Y > 0 > a) \quad & \int_Y^X \frac{\log(x-a)}{x} dx \equiv F(X, Y; a) \\
& = -\mathrm{Li}_2\left(\frac{X}{a}\right) + \mathrm{Li}_2\left(\frac{Y}{a}\right) + \log\left(\frac{X}{Y}\right)\log(-a) \\
(a > X, Y, 0 \ \&\ XY > 0) \quad & \int_Y^X \frac{\log(a-x)}{x} dx \equiv F(-X, -Y; -a) \\
& = -\mathrm{Li}_2\left(\frac{X}{a}\right) + \mathrm{Li}_2\left(\frac{Y}{a}\right) + \log\left(\frac{X}{Y}\right)\log(a) \\
(X, Y > 0) \quad & \int_Y^X \frac{\log(x)}{x} dx \equiv F(X, Y; 0) \\
& = \frac{1}{2}\log^2(X) - \frac{1}{2}\log^2(Y).
\end{aligned}$$

Lemma 3 (Properties of Dilogarithm functions (Zagier, 2007, Section I-2))**.** *The following relations hold*

$$\begin{aligned}
(x < 0) \quad & \mathrm{Li}_2\left(\frac{1}{x}\right) + \mathrm{Li}_2(x) = -\frac{1}{2}\log^2(-x) - \frac{\pi^2}{6} \\
(0 < x < 1) \quad & \mathrm{Li}_2(1-x) + \mathrm{Li}_2(x) = -\log(x)\log(1-x) + \frac{\pi^2}{6}
\end{aligned}$$

$$(0 < x < 1) \quad \mathrm{Li}_2(1-x) + \mathrm{Li}_2\left(1 - \frac{1}{x}\right) = -\frac{1}{2}\log^2(x).$$

Besides, for $x < 1$ and $\varepsilon > 0$ small,

$$\mathrm{Li}_2(x+\varepsilon) = \mathrm{Li}_2(x) - \varepsilon\frac{\log(1-x)}{x} + \varepsilon^2\frac{(1-x)\log(1-x)+x}{2(1-x)x^2} + O(\varepsilon^3).$$

Integral evaluation As in the case where $f(t) = \log(t)$, we shall evaluate the complex integral based on the contour displayed in Figure 6.6. The main difficulty here arises in evaluating the real integrals over the segments I_i^B and I_i^D.

Again, we start from the Equation (6.10). In particular, the integral over I_i^B reads

$$\begin{aligned}
&\frac{1}{2\pi\imath}\int_{I_i^B}\log^2\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz \\
&= 2\int_{\zeta_i+\varepsilon}^{\hat{\lambda}_i-\varepsilon}\log\left(-\frac{\varphi_p(x)}{\psi_p(x)}\right)\left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)}\right)\frac{\psi_p(x)}{c_2}dx \\
&= 2\int_{\zeta_i+\varepsilon}^{\hat{\lambda}_i-\varepsilon}\Bigg(\log(1-c_1) + \log(x) + \sum_{l<i}\log(x-\eta_l) \\
&+ \sum_{l>i}\log(\eta_l - x) + \log(\eta_i - x) - \sum_{l\le i}\log(x-\zeta_l) - \sum_{l>i}\log(\zeta_l - x)\Bigg) \\
&\times\left(\left(\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\sum_{j=1}^{p}\frac{1}{x-\hat{\lambda}_j} + \frac{1-c_2}{c_2}\frac{1}{x} + \frac{c_1+c_2-c_1c_2}{c_1c_2}\sum_{j=1}^{p}\frac{1}{x-\eta_j}\right)dx.
\end{aligned}$$

Note that above we have specifically chosen to write the logarithms in such a way that every integral is a well-defined real integral.

Using now the fact that

$$\int_Y^X \frac{\log(x-a)}{x-b}dx = F(X-b, Y-b; a-b), \quad \int_Y^X \frac{\log(a-x)}{x-b}dx \quad = F(b-X, b-Y; b-a)$$

that we apply repetitively (and very carefully) to the previous equality, we find that the sum of the integral of I_i^B and I_i^D gives

$$\begin{aligned}
&\frac{1}{2\pi\imath}\left[\int_{I_i^B} + \int_{I_i^D}\right]\log^2\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz \\
&= 2\left[\int_{\zeta_i+\varepsilon}^{\hat{\lambda}_i-\varepsilon} + \int_{\hat{\lambda}_i+\varepsilon}^{\eta_i-\varepsilon}\right]\log\left(-\frac{\varphi_p(x)}{\psi_p(x)}\right)\left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)}\right)\frac{\psi_p(x)}{c_2}dx \\
&= 2\left(\frac{1}{p} - \frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\Bigg(\log(1-c_1)\sum_{j\neq i}\log\left|\frac{\eta_i - \hat{\lambda}_j}{\zeta_i - \hat{\lambda}_j}\right|
\end{aligned}$$

$$
\begin{aligned}
&+F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;-\hat\lambda_i)\\
&+\sum_{j\neq i}F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;-\hat\lambda_j)\\
&+\sum_{l<i}\sum_{j\neq(i,l)}F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\eta_l-\hat\lambda_j)-F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\zeta_l-\hat\lambda_j)\\
&+\sum_{l<i}F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\eta_l-\hat\lambda_l)-F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\zeta_l-\hat\lambda_l)\\
&+\sum_{l<i}F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\eta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\eta_l-\hat\lambda_i)\\
&-\sum_{l<i}F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_l-\hat\lambda_i)\\
&+\sum_{l>i}\sum_{j\neq(i,l)}F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\eta_l+\hat\lambda_j)-F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\zeta_l+\hat\lambda_j)\\
&+\sum_{l>i}F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\eta_l+\hat\lambda_l)-F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\zeta_l+\hat\lambda_l)\\
&+\sum_{l>i}F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\eta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\eta_l+\hat\lambda_i)\\
&-\sum_{l>i}F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\zeta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\zeta_l+\hat\lambda_i)\\
&+\sum_{j\neq i}F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\eta_i+\hat\lambda_j)-F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\zeta_i-\hat\lambda_j)\\
&+F(\varepsilon,\hat\lambda_i-\zeta_i-\varepsilon;\hat\lambda_i-\eta_i)+F(\hat\lambda_i-\eta_i+\varepsilon,-\varepsilon;\hat\lambda_i-\eta_i)\\
&-F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_i-\hat\lambda_i)-F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_i-\hat\lambda_i)\Big)\\
&+2\frac{1-c_2}{c_2}\left(\log(1-c_1)\log\left(\frac{\eta_i}{\zeta_i}\right)+F(\eta_i,\zeta_i;0)+F(-\eta_i,-\zeta_i;-\eta_i)-F(\eta_i,\zeta_i;\zeta_i)\right.\\
&\left.+\sum_{l<i}F(\eta_i,\zeta_i;\eta_l)-F(\eta_i,\zeta_i;\zeta_l)+\sum_{l>i}F(-\eta_i,-\zeta_i;-\eta_l)-F(-\eta_i,-\zeta_i;-\zeta_l)\right)\\
&+2\frac{c_1+c_2-c_1c_2}{c_1c_2}\left(\log(1-c_1)\sum_{j\neq i}\log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)+\sum_{j\neq i}F(\eta_i-\eta_j,\zeta_i-\eta_j;-\eta_j)\right.\\
&+\sum_{j\neq(i,l)}\sum_{l<i}F(\eta_i-\eta_j,\zeta_i-\eta_j;\eta_l-\eta_j)+\sum_{j\neq(i,l)}\sum_{l>i}F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\eta_l+\eta_j)\\
&+\sum_{j\neq i}F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\eta_i+\eta_j)-\sum_{j\neq(i,l)}\sum_{l<i}F(\eta_i-\eta_j,\zeta_i-\eta_j;\zeta_l-\eta_j)\\
&\left.-\sum_{j\neq(i,l)}\sum_{l>i}F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\zeta_l+\eta_j)-\sum_{j\neq i}F(\eta_i-\eta_j,\zeta_i-\eta_j;\zeta_i-\eta_j)\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\log(1-c_1)\log\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)+\sum_{l<i}F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\eta_l-\eta_i)-F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\zeta_l-\eta_i)\\
&+\sum_{l>i}F(\varepsilon,\eta_i-\zeta_i-\varepsilon;\eta_i-\eta_l)-F(\varepsilon,\eta_i-\zeta_i-\varepsilon;\eta_i-\zeta_l)\\
&+F(\varepsilon,-\zeta_i+\eta_i-\varepsilon,\varepsilon)-F(-\varepsilon,\zeta_i-\eta_i+\varepsilon,\zeta_i-\eta_i)+F(-\varepsilon,\zeta_i-\eta_i+\varepsilon,-\eta_i)\\
&+\sum_{l<i}F(\eta_i-\eta_l,\zeta_i-\eta_l;0)+\sum_{l>i}F(-\eta_i+\eta_l,-\zeta_i+\eta_l;0)\\
&-\sum_{l<i}F(\eta_i-\eta_l,\zeta_i-\eta_l;\zeta_l-\eta_l)-\sum_{l>i}F(-\eta_i+\eta_l,-\zeta_i+\eta_l;-\zeta_l+\eta_l)\Bigg)\\
&+2\left(\frac{1}{p}-\frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\log(1-c_1)\log\left(\frac{\eta_i-\hat\lambda_i}{\hat\lambda_i-\zeta_i}\right)+o_\varepsilon(1).
\end{aligned}
$$

To retrieve the expression above, particular care was taken on the relative positions of the $\hat\lambda_{i,j,l}$, $\eta_{i,j,l}$ and $\zeta_{i,j,l}$ to obtain the proper form of the F function; besides, to avoid further complications, a small ε approximation was used whenever the F function has a finite limit when $\varepsilon\to 0$ (hence the trailing $o_\varepsilon(1)$ in the formula).

To go further, we now make use of the following additional identities obtained from Lemma 3 (these are easily proved).

Lemma 4 (Properties of the function F)**.** *We have the following properties of the function F:*

$$
\begin{aligned}
(X\ge Y>0)\quad & F(-X,-Y;-X)-F(X,Y;Y)=\frac{1}{2}\log^2\left(\frac{X}{Y}\right)\\
(Y\ge X>0)\quad & F(X,Y;X)-F(-X,-Y;-Y)=\frac{1}{2}\log^2\left(\frac{X}{Y}\right)\\
(X,Y>0)\quad & F(-X+\varepsilon,-\varepsilon;-X)+F(\varepsilon,Y-\varepsilon;-X)\\
&-F(X-\varepsilon,\varepsilon;-Y)-F(-\varepsilon,-Y+\varepsilon;-Y)\\
&=-\frac{\pi^2}{2}+\frac{1}{2}\log^2\left(\frac{X}{Y}\right)+o_\varepsilon(1)\\
(T,Z\ge X,Y>0)\quad & F(-\varepsilon,-Y+\varepsilon;-T)+F(X-\varepsilon,\varepsilon;-T)\\
&-F(-\varepsilon,-Y+\varepsilon;-Z)-F(X-\varepsilon,\varepsilon;-Z)\\
&+F(T,Z;-X)-F(T,Z;Y)\\
&=\log\left(\frac{X}{Y}\right)\log\left(\frac{T}{Z}\right)+o_\varepsilon(1)\\
(T,Z\ge X,Y>0)\quad & F(\varepsilon,-Y-\varepsilon;-T)+F(-X+\varepsilon,-\varepsilon;-T)\\
&-F(\varepsilon,X-\varepsilon;-Z)-F(-Y+\varepsilon,-\varepsilon;-Z)\\
&+F(T,Z;X)-F(T,Z;-Y)\\
&=\log\left(\frac{X}{Y}\right)\log\left(\frac{T}{Z}\right)+o_\varepsilon(1)
\end{aligned}
$$

$$
\begin{aligned}
(X,Y,Z,T>0) \quad & F(X,Y;T)+F(T,Z;X)-F(X,Y;Z)-F(T,Z;Y) \\
& = \log\left(\frac{X}{Y}\right)\log\left(\frac{T}{Z}\right) \\
(XY>0 \ \&\ ZT>0) \quad & F(X,Y;-Z)+F(Z,T;-X)-F(X,Y;-T)-F(Z,T;-Y) \\
& = \log\left(\frac{X}{Y}\right)\log\left(\frac{T}{Z}\right) \\
(Y,Z>0 \ \&\ \varepsilon X>0) \quad & F(\varepsilon,X;-Y)+F(\varepsilon,X;-Z)-F(Y,Z;-X) \\
& = \log\left(\frac{\varepsilon}{X}\right)\log\left(\frac{Y}{Z}\right)+\frac{1}{2}\log^2(Z)-\frac{1}{2}\log^2(Y).
\end{aligned}
$$

Exploiting the relations from the previous lemma, we have the following first result:

$$
\begin{aligned}
&\sum_{i=1}^{p} F(-\varepsilon,\zeta_i-\hat{\lambda}_i+\varepsilon;-\hat{\lambda}_i)+F(\eta_i-\hat{\lambda}_i-\varepsilon,\varepsilon;-\hat{\lambda}_i) \\
&= \sum_{i=1}^{p} \mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_i}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_i}\right)+\log(\hat{\lambda}_i)\log\left(\frac{\eta_i-\hat{\lambda}_i}{\hat{\lambda}_i-\zeta_i}\right)
\end{aligned}
$$

The terms involving double or triple sums (over i,l or i,j,l) are more subtle to handle. By observing that $\sum_i \sum_{l>i} G_{il} = \sum_l \sum_{i<l} G_{il}$ which, up to a switch in the notation (i,l) into (l,i), is the same as $\sum_i \sum_{l<i} G_{li}$, we have that

$$
\sum_i \sum_{l>i} G_{il} + \sum_i \sum_{l<i} G_{il} = \sum_i \sum_{l>i} G_{il} + G_{li}.
$$

