A random matrix framework for large dimensional machine learning and neural networks

Zhenyu Liao

université PARIS-SACLAY

CentraleSupélec

NNT : 2019SACLC068

Thèse de doctorat

Théorie des matrices aléatoires pour l'apprentissage automatique en grande dimension et les réseaux de neurones

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

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

Thèse présentée et soutenue à Gif-sur-Yvette, le 30/09/2019, par

M. ZHENYU LIAO

Composition du Jury :

M. Eric Moulines Professeur, École Polytechnique	Président
M. Julien Mairal Directeur de recherche (INRIA), Grenoble	Rapporteur
M. Stéphane Mallat Professeur, ENS Paris	Rapporteur
M. Arnak Dalalyan Professeur, ENSAE	Examinateur
M. Jakob Hoydis Docteur, ingénieur de recherche, Nokia Bell Labs Saclay	Examinateur
Mme. Mylène Maida Professeure, Université de Lille	Examinateur
M. Michal Valko Chargé de recherche (INRIA), Lille	Examinateur
M. Romain Couillet Professeur, CentraleSupélec	Directeur de thèse
M. Yacine Chitour Professeur, Université Paris-Sud	Co-directeur de thèse

UNIVERSITY PARIS-SACLAY

DOCTORAL THESIS

A Random Matrix Framework for Large Dimensional Machine Learning and Neural Networks

Author:
Zhenyu LIAO

Supervisor:
Romain COUILLET
Yacine CHITOUR

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Laboratoire des Signaux et Systèmes
Sciences et technologies de l'information et de la communication (STIC)

October 15, 2019

Contents

List of Symbols

Mathematical Symbols

$\mathbb{R}$	Set of real numbers.
$\mathbb{C}$	Set of complex numbers, we denote $\mathbb{C}^+$ the set $\{z \in \mathbb{C}, \Im[z] > 0\}$.
$(\cdot)^\mathsf{T}$	Transpose operator.
$\mathrm{tr}(\cdot)$	Trace operator.
$\mathrm{diag}(\cdot)$	Diagonal operator, for $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathrm{diag}(\mathbf{A}) \in \mathbb{R}^n$ is the vector with entries $\{\mathbf{A}_{ii}\}_{i=1}^n$; for $\mathbf{a} \in \mathbb{R}^n$, $\mathrm{diag}(\mathbf{a}) \in \mathbb{R}^{n\times n}$ is the diagonal matrix taking $\mathbf{a}$ as its diagonal.
$\|\cdot\|$	Operator (or spectral) norm of a matrix and Euclidean norm of a vector.
$\|\cdot\|_F$	Frobenius norm of a matrix, $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})}$.
$\|\cdot\|_\infty$	Infinite norm of a matrix, $\|\mathbf{A}\|_\infty = \max_{i,j} \|\mathbf{A}_{ij}\|$.
$\mathrm{supp}(\cdot)$	Support of a (real or complex) function.
$\mathrm{dist}(\cdot)$	Distance between elements in a metric space.
$P(\cdot)$	Probability of an event on the (underlying) probability measure space $(\Omega, \mathcal{F}, P)$.
$\mathbb{E}[\cdot]$	Expectation operator.
$\mathrm{Var}[\cdot]$	Variance operator, $\mathrm{Var}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$ if the first two moments of x exist.
$\xrightarrow{a.s.}$	Almost surely convergence. We say a sequence $x_n \xrightarrow{a.s.} x$ if $P(\lim_{n\to\infty} x_n = x) = 1$.
$Q(x)$	Q-function: the tail distribution of standard Gaussian $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$.
$O(1)$, $o(1)$	A sequence x_n is bounded or converges to zero as $n \to \infty$, respectively.

Vectors and Matrices

$\mathbf{x}_i \in \mathbb{R}^p$	Input data/feature vector.
$\mathbf{z}_i \in \mathbb{R}^p$	Random vector having i.i.d. zero mean and unit variance entries.
$\mathbf{X}$	Data/feature matrix having $\mathbf{x}_i$ as column vectors.
$\mathbf{Z}$	Random matrix having $\mathbf{z}_i$ as column vectors.
$\boldsymbol{\mu}$	Mean vector.
$\mathbf{C}$	Covariance matrix.
$\mathbf{K}$	Kernel matrix.
$\mathbf{w}$, $\mathbf{W}$	Weight vector or matrix (in the context of neural networks).
$\boldsymbol{\beta}$	Regression vector.
$\mathbf{I}_n$	Identity matrix of size n.
$\mathbf{1}_n$	Column vector of size n with all entries equal to one.
$\mathbf{0}_n$	Column vector of size n with all entries equal to zero.
$\mathbf{Q}_\mathbf{A}$	Resolvent of a symmetric real matrix $\mathbf{A}$, see Definition 4.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Prof. Romain Couillet and co-advisor Yacine Chitour, for their continuous support of my Ph.D study and related research, for their great patience, incredible motivation, and immense knowledge. Their guidance helped me in all the three years of my Ph.D. research and in writing of this thesis. I could not imagine having better advisors and mentors than them.

Besides my advisors, I would like to thank the two rapporteurs: Prof. Julien Mairal and Prof. Stéphane Mallat, for their helpful and insightful comments on my manuscript. I would like to thank all jury members for their time and effort in coming to my defense and in evaluating my work.

A very special gratitude goes out to Fondation CentraleSupélec for helping and providing the funding for my research work.

I would also like to express my gratitude to colleagues in both LSS and GIPSA, I have had a wonderful time in both labs. In particular, I am grateful to Christian David, Myriam Baverel, Sylvie Vincourt, Huu Hung Vuong, Stéphanie Douesnard, Luc Batalie, Anne Batalie and Prof. Gilles Duc at LSS and Marielle Di Maria, Virginie Faure at GIPSA-lab, for their unfailing support and assistance during my Ph.D. study and research.

With a special mention to my teammates in Couillet's group: Hafiz, Xiaoyi, Cosme, Malik, Lorenzo, Mohamed, Arun and Cyprien, I really enjoy working and staying with you guys!

I would also like to thank my friends, in France and in China, for their mental support and helpful discussions: these are really important to me.

Last but not the least, I would like to thank my family: my father Liao Yunxia, mother Tao Zhihui, and my wife Huang Qi for supporting me spiritually throughout my three years of Ph.D. study and my life in general.

Introduction (en français)

L'apprentissage automatique étudie la capacité des systèmes d'intelligence artificielle (IA) à acquérir leurs propres connaissances, en extrayant des "modèles" à partir de données. L'introduction de l'apprentissage automatique permet aux ordinateurs d'appréhender le monde réel et de prendre des décisions qui semblent (plus ou moins) subjectives. Les algorithmes d'apprentissage automatique simples, tels que le filtrage collaboratif ou la classification naïve bayésienne, sont utilisés en pratique pour traiter des problèmes aussi divers que la recommandation de films ou la détection de spam dans les courriers électroniques [GBC16].

La performance des algorithmes d'apprentissage automatique repose énormément sur la *représentation* des données. La représentation qui (idéalement) contient les informations les plus cruciales pour effectuer la tâche s'appelle les *caractéristiques* ("features" en anglais) des données et peut varier d'un cas sur l'autre en fonction du problème. Par exemple, les caractéristiques de couleur peuvent jouer un rôle plus important dans la classification des images de chats "noirs versus blancs" que, par exemple, les caractéristiques qui capturent la "forme" des animaux dans les images.

De nombreuses tâches d'intelligence artificielle peuvent être résolues en concevant le bon ensemble de caractéristiques, puis en les transmettant à des algorithmes d'apprentissage simples pour la prise de décision. Cependant, cela est plus facile à dire qu'à faire et pour la plupart des tâches, il est souvent très difficile de savoir quelles caractéristiques doivent être utilisée pour prendre une décision éclairée. À titre d'exemple concret, il est difficile de dire comment chaque pixel d'une image doit peser de sorte que l'image ressemble plus à un chat qu'à un chien. Pendant assez longtemps, trouver ou concevoir les caractéristiques les plus pertinentes avec une expertise humaine, ou "feature engineering", a été considéré comme la clé pour les systèmes d'apprentissage automatique afin d'obtenir de meilleures performances [BCV13].

Les réseaux de neurones, en particulier les réseaux de neurones profonds, tentent d'extraire des caractéristiques de "haut niveau" (plus abstraites) en introduisant des combinaisons non-linéaires des représentations plus simples (de bas niveau) et ont obtenu des résultats impressionnants au cours de la dernière décennie [Sch15]. Malgré tous les succès remportés avec ces modèles, nous n'avons qu'une compréhension assez rudimentaire de pourquoi et dans quels contextes ils fonctionnent bien. De nombreuses questions sur la conception de ces réseaux, telles que la détermination du nombre de couches et de la taille de chaque couche, le type de fonction d'activation à utiliser, restent sans réponse.

Il a été observé empiriquement que les réseaux de neurones profonds présentent un avantage crucial lors du traitement de données **de grande dimension** et **nombreuses**, autrement dit, lorsque la dimension des données p et leur nombre n sont grands. Par exemple, le jeu de données MNIST [LBBH98] contient $n = 70\,000$ d'images de chiffres, de dimension $p = 28 \times 28 = 784$ chacune, réparties en 10 classes (nombres $0 - 9$). Par conséquent, les systèmes d'apprentissage automatique qui traitent ces grands jeux de

données sont également de taille énorme: le nombre de paramètres du modèle N est au moins du même ordre que la dimension p et peuvent parfois même être beaucoup plus nombreux que n.

Plus généralement, les grands systèmes d'apprentissage qui permettent de traiter des jeux de données de très *grandes dimensions* deviennent de plus en plus importants dans l'apprentissage automatique moderne aujourd'hui. Contrairement à l'apprentissage en petite dimension, les algorithmes d'apprentissage en grande dimension sont sujets à divers phénomènes **contre-intuitifs**, qui ne se produisent jamais dans des problèmes de petite dimension. Nous montrerons comment les méthodes d'apprentissage automatique, lorsqu'elles sont appliquées à des données de grande dimension, peuvent en effet se comporter totalement **différemment** des intuitions en petite dimension. Comme nous le verrons avec les exemples de (l'estimation des) matrices de covariance et de matrices de noyau, le fait que n n'est pas *beaucoup plus grand* que p (disons $n \sim 100p$) rend inefficace de nombreux résultats statistiques dans le régime asymptotique "standard", où on suppose $n \to \infty$ avec p fixé. Ce "*comportement perturbateur*" est en effet l'une des difficultés principales qui interdisent l'utilisation de nos intuitions de petite dimension (de l'expérience quotidienne) dans la compréhension et l'amélioration des systèmes d'apprentissage automatique en grande dimension.

Néanmoins, en supposant que la dimension p et le nombre n de données sont à la fois grands et comparables, dans le régime *double asymptotique* où $n, p \to \infty$ avec $p/n \to \bar{c} \in (0, \infty)$, la **théorie des grandes matrices aléatoires (RMT)** nous fournit une approche systématique pour évaluer le comportement statistique de ces grands systèmes d'apprentissage sur des données de grande dimension. Comme nous le verrons plus loin, dans les deux exemples de la matrice de covariance empirique ainsi que la matrices de noyau, RMT nous fournit un accès direct aux performances, et par conséquent une compréhension plus profonde de ces objets clés, ainsi que la direction pour l'amélioration de ces grands systèmes. L'objectif principal de cette thèse est d'aller bien au-delà de ces exemples simples et de proposer une méthodologie complète pour l'analyse des systèmes d'apprentissage plus élaborés et plus pratiques: pour **évaluer** leurs performances, pour mieux les **comprendre** et finalement pour les **améliorer**, afin de mieux gérer les problèmes de grandes dimensions aujourd'hui. La méthodologie proposée est suffisamment souple pour être utilisée avec des méthodes d'apprentissage automatique aussi répandues que les régressions linéaires, les SVM, mais aussi pour les réseaux de neurones simples.

Chapter 1

Introduction

The field studying the capacity of artificial intelligence (AI) systems to acquire their own knowledge, by extracting patterns from raw data is known as machine learning. The introduction of machine learning enables computers to "learn" from the real world and make decisions that appear subjective. Simple machine learning algorithms such as collaborative filtering or naive Bayes are being used in practice to treat as diverse problems as movie recommendation or legitimate e-mails versus spam classification [GBC16].

The performance of machine learning algorithms relies largely on the *representation* of the data they need to handle. The representation that (ideally) contains the crucial information to perform the task is known as the data *feature*, and may vary from case to case depending on the problem at hand. As an example, color features may play a more important role in the classification between black and white cats than, for instance, the features that capture the "look" or the "shape" of the animals.

Many AI tasks can be solved by designing the right set of features and then providing these features to simple machine learning algorithms to make decisions. However, this is easier said than done and for most tasks, it is unclear which feature should be used to make a wise decision. As a concrete example, it is difficult to say how each pixel in a picture should weigh so that the picture looks more like a cat than a dog. For quite a long time, finding or designing the most relevant features with human expertise, or "feature engineering", has been considered the key for machine learning systems to achieve better performance [BCV13].

Neural networks (NNs), in particular, deep neural networks (DNNs), try to extract high-level features by introducing representations that are expressed in terms of other, simpler representations and have obtained impressive achievements in the last decade [KSH12, Sch15]. Yet for all the success won with these models, we have managed only a rudimentary understanding of why and in what contexts they work well. Many questions on the design of these networks, such as how to decide the number of layers or the size of each layer, what kind of activation function should be used, how to train the network more efficiently to avoid overfitting, remain unanswered.

It has been empirically observed that deep neural network models exhibit a major advantage (against "shallow" models) when handling **large dimensional** and **numerous** data, i.e., when both the data dimension p and their number n are large. For example, the popular MNIST dataset [LBBH98] contains $n = 70\,000$ images of handwritten digits, of dimension $p = 28 \times 28 = 784$ each, from 10 classes (numbers $0 - 9$). As a consequence, the machine learning systems to "process" these large dimensional data are also huge in their size: the number of system parameters N is at least in the same order of p (so as to map the input data to a scalar output for example) and can sometimes be even much

larger than n, in the particular case of modern DNNs.[1]

More generally, **large dimensional** data and learning systems are ubiquitous in modern machine learning. As opposed to small dimensional learning, large dimensional machine learning algorithms are prone to various **counterintuitive** phenomena that never occur in small dimensional problems. In the following Section 1.1, we will show how machine learning methods, when applied to large dimensional data, may indeed strikingly **differ** from the low dimensional intuitions upon which they are built. As we shall see with the examples of sample covariance matrices and kernel matrices, the fact that n is not *much larger* than p (say $n \approx 100p$) annihilates many results from standard asymptotic statistics that assume $n \to \infty$ alone with p fixed. This "*disrupting behavior*" is indeed one of the main difficulties that prohibit the use of our low-dimensional intuitions (from everyday experience) in the comprehension and improvement of large-scale machine learning systems.

Nonetheless, by considering the data dimension p and their number n to be both large and comparable and positioning ourselves in the *double asymptotic* regime where $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, **random matrix theory (RMT)** provides a systematic approach to assess the (statistical) behavior of these large learning systems on large dimensional data. As we shall see next, in both examples of sample covariance matrices and kernel matrices, RMT provides direct access to the performance of these objects and thereby allows for a deeper understanding as well as further improvements of these large systems. The major objective of this thesis is to go well beyond these toy examples and to propose a fully-fledged RMT-based framework for more elaborate and more practical machine learning systems: to **assess** their performance, to better **understand** their mechanism and to carefully **refine** them, so as to better handle large dimensional problems in the big data era today. The proposed framework is flexible enough to handle as popular machine learning methods as linear regressions, SVMs, and also to scratch the surface of more involved neural networks.

1.1 Motivation: the Pitfalls of Large Dimensional Statistics

In many modern machine learning problems, the dimension p of the observations is as large as – if not much larger than – their number n. For image classification tasks, it is hardly possible to have more than $100p$ image samples per class: with the MNIST dataset [LBBH98] we have approximately $n = 6\,000$ images of size $p = 784$ for each class; while the popular ImageNet dataset [RDS+15] contains typically no more than $n = 500\,000$ image samples of dimension $p = 256 \times 256 = 65\,536$ in each class.

More generally, in modern signal and data processing tasks we constantly face the situation where n, p are both large and comparable. In genomics, the identification of correlations among hundreds of thousands of genes based on a limited number of independent (and expensive) samples induces an even larger ratio p/n [AGD94]. In statistical finance, portfolio optimization relies on the opposite need to invest in a large number p of assets to reduce volatility but at the same time to estimate the current (rather than past) asset statistics from a relatively small number n of "short-time stationary" asset return records [LCPB00].

As we shall demonstrate later in this section, the fact that in these problems n, p are

[1] Simple DNNs such as the LeNet [LBBH98] (of 5 layers only) can have $N = 60\,000$ parameters. For more involved structures such as the deep residual network model [HZRS16] (of more than 100 layers), N can be as large as hundreds or thousands of millions.

both large and in particular, n is not *much larger* than p inhibits most of the results from standard asymptotic statistics that assume n alone is large [VdV00]. As a rule of thumb, by *much larger* we mean here that n must be at least 100 times as large as p for standard asymptotic statistics to be of practical convenience (see our argument on covariance estimation in Section 1.1.1). Many algorithms in statistics, signal processing, and machine learning are precisely derived from this inappropriate $n \gg p$ assumption. A main primary objective of this thesis is to cast a light on the resulting biases and problems incurred, and then to provide a systematic random matrix framework that helps better understand and improve these algorithms.

Perhaps more importantly, we shall see that the low dimensional intuitions which are at the core of many machine learning algorithms (starting with spectral clustering [NJW02, VL07]) often strikingly fail when applied in a simultaneously large n, p setting. A compelling key disrupting property lies in the notion of "distance" between large dimensional data vectors. Most classification methods in machine learning are rooted in the observation that random vectors arising from a mixture distribution (say Gaussian) gather in "groups" of close-by vectors in Euclidean norm. When dealing with large dimensions, concentration phenomena arise that makes Euclidean distances "non-informative", if not counterproductive: vectors of the same Gaussian mixture class may be further away in Euclidean distance than vectors arising from different classes, while, paradoxically, non-trivial classification of the whole set of $n \sim p$ data may still be doable. This fundamental example of the "curse of dimensionality" phenomenon, as well as its effect on the popular kernel methods and simple nonlinear neural networks, will be discussed at length in Section 1.1.2 and the remainder of the manuscript as well.

1.1.1 Sample covariance matrices in the large n, p regime

Covariance matrices, as a measure of the joint variability between different entries of a random vector, play a significant role in a host of signal processing and machine learning methods. It is particularly efficient in classifying data vectors that are most distinguished through their second order statistics, for instance, in the case of EEG time series [DVFRCA14] or synthetic aperture radar (SAR) images [Cha03, VOPT09].