Using this observation to gather terms together, we find notably from Lemma 4 that

$$
\begin{aligned}
&\sum_i \sum_{l<i} \sum_{j\neq(i,l)} F(\eta_i-\hat{\lambda}_j,\zeta_i-\hat{\lambda}_j;\eta_l-\hat{\lambda}_j)-F(\eta_i-\hat{\lambda}_j,\zeta_i-\hat{\lambda}_j;\zeta_l-\hat{\lambda}_j) \\
&+\sum_i \sum_{l>i} \sum_{j\neq(i,l)} F(-\eta_i+\hat{\lambda}_j,-\zeta_i+\hat{\lambda}_j;-\eta_l+\hat{\lambda}_j)-F(-\eta_i+\hat{\lambda}_j,-\zeta_i+\hat{\lambda}_j;-\zeta_l+\hat{\lambda}_j) \\
&= \sum_i \sum_{l<i} \sum_{j\neq(i,l)} \log\left(\frac{\hat{\lambda}_j-\eta_i}{\hat{\lambda}_j-\zeta_i}\right)\log\left(\frac{\hat{\lambda}_j-\eta_l}{\hat{\lambda}_j-\zeta_l}\right) \\
&= \frac{1}{2}\sum_i \sum_{l\neq i} \sum_{j\neq(i,l)} \log\left(\frac{\hat{\lambda}_j-\eta_i}{\hat{\lambda}_j-\zeta_i}\right)\log\left(\frac{\hat{\lambda}_j-\eta_l}{\hat{\lambda}_j-\zeta_l}\right)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\sum_i \sum_{l<i} F(-\varepsilon,\zeta_i-\hat{\lambda}_i+\varepsilon;\eta_l-\hat{\lambda}_i)+F(\eta_i-\hat{\lambda}_i-\varepsilon,\varepsilon;\eta_l-\hat{\lambda}_i) \\
&-\sum_i \sum_{l<i} F(-\varepsilon,\zeta_i-\hat{\lambda}_i+\varepsilon;\zeta_l-\hat{\lambda}_i)+F(\eta_i-\hat{\lambda}_i-\varepsilon,\varepsilon;\zeta_l-\hat{\lambda}_i)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_i\sum_{l>i} F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\eta_l+\hat\lambda_l)-F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\zeta_l+\hat\lambda_l)\\
&=\sum_i\sum_{l<i} F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\eta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\eta_l-\hat\lambda_i)\\
&-\sum_i\sum_{l<i} F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_l-\hat\lambda_i)\\
&+\sum_i\sum_{l<i} F(-\eta_l+\hat\lambda_i,-\zeta_l+\hat\lambda_i;-\eta_i+\hat\lambda_i)-F(-\eta_l+\hat\lambda_i,-\zeta_l+\hat\lambda_i;-\zeta_i+\hat\lambda_i)\\
&=\sum_i\sum_{l<i}\log\left(\frac{\eta_i-\hat\lambda_i}{\zeta_i-\hat\lambda_i}\right)\log\left(\frac{\hat\lambda_i-\eta_l}{\hat\lambda_i-\zeta_l}\right)+o_\varepsilon(1)
\end{aligned}
$$

and, symmetrically,

$$
\begin{aligned}
&\sum_i\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\eta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\eta_l+\hat\lambda_i)\\
&-\sum_i\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\zeta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\zeta_l+\hat\lambda_i)\\
&+\sum_i\sum_{l<i} F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\eta_l-\hat\lambda_l)-F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\zeta_l-\hat\lambda_l)\\
&=\sum_i\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\eta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\eta_l+\hat\lambda_i)\\
&-\sum_i\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\zeta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\zeta_l+\hat\lambda_i)\\
&+\sum_i\sum_{l>i} F(\eta_l-\hat\lambda_i,\zeta_l-\hat\lambda_i;\eta_i-\hat\lambda_i)-F(\eta_l-\hat\lambda_i,\zeta_l-\hat\lambda_i;\zeta_i-\hat\lambda_i)\\
&=\sum_i\sum_{l>i}\log\left(\frac{\eta_i-\hat\lambda_i}{\zeta_i-\hat\lambda_i}\right)\log\left(\frac{\hat\lambda_i-\eta_l}{\hat\lambda_i-\zeta_l}\right)+o_\varepsilon(1).
\end{aligned}
$$

Also, using Items 1 and 2 of Lemma 4, we find

$$
\begin{aligned}
&\sum_i\sum_{j\neq i} F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\eta_i+\hat\lambda_j)-F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\zeta_i-\hat\lambda_j)\\
&=\sum_i\sum_{j\neq i}\frac{1}{2}\log^2\left(\frac{\hat\lambda_j-\eta_i}{\hat\lambda_j-\zeta_i}\right).
\end{aligned}
$$

Again from Lemma 4, we also have

$$
\begin{aligned}
&F(\varepsilon,\hat\lambda_i-\zeta_i-\varepsilon;\hat\lambda_i-\eta_i)+F(\hat\lambda_i-\eta_i+\varepsilon,-\varepsilon;\hat\lambda_i-\eta_i)\\
&-F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_i-\hat\lambda_i)-F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_i-\hat\lambda_i)\\
&=-\frac{\pi^2}{2}+\frac{1}{2}\log^2\left(\frac{\eta_i-\hat\lambda_i}{\hat\lambda_i-\zeta_i}\right).
\end{aligned}
$$

Before going further, remark that the last four established relations can be assembled to reach

$$
\begin{aligned}
&\sum_{j\neq i} F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;-\hat\lambda_j)+\sum_{l<i}\sum_{j\neq(i,l)} F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\eta_l-\hat\lambda_j)\\
&-F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\zeta_l-\hat\lambda_j)\\
&+\sum_{l<i} F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\eta_l-\hat\lambda_l)-F(\eta_i-\hat\lambda_l,\zeta_i-\hat\lambda_l;\zeta_l-\hat\lambda_l)\\
&+\sum_{l<i} F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\eta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\eta_l-\hat\lambda_i)\\
&-\sum_{l<i} F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_l-\hat\lambda_i)+F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_l-\hat\lambda_i)\\
&+\sum_{l>i}\sum_{j\neq(i,l)} F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\eta_l+\hat\lambda_j)-F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\zeta_l+\hat\lambda_j)\\
&+\sum_{l>i} F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\eta_l+\hat\lambda_l)-F(-\eta_i+\hat\lambda_l,-\zeta_i+\hat\lambda_l;-\zeta_l+\hat\lambda_l)\\
&+\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\eta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\eta_l+\hat\lambda_i)\\
&-\sum_{l>i} F(\varepsilon,-\zeta_i+\hat\lambda_i-\varepsilon;-\zeta_l+\hat\lambda_i)+F(-\eta_i+\hat\lambda_i+\varepsilon,-\varepsilon;-\zeta_l+\hat\lambda_i)\\
&+\sum_{j\neq i} F(-\eta_i+\hat\lambda_j,-\zeta_i+\hat\lambda_j;-\eta_i+\hat\lambda_j)-F(\eta_i-\hat\lambda_j,\zeta_i-\hat\lambda_j;\zeta_i-\hat\lambda_j)\\
&+F(\varepsilon,\hat\lambda_i-\zeta_i-\varepsilon;\hat\lambda_i-\eta_i)+F(\hat\lambda_i-\eta_i+\varepsilon,-\varepsilon;\hat\lambda_i-\eta_i)\\
&-F(-\varepsilon,\zeta_i-\hat\lambda_i+\varepsilon;\zeta_i-\hat\lambda_i)-F(\eta_i-\hat\lambda_i-\varepsilon,\varepsilon;\zeta_i-\hat\lambda_i)\\
&=\frac{1}{2}p\log^2\left(\frac{c_1}{c_2(1-c_1)}\right)+o_\varepsilon(1).
\end{aligned}
$$

Next, we have

$$F(\eta_i,\zeta_i;0)=\frac{1}{2}\log^2(\eta_i)-\frac{1}{2}\log^2(\zeta_i)$$

and, again by Lemma 4,

$$
\begin{aligned}
&\sum_i\sum_{l\neq i} F(\eta_i,\zeta_i;\eta_l)-F(\eta_i,\zeta_i;\zeta_l)\\
&=\sum_{i>l} F(\eta_i,\zeta_i;\eta_l)+F(\eta_l,\zeta_l;\eta_i)-F(\eta_i,\zeta_i;\zeta_l)-F(\eta_l,\zeta_l;\zeta_i)\\
&=\sum_{i>l}\log\left(\frac{\eta_i}{\zeta_i}\right)\log\left(\frac{\eta_l}{\zeta_l}\right)\\
&=\frac{1}{2}\sum_i\sum_{l\neq i}\log\left(\frac{\eta_i}{\zeta_i}\right)\log\left(\frac{\eta_l}{\zeta_l}\right)
\end{aligned}
$$

$$\sum_i F(-\eta_i, -\zeta_i; -\eta_i) - F(\eta_i, \zeta_i; \zeta_i) = \frac{1}{2} \sum_i \log^2 \left(\frac{\eta_i}{\zeta_i} \right)$$

so that

$$\sum_i \sum_{l \neq i} F(\eta_i, \zeta_i; \eta_l) - F(\eta_i, \zeta_i; \zeta_l) + \sum_i F(-\eta_i, -\zeta_i; -\eta_i) - F(\eta_i, \zeta_i; \zeta_i)$$
$$= \frac{1}{2} \left(\sum_i \log \left(\frac{\eta_i}{\zeta_i} \right) \right)^2 .$$

Recall now the already established identity $\sum_i \log(\frac{\eta_i}{\zeta_i}) = -\log((1-c_1)(1-c_2))$ from which

$$\sum_i \sum_{l \neq i} F(\eta_i, \zeta_i; \eta_l) - F(\eta_i, \zeta_i; \zeta_l) + \sum_i F(-\eta_i, -\zeta_i; -\eta_i) - F(\eta_i, \zeta_i; \zeta_i)$$
$$= \frac{1}{2} \log^2((1-c_1)(1-c_2)).$$

Continuing, we also have

$$\sum_{j \neq i} \log \left(\frac{\eta_i - \hat{\lambda}_j}{\zeta_i - \hat{\lambda}_j} \right) = \log \left(-\frac{c_1}{c_2(1-c_1)} \right) - \log \left(\frac{\eta_i - \hat{\lambda}_i}{\zeta_i - \hat{\lambda}_i} \right) = \log \left(\frac{c_1}{c_2(1-c_1)} \frac{\eta_i - \hat{\lambda}_i}{\hat{\lambda}_i - \zeta_i} \right).$$

By Lemma 4 again, we next find

$$\sum_i \sum_{l>i} F(-\eta_i, -\zeta_i; -\eta_l) - F(-\eta_i, -\zeta_i; -\zeta_l) + \sum_{l<i} F(\eta_i, \zeta_i; \eta_l) - F(\eta_i, \zeta_i; \zeta_l)$$
$$= \sum_i \sum_{l<i} \log \left(\frac{\eta_i}{\zeta_i} \right) \log \left(\frac{\eta_l}{\zeta_l} \right)$$

from which we deduce that

$$\sum_i F(-\eta_i, -\zeta_i, -\eta_i) - F(\eta_i, \zeta_i; \zeta_i) + \sum_i \sum_{l>i} F(-\eta_i, -\zeta_i; -\eta_l)$$
$$- F(-\eta_i, -\zeta_i; -\zeta_l) + \sum_{l<i} F(\eta_i, \zeta_i; \eta_l) - F(\eta_i, \zeta_i; \zeta_l)$$
$$= \frac{1}{2} \sum_{i,l} \log \left(\frac{\eta_i}{\zeta_i} \right) \log \left(\frac{\eta_l}{\zeta_l} \right)$$
$$= \frac{1}{2} \log^2 \left((1-c_1)(1-c_2) \right).$$

The next term also simplifies through the definition of φ_p/ψ_p:

$$\sum_i \sum_{j \neq i} \log \left| \frac{\eta_i - \eta_j}{\zeta_i - \eta_j} \right| = \sum_j \log \left(\frac{c_1}{c_1 + c_2 - c_1 c_2} \varphi_p'(\eta_j) \frac{(\zeta_j - \eta_j)}{(1-c_1)\eta_j} \right).$$

Still from Lemma 4 and with the same connection to φ_p/ψ_p, we have

$$\begin{aligned}
&\sum_j \sum_{l<i} F(\eta_i-\eta_j,\zeta_i-\eta_j;\eta_l-\eta_j) + \sum_j \sum_{l>i} F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\eta_l+\eta_j)\\
&+ \sum_{j\neq i} F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\eta_i+\eta_j) - \sum_j \sum_{l<i} F(\eta_i-\eta_j,\zeta_i-\eta_j;\zeta_l-\eta_j)\\
&- \sum_j \sum_{l>i} F(-\eta_i+\eta_j,-\zeta_i+\eta_j;-\zeta_l+\eta_j) - \sum_{j\neq i} F(\eta_i-\eta_j,\zeta_i-\eta_j;\zeta_i-\eta_j)\\
&= \sum_i \sum_{j\neq(i,l)} \sum_{l<i} \log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)\log\left(\frac{\eta_l-\eta_j}{\zeta_l-\eta_j}\right) + \frac{1}{2}\log^2\left(\frac{\eta_j-\eta_i}{\eta_j-\zeta_i}\right)\\
&= \frac{1}{2}\sum_i \sum_{j\neq(i,l)} \sum_{l\neq i} \log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)\log\left(\frac{\eta_l-\eta_j}{\zeta_l-\eta_j}\right) + \frac{1}{2}\log^2\left(\frac{\eta_j-\eta_i}{\eta_j-\zeta_i}\right)\\
&= \frac{1}{2}\sum_i \sum_l \sum_{j\neq(i,l)} \log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)\log\left(\frac{\eta_l-\eta_j}{\zeta_l-\eta_j}\right)\\
&= \frac{1}{2}\sum_j \log^2\left(\frac{c_1}{c_1+c_2-c_1c_2}\varphi'(\eta_j)\frac{(\zeta_j-\eta_j)}{(1-c_1)\eta_j}\right).
\end{aligned}$$

Again from Lemma 4, we next have

$$\begin{aligned}
&\sum_i \sum_{l<i} F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\eta_l-\eta_i) - F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\zeta_l-\eta_i)\\
&- \sum_{l>i} F(-\eta_i+\eta_l,-\zeta_i+\eta_l;-\zeta_l+\eta_l)\\
&= \sum_i \sum_{l<i} \log\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)\log\left(\frac{\eta_i-\eta_l}{\eta_i-\zeta_l}\right) - \frac{1}{2}\log^2(\eta_i-\eta_l) + \frac{1}{2}\log^2(\eta_i-\zeta_l) + o_\varepsilon(1).
\end{aligned}$$

We also have the following relations

$$\begin{aligned}
&\sum_i \sum_{l>i} F(\varepsilon,\eta_i-\zeta_i-\varepsilon;\eta_i-\eta_l) - F(\varepsilon,\eta_i-\zeta_i-\varepsilon;\eta_i-\zeta_l)\\
&- \sum_{l<i} F(\eta_i-\eta_l,\zeta_i-\eta_l;\zeta_l-\eta_l)\\
&= \sum_i \sum_{l>i} \frac{1}{2}\log^2(\zeta_l-\eta_i) - \frac{1}{2}\log^2(\eta_l-\eta_i) + \log\left(\frac{\eta_l-\eta_i}{\zeta_l-\eta_i}\right)\log\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right) + o_\varepsilon(1)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{l<i} F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\eta_l-\eta_i) - F(-\varepsilon,\zeta_i-\eta_i+\varepsilon;\zeta_l-\eta_i)\\
&- \sum_{l>i} F(-\eta_i+\eta_l,-\zeta_i+\eta_l;-\zeta_l+\eta_l)
\end{aligned}$$

$$= \sum_i \sum_{l<i} \log\left(\frac{\varepsilon}{\eta_i - \zeta_i}\right) \log\left(\frac{\eta_i - \eta_l}{\eta_i - \zeta_l}\right) - \frac{1}{2}\log^2(\eta_i - \eta_l) + \frac{1}{2}\log^2(\eta_i - \zeta_l) + o_\varepsilon(1)$$

which together gives

$$\begin{aligned}
&\sum_i \sum_{l>i} F(\varepsilon, \eta_i - \zeta_i - \varepsilon; \eta_i - \eta_l) - F(\varepsilon, \eta_i - \zeta_i - \varepsilon; \eta_i - \zeta_l) \\
&- \sum_{l<i} F(\eta_i - \eta_l, \zeta_i - \eta_l; \zeta_l - \eta_l) + \sum_{l<i} F(-\varepsilon, \zeta_i - \eta_i + \varepsilon; \eta_l - \eta_i) \\
&- F(-\varepsilon, \zeta_i - \eta_i + \varepsilon; \zeta_l - \eta_i) - \sum_{l>i} F(-\eta_i + \eta_l, -\zeta_i + \eta_l; -\zeta_l + \eta_l) \\
&= \sum_i \sum_{l\neq i} \log\left(\frac{\varepsilon}{\eta_i - \zeta_i}\right) \log\left(\frac{\eta_i - \eta_l}{\eta_i - \zeta_l}\right).
\end{aligned}$$

The next term is

$$F(\varepsilon, -\zeta_i + \eta_i - \varepsilon, \varepsilon) - F(-\varepsilon, \zeta_i - \eta_i + \varepsilon, \zeta_i - \eta_i) = \frac{1}{2}\log^2\left(\frac{\varepsilon}{\eta_i - \zeta_i}\right)$$

and finally the last term gives

$$\begin{aligned}
&\sum_{l<i} F(\eta_i - \eta_l, \zeta_i - \eta_l; 0) + \sum_{l>i} F(-\eta_i + \eta_l, -\zeta_i + \eta_l; 0) \\
&= \frac{1}{2}\sum_i \sum_{l\neq i} \log^2|\eta_i - \eta_l| - \log^2|\zeta_i - \eta_l|.
\end{aligned}$$