Let us consider the following illustrating example which shows a first elementary, yet counterintuitive, result: for simultaneously large n, p, sample covariance matrices $\hat{\mathbf{C}} \in \mathbb{R}^{p\times p}$ based on n samples $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ are jointly *entry-wise* consistent estimators of the population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ (in particular, $\|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \to 0$ as $n, p \to \infty$ for $\|\mathbf{C}\|_\infty \equiv \max_{ij} |\mathbf{C}_{ij}|$), while overall being extremely poor estimators for a majority of covariance-based methods (i.e., $\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0$ with here $\|\cdot\|$ the operator norm). This brings forth a first counterintuitive large dimensional observation: matrix norms are *not* equivalent from a large n, p standpoint.

Let us detail this claim, in the simplest case where $\mathbf{C} = \mathbf{I}_p$. Consider a data set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ of n independent and identically distributed (i.i.d.) observations from a p-dimensional Gaussian distribution, i.e., $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ for $i = 1, \ldots, n$. We wish to estimate the population covariance matrix $\mathbf{C} = \mathbf{I}_p$ from the n available samples. The maximum likelihood estimator in this zero-mean Gaussian setting is the sample covariance matrix $\hat{\mathbf{C}}$ defined by

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T}. \tag{1.1}$$

By the strong law of large numbers, for fixed p, $\hat{\mathbf{C}} \to \mathbf{I}_p$ almost surely as $n \to \infty$, so

that $\|\hat{\mathbf{C}} - \mathbf{I}_p\| \xrightarrow{a.s.} 0$ holds for any standard matrix norm and in particular for the operator norm.

One must be more careful when dealing with the regime $n, p \to \infty$ with ratio $p/n \to c \in (0, \infty)$ (or, from a practical standpoint, n is not much larger than p). First, note that the entry-wise convergence still holds since, invoking the law of large numbers again

$$\hat{\mathbf{C}}_{ij} = \frac{1}{n} \sum_{l=1}^{n} \mathbf{X}_{il} \mathbf{X}_{jl} \xrightarrow{a.s.} \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

Besides, by a concentration inequality argument, it can even be shown that

$$\max_{1 \leq i,j \leq p} \left| (\hat{\mathbf{C}} - \mathbf{I}_p)_{ij} \right| \xrightarrow{a.s.} 0$$

which holds as long as p is no larger than a polynomial function of n, and thus

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_\infty \xrightarrow{a.s.} 0.$$

Consider now the (very special) case of n, p both large but with $p > n$. Since $\hat{\mathbf{C}}$ is the sum of n rank one matrices (as per the form (1.1)), the rank of $\hat{\mathbf{C}}$ is *at most* equal to n and thus, being a $p \times p$ matrix with $p > n$, the sample covariance matrix $\hat{\mathbf{C}}$ is a singular matrix having at least $p - n > 0$ null eigenvalues. As a consequence,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\| \not\to 0$$

for $\|\cdot\|$ the matrix operator (or spectral) norm. This claim, derived from the case of $\mathbf{C} = \mathbf{I}_p$ with $p > n$, actually holds true in the general case where $n, p \to \infty$ with $p/n \to c > 0$. As such, as claimed at the beginning of this subsection, matrix norms cannot be considered equivalent in the regime where p is not negligible compared to n. This follows from the fact that the equivalence factors depend on the matrix size p; here for instance, $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$.

Unfortunately, in practice, the (non-converging) operator norm is of more practical interest than the (converging) infinity norm.

Remark 1.1 (On the importance of operator norm)**.** *For practical purposes, this loss of norm equivalence raises the question of the relevant matrix norm to be considered in any given application. For many applications in machine learning, the operator (or spectral) norm turns out to be much more relevant than the infinity norm. First, the operator norm is the matrix norm induced by the Euclidean norm of vectors. Thus, the study of regression vectors or label/score vectors in classification is naturally attached to the spectral study of matrices. Besides, we will often be interested in the asymptotic* equivalence *of families of large dimensional matrices. If $\|\mathbf{A}_p - \mathbf{B}_p\| \to 0$ for matrix sequences $\{\mathbf{A}_p\}$ and $\{\mathbf{B}_p\}$, indexed by their dimension p, according to Weyl's inequality (e.g., Lemma 2.10 or [HJ12, Theorem 4.3.1]),*

$$\max_i \left| \lambda_i(\mathbf{A}_p) - \lambda_i(\mathbf{B}_p) \right| \to 0$$

for $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots$ the eigenvalues of $\mathbf{A}$ in a decreasing order. Besides, for $\mathbf{u}_i(\mathbf{A}_p)$ an eigenvector of $\mathbf{A}_p$ associated with an isolated *eigenvalue $\lambda_i(\mathbf{A}_p)$ (i.e., such that $\min(|\lambda_{i+1}(\mathbf{A}_p) - \lambda_i(\mathbf{A}_p)|, |\lambda_i(\mathbf{A}_p) - \lambda_{i-1}(\mathbf{A}_p)|) > \epsilon$ for some $\epsilon > 0$ uniformly on p),*

$$\left\| \mathbf{u}_i(\mathbf{A}_p) - \mathbf{u}_i(\mathbf{B}_p) \right\| \to 0.$$

These results ensure that, as far as spectral properties are concerned, $\mathbf{A}_p$ can be studied equivalently through $\mathbf{B}_p$. We will often use this argument to examine intractable random matrices $\mathbf{A}_p$ by means of a tractable ersatz $\mathbf{B}_p$, which is the main approach to handle the subtle nonlinearity in many random matrix models of interest in machine learning.

The pitfall that consists in assuming that $\hat{\mathbf{C}}$ is a valid estimator of $\mathbf{C}$ since $\|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \xrightarrow{a.s.} 0$ may thus have deleterious practical consequences when n is not significantly larger than p.

Resuming on our norm convergence discussion, it is now natural to ask whether $\hat{\mathbf{C}}$, which badly estimates $\mathbf{C}$, has a controlled asymptotic behavior. There precisely lay the first theoretical interests of random matrix theory. While $\hat{\mathbf{C}}$ itself does not converge in any useful way, its eigenvalue distribution does exhibit a traceable limiting behavior [MP67, SB95, BS10]. The seminal result in this direction, due to Marčenko and Pastur, states that, for $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one, the (random) *discrete empirical spectral distribution* (see also Definition 5)

$$\mu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\hat{\mathbf{C}})}$$

converges in law to a non-random *smooth* limit, today referred to as the "Marčenko-Pastur law" [MP67]

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x-a)^+ (b-x)^+} dx \tag{1.2}$$

where $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$.

Figure 1.1 compares the empirical spectral distribution of $\hat{\mathbf{C}}$ to the limiting Marčenko-Pastur law given in (1.2), for $p = 500$ and $n = 50\,000$.

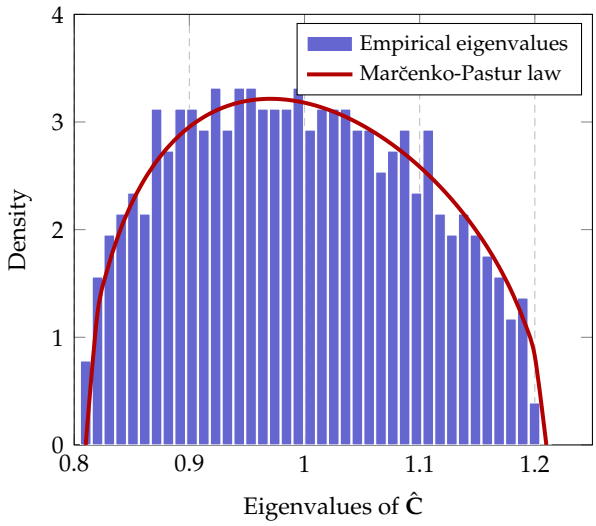


Figure 1.1: Histogram of the empirical spectral distribution of $\hat{\mathbf{C}}$ versus the Marčenko-Pastur law, for $p = 500$ and $n = 50\,000$.

The elementary Marčenko-Pastur result is already quite instructive and insightful.

Remark 1.2 (When is one under the random matrix regime?)**.** *Equation* (1.2) *reveals that the eigenvalues of $\hat{\mathbf{C}}$, instead of concentrating at $x = 1$ as a large-n alone analysis would suggest, spread from $(1 - \sqrt{c})^2$ to $(1 + \sqrt{c})^2$. As such, the eigenvalues span on a range*

$$(1 + \sqrt{c})^2 - (1 - \sqrt{c})^2 = 4\sqrt{c}$$

which is indeed a slow decaying behavior with respect to $c = \lim p/n$. *In particular, for* $n = 100p$, *where one would expect a sufficiently large number of samples for* $\hat{\mathbf{C}}$ *to properly estimate* $\mathbf{C} = \mathbf{I}_p$, *one has* $4\sqrt{c} = 0.4$ *which is a large spread around the mean (and true) eigenvalue* 1. *This is visually confirmed by Figure 1.1 for* $p = 500$ *and* $n = 50\,000$, *where the histogram of the eigenvalues is nowhere near concentrated at* $x = 1$. *As such, random matrix results will be largely more accurate than classical asymptotic statistics even when* $n \sim 100p$. *As a telling example, estimating the covariance matrix of each digit from the popular MNIST dataset [LBBH98] using the sample covariance, made of no more than* 60 000 *training samples (and thus about* $n = 6\,000$ *samples per digit) of size* $p = 28 \times 28 = 784$, *is likely a hazardous undertaking.*

Remark 1.3 (On universality)**.** *Although introduced here in the context of Gaussian distributions for* $\mathbf{x}_i$, *the Marčenko-Pastur law applies to much more general cases. Indeed, the result remains valid so long that the* $\mathbf{x}_i$*'s have i.i.d. normalized entries of zero mean and unit variance (and even beyond this setting [A*$^+$*11, EK09, LC19c]). Similar to the law of large numbers in standard statistics, this* universality *phenomenon commonly arises in random matrix theory and high dimensional statistics and helps to justify the wide applicability of the presented theoretical results, even to real datasets.*

We have seen in this subsection that the sample covariance matrix $\hat{\mathbf{C}}$, despite being a consistent estimator of the population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ for fixed p as $n \to \infty$, fails to provide a precise prediction on the eigenspectrum of $\mathbf{C}$ even with $n \sim 100p$. The fact that we are dealing with large dimensional data vectors has an impact not only on the (direct) estimation of the data covariance or correlation but also on various statistics commonly used in machine learning methods. In the following subsection, we will discuss how popular kernel methods behave differently (from our low dimensional intuition) in large dimensional problems, due to the "curse of dimensionality" phenomenon.

1.1.2 Kernel matrices of large dimensional data

Another less known but equally important example of the curse of dimensionality in machine learning involves the loss of relevance of the notion of Euclidean distance between large dimensional data vectors. To be more precise, we will see in this subsection that, under an asymptotically *non-trivial* classification setting (that is, ensuring that asymptotic classification is neither too simple nor impossible), large and numerous data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ extracted from a few-class mixture model tend to be asymptotically at *equal* distance from one another, irrespective of their mixture class. Roughly speaking, in this non-trivial setting and under reasonable statistical assumptions on the $\mathbf{x}_i$'s, we have

$$\max_{1\le i\ne j\le n} \left\{ \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \to 0 \tag{1.3}$$

for some constant $\tau > 0$ as $n, p \to \infty$, *independently* of the classes (*same or different*) of $\mathbf{x}_i$ and $\mathbf{x}_j$ (here the distance normalization by p is used for compliance with the notations in the remainder of the manuscript but has no particular importance).

This asymptotic behavior is extremely counterintuitive and conveys the idea that classification by standard methods ought not to be doable in this large n, p regime. Indeed, in the conventional low dimensional intuition that forged many of the leading machine learning algorithms of everyday use (such as spectral clustering [NJW02, VL07]), two data points belong to the same class if they are "close" in Euclidean distance. Here we claim that, when p is large, *data pairs are neither close nor far* from each other, regardless of their belonging to the same class or not. Despite this troubling loss of individual

discriminative power between data pairs, we subsequently show that thanks to a *collective behavior* of all data belonging to the same (few and thus of large size each) classes, asymptotic data classification or clustering is still achievable. Better, we shall see that, while many conventional methods devised from small dimensional intuitions do fail in the large dimensional regime, some popular approaches (such as the Ng-Jordan-Weiss spectral clustering method [NJW02] or the PageRank semi-supervised learning approach [AMGS12]) still function. But the core reasons for their functioning are strikingly different from the reasons for their initial designs, and they often operate far from optimally.

The non-trivial classification regime[2]

To get a clear picture of the source of Equation (1.3), we first need to clarify what we refer to as the "asymptotically non-trivial" classification setting. Consider the simplest setting of a binary Gaussian mixture classification. We give ourselves a training set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ of n samples independently drawn from the two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) Gaussian mixture

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$$
$$\mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$$

each with probability $1/2$, for some deterministic $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$, positive definite $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{p \times p}$, and assume

Assumption 1 (Covariance scaling)**.** *As $p \to \infty$, we have for $a \in \{1, 2\}$ that*

$$\max\{\|\mathbf{C}_a\|, \|\mathbf{C}_a^{-1}\|\} = O(1).$$

In the ideal case where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\mathbf{C}_1, \mathbf{C}_2$ are perfectly known, one can devise a (decision optimal) Neyman-Pearson test. For an unknown $\mathbf{x}$, genuinely belonging to $\mathcal{C}_1$, the Neyman-Pearson test to decide on the class of $\mathbf{x}$ reads

$$(\mathbf{x} - \boldsymbol{\mu}_2)^\mathsf{T} \mathbf{C}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)^\mathsf{T} \mathbf{C}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \log \frac{\det(\mathbf{C}_1)}{\det(\mathbf{C}_2)}. \tag{1.4}$$

Writing $\mathbf{x} = \boldsymbol{\mu}_1 + \mathbf{C}_1^{\frac{1}{2}} \mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, the above test is equivalent to

$$T(\mathbf{x}) \equiv \frac{1}{p} \mathbf{z}^\mathsf{T} (\mathbf{C}_1^{\frac{1}{2}} \mathbf{C}_2^{-1} \mathbf{C}_1^{\frac{1}{2}} - \mathbf{I}_p) \mathbf{z} + \frac{2}{p} \Delta\boldsymbol{\mu}^\mathsf{T} \mathbf{C}_2^{-1} \mathbf{C}_1^{\frac{1}{2}} \mathbf{z} + \frac{1}{p} \Delta\boldsymbol{\mu}^\mathsf{T} \mathbf{C}_2^{-1} \Delta\boldsymbol{\mu} - \frac{1}{p} \log \frac{\det(\mathbf{C}_1)}{\det(\mathbf{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} 0. \tag{1.5}$$

where we denote $\Delta\boldsymbol{\mu} \equiv \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ that is then normalized by $1/p$. Since $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{U}^\mathsf{T} \mathbf{z}$ follows the same distribution as $\mathbf{z}$ for $\mathbf{U} \in \mathbb{R}^{p \times p}$ an eigenvector basis of $\mathbf{C}_1^{\frac{1}{2}} \mathbf{C}_2^{-1} \mathbf{C}_1^{\frac{1}{2}} - \mathbf{I}_p$. As such, the random variable $T(\mathbf{x})$ can be written as the sum of p independent random variables. By Lyapunov's central limit theorem (e.g., [Bil12, Theorem 27.3]), we have, as $p \to \infty$,

$$V_T^{-\frac{1}{2}} (T(\mathbf{x}) - \bar{T}) \xrightarrow{d} \mathcal{N}(0, 1)$$

where

$$\bar{T} \equiv \frac{1}{p} \operatorname{tr}(\mathbf{C}_1 \mathbf{C}_2^{-1}) - 1 + \frac{1}{p} \Delta\boldsymbol{\mu}^\mathsf{T} \mathbf{C}_2^{-1} \Delta\boldsymbol{\mu} - \frac{1}{p} \log \frac{\det(\mathbf{C}_1)}{\det(\mathbf{C}_2)},$$
$$V_T \equiv \frac{2}{p^2} \|\mathbf{C}_1^{\frac{1}{2}} \mathbf{C}_2^{-1} \mathbf{C}_1^{\frac{1}{2}} - \mathbf{I}_p\|_F^2 + \frac{4}{p^2} \Delta\boldsymbol{\mu}^\mathsf{T} \mathbf{C}_2^{-1} \mathbf{C}_1 \mathbf{C}_2^{-1} \Delta\boldsymbol{\mu}.$$

[2]This subsection is extracted from our contribution [CLM18].

As a consequence, the classification performance of $\mathbf{x} \in \mathcal{C}_1$ is asymptotically non-trivial (i.e., the classification error neither goes to 0 nor 1 as $p \to \infty$) if and only if $\bar{T}$ is of the same order of $\sqrt{V_T}$ with respect to p. Let us focus on the difference in means $\Delta\boldsymbol{\mu}$ by considering the worst case scenario where the two classes share the same covariance $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$. In this setting, one has from Assumption 1 that

$$\bar{T} = \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} = O(\|\Delta\boldsymbol{\mu}\|^2 p^{-1}), \quad \sqrt{V_T} = \frac{2}{p}\sqrt{\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\mathbf{C}_2^{-1}\Delta\boldsymbol{\mu}} = O(\|\Delta\boldsymbol{\mu}\|p^{-1})$$

so that one must have $\|\Delta\boldsymbol{\mu}\| \geq O(1)$ with respect to p (indeed, if $\|\Delta\boldsymbol{\mu}\| = o(1)$, the classification of $\mathbf{x}$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$ is asymptotically impossible).

Under the critical constraint $\|\Delta\boldsymbol{\mu}\| = O(1)$ we move on to considering the case of different covariances $\mathbf{C}_1 \neq \mathbf{C}_2$ in such a way that their difference $\Delta\mathbf{C} \equiv \mathbf{C}_1 - \mathbf{C}_2$ satisfies $\|\Delta\mathbf{C}\| = o(1)$. In this situation, by a Taylor expansion of both $\mathbf{C}_2^{-1}$ and $\det(\mathbf{C}_2)$ around $\mathbf{C}_1 = \mathbf{C}$ we obtain

$$\bar{T} = \frac{1}{p}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} + \frac{1}{2p}\|\mathbf{C}^{-1}\Delta\mathbf{C}\|_F^2 + o(p^{-1}), \quad V_T = \frac{4}{p^2}\Delta\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\Delta\boldsymbol{\mu} + \frac{2}{p^2}\|\mathbf{C}^{-1}\Delta\mathbf{C}\|_F^2 + o(p^{-2}),$$

which demands $\|\Delta\mathbf{C}\|$ to be at least of order $O(p^{-\frac{1}{2}})$, so that $\|\mathbf{C}^{-1}\Delta\mathbf{C}\|_F^2$ is of order $O(1)$ (as $\|\boldsymbol{\mu}\|$) and can have discriminative power, since $\|\mathbf{C}^{-1}\Delta\mathbf{C}\|_F^2 \leq p\|\mathbf{C}^{-1}\Delta\mathbf{C}\|^2 \leq p\|\Delta\mathbf{C}\|^2$, with equality if and only if both $\mathbf{C}$ and $\Delta\mathbf{C}$ are proportional to identity, i.e., $\mathbf{C} = \epsilon_1\mathbf{I}_p$ and $\Delta\mathbf{C} = \epsilon_2\mathbf{I}_p$. Also, by the Cauchy–Schwarz inequality, we have $|\operatorname{tr}\Delta\mathbf{C}| \leq \sqrt{\operatorname{tr}(\Delta\mathbf{C}^2)\cdot\operatorname{tr}\mathbf{I}_p} = O(\sqrt{p})$, with equality (again) if and only if $\Delta\mathbf{C} = \epsilon\mathbf{I}_p$, and we must therefore have $|\operatorname{tr}\Delta\mathbf{C}| \geq O(\sqrt{p})$. This allows us to conclude on the following non-trivial classification conditions

$$\|\Delta\boldsymbol{\mu}\| \geq O(1), \quad \|\Delta\mathbf{C}\| \geq O(p^{-1/2}), \quad |\operatorname{tr}\Delta\mathbf{C}| \geq O(\sqrt{p}), \quad \|\Delta\mathbf{C}\|_F^2 \geq O(1). \tag{1.6}$$

These are the minimal conditions for classification in the case of perfectly known means and covariances in the following sense: i) if none of the inequalities hold (i.e., if means and covariances from both classes are too close), asymptotic classification must fail; and ii) if at least one of the inequalities is not tight (say if $\|\Delta\boldsymbol{\mu}\| \geq O(\sqrt{p})$), asymptotic classification becomes (asymptotically) trivially easy.