Putting all results above together, we obtain

$$\begin{aligned}
&\sum_{i=1}^{p} 2\left[\int_{\zeta_i+\varepsilon}^{\hat\lambda_i-\varepsilon} + \int_{\hat\lambda_i+\varepsilon}^{\eta_i-\varepsilon}\right] \log\left(-\frac{\varphi_p(x)}{\psi_p(x)}\right)\left(\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)}\right)\frac{\psi_p(x)}{c_2}dx \\
&= 2\left(\frac{1}{p} - \frac{c_1 + c_2 - c_1c_2}{c_1c_2}\right)\left(p\log(1-c_1)\log\left(\frac{c_1}{c_2(1-c_1)}\right) + \sum_i \mathrm{Li}_2\left(1 - \frac{\zeta_i}{\hat\lambda_i}\right)\right. \\
&-\mathrm{Li}_2\left(1 - \frac{\eta_i}{\hat\lambda_i}\right) + \log(\hat\lambda_i)\log\left(\frac{\eta_i - \hat\lambda_i}{\hat\lambda_i - \zeta_i}\right) \\
&\left.+ \sum_i \sum_{j\neq i} F(\eta_i - \hat\lambda_j, \zeta_i - \hat\lambda_j; -\hat\lambda_j) + \frac{1}{2}p\log^2\left(\frac{c_1}{c_2(1-c_1)}\right) - \frac{\pi^2}{2}p\right) \\
&+ 2\frac{1-c_2}{c_2}\left(-\log(1-c_1)\log\left((1-c_1)(1-c_2)\right) + \sum_i \frac{1}{2}\log^2(\eta_i)\right. \\
&\left.-\frac{1}{2}\log^2(\zeta_i) + \frac{1}{2}\log^2\left((1-c_1)(1-c_2)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&+2\frac{c_1+c_2-c_1c_2}{c_1c_2}\left(\log(1-c_1)\sum_i\sum_{j\neq i}\log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)+\sum_i\sum_{j\neq i}F(\eta_i-\eta_j,\zeta_i-\eta_j;-\eta_j)\right.\\
&+\frac{1}{2}\sum_i\sum_l\sum_{j\neq(i,l)}\log\left(\frac{\eta_i-\eta_j}{\zeta_i-\eta_j}\right)\log\left(\frac{\eta_l-\eta_j}{\zeta_l-\eta_j}\right)+\sum_i\log(1-c_1)\log\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)\\
&+\sum_i\sum_{l\neq i}\log\left|\frac{\eta_l-\eta_i}{\zeta_l-\eta_i}\right|\log\left|\frac{\varepsilon}{\eta_i-\zeta_i}\right|\\
&\left.+\sum_i\mathrm{Li}_2\left(1-\frac{\zeta_i}{\eta_i}\right)+\log\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)\log(\eta_i)+\sum_i\frac{1}{2}\log^2\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)-p\frac{\pi^2}{6}\right)+o_\varepsilon(1).
\end{aligned}
$$

The integral over the contour I_i^E can be computed using the same reasoning as for the $f(t)=\log(t)$ function and is easily obtained as

$$
\begin{aligned}
\frac{1}{2\pi\imath}\int_{I_i^E}=\frac{c_1+c_2-c_1c_2}{c_1c_2}&\left(-\log^2\left(\varphi_p'(\eta_i)\frac{c_1}{c_1+c_2-c_1c_2}\right)-\log^2(\varepsilon)\right.\\
&\left.-2\log(\varepsilon)\left[\left(\frac{\varphi}{\psi}\right)'(\eta_i)\right]-\frac{\pi^2}{3}\right)+O(\varepsilon\log^2(\varepsilon)).
\end{aligned}
$$

Adding up the "residue" at $\hat{\lambda}_i$ (i.e., the integral over I_i^C), we end up with the following expression for the sought-for integral

$$
\begin{aligned}
&\frac{1}{2\pi\imath}\oint_\Gamma\log^2\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz\\
&=-2\left(\frac{1}{p}-\frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\left(p\log(1-c_1)\log\left(\frac{c_1}{c_2(1-c_1)}\right)\right.\\
&+\sum_i\mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_i}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_i}\right)+\log(\hat{\lambda}_i)\log\left(\frac{\eta_i-\hat{\lambda}_i}{\hat{\lambda}_i-\zeta_i}\right)\\
&\left.+\sum_i\sum_{j\neq i}F(\eta_i-\hat{\lambda}_j,\zeta_i-\hat{\lambda}_j;-\hat{\lambda}_j)+p\log^2\left(\frac{c_1}{c_2(1-c_1)}\right)-\frac{\pi^2}{2}p\right)\\
&-2\frac{1-c_2}{c_2}\left(-\log(1-c_1)\log\left((1-c_1)(1-c_2)\right)+\sum_i\frac{1}{2}\log^2(\eta_i)-\frac{1}{2}\log^2(\zeta_i)\right.\\
&\left.+\frac{1}{2}\log^2\left((1-c_1)(1-c_2)\right)\right)\\
&-2\frac{c_1+c_2-c_1c_2}{c_1c_2}\left(\log(1-c_1)\sum_j\log\left(\frac{c_1}{c_1+c_2-c_1c_2}\varphi'(\eta_j)\frac{(\zeta_j-\eta_j)}{(1-c_1)\eta_j}\right)\right.\\
&\left.+\sum_i\sum_{j\neq i}F(\eta_i-\eta_j,\zeta_i-\eta_j;-\eta_j)\right.
\end{aligned}
$$

$$\frac{1}{2}\sum_j \log^2\left(\frac{c_1}{c_1+c_2-c_1c_2}\varphi'(\eta_j)\frac{(\zeta_j-\eta_j)}{(1-c_1)\eta_j}\right)+\sum_i \mathrm{Li}_2\left(1-\frac{\zeta_i}{\eta_i}\right)$$
$$+\sum_i \log\left[\left(\frac{\varphi}{\psi}\right)'(\eta_i)\right]\log\left|\frac{\varepsilon}{\eta_i-\zeta_i}\right|+\log(\eta_i-\zeta_i)\log\left|\frac{\varepsilon}{\eta_i-\zeta_i}\right|$$
$$+\sum_i \frac{1}{2}\log^2\left(\frac{\varepsilon}{\eta_i-\zeta_i}\right)-\frac{\pi^2}{6}\Bigg)+o_\varepsilon(1)$$
$$+\sum_i \frac{c_1+c_2-c_1c_2}{c_1c_2}\Bigg(-\log^2\left(\varphi_p'(\eta_i)\frac{c_1}{c_1+c_2-c_1c_2}\right)-\log^2(\varepsilon)$$
$$-2\log(\varepsilon)\log\left[\left(\frac{\varphi}{\psi}\right)'(\eta_i)\right]-\frac{\pi^2}{3}\Bigg)$$
$$+\sum_i\left(\log^2\left(\frac{c_1}{c_2}\hat{\lambda}_i\right)-\pi^2\right)\left[\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right]+o_\varepsilon(1).$$

which, in the limit of small ε, can be simplified as

$$\frac{1}{2\pi\imath}\oint_\Gamma \log^2\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz$$
$$=-2\left(\frac{1}{p}-\frac{c_1+c_2-c_1c_2}{c_1c_2}\right)\Bigg(p\log(1-c_1)\log\left(\frac{c_1}{c_2(1-c_1)}\right)$$
$$+\sum_i \mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_i}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_i}\right)+\log(\hat{\lambda}_i)\log\left(\frac{\eta_i-\hat{\lambda}_i}{\hat{\lambda}_i-\zeta_i}\right)$$
$$+\sum_i\sum_{j\neq i}F(\eta_i-\hat{\lambda}_j,\zeta_i-\hat{\lambda}_j;-\hat{\lambda}_j)+p\log^2\left(\frac{c_1}{c_2(1-c_1)}\right)\Bigg)$$
$$-2\frac{1-c_2}{c_2}\Bigg(-\log(1-c_1)\log\left((1-c_1)(1-c_2)\right)+\sum_i\frac{1}{2}\log^2(\eta_i)-\frac{1}{2}\log^2(\zeta_i)$$
$$+\frac{1}{2}\log^2\left((1-c_1)(1-c_2)\right)\Bigg)$$
$$-2\frac{c_1+c_2-c_1c_2}{c_1c_2}\Bigg(\log(1-c_1)\sum_j\log\left(\frac{c_1}{c_1+c_2-c_1c_2}\varphi'(\eta_j)\frac{(\zeta_j-\eta_j)}{(1-c_1)\eta_j}\right)$$
$$+\sum_i\sum_{j\neq i}F(\eta_i-\eta_j,\zeta_i-\eta_j;-\eta_j)$$
$$-\sum_i\log\left[\left(\frac{\varphi'}{\psi}(\eta_i)\right)\right]\log\left(\eta_i-\zeta_i\right)+\sum_i\mathrm{Li}_2\left(1-\frac{\zeta_i}{\eta_i}\right)-\frac{1}{2}\log^2\left(\eta_i-\zeta_i\right)\Bigg)$$
$$-\sum_i\frac{c_1+c_2-c_1c_2}{c_1c_2}\Bigg(-\log^2\left(\frac{\eta_i-\zeta_i}{(1-c_1)\eta_i}\right)$$

$$+2\log\left(\frac{\eta_i-\zeta_i}{(1-c_1)\eta_i}\right)\log\left(\frac{c_1}{c_1+c_2-c_1c_2}\varphi'(\eta_i)\frac{(\zeta_i-\eta_i)}{(1-c_1)\eta_i}\right)\Bigg)$$
$$-\sum_i \log^2\left(\frac{c_1}{c_2}\hat{\lambda}_i\right)\left[\frac{c_1+c_2-c_1c_2}{c_1c_2}-\frac{1}{p}\right].$$

After further book-keeping and simplifications, we ultimately find:

$$\begin{aligned}
&\frac{1}{2\pi\imath}\oint_\Gamma \log^2\left(\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz \qquad (6.16)\\
&=\frac{c_1+c_2-c_1c_2}{c_1c_2}\left[\sum_{i=1}^{p}\left\{\log^2\left((1-c_1)\eta_i\right)-\log^2\left((1-c_1)\hat{\lambda}_i\right)\right\}\right.\\
&\left.+2\sum_{1\le i,j\le p}\left\{\mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_j}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_j}\right)+\mathrm{Li}_2\left(1-\frac{\eta_i}{\eta_j}\right)-\mathrm{Li}_2\left(1-\frac{\zeta_i}{\eta_j}\right)\right\}\right]\\
&-\frac{1-c_2}{c_2}\left[\log^2(1-c_2)-\log^2(1-c_1)+\sum_{i=1}^{p}\left\{\log^2(\eta_i)-\log^2(\zeta_i)\right\}\right]\\
&-\frac{1}{p}\left[2\sum_{1\le i,j\le p}\left\{\mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_j}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_j}\right)\right\}-\sum_{i=1}^{p}\log^2\left((1-c_1)\hat{\lambda}_i\right)\right]
\end{aligned}$$

which provides an exact, yet rather impractical (the expression involves the evaluation of $O(p^2)$ dilogarithm terms which may be computationally intense for large p), final expression for the integral.

At this point, it is also not easy to fathom why the retrieved expression would remain of order $O(1)$ with respect to p. In order to both simplify the expression and retrieve a visually clear $O(1)$ estimate, we next proceed to a large p Taylor expansion of the above result. In particular, using the last item in Lemma 3, we perform a (second-order) Taylor expansion of all terms of the type $\mathrm{Li}_2(1-X)$ above in the vicinity of $\hat{\lambda}_i/\hat{\lambda}_j$. This results in the following two relations

$$\begin{aligned}
&\sum_{i,j}\mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_j}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_j}\right)+\mathrm{Li}_2\left(1-\frac{\eta_i}{\eta_j}\right)-\mathrm{Li}_2\left(1-\frac{\zeta_i}{\eta_j}\right)\\
&=(\Delta_\zeta^\eta)^T M(\Delta_{\hat{\lambda}}^\eta)+o_p(1)\\
&\frac{1}{p}\sum_{i,j}\mathrm{Li}_2\left(1-\frac{\zeta_i}{\hat{\lambda}_j}\right)-\mathrm{Li}_2\left(1-\frac{\eta_i}{\hat{\lambda}_j}\right)\\
&=-\frac{1}{p}(\Delta_\zeta^\eta)^T N 1_p+o_p(1)
\end{aligned}$$

with Δ_a^b, M and N defined in the statement of Corollary 5.

With these developments, we deduce the final approximation

$$\begin{aligned}
&\frac{1}{2\pi\imath}\oint_\Gamma \log^2\left(-\frac{\varphi_p(z)}{\psi_p(z)}\right)\left(\frac{\varphi_p'(z)}{\varphi_p(z)}-\frac{\psi_p'(z)}{\psi_p(z)}\right)\frac{\psi_p(z)}{c_2}dz\\
&=2\frac{c_1+c_2-c_1c_2}{c_1c_2}\left(\left(\Delta_\zeta^\eta\right)^T M\left(\Delta_{\hat\lambda}^\eta\right)+\sum_i\frac{\log\left((1-c_1)\hat\lambda_i\right)}{\hat\lambda_i}\left(\eta_i-\hat\lambda_i\right)\right)\\
&-\frac{2}{p}\left(\Delta_\zeta^\eta\right)^T N1_p+\frac{1}{p}\sum_i\log^2\left((1-c_1)\hat\lambda_i\right)\\
&-2\frac{1-c_2}{c_2}\left(\frac{1}{2}\log^2(1-c_2)-\frac{1}{2}\log^2(1-c_1)+\sum_i(\eta_i-\zeta_i)\frac{\log(\hat\lambda_i)}{\hat\lambda_i}\right)+o_p(1).
\end{aligned}$$

For symmetry, it is convenient to finally observe that $\log(1-c_1)\sum_i(\eta_i-\zeta_i)/\hat\lambda_i \sim \log(1-c_1)\sum_i\log(\eta_i/\zeta_i) = -\log^2(1-c_1)$; replacing in the last parenthesis provides the result of Corollary 5 for $c_1>0$.

To determine the limit as $c_1\to 0$, it suffices to remark that in this limit $\eta_i = \hat\lambda_i+\frac{c_1}{p}\hat\lambda_i+o(c_1)$ (this can be established using the functional relation $\varphi_p(\eta_i)=0$ in the small c_1 limit). Thus it suffices to replace in the above expression the vector $\eta-\zeta$ by the vector $\eta-\hat\lambda$, the vector $\frac{c_1+c_2-c_1c_2}{c_1c_2}(\eta-\hat\lambda)$ by the vector $\frac{1}{p}\hat\lambda$, and taking $c_1=0$ in all other instances (where the limits for $c_1\to 0$ are well defined).

6.2.4 Gradient calculus

The main concern is to find an analytical expression of the gradient defined as:

$$\nabla h_X(M) = \frac{-\hat D(M,X)}{\pi\imath p}\oint_{\hat\Gamma} g\left(-m_{\tilde\mu_p}(z;M)\right)\mathrm{sym}\left(\hat\Sigma(M^{-1}\hat\Sigma-zI_p)^{-2}\right)dz. \tag{6.17}$$

Equation (6.17) can be written as:

$$\nabla h_X(M) = \frac{-\hat D(M,X)}{\pi\imath p}\mathrm{sym}\left(\hat\Sigma U\left(\oint_{\hat\Gamma} g\left(-m_{\tilde\mu_p}(z;M)\right)(\Lambda-zI_p)^{-2}dz\right)U^\mathsf{T}\right). \tag{6.18}$$

where $M^{-1}\hat\Sigma = U\Lambda U^\mathsf{T}$ in its spectral decomposition. Our main focus is on the diagonal matrix

$$A\equiv\frac{1}{2\imath\pi}\oint_{\hat\Gamma} g\left(-m_{\tilde\mu_p}(z;M)\right)(\Lambda-zI_p)^{-2}dz$$

and particularly on its k-th diagonal element

$$A_{kk} = \frac{1}{2\imath\pi}\oint_{\hat\Gamma}\frac{g\left(-m_{\tilde\mu_p}(z;M)\right)}{(\hat\lambda_k-z)^2}dz \tag{6.19}$$

with $\Lambda = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$.

To solve (6.19), we use elementary properties of the rational function $m_{\tilde{\mu}_p}(z; M)$ that we recall is defined as

$$m_{\tilde{\mu}_p}(z; M) = \frac{c}{p} \sum_{i=1}^{p} \frac{1}{\hat{\lambda}_i - z} + \frac{1-c}{z}.$$

Remarking that the poles of $m_{\tilde{\mu}_p}(z; M)$ are $\{\hat{\lambda}_i\}_{i=1}^p$ and 0, and that $\{\xi_i\}_{i=1}^p$ are the zeros of $m_{\tilde{\mu}_p}(z; M)$ (see Section 6.2.3 of the appendix for details), we have :

$$m_{\tilde{\mu}_p}(z; M) = \frac{\prod_{i=1}^{p} (z - \xi_i)}{z \prod_{i=1}^{p} \left(z - \hat{\lambda}_i\right)}.$$

With these ingredients, we can evaluate A_{kk} for various functions f (recall that $g(z) = f(1/z)$).

Case $f(t) = t$ For this case, $g(t) = \frac{1}{t}$, and thus the integrand of A_{kk} is

$$I = \frac{z \prod_{\substack{i=1 \\ i \neq k}}^{p} \left(z - \hat{\lambda}_i\right)}{(z - \hat{\lambda}_k) \prod_{i=1}^{p} (z - \xi_i)}.$$

Under this rational form, A_{kk} is easy to evaluate since it only requires to evaluate the residue for each pole of I:

- the first-order pole $\hat{\lambda}_k$ for which the residue R_1 is given by

$$\begin{aligned} R_1 &= \lim_{z \to \hat{\lambda}_k} \frac{z \prod_{\substack{i=1 \\ i \neq k}}^{p} \left(z - \hat{\lambda}_i\right)}{\prod_{i=1}^{p} (z - \xi_i)} \\ &= \lim_{z \to \hat{\lambda}_k} \frac{1}{(z - \hat{\lambda}_k) m_{\tilde{\mu}_p}(z)} \\ &= -\frac{p}{c} \end{aligned}$$

- and the first-order poles ξ_j, $j \in \{1, \ldots, p\}$ for which the residue R_2 is given by:

$$\begin{aligned} R_2 &= \sum_{j=1}^{p} \lim_{z \to \xi_j} \frac{z \prod_{i=1}^{p} \left(z - \hat{\lambda}_i\right)}{(z - \hat{\lambda}_k)^2 \prod_{\substack{i=1 \\ i \neq j}}^{p} (z - \xi_i)} \\ &= \sum_{j=1}^{p} \frac{1}{(\xi_j - \hat{\lambda}_k)^2} \lim_{z \to \xi_j} \frac{z - \xi_j}{m_{\tilde{\mu}_p}(z)} \\ &= \sum_{j=1}^{p} \frac{1}{(\xi_j - \hat{\lambda}_k)^2 m'_{\tilde{\mu}_p}(\xi_j)} \end{aligned}$$

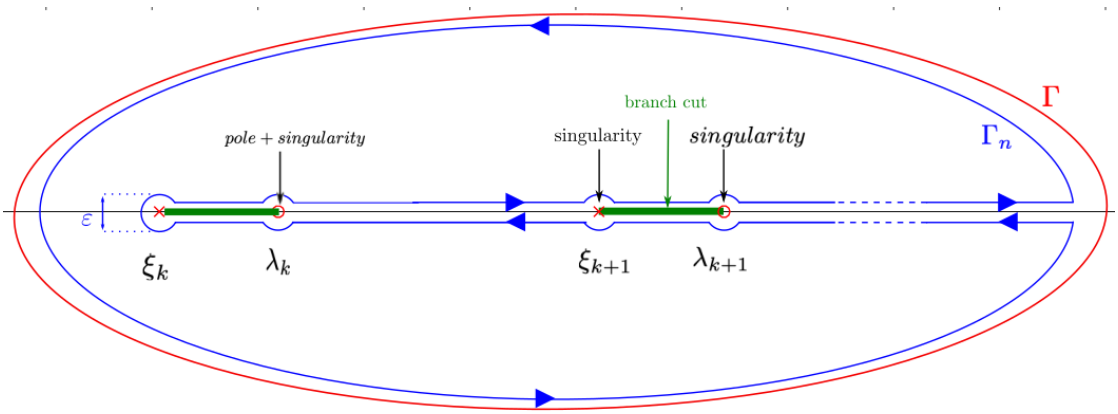

Figure 6.8: Contour deformation

Putting the $p+1$ residues together then yields:

$$A_{kk} = -\frac{p}{c} + \sum_{j=1}^{p} \frac{1}{(\xi_j - \hat{\lambda}_k)^2 m'_{\tilde{\mu}_p}(\xi_j)}.$$

Case $f(t) = \log(t)$ For this case, $g(t) = -\log(t)$ and therefore the integrand of A_{kk} becomes

$$I = -\frac{\log\left(\frac{\prod_{i=1}^{p}(z-\xi_i)}{z\prod_{i=1}^{p}(z-\hat{\lambda}_i)}\right)}{(\hat{\lambda}_k - z)^2}$$

Elementary functional analysis allows us to find the discontinuities of this multi-valued function (the z's for which the argument of the logarithm function is negative). This set of points, or branch cuts, are exactly the segments $[\xi_i, \hat{\lambda}_i]$, $i = 1, \ldots, p$. These segments lie inside the integration contour Γ, that needs be modified for proper integration; the new contour, denoted Γ_n is depicted in Figure 6.8.

The complex integral over the contour Γ_n, is the sum of several integrals, subdivided in four types:

- integrals over the circles surrounding $\{\xi_j\}_{j=1}^{p}$ which, thanks to the variable change $z = \xi_j + \epsilon\, \mathrm{e}^{\imath\theta}$, reduce to

$$\lim_{\epsilon\to 0} -\int_{\epsilon}^{2\pi-\epsilon} \frac{\log\left(\epsilon\, \mathrm{e}^{\imath\theta} \frac{\prod_{\substack{i=1\\ i\neq j}}^{p}(\xi_j+\epsilon\, \mathrm{e}^{\imath\theta}-\xi_i)}{(\xi_j+\epsilon\, \mathrm{e}^{\imath\theta})\prod_{i=1}^{p}(\xi_j+\epsilon\, \mathrm{e}^{\imath\theta}-\hat{\lambda}_i)}\right)}{(\hat{\lambda}_k - \xi_j - \epsilon\, \mathrm{e}^{\imath\theta})^2} \imath\epsilon\, \mathrm{e}^{\imath\theta} = 0$$

- integrals over the circles surrounding $\{\hat{\lambda}_i\}_{\substack{i=1\\ i\neq k}}^{p}$ which are null following the same line of reasoning as for ξ_i.

- real integrals over the segments $[\xi_i, \hat{\lambda}_i]$ which can be computed by remarking that the log function has a discontinuity of $2\imath\pi$ at the branch cut.

$$\begin{aligned}
&\frac{1}{2\imath\pi}\sum_{j=1}^{p}\int_{\xi_j+\epsilon}^{\hat{\lambda}_j-\epsilon}\frac{\log\left(|m_{\tilde{\mu}_p}|\right)+\imath\pi-\log\left(|m_{\tilde{\mu}_p}|\right)+\imath\pi}{(\hat{\lambda}_k-z)^2}\\
&=\sum_{j=1}^{p}\int_{\xi_j+\epsilon}^{\hat{\lambda}_j-\epsilon}\frac{1}{(\hat{\lambda}_k-z)^2}\\
&=\sum_{j=1}^{p}\lim_{z\to\hat{\lambda}_j}\frac{1}{(\hat{\lambda}_k-z)}-\frac{1}{(\hat{\lambda}_k-\xi_j)}\\
&=\sum_{\substack{j=1\\ j\neq k}}^{p}\frac{1}{(\hat{\lambda}_k-\hat{\lambda}_j)}-\sum_{j=1}^{p}\frac{1}{(\hat{\lambda}_k-\xi_j)}+\lim_{z\to\hat{\lambda}_k}\frac{1}{(\hat{\lambda}_k-z)}
\end{aligned}$$

- the integral over the circle surrounding $\hat{\lambda}_k$ computed by remarking that $\hat{\lambda}_k$ is a second-order pole

$$\begin{aligned}
&\lim_{z\to\hat{\lambda}_k}\frac{\partial}{\partial z}\left(\log\left(-m_{\tilde{\mu}_p}(z;M)\right)\right)\\
&=\lim_{z\to\hat{\lambda}_k}\sum_{j=1}^{p}\frac{1}{z-\xi_j}-\frac{1}{z}-\sum_{j=1}^{p}\frac{1}{z-\hat{\lambda}_j}\\
&=\sum_{j=1}^{p}\frac{1}{\hat{\lambda}_k-\xi_j}-\frac{1}{\hat{\lambda}_k}-\sum_{\substack{j=1\\ j\neq k}}^{p}\frac{1}{\hat{\lambda}_k-\hat{\lambda}_j}-\lim_{z\to\hat{\lambda}_k}\frac{1}{z-\hat{\lambda}_k}
\end{aligned}$$

where the second line is obtained remarking that:

$$\frac{m'_{\tilde{\mu}_p}(z;M)}{m_{\tilde{\mu}_p}(z;M)}=\sum_{j=1}^{p}\frac{1}{z-\xi_j}-\frac{1}{z}-\sum_{j=1}^{p}\frac{1}{z-\hat{\lambda}_j}.$$

Combining these integrals then yields to the solution of the integral:

$$A_{kk}=-\frac{1}{\hat{\lambda}_k}.$$

Case $f(t)=\log(1+st)$ For this case, the integrand of A_{kk} can be derived similarly as in the case of the logarithm by noting that the argument of the logarithm $(1-s/m_{\tilde{\mu}_p}(z))$ is a polynomial for which the poles are $\hat{\lambda}_i$ and 0. The zeros are in number $p+1$ and denoted κ_i, $i=0,\ldots,p$ with $\kappa_0<0<\kappa_1<\ldots<\kappa_p$; in particular, only $\kappa_1,\ldots,\kappa_p$ are

inside the integration contour (see Section 6.2.3 of the appendix for details). Therefore, the integrand is written similarly as for the log function as:

$$I = -\frac{\log\left(\frac{\prod_{i=0}^{p}(z-\kappa_i)}{z\prod_{i=1}^{p}(z-\hat{\lambda}_i)}\right)}{(\hat{\lambda}_k - z)^2}.$$

The integration contour can be deformed as for the log function. Using similar integration techniques, the calculus then yields to the solution derived in Section 3.5.3.

Case $f(t) = \log^2(t)$ Here $g(t) = f(t) = \log^2(t)$ and the integrand for this case is simply

$$I = -\frac{\log^2\left(\frac{\prod_{i=1}^{p}(z-\xi_i)}{z\prod_{i=1}^{p}(z-\hat{\lambda}_i)}\right)}{(\hat{\lambda}_k - z)^2}.$$

Again, we use here exactly the same line of work performed on the $\log(t)$ and $\log(1+st)$ functions. Technical difficulties however arise when addressing the real integrals which involve products of logarithms and rational functions. These difficulties are mostly cumbersome calculus which are addressed similar to 6.2.3.

6.3 Appendix for Chapter 4

6.3.1 Solution of MTL LS-SVM

The Lagrangian of the constrained optimization problem using the relatedness assumption ($W_i = W_0 + V_i$) reads:

$$\begin{aligned}\mathscr{L}(\omega_0, v_i, \xi_i, \alpha_i, b_i) = \frac{1}{2\lambda}\mathrm{tr}\left(W_0^\mathsf{T} W_0\right) + \frac{1}{2}\sum_{i=1}^{k}\frac{\mathrm{tr}\left(V_i^\mathsf{T} V_i\right)}{\gamma_i} + \frac{1}{2}\sum_{i=1}^{k}\mathrm{tr}\left(\xi_i^\mathsf{T}\xi_i\right)\\ + \sum_{i=1}^{k}\mathrm{tr}\left(\alpha_i^\mathsf{T}\left(Y_i - \frac{\mathring{X}_i^\mathsf{T} W_0}{\sqrt{kp}} - \frac{\mathring{X}_i^\mathsf{T} V_i}{\sqrt{kp}} - \mathbb{1}_{n_i} b_i^\mathsf{T} - \xi_i\right)\right)\end{aligned}$$

with $\alpha_i \in \mathbb{R}^{n_i \times m}$ the Lagrangian parameter attached to task i.

Differentiating with respect to the unknowns W_0, V_i, ξ_i, α_i, and b_i leads to the following system of equations:

$$\frac{1}{\lambda}W_0 - \sum_{i=1}^{k}\frac{\mathring{X}_i}{\sqrt{kp}}\alpha_i = 0 \tag{6.20}$$

$$\frac{1}{\gamma_i}V_i - \frac{\mathring{X}_i}{\sqrt{kp}}\alpha_i = 0 \tag{6.21}$$

$$\xi_i - \alpha_i = 0 \tag{6.22}$$

$$Y_i - \frac{\mathring{X}_i^\mathsf{T} W_0}{\sqrt{kp}} - \frac{\mathring{X}_i^\mathsf{T} V_i}{\sqrt{kp}} - \mathbb{1}_{n_i} b_i^\mathsf{T} - \xi_i = 0 \tag{6.23}$$

$$\alpha_i^\mathsf{T} \mathbb{1}_{n_i} = 0. \tag{6.24}$$

Plugging the expression of W_0 (Equation (6.20)), V_i (Equation (6.21)) and ξ_i (Equation (6.22)) into Equation (6.23) leads to:

$$Y_i = (\lambda + \gamma_i) \frac{\mathring{X}_i^\mathsf{T} \mathring{X}_i}{kp} \alpha_i + \lambda \sum_{j \neq i} \frac{\mathring{X}_i^\mathsf{T} \mathring{X}_j}{kp} \alpha_j + \mathbb{1}_{n_i} b_i^\mathsf{T} + \alpha_i$$
$$\mathbb{1}_{n_i}^\mathsf{T} \alpha_i = 0.$$

With $Y = [Y_1^\mathsf{T}, \ldots, Y_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^n$, $\alpha = [\alpha_1^\mathsf{T}, \ldots, \alpha_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^n$, $Z = \sum_{i=1}^k e_i^{[k]} {e_i^{[k]}}^\mathsf{T} \otimes \mathring{X}_i \in \mathbb{R}^{kp \times n}$ and $P \in \mathbb{R}^{n \times k}$ such that the j-th column is $P_{\cdot j} = [\mathbf{0}_{n_1+\ldots+n_{j-1}}^\mathsf{T}, \mathbb{1}_{n_j}^\mathsf{T}, \mathbf{0}_{n_{j+1}+\ldots+n_k}^\mathsf{T}]^\mathsf{T}$, this system of equations can be written under the following compact matrix form:

$$Pb + Q^{-1}\alpha = Y$$
$$P^\mathsf{T}\alpha = \mathbf{0}_k$$

with $Q = \left(\frac{Z^\mathsf{T} A Z}{kp} + I_n\right)^{-1} \in \mathbb{R}^{n \times n}$, and $A = \left(\mathscr{D}_\gamma + \lambda \mathbb{1}_k \mathbb{1}_k^\mathsf{T}\right) \otimes I_p \in \mathbb{R}^{kp \times kp}$.