We shall subsequently see that (1.6) precisely induces the asymptotic loss of distance discrimination raised in (1.3) but that in the meantime standard spectral clustering methods based on $n \sim p$ data remain valid in practice.

Asymptotic loss of pairwise distance discrimination

Under the equality case for the conditions in (1.6), the (normalized) Euclidean distance between two distinct data vectors $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b$, $i \neq j$, is given by

$$\begin{aligned}\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 &= \frac{1}{p}\|\mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_i - \mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j\|^2 - \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_i - \mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j) + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 \\ &= \frac{1}{p}(\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i + \mathbf{z}_j^\mathsf{T}\mathbf{C}_b\mathbf{z}_j) - \underbrace{\frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a^{\frac{1}{2}}\mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j}_{O(p^{-1/2})} - \underbrace{\frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_i - \mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j)}_{O(p^{-1})} + \underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2}_{O(p^{-1})}\end{aligned}$$

where we see again from Lyapunov's CLT that the second and third terms are of order $O(p^{-1/2})$ and $O(p^{-1})$ respectively, while the last term is of order $O(p^{-1})$, following directly from the critical condition of (1.6). Denote the average covariance of the two classes $\mathbf{C}^\circ \equiv \frac{1}{2}(\mathbf{C}_1 + \mathbf{C}_2)$ so that $\|\mathbf{C}^\circ\| = O(1)$. Note that

$$\frac{1}{p}\mathbb{E}[\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i] = \frac{1}{p}\mathbb{E}[\mathrm{tr}(\mathbf{C}_a\mathbf{z}_i\mathbf{z}_i^\mathsf{T})] = \frac{1}{p}\mathrm{tr}\,\mathbf{C}_a = \frac{1}{p}\mathrm{tr}\,\mathbf{C}^\circ + \frac{1}{p}\mathrm{tr}(\mathbf{C}_a - \mathbf{C}^\circ) \equiv \frac{1}{2}\tau + \frac{1}{p}\mathrm{tr}\,\mathbf{C}_a^\circ \tag{1.7}$$

where we introduce $\tau \equiv \frac{2}{p}\mathrm{tr}\,\mathbf{C}^\circ = O(1)$ and $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$ for $a \in \{1, 2\}$, the operator norm of which is of order $O(p^{-1/2})$ under the critical condition of (1.6).

As such, it is convenient to further write

$$\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i = \frac{1}{p}\mathrm{tr}\,\mathbf{C}_a + \left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i - \frac{1}{p}\mathrm{tr}\,\mathbf{C}_a\right) \equiv \frac{\tau}{2} + \underbrace{\frac{1}{p}\mathrm{tr}\,\mathbf{C}_a^\circ}_{O(p^{-1/2})} + \underbrace{\psi_i}_{O(p^{-1/2})}$$

with $\psi_i \equiv \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i - \frac{1}{p}\mathrm{tr}\,\mathbf{C}_a = O(p^{-1/2})$ again by the CLT and $\frac{1}{p}\mathrm{tr}\,\mathbf{C}_a^\circ = O(p^{-1/2})$ under (1.6). A similar result holds for $\frac{1}{p}\mathbf{z}_j^\mathsf{T}\mathbf{C}_b\mathbf{z}_j$ and

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \tau + \underbrace{\frac{1}{p}\mathrm{tr}(\mathbf{C}_a^\circ + \mathbf{C}_b^\circ) + \psi_i + \psi_j - \frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a^{\frac{1}{2}}\mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j}_{O(p^{-1/2})}$$
$$\underbrace{- \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_i - \mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j) + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2}_{O(p^{-1})}.$$

This holds *regardless* of the values taken by a, b. Indeed, one can show with some further refinements that

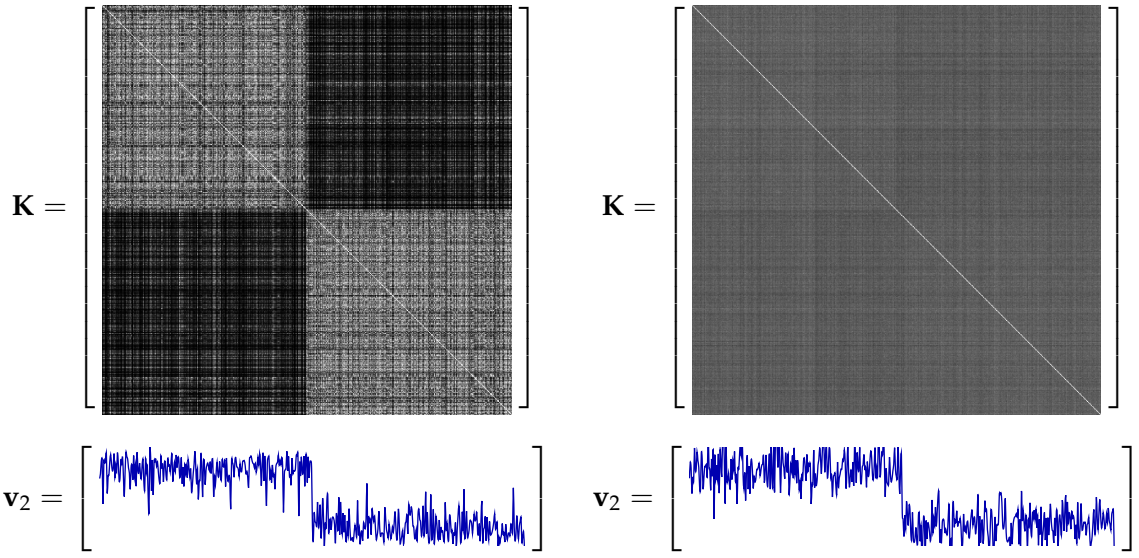
$$\max_{1\le i\neq j\le n}\left\{\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau\right\} \to 0$$

almost surely as $n, p \to \infty$, as previously claimed in (1.3).

To visually confirm the joint convergence of the data distances, Figure 1.2 displays the content of the Gaussian heat kernel matrix $\mathbf{K}$ with $\mathbf{K}_{ij} = \exp\left(-\frac{1}{2p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ (which is, therefore, a Euclidean distance-based similarity measure between data points) and the associated second top eigenvector $\mathbf{v}_2$ for a two-class Gaussian mixture $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ with $\boldsymbol{\mu} = [2;\ \mathbf{0}_{p-1}]$. For a constant $n = 500$, we take $p = 5$ in Figure 1.2a and $p = 250$ in Figure 1.2b.

While the "block-structure" in Figure 1.2a agrees with the low dimensional intuition that data vectors from the same class are "closer" to one another, corresponding to diagonal blocks with larger values (since $\exp(-x/2)$ decreases with the distance) than in non-diagonal blocks, this intuition collapses when large dimensional data vectors are considered. Indeed, in the large data setting of Figure 1.2b, all entries (but obviously on the diagonal) of $\mathbf{K}$ have approximately the same value, which we now know from (1.3) is $\exp(-1)$.

This is no longer surprising to us. However, what remains surprising at this stage of our analysis is that the eigenvector $\mathbf{v}_2$ of $\mathbf{K}$ is not affected by the asymptotic loss of class-wise discrimination of individual distances. Thus spectral clustering seems to work equally well for $p = 5$ or $p = 250$, despite the radical and intuitively destructive change



(a) $p = 5$, $n = 500$. (b) $p = 250$, $n = 500$.

Figure 1.2: Kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for small and large dimensional data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{x}_1, \ldots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$.

in the behavior of $\mathbf{K}$ for $p = 250$. An answer to this question will be given in Section 3.1.1. We will also see that not all kernel choices can reach the same (non-trivial) classification rates, in particular, the popular Gaussian kernel will be shown to be sub-optimal in this respect.

1.1.3 Summary of Section 1.1

In this section, we discussed two simple, yet counterintuitive, examples of common pitfalls in handling large dimensional data.

In the sample covariance matrix example in Section 1.1.1, we made the important remark of the loss of equivalence between matrix norms in the *random matrix regime* where the data/feature dimension p and their number n are both large and comparable, which is at the source of many intuition errors. We in particular insist that, for matrices $\mathbf{A}_n, \mathbf{B}_n \in \mathbb{R}^{n \times n}$ of large sizes

$$\forall i, j, \ (\mathbf{A}_n - \mathbf{B}_n)_{ij} \to 0 \not\Rightarrow \|\mathbf{A}_n - \mathbf{B}_n\| \to 0 \tag{1.8}$$

in operator norm.