Solving for α and b then gives:

$$\alpha = Q(Y - Pb)$$
$$b = (P^\mathsf{T} Q P)^{-1} P^\mathsf{T} Q Y.$$

Moreover, using $W_i = W_0 + V_i$ and Equations (6.20) and (6.21), the expression of W_i becomes:

$$W_i = \left({e_i^{[k]}}^\mathsf{T} \otimes I_p\right) A \frac{Z}{\sqrt{kp}} \alpha.$$

6.3.2 Calculus of deterministic equivalents

Lemma 5 (Deterministic equivalents)**.** *Define, for class $\mathscr{C}_j$ in Task i, the data deterministic matrices*

$$M = \left(e_1^{[k]} \otimes [\mu_{11}, \ldots, \mu_{1m}], \ldots, e_k^{[k]} \otimes [\mu_{k1}, \ldots, \mu_{km}]\right)$$
$$C_{ij} = A^{\frac{1}{2}} \left(e_i^{[k]} {e_i^{[k]}}^\mathsf{T} \otimes (\Sigma_{ij} + \mu_{ij}\mu_{ij}^\mathsf{T})\right) A^{\frac{1}{2}}.$$

Then we have the deterministic equivalents of first-order

$$\tilde{Q} \leftrightarrow \bar{\tilde{Q}} \equiv \left(\sum_{i=1}^k \sum_{j=1}^m \delta_{ij}^{[mk]} C_{ij} + I_{kp}\right)^{-1}$$

$$A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z \leftrightarrow A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\delta J^\mathsf{T}$$

and of second-order

$$\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q} \leftrightarrow B_{ij}$$
$$Z^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z \leftrightarrow JM_\delta^\mathsf{T}A^{\frac{1}{2}}(B_{ij}A^{\frac{1}{2}}M_\delta J^\mathsf{T} - \bar{\tilde{Q}}A^{\frac{1}{2}}M_\delta W_{ij}) + F_{ij}$$

in which we defined

$$W_{ij} = [w_{11}, \ldots, w_{km}]^\mathsf{T}, \quad w_{sl} = \left[\boldsymbol{0}^\mathsf{T}_{n_{11}+\ldots+n_{(s-1)l}}, \frac{2\delta_{sl}^{[mk]}\mathrm{tr}\,(B_{ij}C_{sl})}{n_{sl}}\mathbb{1}^\mathsf{T}_{n_{sl}}, \boldsymbol{0}^\mathsf{T}_{n_{(s+1)l}+\ldots+n_{km}}\right]^\mathsf{T}$$

$$F_{ij} = \sum_{i',j'}\frac{c_0^2{\delta_{i'j'}^{[mk]}}^2}{c_{i'j'}^2}\mathrm{tr}(C_{i'j'}B_{ij})e_{i'j'}^{[mk]}{e_{i'j'}^{[mk]}}^\mathsf{T}$$

$$B_{ij} = \bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}} + \sum_{i'=1}^{k}\sum_{j'=1}^{2}d_{i'j'}T_{ij,i'j'}[\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}}]$$

$$D = \sum_{i,j}d_{ij}e_{ij}^{[mk]}{e_{ij}^{[mk]}}^\mathsf{T}, \; d_{ij} = \frac{c_0}{c_{ij}}{\delta_{ij}^{[mk]}}^2$$

$$J = [j_{11}, \ldots, j_{km}],$$

$$j_{lm} = \left(0^\mathsf{T}_{n_{11}+\ldots+n_{(i-1)m}}, \mathbb{1}^\mathsf{T}_{n_{ij}}, 0^\mathsf{T}_{n_{(i+1)1}+\ldots+n_{km}}\right)^\mathsf{T},$$

$$M_\delta = M\sum_{ij}\frac{c_0}{c_{ij}}\delta_{ij}^{[mk]}e_{ij}^{[mk]}{e_{ij}^{[mk]}}^\mathsf{T}$$

$$S_{ij} = e_i^{[k]}{e_i^{[k]}}^\mathsf{T} \otimes \Sigma_{ij}$$

$$T = \bar{T}(I_k - D\mathscr{T})^{-1}, \; \mathscr{T}_{ij,i'j'} = \frac{1}{kp}\mathrm{tr}(C_{ij}\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}}), \bar{T}_{ij,i'j'} = \frac{1}{kp}tr\left(C_{i'j'}\bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}}\right)$$

and the $(\delta_{11}^{[mk]}, \ldots, \delta_{km}^{[mk]})$ *are the unique positive solutions of*

$$\delta_{ij}^{[mk]} = \frac{c_{ij}}{c_0\left(1 + \frac{1}{kp}\mathrm{tr}(C_{ij}\bar{\tilde{Q}})\right)}, \; \forall i, j.$$

6.3.3 Proof of Lemma 5

First-order deterministic equivalent. A deterministic equivalent for $\tilde{Q}$ is retrieved similarly as provided in (Louart & Couillet, 2018). Our objective is then to find, based on this result, a deterministic equivalent for the random matrix $A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z$. To this end, we evaluate the scalar quantity $\mathbb{E}[u^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv]$ for any deterministic vector $u \in \mathbb{R}^{kp}$ and $v \in \mathbb{R}^n$ such that $\|u\| = 1$ and $\|v\| = 1$, which we can write

$$\mathbb{E}\left[u^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv\right] = \sum_{i=1}^{n}v_i\mathbb{E}\left[u^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}z_i\right]. \tag{6.25}$$

Furthermore, let us define for convenience the matrix Z_{-i}, which is the matrix Z with a vector of $\mathbf{0}$ on its i-th column such that $ZZ^\mathsf{T} = Z_{-i}Z_{-i}^\mathsf{T} + z_i z_i^\mathsf{T}$. Using the Sherman-Morrison matrix inversion lemma (i.e., $(A+uv^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1+v^\mathsf{T}A^{-1}u}$), we find:

$$\tilde{Q} = \left(\frac{A^{\frac{1}{2}}ZZ^\mathsf{T}A^{\frac{1}{2}}}{kp} + I_{kp}\right)^{-1} = \tilde{Q}_{-i} - \frac{1}{kp}\frac{\tilde{Q}_{-i}A^{\frac{1}{2}}z_i z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}}{1+\frac{1}{kp}z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i} \tag{6.26}$$

with $\tilde{Q}_{-i} = (\frac{A^{\frac{1}{2}}Z_{-i}Z_{-i}^\mathsf{T}A^{\frac{1}{2}}}{kp} + I_{kp})^{-1}$. Furthermore,

$$\tilde{Q}A^{\frac{1}{2}}z_i = \frac{\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1+\frac{1}{kp}z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}. \tag{6.27}$$

Plugging Equation (6.27) into Equation (6.25) leads to

$$\mathbb{E}\left[u^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv\right] = \sum_{i=1}^n v_i\mathbb{E}\left[u^\mathsf{T}\frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1+\frac{1}{kp}z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}\right]. \tag{6.28}$$

Moreover, following the same line of reasoning as in (Seddik et al., 2020, Proposition A.3), based on Assumption 2 and tools from the concentration of measure theory (see also (Ledoux, 2001; Louart et al., 2018)), one can show that:

$$\sum_{i=1}^n v_i\mathbb{E}\left[u^\mathsf{T}\frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1+\frac{1}{kp}z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}\right] = \sum_{i=1}^n v_i\mathbb{E}\left[u^\mathsf{T}\frac{A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i}{1+\delta_{ij}}\right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \tag{6.29}$$

with $\delta_{ij} \equiv \mathbb{E}\left[\frac{1}{kp}z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i\right]$. Note that δ_{ij} can be estimated as the solution of the fixed point equation

$$\delta_{ij} = \frac{1}{kp}\mathbb{E}\left[\mathrm{tr}\left(A^{\frac{1}{2}}z_i z_i^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i}\right)\right] = \frac{1}{kp}\mathrm{tr}\left(C_{ij}\bar{\tilde{Q}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$$

since z_i's are independent of $\tilde{Q}_{-i}$.

We then conclude that:

$$\mathbb{E}\left[u^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Zv\right] = \sum_{i=1}^n v_i u^\mathsf{T}\frac{\mathbb{E}\left[A^{\frac{1}{2}}\tilde{Q}_{-i}A^{\frac{1}{2}}z_i\right]}{1+\delta_{ij}} + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) = u^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\delta v + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)$$

where in the last equality, we used the fact that $\tilde{Q}_{-i}$ is independent from z_i. This concludes the proof.

Second-order deterministic equivalent We aim in the following section to prove that

$$Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z \leftrightarrow J M_\delta^\mathsf{T} A^{\frac{1}{2}} (B_{ij} A^{\frac{1}{2}} M_\delta J^\mathsf{T} - \bar{\tilde{Q}} A^{\frac{1}{2}} M_\delta W_{ij}) + F_{ij}.$$

Let us define for convenience $\mathscr{C}(i)$ the class of the i-th sample. Similarly as done for the first-order deterministic equivalents, the focus will be on $\mathbb{E}[u^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z v]$.

In order to obtain an estimate of this bilinear form, or equivalently here a deterministic equivalent for $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z$, one must isolate the contribution of the off-diagonal versus diagonal elements of the latter matrix. Starting with the off-diagonal elements, using successively Equation (6.26) and Equation (6.29) on i and j , we have

$$\begin{aligned}
&\sum_{\substack{i',j'=1\\ i'\neq j'}}^{n} u_{i'} v_{j'} \mathbb{E}\left[z_i^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} z_{j'} \right] \\
&= \sum_{\substack{i',j'=1\\ i'\neq j'}}^{n} u_{i'} v_{j'} \mathbb{E}\left[\frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} \right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}} \right) \\
&= \sum_{\substack{i',j'=1\\ i'\neq j'}}^{n} u_{i'} v_{j'} \mathbb{E}\left[\frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} - \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{i'} z_{i'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} \right. \\
&\left. - \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} z_{j'} z_{j'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} \right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}} \right) \\
&= \sum_{\substack{i',j'=1\\ i'\neq j'}}^{n} u_{i'} v_{j'} \mathbb{E}\left[\frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} - \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{i'} z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(i')})} \right. \\
&- \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} z_{j'} z_{j'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(j')})} \\
&\left. + \frac{1}{(kp)^2} \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} z_{j'} z_{j'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{i'} z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(i')})} \right] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}} \right) \\
&= \sum_{\substack{i',j'=1\\ i'\neq j'}}^{n} u_{i'} v_{j'} \mathbb{E}\left[\frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})} - \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-j'}{\tilde{Q}_{-i'}} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{i'} z_{i'}^\mathsf{T} A^{\frac{1}{2}} \underset{-i'}{\tilde{Q}_{-j'}} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(i')})} \right.
\end{aligned}$$

$$- \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} z_{j'} z_{j'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{j'}}{kp(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(j')})} \Bigg] + \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right).$$

where the term

$$\frac{1}{(kp)^2} \frac{z_{i'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} z_{j'} z_{j'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-i'} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{i'} z_{i'}^\mathsf{T} A^{\frac{1}{2}} \tilde{Q}_{-j'} A^{\frac{1}{2}} z_{j'}}{(1+\delta_{\mathscr{C}(i')})(1+\delta_{\mathscr{C}(j')})(1+\delta_{\mathscr{C}(i')})}$$

is proved to be of order $\mathcal{O}(\frac{1}{\sqrt{p}})$ using (Seddik et al., 2020, Lemma A.2).

As such, the "sub-deterministic equivalent" for the matrix $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z$ with diagonal elements discarded is $J M_\delta^\mathsf{T} A^{\frac{1}{2}} B_{ij} A^{\frac{1}{2}} M_\delta J^\mathsf{T} - J M_\delta^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} A^{\frac{1}{2}} M_\delta W_{ij}$, with

$$A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} \leftrightarrow B_{ij}$$

$$W_{ij} = [w_{11}, \ldots, w_{km}]^\mathsf{T}, \quad w_{sl} = \left[\mathbf{0}_{n_{11}+\ldots+n_{(s-1)l}}^\mathsf{T}, \frac{2\mathrm{tr}\,(B_{ij} C_{sl})}{kp(1+\delta_{sl})} \mathbb{1}_{n_{sl}}^\mathsf{T}, \mathbf{0}_{n_{(s+1)l}+\ldots+n_{km}}^\mathsf{T} \right]^\mathsf{T}$$

(note that this matrix estimator of the off-diagonal elements is not zero on the diagonal; however its diagonal elements vanish as $n, p \to \infty$ and may thus be maintained without affecting the final result).

We next need to handle the contribution of the diagonal elements. These are obtained similarly as the off-diagonal elements and lead to the deterministic diagonal matrix equivalent

$$F_{ij} = \sum_{i',j'} \frac{\mathrm{tr}(C_{i'j'} B_{ij})}{(1+\delta_{i'j'})^2} e_{i'j'}^{[mk]} e_{i'j'}^{[mk]\mathsf{T}}.$$

Put together, the complete deterministic equivalent is then:

$$J M_\delta^\mathsf{T} A^{\frac{1}{2}} B_{ij} A^{\frac{1}{2}} M_\delta J^\mathsf{T} - J M_\delta^\mathsf{T} A^{\frac{1}{2}} \bar{\tilde{Q}} A^{\frac{1}{2}} M_\delta W_{ij} + \sum_{i',j'} \frac{\mathrm{tr}(C_{i'j'} B_{ij})}{(1+\delta_{i'j'})^2} e_{i'j'}^{[mk]} e_{i'j'}^{[mk]\mathsf{T}}.$$

This proves that $Z^\mathsf{T} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} A^{\frac{1}{2}} Z \leftrightarrow J M_\delta^\mathsf{T} A^{\frac{1}{2}} (B_{ij} A^{\frac{1}{2}} M_\delta J^\mathsf{T} - \bar{\tilde{Q}} A^{\frac{1}{2}} M_\delta W_{ij}) + F_{ij}$.