We also realized, from a basic reading of the Marčenko-Pastur theorem, that the random matrix regime arises more often than one may think: while $n/p \sim 100$ may seem large enough a ratio for classical asymptotic statistics to be accurate, random matrix theory is, in general, a far more appropriate tool (with as much as 20% gain in precision for the estimation of the covariance eigenvalues).

In Section 1.1.2, we gave a concrete machine learning classification example of the message in (1.8) above. We saw that, in the practically most relevant scenario of non-trivial (not too easy, not too hard) large data classification tasks, the distance between any two data vectors "concentrates" around a constant (1.3), regardless of their respective classes. Yet, since again $(\mathbf{K}_n)_{ij} \to \tau$ does not imply that $\|\mathbf{K}_n - \tau \mathbf{1}_n \mathbf{1}_n^\mathsf{T}\| \to 0$ in

operator norm, we understood that, thanks to a collective effect of the small but similarly "oriented" (informative) fluctuations, kernel spectral clustering remains valid for large dimensional problems.

In a nutshell, the fundamental counterintuitive yet mathematically addressable changes in behavior of large dimensional data have two major consequences to statistics and machine learning: i) most algorithms, originally developed under a low dimensional intuition, are likely to fail or at least to perform inefficiently, yet ii) by benefiting from the extra degrees of freedom offered by large dimensional data, random matrix theory is apt to analyze, improve, and evoke a whole new paradigm for large dimensional learning.

In the following sections of this chapter, we introduce the basic settings under which we work, the machine learning models that we are interested in, as well as the major challenges that we face. In Section 1.2 we discuss three closely related topics: random kernel matrices, random (nonlinear) feature maps, and simple random neural networks. Under a natural and rather general mixture model, we characterize the eigenspectrum of a large kernel matrix and predict the performance of the kernel ridge regressor (which is an explicit function of the kernel function). Random feature maps that are designed to approximate large kernel matrices, take exactly the same form as a single-hidden-layer NN model with random weights. As a consequence, studying such a network is equivalent to assess the performance of a random feature-based kernel ridge regression, and is thus closely connected to the kernel ridge regression model.

1.2 Random Kernels, Random Neural Networks and Beyond

Randomness is intrinsic to machine learning, as the probabilistic approach is one of the most common tools used to study the performances of machine learning methods. Many machine learning algorithms are designed to estimate some unknown probability distribution, to separate the mixture of several different distribution classes, or to generate new samples following some underlying distribution.

Randomness has various sources in machine learning: it may come from the basic statistical assumption that the data are (independently or not) drawn from some probability distribution, from a possibly random search of the hyperparameters of the model (the popular dropout technique in training DNNs [SHK+14]) or from the stochastic nature of the optimization methods applied (e.g., stochastic gradient descent method and its variants). There are also numerous machine learning methods that are intrinsically random, where randomness is not used to enhance a particular part of the model but is indeed the basis of the model itself, for example in the case of random weights neural networks (as we shall discuss later in Section 1.2.2) as well as many tree-based models such as random forests [Bre01] and extremely randomized trees [GEW06].

Let us first focus on the randomness from data by introducing the following multivariate mixture modeling for the input data, which will be considered throughout this manuscript.

Definition 1 (Multivariate mixture model)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n random vectors drawn independently from a K-class mixture model $\mathcal{C}_1, \ldots, \mathcal{C}_K$ so that each class $\mathcal{C}_a$ has cardinality n_a for $a = \{1, \ldots, K\}$ and $\sum_{a=1}^{K} n_a = n$. We say $\mathbf{x}_i \in \mathcal{C}_a$ if*

$$\mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{C}_a^{\frac{1}{2}} \mathbf{z}_i$$

for some deterministic $\boldsymbol{\mu}_a \in \mathbb{R}^p$*, symmetric and positive definite* $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ *and some random vector* $\mathbf{z}_i$ *with i.i.d. zero mean and unit variance entries. In particular, we say* $\mathbf{x}_i$ *follows a K-class* Gaussian mixture model (GMM) *if* $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

The assumption that the data $\mathbf{x}_i$ are a linear or affine transformation of i.i.d. random vectors $\mathbf{z}_i$ characterizes the first and second order statistics of the underlying data distribution. Many RMT results such as the popular Marčenko-Pastur law in (1.2), as stated in Remark 1.3, hold *universally* with respect to the distribution of the independent entries. As such, it often suffices to study the (most simple) Gaussian case to reach a universal conclusion. We will come back to this point with some fundamental RMT results in Section 2.2 and discuss some possible limitations of this "universality" in RMT in Chapter 5.

We insist here that this multivariate mixture model setting for the input data/feature makes the crucial difference between our works and other existing RMT-based analyses of machine learning systems, such as those of L. Zdeborova and F. Krzakala [KMM+13, SKZ14], or those of J. Pennington [PW17, PW17, PSG17]. In all these contributions, the authors are routinely interested in very homogeneous problems with simple data structures, e.g., $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, which makes their analyses more tractable and leads to very simple and more straightforward intuitions or conclusions. On the opposite, we take the more natural choice of **mixture model** with more involved structures, as in the case of GMM in Definition 1. This enhances the practical applicability of our results, as shall be demonstrated throughout this manuscript: when compared to experiments on commonly used datasets, our theoretical results constantly exhibit an extremely close match to practice. This observation, together with the underlying "universality" argument from RMT, conveys a strong applicative motivation for our works.

As we shall always work in the regime where both n, p are large and comparable, we will, according to the discussions in Section 1.1.2, position ourselves under the following non-trivial regime where the separation of the K-class mixture above is neither too easy nor impossible as $n, p \to \infty$.

Assumption 2 (Non-trivial classification)**.** *As* $n \to \infty$*, for* $a \in \{1, \ldots, K\}$

1. $p/n = c \to \bar{c} \in (0, \infty)$ *and* $n_a/n = c_a \to \bar{c}_a \in (0, 1)$.
2. $\|\boldsymbol{\mu}_a\| = O(1)$ *and* $\max\{\|\mathbf{C}_a\|, \|\mathbf{C}_a^{-1}\|\} = O(1)$ *with* $|\operatorname{tr} \mathbf{C}_a^\circ| = O(\sqrt{p})$, $\|\mathbf{C}_a^\circ\|_F^2 = O(\sqrt{p})$ *for* $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \sum_{i=1}^K \frac{n_i}{n} \mathbf{C}_i$.

Assumption 2 is a natural extension of the non-trivial classification condition in (1.6) to the general K-class setting.

Under a GMM for the input data that satisfies Assumption 2, we aim to investigate the (asymptotic) performance of any machine learning method of interest, in the large dimensional regime where $n, p \to \infty$ and $p/n \to \bar{c} \in (0, \infty)$. However, in spite of a huge number of well-established RMT results, the "exact" performance (as a function of the data statistics $\boldsymbol{\mu}, \mathbf{C}$ and the problem dimensionality n, p) still remains technically out of reach, for most machine learning algorithms.

The major technical difficulty that prevents existing RMT results from being applied directly to understand these machine learning systems is the *nonlinear* and sometimes *implicit* nature of these models. Powerful machine learning methods are commonly built on top of highly complex and nonlinear features, which make the evaluation of these methods less immediate than linear and explicit methods. As a consequence, we need a proper adaptation of the matrix-based classical RMT results to handle the entry-wise

nonlinearity (e.g., the kernel function or the activation function in neural networks) or the implicit solutions arising from optimization problems (of logistic regression for instance).

In this section, in pursuit of a satisfying answer to these aforementioned technical difficulties, we review some recent advances in RMT in this vein, starting from large kernel matrices.

1.2.1 Random matrix consideration of large kernel matrices

Large kernel matrices and their spectral properties

In a broad sense, kernel methods are at the core of many, if not most, machine learning algorithms. For a given data set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, most learning mechanisms aim to extract the structural information of the data (to perform classification for instance) by assessing the pairwise comparison $k(\mathbf{x}_i, \mathbf{x}_j)$ for some *affinity metric* $k(\cdot, \cdot)$ to obtain the matrix

$$\mathbf{K} \equiv \left\{ k(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1}^{n}. \tag{1.9}$$

The "cumulative" effect of these comparisons for numerous ($n \gg 1$) data is at the heart of a broad range of applications: from supervised learning methods of maximum margin classifier such as support vector machines (SVMs), kernel ridge regression or kernel Fisher discriminant, to kernel spectral clustering and kernel PCA that work in a totally unsupervised manner.

The choice of the affinity function $k(\cdot, \cdot)$ is central to a good performance of the learning method. A typical viewpoint is to assume that data $\mathbf{x}_i$ and $\mathbf{x}_j$ are not directly comparable (e.g., not linearly separable) in their ambient space but that there exists a convenient *feature mapping* $\phi : \mathbb{R}^p \to \mathcal{H}$ that "projects" the data to some (much higher or even infinite dimensional) feature space $\mathcal{H}$ where $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ are more amenable to comparison, for example, can be linearly separable by a hyperplane in $\mathcal{H}$.

While the dimension of the feature space $\mathcal{H}$ can be much higher than that of the data (p), with the so-called "kernel trick" [SS02] one can avoid the evaluation of $\phi(\cdot)$ and instead work only with the associated *kernel function* $k : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ that is uniquely determined by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{x}_j) \tag{1.10}$$

via Mercer's theorem [SS02]. Some commonly used kernel functions are the radial basis function (RBF, or Gaussian, or heat) kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$, the sigmoid kernel $\tanh(\mathbf{x}_i^\mathsf{T} \mathbf{x}_j + c)$ and polynomial kernels $(\mathbf{x}_i^\mathsf{T} \mathbf{x}_j + c)^d$. These kernel-based learning algorithms have been extensively studied both theoretically and empirically for their feature extraction power in highly nonlinear data manifolds, before the recent "rebirth" of neural networks.

But only very recently has the large dimensional ($p \sim n \gg 1$) nature of the data started to be taken into consideration for kernel methods. We have seen in Section 1.1.2 empirical evidence showing that the kernel matrix $\mathbf{K}$ behaves dramatically differently from its low dimensional counterpart. In the following, we review some theoretical explanations of this counterintuitive behavior provided by RMT analyses.

To assess the eigenspectrum behavior of $\mathbf{K}$ for n, p both large and comparable, the author of [EK10b] considered kernel matrices $\mathbf{K}$ with "shift-invariant" type kernel functions $k(\mathbf{x}_i, \mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ or "inner-product" kernel $f(\mathbf{x}_i^\mathsf{T} \mathbf{x}_j/p)$ for some nonlinear and locally smooth function f. It was shown that the kernel matrix $\mathbf{K}$ is asymptotically

equivalent to a more tractable random matrix $\tilde{\mathbf{K}}$ in the sense of operator norm, as stated in the following theorem.

Theorem 1.1 (Theorem 2.1 in [EK10b]). *For inner-product kernel $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ and $\mathbf{x}_i = \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i$ with positive definite $\mathbf{C} \in \mathbb{R}^{p\times p}$ satisfying Assumption 1 and random vector $\mathbf{z}_i \in \mathbb{R}^p$ with i.i.d. zero mean, unit variance and finite $4+\epsilon$ (absolute) moment entries and f three-times differentiable in a neighborhood of 0, we have*

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$$

in probability as $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, where

$$\tilde{\mathbf{K}} = \left(f(0) + \frac{f''(0)}{2p^2}\operatorname{tr}(\mathbf{C}^2)\right)\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{f'(0)}{p}\mathbf{X}^\mathsf{T}\mathbf{X} + \left(f\left(\frac{\tau}{2}\right) - f(0) - \frac{\tau}{2}f'(0)\right)\mathbf{I}_n \quad (1.11)$$

with $\tau \equiv \frac{2}{p}\operatorname{tr}\mathbf{C} > 0$.

We provide here only the intuition for this result in the inner-product kernel $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ case with Gaussian $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The shift-invariant case can be treated similarly. To prove the universality with respect to the distribution of entries of $\mathbf{z}_i$, a more cumbersome combinatoric approach was adopted in [EK10b], which is likely unavoidable.

Intuition of Theorem 1.1. The proof is based on an entry-wise Taylor expansion of the kernel matrix $\mathbf{K}$, which naturally demands the kernel function f to be locally smooth around zero. More precisely, we obtain, using Taylor expansion of $f(x)$ around $x = 0$ that, for $i \neq j$,

$$\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p) = f(0) + \frac{f'(0)}{p}\mathbf{x}_i^\mathsf{T}\mathbf{x}_j + \frac{f''(0)}{p^2}(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2 + T_{ij}$$

with some higher order (≥ 3) terms T_{ij} that contain higher order derivatives of f. Note that T_{ij} typically contains terms of the form $(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)^\kappa$ for $\kappa \geq 3$ so that T_{ij} is of order $O(p^{-3/2})$, uniformly on $i \neq j$, as a result of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p = O(p^{-1/2})$ by the CLT. As such, the matrix $\{T_{ij}\,\delta_{i\neq j}\}_{1\leq i,j\leq n}$ (of size $n \times n$ and with zeros on its diagonal) can be shown to have a vanishing operator norm as $n, p \to \infty$ since

$$\|\{T_{ij}\,\delta_{i\neq j}\}_{1\leq i,j\leq n}\| \leq n\|\{T_{ij}\delta_{i\neq j}\}_{1\leq i,j\leq n}\|_\infty = o(1)$$

as $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$. We then move on to the second order term $\frac{f''(0)}{p^2}(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2$, the expectation of which is given by

$$\frac{f''(0)}{p^2}\mathbb{E}(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2 = \frac{f''(0)}{p^2}\mathbb{E}_{\mathbf{x}_i}\operatorname{tr}\left(\mathbf{x}_i^\mathsf{T}\mathbb{E}_{\mathbf{x}_j}[\mathbf{x}_j\mathbf{x}_j^\mathsf{T}]\mathbf{x}_i\right) = \frac{f''(0)}{p^2}\operatorname{tr}(\mathbf{C}^2) = O(p^{-1})$$

where we use the linearity of the trace operator to push the expectation inside, together with the fact that $\|\mathbf{C}\| = O(1)$. In matrix form this gives,

$$\frac{f''(0)}{p^2}\mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)^2\delta_{i\neq j}]_{1\leq i,j\leq n} = \frac{f''(0)}{p^2}\operatorname{tr}(\mathbf{C}^2)(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} - \mathbf{I}_n)$$

which is of operator norm of order $O(1)$, even without the diagonal term $\frac{f''(0)}{p^2}\operatorname{tr}(\mathbf{C}^2)\mathbf{I}_n$ (the operator norm of which is of order $O(p^{-1})$ and thus vanishing, also note that we leave the diagonal untreated for the moment). With concentration arguments we can show the fluctuation (around the expectation) of this second order term also has a vanishing operator norm as $n, p \to \infty$ with $p/n \to \bar{c} > 0$, which, together with easy treatment of the diagonal terms, concludes the proof. □

This operator norm consistent approximation $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$, as discussed in Remark 1.1, provides a direct access to the limiting spectral measure (if it exists), the isolated eigenvalues and associated eigenvectors that are of central interest in spectral clustering or PCA applications, as well as the regression or label/score vectors, of (functions of) the intractable kernel matrix $\mathbf{K}$, via the study of the more tractable random matrix model $\tilde{\mathbf{K}}$, according to Weyl's inequality (see Lemma 2.10).

The asymptotic equivalent kernel matrix $\tilde{\mathbf{K}}$ is the sum of three matrices: i) a rank one matrix proportional to $\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, ii) the identity matrix that makes a constant shift of all eigenvalues and iii) a rescaled version of the standard Gram matrix $\frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X}$. The eigenvalue distribution of the Gram matrix, or equivalently of the sample covariance matrix model $\frac{1}{p}\mathbf{X}\mathbf{X}^\mathsf{T}$ (via Sylvester's identity in Lemma 2.3) has been intensively studied in the random matrix literature [SB95], and we shall talk about this model at length in Section 2.2.3. The low rank (rank one here) perturbation of a random matrix model, is known under the name of "spiked model" and has received considerable research attention in the RMT community [BAP05]; we will come back to this point in more details in Section 2.3.

On closer inspection of (1.11), we see that $\tilde{\mathbf{K}}$ is in essence a local linearization of the nonlinear $\mathbf{K}$, in the sense that the nonlinear function $f(x)$ is evaluated solely in the neighborhood of $x = 0$. This is because by the CLT we have $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p = O(p^{-\frac{1}{2}})$ for $i \neq j$, and entry-wise speaking, all off-diagonal entries are constantly equal to $f(0)$ to the first order. As such, the nonlinear function f acts only in a "local" manner around 0: all smooth nonlinear f with the same values of $f(0)$, $f'(0)$ and $f''(0)$ have the same expression of $\tilde{\mathbf{K}}$, up to a constant shift of all eigenvalue due to $f(\tau/2)\mathbf{I}_n$, and consequently have the *same* (asymptotic) performance on all kernel-based spectral algorithms.

In [EK10a] the author considered an "information-plus-noise" model for the data, where each observation consists of two parts: one from a "low dimensional" structure $\mathbf{y}_i \in \mathbb{R}^p$ (e.g., with its ℓ_0 norm $\|\mathbf{y}_i\|_0 = O(1)$) that is considered the "signal" as well as the informative part of the data and the other being high dimensional noise. This modeling indeed assumes that data are randomly drawn from, not exactly a fixed and low dimensional manifold, but somewhere "nearby" such that the resulting observations are perturbed with additive high dimensional noise, more precisely

$$\mathbf{x}_i = \mathbf{y}_i + \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i$$

where $\mathbf{y}_i$ denotes the random "signal" part of the observation and $\mathbf{z}_i$ the high dimension noise part that is independent of $\mathbf{y}_i$. The proof techniques are basically the same as in [EK10b]. However, by assuming that all $\mathbf{x}_i$'s drawn from the *same* distribution, this result is not sufficient for the understanding of $\mathbf{K}$ in a more practical classification context.

More recently in [CBG16], the authors considered the shift-invariant kernel $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$, under a more involved K-class Gaussian mixture model (GMM, see Definition 1) for the data, that is of more practical interest from a machine learning standpoint, and investigated more precisely the eigenspectrum behavior for a kernel spectral clustering purpose. Not only the kernel matrix $\mathbf{K}$ but also the associated (normalized) Laplacian matrix $\mathbf{L}$ defined as

$$\mathbf{L} \equiv n\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$$

are considered in [CBG16], where $\mathbf{D} \equiv \mathrm{diag}(\mathbf{K}\mathbf{1}_n)$ denotes the so-called "degree matrix" of $\mathbf{K}$. While the authors followed the same technical approach as in [EK10b], the fact that they considered a Gaussian mixture modeling provides rich insights into the impact of nonlinearity in kernel spectral clustering applications. For the first time, the exact

stochastic characterization of the isolated eigenpairs are given, as a function of the data statistics $(\boldsymbol{\mu}, \mathbf{C})$, problem dimensionality n, p, as well as the nonlinear f in a local manner.

Built upon the investigation of $\mathbf{K}$ in [CBG16], we evaluate the exact performance of the kernel ridge regression (or least squares support vector machine, LS-SVM) in classifying a two-class GMM, in the following contributions.