Calculus of B_{ij}. To conclude the proof of Lemma 5, it then remains to find a deterministic equivalent for $\tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q}$ which we denote by B_{ij}. Similar derivations and results are provided in detail in (Louart et al., 2018). For conciseness, we sketch the most important elements of the proof. The interested reader can refer to (Louart et al., 2018, Section 5.2.3). Let us evaluate $\mathbb{E}[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} (\tilde{Q} - \bar{\tilde{Q}}) v]$ for any deterministic vector $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$ such that $\|u\| = 1$ and $\|v\| = 1$ by using successively Equations (6.29) and (6.26):

$$\mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} (\tilde{Q} - \bar{\tilde{Q}}) v \right] = \mathbb{E}\left[u^\mathsf{T} \tilde{Q} A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}} \tilde{Q} (-\frac{A^{\frac{1}{2}} Z Z^\mathsf{T} A^{\frac{1}{2}}}{kp} + C_\delta) \bar{\tilde{Q}} v \right]$$

$$
\begin{aligned}
&= -\frac{1}{kp}\sum_{i'}\mathbb{E}\left[\frac{u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}v}{1+\delta_{i'j}}\right]\\
&+\mathbb{E}\left[u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}C_\delta\bar{\tilde{Q}}v\right]\\
&-\frac{1}{kp}\mathbb{E}\left[u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}C_\delta\bar{\tilde{Q}}v\right]+\mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)
\end{aligned}
$$

where $C_\delta = \sum_{ij}\frac{c_{ij}}{c_0}\frac{C_{ij}}{1+\delta_{ij}}$. Using Assumption 2 and following the work of (Louart & Couillet, 2018),

$$
\frac{1}{kp}\mathbb{E}\left[u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}C_\delta\bar{\tilde{Q}}v\right]=\mathcal{O}(\frac{1}{p}).
$$

Furthermore,

$$
\begin{aligned}
\mathbb{E}\left[u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}(\tilde{Q}-\bar{\tilde{Q}})v\right] &= -\frac{1}{kp}\sum_{i'}\mathbb{E}\left[\frac{u^\mathsf{T}\tilde{Q}_{-i'}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}v}{1+\delta_{i'j}}\right]\\
&+\frac{1}{kp}\sum_{i'}\mathbb{E}\left[\frac{u^\mathsf{T}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}A^{\frac{1}{2}}z_{i'}z_{i'}^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}v}{kp(1+\delta_{i'j})^2}\right]\\
&+\mathbb{E}\left[u^\mathsf{T}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}Q_{-i'}C_\delta\bar{\tilde{Q}}v\right]+\mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)\\
&= \frac{1}{kp}\sum_{i'}\mathbb{E}\frac{\mathrm{tr}\left(C_{\mathscr{C}(i')}\tilde{Q}A^{\frac{1}{2}}S_{i'j}A^{\frac{1}{2}}\tilde{Q}\right)}{(1+\delta_{\mathscr{C}(i')})^2}\mathbb{E}\left[u^\mathsf{T}\bar{\tilde{Q}}C_{\mathscr{C}(i')}\bar{\tilde{Q}}v\right]+\mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)
\end{aligned}
$$

where $-\frac{1}{kp}\sum_i\mathbb{E}[\frac{u^\mathsf{T}\tilde{Q}_{-i'}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}_{-i'}z_iz_i^\mathsf{T}\bar{\tilde{Q}}v}{1+\delta_{i'j}}]+\mathbb{E}[u^\mathsf{T}\tilde{Q}_{-i'}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}QC_\delta\bar{\tilde{Q}}v]=\mathcal{O}\left(\frac{1}{\sqrt{p}}\right)$, following again (Louart & Couillet, 2018).

Let us next denote $d_{ab}=\frac{n_{ab}}{kp(1+\delta_{ab})^2}$. We then have the following identity for $\mathbb{E}[\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}]$:

$$
\mathbb{E}[\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}]=\bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}}+\sum_{i'=1}^{k}\sum_{j'=1}^{m}\frac{d_{i'j'}}{kp}\mathbb{E}\left[\mathrm{tr}\left(C_{i'j'}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}\right)\right]\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}}+\mathcal{O}_{\|\cdot\|}\left(\sqrt{\frac{\log p}{p}}\right) \tag{6.30}
$$

Further introduce the two matrices $\bar{T}$ and T defined as: $\bar{T}_{ab,ij}=\frac{1}{kp}\mathrm{tr}(C_{ab}\bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}})$ and

$T_{ij,i'j'}=\frac{1}{kp}\mathbb{E}[\mathrm{tr}\left(C_{i'j'}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}\right)]$. These satisfy the following equations (i.e., by right multiplying Equation (6.30) by $C_{i'j'}$ and taking the trace)

$$
T_{i'j'}^{(ij)}=\bar{T}_{ij,i'j'}+\sum_{e=1}^{k}\sum_{f=1}^{m}d_{ef}T_{ef,ij}\mathcal{T}_{i'j',ef},
$$

so that $T = \bar{T}(I_k - D\mathscr{T})^{-1}$ where $D = \mathscr{D}_{[d_{11},\ldots,d_{km}]^\intercal}$ and $\mathscr{T}_{ef,i'j'} = \frac{1}{kp}\mathrm{tr}(C_{ef}\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}})$.

Finally,

$$\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q} \leftrightarrow \bar{\tilde{Q}}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\bar{\tilde{Q}} + \sum_{i'=1}^{k}\sum_{j'=1}^{m} d_{i'j'}T_{i'j'}^{(ij)}\mathbb{E}[\bar{\tilde{Q}}C_{i'j'}\bar{\tilde{Q}}] \tag{6.31}$$

with $T = \bar{T}(I_k - D\mathscr{T})^{-1}$.

6.3.4 Proof of Theorem 9

Proof of the convergence in distribution. Under a Gaussian mixture assumption for the input data X, the convergence in distribution of the statistics of the classification score $g_i(\mathbf{x})$ is identical to the central limit theorem derived in (Liao & Couillet, 2019, Appendix B) by writing the classification score $g_i(\mathbf{x})$ in polynomial form of a Gaussian vector and by resorting to the Lyapounov central limit theorem (Billingsley, 2008).

Since conditionally on the training data X, the classification score $g(x)$ is expressed as the projection of the deterministic vector W on the concentrated random vector $\mathbf{x}$, the CLT for concentrated vector unfolds by proving that projections of deterministic vector on concentrated random vector is asymptotically gaussian. This is ensured by the following result.

Theorem 10 (CLT for concentrated vector (Klartag, 2007; Fleury et al., 2007))**.** *If* $\mathbf{x}$ *is a concentrated random vector with* $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\intercal] = I_p$ *with an observable diameter of order* $\mathcal{O}(1)$ *and* σ *be the uniform measure on the sphere* $\mathcal{S}^{p-1} \subset \mathbb{R}^p$ *of radius* 1*, then for any integer* k*, small compared to* p*, there exist two constants* C, c *and a set* $\Theta \subset (\mathcal{S}^{p-1})^k$ *such that* $\underbrace{\sigma \otimes \ldots \otimes \sigma}_{k}(\Theta) \geq 1 - \sqrt{p}Ce^{-c\sqrt{p}}$ *and* $\forall \theta = (\theta_1, \ldots, \theta_k) \in \Theta$*,*

$$\forall a \in \mathbb{R}^k : \sup_{t\in\mathbb{R}} |\mathbb{P}(a^\intercal\theta^\intercal\mathbf{x} \geq t) - G(t)| \leq Cp^{-\frac{1}{4}}.$$

with $G(t)$ *the cumulative distribution function of* $\mathcal{N}(0,1)$

Then the result unfolds naturally.

Statistical mean of the classification scores. Using the definition of the score in (4.2), the average output score $g_i(\mathbf{x})$ for $\mathbf{x} \in \mathscr{C}_j$ is

$$\mathbb{E}[g_i(\mathbf{x})] = \mathbb{E}\left[\frac{1}{kp}\left(e_i^{[k]} \otimes \mu_{ij}\right)^\intercal A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z(Y - Pb)\right] + b_i.$$

Using Lemma 5, this can be further developed as:

$$\mathbb{E}[g_i(\mathbf{x})] = \frac{1}{kp}\left(e_i^{[k]} \otimes \mu_{ij}\right)^\intercal A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\delta J^\intercal(Y - P\bar{b}) + b_i + o(1). \tag{6.32}$$

Since $C_{ij} = A^{\frac{1}{2}}(e_i^{[k]}{e_i^{[k]}}^\mathsf{T} \otimes (\Sigma_{ij} + \mu_{ij}\mu_{ij}^\mathsf{T}))A^{\frac{1}{2}}$ is a finite rank update of Σ_{ij}, one can further use Woodbury identity matrix (i.e., $(A+UCV)^{-1} = A^{-1} + A^{-1}UC(I+VA^{-1}U)VA^{-1}$ for invertible square A) to write $\bar{\tilde{Q}} = \bar{\tilde{Q}}_0 - \bar{\tilde{Q}}_0\mathbb{M}(I_{km} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M})^{-1}\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0$, with

$$\bar{\tilde{Q}}_0 = \left[\sum_{i=1}^{k}\sum_{j=1}^{m} (\mathscr{D}_\gamma + \lambda \mathbb{1}_k\mathbb{1}_k)^{\frac{1}{2}} e_i e_i^\mathsf{T} (\mathscr{D}_\gamma + \lambda \mathbb{1}_k\mathbb{1}_k)^{\frac{1}{2}} \otimes \delta_{ij}^{[mk]}\Sigma_{ij} + I_{kp}\right]^{-1}$$
$$\mathbb{M} = A^{\frac{1}{2}} M \mathscr{D}_{\delta^{[mk]}}^{\frac{1}{2}}$$

with $\delta^{[mk]} = [\delta_{i1}^{[mk]}, \ldots, \delta_{mk}^{[mk]}]$ for $\delta_{ij}^{[mk]} = \frac{c_{ij}}{c_0(1+\delta_{ij}^{[mk]})}$. Plugging the expression of $\bar{\tilde{Q}}$ in Equation (6.32), we obtain

$$\begin{aligned}\mathbb{E}[g_i(\mathbf{x})] &= e_{ij}^\mathsf{T}\mathscr{D}_{\delta^{[mk]}}^{-\frac{1}{2}}\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}\mathbb{M}\mathscr{D}_{\delta^{[mk]}}^{\frac{1}{2}}\mathring{\mathscr{Y}} + b_i + o(1)\\ &= e_{ij}^\mathsf{T}\mathscr{D}_{\delta^{[mk]}}^{-\frac{1}{2}}(I_{mk} - \Gamma)\mathscr{D}_{\delta^{[mk]}}^{\frac{1}{2}}\mathring{\mathscr{Y}} + b_i + o(1)\end{aligned}$$

with $\Gamma = (I_{mk} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M})^{-1}$ and $e_{ij}^{[mk]}$ is the canonical vector. Finally, to be exhaustive without going into the technical details,[1] let us conclude by remarking that one can show using the deterministic equivalent for Q provided in (Louart & Couillet, 2018) that $b_i = \frac{\mathbb{1}_{n_i}^\mathsf{T} Y_i}{n_i} + \mathcal{O}(p^{-\frac{1}{2}}) = \mathscr{Y} - \mathring{\mathscr{Y}} + \mathcal{O}(p^{-\frac{1}{2}})$.

Finally, letting m_{ij} be the above expression of $\mathbb{E}[g_i(\mathbf{x})]$ without the trailing $o(1)$ and $m = [m_{11}, \ldots, m_{km}]^\mathsf{T}$, one concludes using the notations of Theorem 9 that

$$m = \mathscr{Y} - \mathscr{D}_{\delta^{[mk]}}^{-\frac{1}{2}}\Gamma\mathscr{D}_{\delta^{[mk]}}^{\frac{1}{2}}\mathring{\mathscr{Y}}$$

as desired.

Variance of the classification score. Using Equation (4.2), for $\mathbf{x} \in \mathscr{C}_j$, the covariance of the score $g_i(\mathbf{x})$ is given by

$$\mathrm{Cov}[g_i(\mathbf{x})] = \mathbb{E}\left[\frac{1}{(kp)^2}(Y - Pb)^\mathsf{T} Z^\mathsf{T} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} S_{ij} A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}} Z(Y - Pb)\right]$$

Using the deterministic equivalent of $Z^\mathsf{T}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}}\tilde{Q}A^{\frac{1}{2}}Z$ in Lemma 1, the expression further reads

$$\begin{aligned}\mathrm{Cov}[g_i(\mathbf{x})] = = \frac{1}{(kp)^2}(Y - P\bar{b})^\mathsf{T}\left(JM_\delta^\mathsf{T}A^{\frac{1}{2}}B_{ij}A^{\frac{1}{2}}M_\delta J + F_{ij}\right)(Y - P\bar{b})\\ - \frac{1}{p^2}(Y - P\bar{b})^\mathsf{T}JM_\delta^\mathsf{T}A^{\frac{1}{2}}\bar{\tilde{Q}}A^{\frac{1}{2}}M_\delta W_{ij}(Y - P\bar{b}).\end{aligned}$$

[1] Due to Remark 4, b_i can take any arbitrary value since only the decision threshold but not the performance is sensitive to a shift of Y.

Similarly to the calculus performed for $\mathbb{E}[g_i(\mathbf{x})]$, using again $\bar{\tilde{Q}} = \bar{\tilde{Q}}_0 - \bar{\tilde{Q}}_0\mathbb{M}(I_{kp} + \mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M})^{-1}\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0$, similar algebraic manipulations lead to:

$$\begin{aligned}
\mathrm{Cov}[g_i(\mathbf{x})] = &= \mathring{\mathscr{Y}}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathbb{M}^\mathsf{T}B_{ij}\mathbb{M}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathring{\mathscr{Y}} + \mathring{\mathscr{Y}}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathscr{D}_{\kappa_{ij,\cdot}}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathring{\mathscr{Y}} - \mathring{\mathscr{Y}}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}\mathbb{M}\mathscr{D}_{\frac{\kappa_{ij,\cdot}}{\delta^{[mk]}}}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathring{\mathscr{Y}} \\
&= \mathscr{Y}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\Gamma\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathbb{M}\Gamma\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathscr{Y} + \mathscr{Y}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}(I-\Gamma)\mathscr{D}_{\kappa_{ij,\cdot}}(I-\Gamma) + \\
&\mathring{\mathscr{Y}}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathscr{D}_{\kappa_{ij,\cdot}}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathring{\mathscr{Y}} - 2\mathring{\mathscr{Y}}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}(I-\Gamma)\mathscr{D}_{\kappa_{ij,\cdot}}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathring{\mathscr{Y}} \\
&= \mathscr{Y}^\mathsf{T}\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\left(\Gamma\mathscr{D}_{\kappa_{ij,\cdot}}\Gamma + \Gamma\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{V}_{ij}\bar{\tilde{Q}}_0\mathbb{M}\Gamma\right)\mathscr{D}^{\frac{1}{2}}_{\delta^{[mk]}}\mathscr{Y}
\end{aligned}$$

with $\mathbb{V}_{ij} = A^{\frac{1}{2}}S_{ij}A^{\frac{1}{2}} + \sum_{i'=1}^{k}\sum_{j'=1}^{m}\delta^{[mk]}_{i'j'}\kappa_{ij,i'j'}A^{\frac{1}{2}}S_{i'j'}A^{\frac{1}{2}}$ and $\kappa_{ij,\cdot} = [\kappa_{ij,11},\ldots,\kappa_{ij,k2}]$ with $\kappa_{ij,i'j'} = d_{i'j'}T_{ij,i'j'}/\delta^{[mk]}_{i'j'}$.

Particular Case

In the case of binary classification ($m = 2$) and for $\Sigma_{ij} = I_p$, we have the simplification:

$$\begin{aligned}
\mathbb{M} &= \sum_{i,j}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{\frac{1}{2}}e_i^{[k]}e_i^{[k]\mathsf{T}} \otimes \sqrt{\tilde{\delta}_i}\mathring{\mu}_{ij} \\
&= \sum_{i}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{\frac{1}{2}}e_i^{[k]}e_i^{[k]\mathsf{T}} \otimes \left(\frac{\left[\frac{c_{i2}\sqrt{c_{i1}}}{c_i}, -\frac{c_{i1}\sqrt{c_{i2}}}{c_i}\right]}{c_0(1+\delta_i)} \otimes \Delta\mu_i\right).
\end{aligned}$$

Moreover, $\bar{\tilde{Q}}_0 = [(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T})^{\frac{1}{2}}\mathscr{D}_{\tilde{\delta}}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{\frac{1}{2}} + I_k]^{-1} \otimes I_p$, so that

$$\begin{aligned}
\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0\mathbb{M} &= \sum_{i,j}e_i^{[k]}e_i^{[k]\mathsf{T}}\left[I_k + \mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{-1}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\right]^{-1}e_j^{[k]}e_j^{[k]\mathsf{T}}\Delta\mu_i^\mathsf{T}\Delta\mu_j \otimes \mathbb{c}_i\mathbb{c}_j^\mathsf{T} \\
&= \sum_{i,j}\mathscr{A}_{ij}\Delta\mu_i^\mathsf{T}\Delta\mu_j e_i^{[k]}e_j^{[k]\mathsf{T}} \otimes \mathbb{c}_i\mathbb{c}_j^\mathsf{T} \\
&= \left(\mathscr{A} \otimes \mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right) \odot \mathscr{M}
\end{aligned}$$

with

$$\begin{aligned}
\mathscr{M} &\equiv \sum_{i,j}\Delta\mu_i^\mathsf{T}\Delta\mu_j\left(E_{ij}^{[k]} \otimes \mathbb{c}_i\mathbb{c}_j^\mathsf{T}\right) \\
\mathbb{c}_i &\equiv \begin{bmatrix}\frac{c_{i2}}{c_i}\sqrt{\frac{c_{i1}}{c_i}} \\ -\frac{c_{i1}}{c_i}\sqrt{\frac{c_{i2}}{c_i}}\end{bmatrix} \\
\mathscr{A} &\equiv \left[I_k + \mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{-1}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\right]^{-1}.
\end{aligned}$$

As for the covariance terms,

$$\begin{aligned}
\mathbb{M}^\mathsf{T}\bar{\tilde{Q}}_0 V_{ij}\bar{\tilde{Q}}_0\mathbb{M} &= \sum_{i,j} e_i^{[k]}e_i^{[k]\mathsf{T}}\left[I_k + \mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(\mathscr{D}_\gamma + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{-1}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\right]^{-1}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(e_i^{[k]}e_i^{[k]\mathsf{T}} + \mathscr{D}_{\kappa_i\odot\tilde{\delta}}\right)\times \\
&\quad \mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left[I_k + \mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(I_k + \lambda\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)^{-1}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\right]^{-1} e_j^{[k]}e_j^{[k]\mathsf{T}}\Delta\mu_i^\mathsf{T}\Delta\mu_j\otimes\mathbb{c}_i\mathbb{c}_j^\mathsf{T} \\
&= \sum_{i,j}\left[\mathscr{A}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(e_i^{[k]}e_i^{[k]\mathsf{T}} + \mathscr{D}_{\kappa_i\odot\tilde{\delta}}\right)\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\mathscr{A}\right]_{ij}\Delta\mu_i^\mathsf{T}\Delta\mu_j e_i^{[k]}e_j^{[k]\mathsf{T}}\otimes\mathbb{c}_i\mathbb{c}_j^\mathsf{T} \\
&= \left(\mathscr{A}\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\left(e_i^{[k]}e_i^{[k]\mathsf{T}} + \mathscr{D}_{\kappa_i\odot\tilde{\delta}}\right)\mathscr{D}_{\tilde{\delta}}^{-\frac{1}{2}}\mathscr{A}\otimes\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)\odot\mathscr{M} \\
&= \frac{1}{\delta_i^{[k]}}\left(\mathscr{A}\mathscr{D}_{\mathscr{K}_i+e_i^{[k]}}\mathscr{A}\otimes\mathbb{1}_k\mathbb{1}_k^\mathsf{T}\right)\odot\mathscr{M}
\end{aligned}$$

with $\mathscr{K}_{ia} = \tilde{\delta}_i\kappa_{ia}$. Using Equation (6.31), after algebraic manipulations, we finally obtain the compact form

$$\mathscr{K} = \frac{c_0}{k}[\mathscr{A}\odot\mathscr{A}]\left(\mathscr{D}_c - \frac{c_0}{k}[\mathscr{A}\odot\mathscr{A}]\right)^{-1}. \tag{6.33}$$

6.3.5 Proof of Propositions 6–7

One-versus-all

The probability of correct classification for Task i and for a test data $\mathbf{x}\in\mathscr{C}_j$ reads

$$\mathbb{P}\left(g_i^{\text{bin}}(\mathbf{x};j) > \max_{j'\neq j}\{g_i^{\text{bin}}(\mathbf{x};j')\}\right) = \mathbb{P}\left(g_i^{\text{bin}}(\mathbf{x};j) - \max_{j'\neq j}\{g_i^{\text{bin}}(\mathbf{x};j')\} > 0\right).$$

Since by definition (Equation (4.2))

$$g_i^{\text{bin}}(\mathbf{x};j) = \frac{1}{kp}\mathring{y}(j)^\mathsf{T}J^\mathsf{T}QZ^\mathsf{T}A\left(e_i^{[k]}\otimes\mathring{\mathbf{x}}\right) + b_i, \tag{6.34}$$

we have that $g_i^{\text{bin}}(\mathbf{x},j)\mathbf{1}_{m-1} - \left\{g_i^{\text{bin}}(\mathbf{x};j')\right\}_{j'\neq j} = \frac{1}{kp}\mathscr{Y}_{-j}J^\mathsf{T}QZ^\mathsf{T}A\left(e_k^{[k]}\otimes\mathring{\mathbf{x}}\right)$, where $\mathscr{Y}_{-j} = (\mathring{y}(j)^\mathsf{T} - \left[\mathring{y}(j')^\mathsf{T}\right]_{j'\neq j}) \in \mathbb{R}^{(m-1)\times km}$. Using Theorem 9 with $\mathscr{Y}$ replaced by $\mathscr{Y}_{-j}$, $g_i^{\text{bin}}(\mathbf{x},j)\mathbf{1}_{m-1} - g_i^{\text{bin}}(\mathbf{x};j')_{j'\neq j} \in \mathbb{R}^{m-1}$ is asymptotically a multivariate Gaussian random vector with statistics detailed in the theorem statement. Proposition 6 then unfolds trivially by remarking that $g_i^{\text{bin}}(\mathbf{x};j) > \max_{j'\neq j} g_i^{\text{bin}}(\mathbf{x};j') \Leftrightarrow \forall j'\neq j,\ g_i^{\text{bin}}(\mathbf{x};j) - g_i^{\text{bin}}(\mathbf{x};j') \geq 0$.

One Hot encoding

The proof is similar to the one-versus-all case.

The probability of correct classification for a test data $\mathbf{x}\in\mathscr{C}_j$ is

$$\mathbb{P}\left(g_i^{\text{bin}}(\mathbf{x};j) > \max_{j'\neq j}\{g_i^{\text{bin}}(\mathbf{x};j')\}\right) = \mathbb{P}\left(g_i^{\text{bin}}(\mathbf{x};j) - \max_{j'\neq j}\{g_i^{\text{bin}}(\mathbf{x};j')\} > 0\right)$$

where

$$g_i^{\text{bin}}(\mathbf{x}; j) = \frac{1}{kp} e_j^{[k]\mathsf{T}} \mathring{\mathscr{Y}}^{\mathsf{T}} J^{\mathsf{T}} Q Z^{\mathsf{T}} A \left(e_i^{[k]} \otimes \mathring{\mathbf{x}} \right) + b_i. \tag{6.35}$$

Therefore $g_i^{\text{bin}}(\mathbf{x}; j)\mathbf{1}_{m-1} - \{(g_i(\mathbf{x}; j'))\}_{j' \neq j} = \frac{1}{kp} \mathscr{E}_j \mathring{\mathscr{Y}}^{\mathsf{T}} J^{\mathsf{T}} Q Z^{\mathsf{T}} A (e_k^{[k]} \otimes \mathring{\mathbf{x}})$, with $\mathscr{E}_j = \{(e_j^{(m)} - e_{j'}^{(m)})^{\mathsf{T}}\}_{j \neq j'} \in \mathbb{R}^{(m-1) \times m}$. By Theorem 9 with $\mathring{\mathscr{Y}}$ replaced by $\mathscr{E}_j^{\mathsf{T}} \mathring{\mathscr{Y}}^{\mathsf{T}}$, this vector is asymptotically normally distributed and Proposition 7 unfolds immediately using again the fact that $g_i^{\text{bin}}(\mathbf{x}; j) > \max_{j' \neq j} g_i^{\text{bin}}(\mathbf{x}; j') \Leftrightarrow \forall j' \neq j,\ g_i^{\text{bin}}(\mathbf{x}; j) - g_i^{\text{bin}}(\mathbf{x}; j'j) \geq 0$.

6.4 Synthèse de la thèse en français

La plupart des méthodes de traitement du signal et d'apprentissage automatique (par exemple, les tests statistiques, l'estimation de paramètres, la classification, la régression, etc.) sont basées sur des fonctionnelles non triviales de n vecteurs aléatoires de dimension p. Dans l'hypothèse où le nombre n de ces données disponibles est largement supérieur à leur dimension p, certains résultats théoriques peuvent être obtenus, car un comportement déterministe apparaît parfois lorsque $n \to \infty$, ce qui simplifie le problème comme l'illustre par exemple la célèbre loi des grands nombres et le théorème central limite. Cependant, comme déjà montré dans la littérature (Wigner, 1958; Marčenko & Pastur, 1967), certains résultats et intuitions basés sur des asymptotiques classiques ($n \to \infty$ et dimension des données p fixe) s'effondrent lorsque la taille des données est comparable à la dimension des données, un problème souvent apparenté à la "malédiction de la dimensionnalité". Dans le cadre du paradigme actuel des données grandissantes, où des masses de données sont produites, échangées et stockées, nous sommes constamment confrontés à une situation où non seulement la taille mais aussi la dimension des données sont importantes.

D'un autre côté, la dernière décennie a vu une augmentation considérable de la diversité des applications auxquelles l'apprentissage automatique est appliqué (apprentissage par transfert, confidentialité des données, équité, etc). Par conséquent, l'apprentissage automatique n'est plus seulement utilisé pour le placement des publicités et l'implémentation de filtres antispam : il est désormais utilisé pour filtrer les demandes de prêt (Fernández, 2019), déployer les forces de police (Rudin, 2013), prendre des décisions dans le domaine de la justice pénale (Berk & Hyatt, 2015), etc., faisant ainsi rapidement son entrée dans les systèmes socio-techniques. Les biais algorithmiques et les algorithmes mal compris sont l'un des plus grands risques d'échec car ils compromettent l'objectif même de l'apprentissage automatique depuis que l'opinion publique s'inquiète de l'impact de la technologie sur la société.

Il est donc important de considérer des asymptotiques où le nombre de données n et leur dimension p sont du même ordre de grandeur tout en fournissant un ensemble d'outils adaptés dans ce régime pour analyser, comprendre et améliorer les problèmes de l'apprentissage automatique et de traitement du signal. Cette réflexion a permis l'émergence de nouvelles théories généralement encapsulées sous la dénomination de "statistiques en grande dimension" dont fait partie la théorie des matrices aléatoires. Un

ensemble d'outils ont été développés pour analyser les objets statistiques communément rencontrés dans le domaine des télécommunications (Couillet & Debbah, 2011), de la finance (Bouchaud & Potters, 2009), de la physique (Guhr et al., 1998) pour ne citer que quelques-uns. Seulement récemment, des études utilisant la théorie des matrices aléatoires (Ali, 2018; Mai & Couillet, 2018; Liao & Couillet, 2019; Seddik et al., 2020; Deng et al., 2019; Elkhalil, 2019) se sont focalisées sur des algorithmes d'apprentissage automatique. De façon spécifique, ces travaux ont permis d'analyser, de comprendre et d'améliorer certains algorithmes simples (machine à vecteurs de support, regroupement spectral, etc). Ces travaux ont mis en particulier en évidence le rôle prépondérant de la moyenne et de la matrice de covariance des données, ce qui en fait des objets qui méritent une compréhension approfondie. Cette thèse s'appuie sur cette importante conclusion pour montrer comment un traitement empirique de la matrice de covariance des données peut avoir des conséquences dramatiques sur des applications de traitement du signal. Nous montrons ensuite comment la moyenne et la matrice de covariance des données interviennent dans la performance d'algorithmes d'apprentissage d'intérêt moderne comme l'apprentissage par transfert et l'apprentissage multi-tâche.

De façon plus détaillée, dans une première partie, la thèse montre en se reposant sur des outils avancés de la théorie des matrices aléatoires comment les estimateurs classiques de distance entre matrices de covariance induisent des biais importants et fournit des estimateurs consistents pour une grande famille de métrique. Ainsi, si on définit par Σ_1 et Σ_2 deux matrices de covariance de taille $p \times p$, on constate que la plupart des distances usuelles, que l'on dénotera génériquement $D(\Sigma_1, \Sigma_2)$, s'expriment comme des fonctionnelles des valeurs propres de la matrice $\Sigma_1^{-1}\Sigma_2$ (distance de Fisher, distance de Bhattacharyya, divergence de Kullback Leibler ou de Rényi entre gaussiennes centrées) ou des valeurs propres de la matrice $\Sigma_1\Sigma_2$ (distance de Wasserstein entre deux gaussiennes centrées). Dans l'hypothèse où le nombre d'échantillons n_1 et n_2 de données ayant Σ_1 et Σ_2 pour covariance est très grand devant p, la loi des grands nombres garantit que $D(\hat{\Sigma}_1, \hat{\Sigma}_2)$ est un estimateur consistent pour $D(\Sigma_1, \Sigma_2)$ avec $\hat{\Sigma}_a = \frac{1}{n_a}\sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)T}$ pour $a \in \{1, 2\}$ la matrice de covariance empirique des n_a échantillons centrés $x_i^{(a)}$. Cependant, cet estimateur est fortement biaisé lorsque $n_1, n_2 \sim p$. À l'aide d'outils de la théorie des matrices aléatoires, cette thèse propose une formule générale d'un estimateur "universel" des distances $D(\Sigma_1, \Sigma_2)$ qui est consistent dans la limite où $p, n_1, n_2 \to \infty$ avec $p/n_1 \to c_1 > 0$ et $p/n_2 \to c_2 > 0$. Ces résultats s'inspirent des travaux de Mestre Mestre (2008a) sur l'estimation de fonctionnelles des valeurs propres $\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(\Sigma))$ de matrices de covariance Σ. La procédure que nous suivons ici est la suivante: (i) la quantité d'intérêt $D(\Sigma_1, \Sigma_2)$ est exprimée comme une intégrale complexe faisant intervenir la transformée de Stieltjes de la mesure des valeurs propres de $\Sigma_1^{-1}\Sigma_2$ (ou $\Sigma_1\Sigma_2$); (ii) cette mesure est reliée asymptotiquement à la mesure limite des valeurs propres de $\hat{\Sigma_1}^{-1}\hat{\Sigma}_2$ (ou $\hat{\Sigma}_1\hat{\Sigma}_2$) en utilisant les travaux de Silverstein & Bai (1995); (iii) on obtient alors un estimateur sous forme d'intégrale complexe qu'il s'agit d'évaluer pour chaque fonction f d'intérêt. Cependant, au contraire de Mestre (2008a) qui s'intéresse à des fonctions f simples, les distances $D(\Sigma_1, \Sigma_2)$ d'intérêt ici impliquent des logarithmes et racines carrées

qui demandent un travail fin d'analyse complexe (notamment un traitement précis des "coupures").

La deuxième partie de la thèse montre encore une fois que les statistiques de premier ordre (moyenne et covariance des données) sont les statistiques suffisantes d'algorithmes plus complexes et d'intérêt plus pratique comme l'apprentissage multi-tâche et par transfert. L'analyse théorique d'un algorithme d'apprentissage multi-tâche basé sur les machines à vecteurs de support révèle d'abord les "statistiques suffisantes" exploitées par l'algorithme et leur interaction. Ces résultats démontrent, que l'approche standard de l'étiquetage des données est largement sous-optimale, peut conduire à de graves effets de transfert négatif, mais que ces déficiences sont facilement corrigées. Ces corrections sont transformées en un algorithme amélioré qui ne fait que bénéficier de données et tâches supplémentaires, et dont la performance théorique est également parfaitement comprise. Comme cela a été démontré et soutenu théoriquement dans de nombreux travaux récents, ces résultats de grande dimension sont robustes à de larges gammes de distributions de données, ce que nos expériences corroborent. Plus précisément, l'étude fait état d'un comportement systématiquement proche entre les performances théoriques et empiriques sur des bases de données populaires, ce qui suggère fortement l'applicabilité de la méthode proposée, soigneusement réglée, à une grande variété de données réelles. Ce réglage fin est entièrement basé sur l'analyse théorique et ne nécessite en particulier aucune procédure de validation croisée. En outre, les performances rapportées sur des ensembles de données réelles surpassent presque systématiquement les méthodes d'apprentissage multi-tâches et de transfert de l'état de l'art, beaucoup plus élaborées et moins intuitives.

Bibliography

B Abdullah-Al-Zubaer Imran and Demetri Terzopoulos. Semi-supervised multi-task learning with chest x-ray images. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*, volume 11861, pp. 151. Springer Nature, 2019.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Arvind Agarwal, Samuel Gerber, and Hal Daume. Learning multiple tasks using manifold regularization. In *Advances in neural information processing systems*, pp. 46–54, 2010.

Murray Aitkin and Nicholas Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General)*, 149(1):1–26, 1986.

Hafiz Tiomoko Ali. *Nouvelles méthodes pour l'apprentissage non-supervisé en grandes dimensions.* PhD thesis, Université Paris-Saclay, 2018.

Greg M Allenby and Peter E Rossi. Marketing models of consumer heterogeneity. *Journal of econometrics*, 89(1-2):57–78, 1998.

Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in neural information processing systems*, pp. 41–48, 2007.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Z. Bai and J. W. Silverstein. Spectral Analysis of Large Dimensional Random Matrices. *Springer Series in Statistics*, 2009.

Z. D. Bai and J. W. Silverstein. No Eigenvalues Outside the Support of the Limiting Spectral Distribution of Large Dimensional Sample Covariance Matrices. *Annals of Probability*, 26(1):316–345, January 1998a.

Zhi-Dong Bai and Jack W Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998b.

Zhidong D Bai and Jack W Silverstein. Clt for linear spectral statistics of large-dimensional sample covariance matrices. In *Advances In Statistics*, pp. 281–333. World Scientific, 2008.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Richard Berk and Jordan Hyatt. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, 27(4):222–228, 2015.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Robert J Boik. Common principal components and related multivariate models. *Statistical Foundations of Econometric Modelling*, 259, 1988.

Jean-Philippe Bouchaud and Marc Potters. Financial applications of random matrix theory: a short review. *arXiv preprint arXiv:0910.1205*, 2009.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Biao Cai, Gemeng Zhang, Aiying Zhang, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu-Ping Wang. Capturing dynamic connectivity from resting state fmri using time-varying graphical lasso. *IEEE Transactions on Biomedical Engineering*, 66 (7):1852–1862, 2018.

Kevin Michael Carter. Dimensionality reduction on statistical manifolds. 2009.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Rich Caruana, Shumeet Baluja, and Tom Mitchell. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in neural information processing systems*, pp. 959–965, 1996.

Chein-I Chang. *Hyperspectral imaging: techniques for spectral detection and classification*, volume 1. Springer Science & Business Media, 2003.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

Sueli IR Costa, Sandra A Santos, and João E Strapasson. Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

R. Couillet and M. Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, New York, NY, USA, first edition, 2011.

R. Couillet, J. W. Silverstein, and M. Debbah. Eigen-Inference for Energy Estimation of Multiple Sources. 2011. URL `http://arxiv.org/abs/1001.3934`.

Romain Couillet, Florent Benaych-Georges, et al. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

Romain Couillet, Malik Tiomoko, Steeve Zozor, and Eric Moisan. Random matrix-improved estimation of covariance matrix distances. *arXiv preprint arXiv:1810.04534*, 2018.

Romain Couillet, Florent Chatelain, and Nicolas Le Bihan. Two-way kernel matrix puncturing: towards resource-efficient pca and spectral clustering. *arXiv preprint arXiv:2102.12293*, 2021.

George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–39, 1983.

Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. A unified framework for spectral clustering in sparse graphs. *arXiv preprint arXiv:2003.09198*, 2020.

Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. Nishimori meets bethe: a spectral method for node classification in sparse weighted graphs. *arXiv preprint arXiv:2103.03561*, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Ming Dong, Dong Yang, Yan Kuang, David He, Serap Erdal, and Donna Kenski. Pm2. 5 concentration prediction using hidden semi-markov model-based times series data mining. *Expert Systems with Applications*, 36(5):9046–9055, 2009.

Khalil Elkhalil. *Random Matrix Theory: Selected Applications from Statistical Signal Processing and Machine Learning*. PhD thesis, 2019.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117. ACM, 2004.

Ana Fernández. Artificial intelligence in financial services. *Banco de Espana Article*, 3: 19, 2019.

Bruno Fleury, Olivier Guédon, and Grigoris Paouris. A stability result for mean width of lp-centroid bodies. *Advances in Mathematics*, 214(2):865–877, 2007.

Joel N Franklin. *Matrix theory*. Courier Corporation, 2012.

Samuel Gerber, Tolga Tasdizen, P Thomas Fletcher, Sarang Joshi, Ross Whitaker, Alzheimers Disease Neuroimaging Initiative, et al. Manifold modeling for brain population analysis. *Medical image analysis*, 14(5):643–653, 2010.

V. L. Girko. Introduction to general statistical analysis. *Theory of Probability and its Applications*, 32:229–242, 1987.

Pinghua Gong, Jieping Ye, and Chang-shui Zhang. Multi-stage multi-task feature learning. In *Advances in neural information processing systems*, pp. 1988–1996, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

William H Greene. Econometric analysis 4th edition. *International edition, New Jersey: Prentice Hall*, pp. 201–215, 2000.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. *technical report*, 2007.

Thomas Guhr, Axel Müller-Groeling, and Hans A Weidenmüller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(4-6):189–425, 1998.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Learning an invariant hilbert space for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3845–3854, 2017.

Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013.

Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5081–5090, 2016.

Abdullah-Al-Zubaer Imran, Chao Huang, Hui Tang, Wei Fan, Yuan Xiao, Dingjun Hao, Zhen Qian, and Demetri Terzopoulos. Partly supervised multitask learning. *arXiv preprint arXiv:2005.02523*, 2020.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 99(2):295–327, 2001.

Abla Kammoun, Romain Couillet, Jamal Najim, and Mérouane Debbah. Performance of mutual information inference methods under unknown interference. *IEEE Transactions on Information Theory*, 59(2):1129–1148, 2013.

N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6):2757–2790, December 2008.

Bo'az Klartag. A central limit theorem for convex sets. *Inventiones mathematicae*, 168 (1):91–131, 2007.

Sajja Tulasi Krishna and Hemantha Kumar Kalluri. Deep learning and transfer learning approaches for image classification. *International Journal of Recent Technology and Engineering (IJRTE)*, 7(5S4):427–432, 2019.

WJ Krzanowski. Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703–707, 1979.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139: 360–384, 2015.

Olivier Ledoit and Michael Wolf. Numerical implementation of the quest function. *Computational Statistics & Data Analysis*, 115:199–223, 2017.

Olivier Ledoit and Michael Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065, 2020.

Olivier Ledoit, Michael Wolf, et al. Optimal estimation of a large-dimensional covariance matrix under stein's loss. *Bernoulli*, 24(4B):3791–3832, 2018.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89 in Mathematical surveys and monographs. American Mathematical Soc., 2001. ISBN 0821837923.

Marc Lelarge and Léo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 639–643. IEEE, 2019.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1805–1818, 2013.

Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*, 2018.

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *The Journal of Machine Learning Research*, 19(1):3074–3100, 2018.

Xiaoyi Mai and Zhenyu Liao. High dimensional classification via empirical risk minimization: Improvements and optimality. *arXiv preprint arXiv:1905.13742*, 2019.

Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361. IEEE, 2019.

V. A. Marčenko and L. A. Pastur. Distributions of eigenvalues for some sets of random matrices. *Math USSR-Sbornik*, 1(4):457–483, April 1967.

Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International conference on machine learning*, pp. 343–351, 2013.

X. Mestre. On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. 56(11):5353–5368, November 2008a.

X. Mestre. Improved estimation of eigenvalues of covariance matrices and their associated subspaces using their sample estimates. 54(11):5113–5129, November 2008b.

X. Mestre and M. Lagunas. Modified Subspace Algorithms for DoA Estimation With Large Arrays. 56(2):598–614, February 2008.

James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2(2.2):2, 2006.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

Shibin Parameswaran and Kilian Q Weinberger. Large margin multi-task metric learning. In *Advances in neural information processing systems*, pp. 1867–1875, 2010.

Leonid Andreevich Pastur and Mariya Shcherbina. *Eigenvalue distribution of large random matrices*. Number 171 in Mathematical Surveys and Monographs. American Mathematical Soc., 2011.

Gabriel Peyré. Manifold models for signals and images. *Computer vision and image understanding*, 113(2):249–260, 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Jonas Richiardi, Sophie Achard, Horst Bunke, and Dimitri Van De Ville. Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal Processing Magazine*, 30(3):58–70, 2013.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

Anderson Rocha and Siome Klein Goldenstein. Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):289–302, 2013.

Pedro Luiz Coelho Rodrigues, Florent Bouchard, Marco Congedo, and Christian Jutten. Dimensionality reduction for bci classification using riemannian geometry. In *7th Graz Brain-Computer Interface Conference (BCI 2017)*, 2017.

Pedro Luiz Coelho Rodrigues, Marco Congedo, and Christian Jutten. Multivariate time-series analysis via manifold learning. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 573–577. IEEE, 2018.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, volume 898, pp. 1–4, 2005.

Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine, August*, 2013.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

James R Schott. Some tests for common principal component subspaces in several groups. *Biometrika*, 78(4):771–777, 1991.

Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain 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)*, pp. 7480–7484. IEEE, 2019.

Mohamed El Amine Seddik, Romain Couillet, and Mohamed Tamaazousti. A random matrix analysis of learning with α-dropout. In *ICML (Artemiss Workshop)*, 2020.

J. W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

J. W. Silverstein and S. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Santosh Srivastava and Maya R Gupta. Bayesian estimation of the entropy of the multivariate gaussian. In *2008 IEEE International Symposium on Information Theory*, pp. 1103–1107. IEEE, 2008.

Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2008.

Alaa Tharwat. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2):145–180, 2016.

Malik Tiomoko and Romain Couillet. Estimation of covariance matrix distances in the high dimension low sample size regime. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 341–345. IEEE, 2019a.

Malik Tiomoko and Romain Couillet. Random matrix-improved estimation of the wasserstein distance between two centered gaussian distributions. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019b.

Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Vladimir Vapnik. Universal learning technology: Support vector machines. *NEC Journal of Advanced Technology*, 2(2):137–144, 2005.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pp. 555–586, 2008.

Jie Wang and Jieping Ye. Safe screening for multi-task feature learning with multiple data matrices. *arXiv preprint arXiv:1505.04073*, 2015.

E. Wigner. On the distribution of roots of certain symmetric matrices. *The Annals of Mathematics*, 67(2):325–327, March 1958.

J. Wishart. The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, 20(1-2):32–52, December 1928.

Shuo Xu, Xin An, Xiaodong Qiao, Lijun Zhu, and Lin Li. Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, 34:1078–1084, 07 2013. doi: 10.1016/j.patrec.2013.01.015.

Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8(Jan):35–63, 2007.

Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. *Transfer learning*. Cambridge University Press, 2020.

Jianfeng Yao, Romain Couillet, Jamal Najim, and Merouane Debbah. Fluctuations of an improved population eigenvalue estimator in sample covariance matrix models. *IEEE transactions on information theory*, 59(2):1149–1163, 2012.

Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd international conference on Machine learning*, pp. 1012–1019, 2005.

Shou-cheng Yuan, Jie Zhou, Jian-xin Pan, and Jie-qiong Shen. Sphericity and identity test for high-dimensional covariance matrix using random matrix theory. *Acta Mathematicae Applicatae Sinica, English Series*, 37(2):214–231, 2021.

Don Zagier. The dilogarithm function. In *Frontiers in number theory, physics, and geometry II*, pp. 3–65. Springer, 2007.

Wenlu Zhang, Rongjian Li, Tao Zeng, Qian Sun, Sudhir Kumar, Jieping Ye, and Shuiwang Ji. Deep model based transfer and multi-task learning for biological image analysis. *IEEE transactions on Big Data*, 2016.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012.

Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):1–31, 2014.

Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5):1017–1037, 2015.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Méthodes avancées de la théorie des matrices aléatoires pour l'apprentissage automatique.

Mots clés: Théorie des Matrices Aléatoires, Matrices de covariance, Apprentissage par transfert, Apprentissage Multi-tâche, Compréhension théorique de l'apprentissage automatique.

Résumé: L'apprentissage automatique a permis de résoudre de nombreuses applications du monde réel, allant des tâches supervisées à des tâches non supervisées, grâce au développement d'algorithmes puissants (machine à vecteurs de support, réseaux neuronaux profonds, regroupement spectral, etc). Ces algorithmes sont basés sur des méthodes d'optimisation motivées par des intuitions de petites dimensions qui s'effondrent en grande dimension, un phénomène connu sous le nom de "malédiction de la dimensionnalité". Néanmoins, en supposant que la dimension des données et leur nombre sont à la fois grands et comparables, la théorie des matrices aléatoires fournit une approche systématique pour évaluer le comportement (statistique) de ces grands systèmes d'apprentissage, afin de bien les comprendre et de les améliorer lorsqu'ils sont appliqués à des données de grande dimension.

Les analyses précédentes de la théorie des matrices aléatoires (Mai & Couillet, 2018; Liao & Couillet, 2019; Deng et al., 2019) ont montré que les performances asymptotiques de la plupart des méthodes d'apprentissage automatique et de traitement du signal ne dépendent que des statistiques de premier et de second ordre (moyennes et matrices de covariance des données). Ceci fait des matrices de covariance des objets extrêmement riches qui doivent être "bien traités et compris". La thèse démontre d'abord comment un traitement empirique et naïf de la matrice de covariance peut détruire le comportement d'algorithmes d'apprentissage automatique en introduisant des biais difficiles à supprimer, alors qu'une estimation cohérente des fonctionnelles d'intérêt en utilisant la théorie des matrices aléatoires évite les biais. Nous montrons ensuite comment les moyennes et les matrices de covariance sont suffisantes (par le biais de fonctionnelles simples) pour traiter le comportement d'algorithmes d'intérêt moderne, tels que les méthodes d'apprentissage multi-tâches et par transfert.

Title: Advanced Random Matrix Methods for Machine Learning.

Keywords: Random Matrix Theory, Covariance matrices, Transfer Learning, Multi-Task Learning, Theory of machine learning.

Abstract: ML has been quite successful to solve many real-world applications going from supervised to unsupervised tasks due to the development of powerful algorithms (SVM, Deep Neural Network, Spectral Clustering, etc). These algorithms are based on optimization schemes motivated by low dimensional intuitions which collapse in high dimension, a phenomenon known as the "*curse of dimensionality*". Nonetheless, by assuming the data dimension and their number to be both large and comparable, RMT provides a systematic approach to assess the (statistical) behavior of these large learning systems, to properly understand and improve them when applied to large dimensional data. Previous random matrix analyses (Mai & Couillet, 2018; Liao & Couillet, 2019; Deng et al., 2019) have shown that asymptotic performances of most machine learning and signal processing methods depend only on first and second-order statistics (means and covariance matrices of the data). This makes covariance matrices extremely rich objects that need to be "well treated and understood". The thesis demonstrates first how poorly naive covariance matrix processing can destroy machine learning algorithms by introducing biases that are difficult to clean, whereas consistent random-matrix estimation of the functionals of interest avoids biases. We then exemplify how means and covariance matrix statistics of the data are sufficient (through simple functionals) to handle the statistical behavior of even quite involved algorithms of modern interest, such as multi-task and transfer learning methods. The large dimensional analysis allows furthermore for an improvement of multi-task and transfer learning schemes