(C1) **Zhenyu Liao** and Romain Couillet. Random matrices meet machine learning: a large dimensional analysis of LS-SVM. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2397–2401, 2017.

(J1) **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.

In a nutshell, we consider the following (soft) output decision function from the LS-SVM formulation

$$h(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T} \mathbf{k}(\mathbf{x}) + b, \quad \boldsymbol{\beta} = \mathbf{Q}(\mathbf{y} - b\mathbf{1}_n), \quad b = \frac{\mathbf{1}_n^\mathsf{T} \mathbf{Q} \mathbf{y}}{\mathbf{1}_n^\mathsf{T} \mathbf{Q} \mathbf{1}_n} \tag{1.12}$$

for $\mathbf{k}(\mathbf{x}) \equiv \{f(\|\mathbf{x} - \mathbf{x}_i\|^2/p)\}_{i=1}^n \in \mathbb{R}^n$ and $\mathbf{Q} = \left(\mathbf{K} + \frac{n}{\lambda}\mathbf{I}_n\right)^{-1}$ the so-called *resolvent* of the kernel matrix $\mathbf{K}$ (see Definition 4). By showing that $h(\mathbf{x})$ can be (asymptotically) well approximated by a normally distributed random variable, we obtain the *exact* asymptotic classification performance of LS-SVM. The minimum orders of magnitude distinguishable by an LS-SVM classifier are given by

$$\|\Delta\boldsymbol{\mu}\| = O(1), \quad |\operatorname{tr} \Delta\mathbf{C}| = O(\sqrt{p}), \quad \|\Delta\mathbf{C}\|_F^2 = O(p)$$

for $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\Delta\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2$. This is very close to the theoretical optimum in the oracle setting in (1.6) and is the best rate reported in [CLM18].

Perhaps more importantly, when tested on several real-world datasets such as the MNIST [LBBH98] and Fashion MNIST [XRV17] image datasets, a surprisingly close match is observed between theory and practice, thereby conveying a strong applicative motivation for the study of simple GMM in a high dimensional setting. We will continue this discussion in more details and justify that this observation is in fact not that surprising in Section 3.1.1.

All works above are essentially based on a local expansion of the kernel function f, which follows from the "concentration" of the similarity measure $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$ around a *single* value of the smooth domain of f. More precisely, due to the independence between two data vectors $\mathbf{x}_i, \mathbf{x}_j$, the diagonal and off-diagonal entries of $\mathbf{K}$ behave in a totally different manner. Consider $\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$; we have roughly $f(1)$ on and $f(0)$ off the diagonal of $\mathbf{K}$, so that most entries (of order $\mathrm{O}(n^2)$) of $\mathbf{K}$ evaluate the nonlinear function f around zero. This leads to the "local linearization" of f in the expression of $\tilde{\mathbf{K}}$ in Theorem 1.1 and thereby disregards most of the domain of f.

On the other hand, since $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} \to \mathcal{N}(0,1)$ in distribution as $p \to \infty$ (and is thus of order $O(1)$ with high probability), it appears that $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ is a more natural scaling to avoid an asymptotic linearization of the nonlinear $\mathbf{K}$. Nonetheless, in this case, all diagonal entries become $\|\mathbf{x}_i\|^2/\sqrt{p} = O(\sqrt{p})$ and thus evaluate f asymptotically at infinity, which is again another "improper scaling", but only on the diagonal of $\mathbf{K}$.

In the dot product model $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$, the entries will not "*concentrate*" at $f(0)$. Alternative tool is thus needed in the place of the Taylor expansion, so as to handle the nonlinear function f in this case. This more naturally scaling model was studied in [CS13], where the authors considered the dot product kernel matrix $\mathbf{K}$, the (i,j) entry of which is given by

$$\mathbf{K}_{ij} = \begin{cases} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}. \tag{1.13}$$

The more natural $\sqrt{p}$ normalization inside f helps, as discussed above, avoid the nonlinear $\mathbf{K}$ (asymptotically) acting as a linear model $\mathbf{X}^\mathsf{T}\mathbf{X}$, while the $\sqrt{p}$ scaling outside is simply convenient to ensure the operator norm $\|\mathbf{K}\|$ is order $O(1)$ for n, p large. Note that the diagonal elements are discarded since they are now "improperly scaled" for the evaluation by f.

Again, we are interested in the spectral property of the kernel matrix $\mathbf{K}$ defined in (1.13). For $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, the *empirical spectral measure* $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\mathbf{K})}$ (see Definition 5) of $\mathbf{K}$ has an asymptotically deterministic behavior as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$. This limiting spectral measure of $\mathbf{K}$ was first characterized in [CS13] and then generalized in [DV13] to handle random vectors $\mathbf{x}_i$ with i.i.d. entries of zero mean, unit variance and finite higher order moments, as summarized in the following theorem.

Theorem 1.2 (Theorem 3.4 in [CS13], Theorem 3 in [DV13])**.** *Under some regularity condition for the kernel function f (see Assumption 3 below), the empirical spectral measure of $\mathbf{K}$ converges weakly and almost surely, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, to a probability measure μ. The latter is uniquely defined through its Stieltjes transform $m : \mathbb{C}^+ \to \mathbb{C}^+$, $z \mapsto \int (t - z)^{-1}\mu(dt)$ (see also Definition 6 in Section 2.2), given as the unique solution in $\mathbb{C}^+$ of the cubic equation*[3]

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z)$$

with $a_1 = \mathbb{E}[\xi f(\xi)]$ and $\nu = \mathrm{Var}[f(\xi)] \geq a_1^2$ for standard Gaussian $\xi \sim \mathcal{N}(0, 1)$.

The technical approach to achieve Theorem 1.2 is rather different from that of Theorem 1.1. Instead of performing a local Taylor expansion of f, the authors of [CS13, DV13] rely on the theory of orthogonal polynomials, in particular, of the class of Hermite polynomials defined with respect to the standard Gaussian distribution [AS65, AAR00]. This thus allows for a polynomial approximation of the nonlinear function f as long as it is square-integrable, and thus covers naturally non-differentiable kernel functions. Some useful concepts from the theory of orthogonal polynomial are recalled as follows.

For a probability measure μ, we denote the set of orthogonal polynomials with respect to the scalar product $\langle f, g\rangle = \int fg d\mu$ as $\{P_l(x), l = 0, 1, \ldots\}$, obtained from the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$ such that $P_0(x) = 1$, P_l is of degree l and $\langle P_{l_1}, P_{l_2}\rangle = \delta_{l_1 - l_2}$. By the Riesz-Fischer theorem [Rud64, Theorem 11.43], for any function $f \in L^2(\mu)$, the set of squared integrable functions with respect to $\langle \cdot, \cdot \rangle$, one can formally expand f as

$$f(x) \sim \sum_{l=0}^{\infty} a_l P_l(x), \quad a_l = \int f(x) P_l(x) d\mu(x) \tag{1.14}$$

[3] $\mathbb{C}^+ \equiv \{z \in \mathbb{C},\ \Im[z] > 0\}$. We also recall that, for $m(z)$ the Stieltjes transform of a measure μ, μ can be obtained from $m(z)$ via $\mu([a, b]) = \lim_{\epsilon \downarrow 0} \frac{1}{\pi}\int_a^b \Im[m(x + \imath\epsilon)]dx$ for all $a < b$ continuity points of μ. See Section 2.1.2 for more details.

where "$f \sim \sum_{l=0}^{\infty} P_l$" indicates that $\|f - \sum_{l=0}^{N} P_l\| \to 0$ as $N \to \infty$ (and $\|f\|^2 = \langle f, f \rangle$).

To ensure the polynomial approximation is accurate when truncated at a large but finite degree L, we assume the following assumption holds.

Assumption 3. *For each p, let $\xi_p = \mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ and let $\{P_{l,p}(x),\, l \geq 0\}$ be the set of orthonormal polynomials with respect to the probability measure μ_p of ξ_p. For $f \in L^2(\mu_p)$ for each p, i.e.,*

$$f(x) \sim \sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x)$$

for $a_{l,p}$ defined in (1.14)*, we demand that*

1. *$\sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x)\mu_p(dx)$ converges in $L^2(\mu_p)$ to $f(x)$ uniformly over large p,*
2. *as $p \to \infty$, $\sum_{l=1}^{\infty} |a_{l,p}|^2 \to \nu \in [0, \infty)$. Moreover, for $l = 0, 1, 2$, $a_{l,p}$ converges and we denote a_0, a_1 and a_2 their limits, respectively.*
3. *$a_0 = 0$.*

Since $\xi_p \to \mathcal{N}(0,1)$ as $p \to \infty$, the limiting parameters a_0, a_1, a_2 and ν are simply (generalized) moments of the standard Gaussian measure involving f. Precisely,

$$a_0 = \mathbb{E}[f(\xi)],\ a_1 = \mathbb{E}[\xi f(\xi)],\ a_2 = \frac{\mathbb{E}[(\xi^2 - 1)f(\xi)]}{\sqrt{2}} = \frac{\mathbb{E}[\xi^2 f(\xi)] - a_0}{\sqrt{2}},\ \nu = \mathrm{Var}[f(\xi)]$$

for $\xi \sim \mathcal{N}(0,1)$. These parameters are of crucial significance in determining the eigenspectrum behavior of $\mathbf{K}$. For example, the limiting spectral measure is uniquely determined by a_1 and ν as per Theorem 1.2. The last assumption $a_0 = 0$ is demanded here mainly for technical convenience and will not affect the classification performance in practice, as it adds a non-informative rank one perturbation of the form $a_0(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} - \mathbf{I}_n)/\sqrt{p}$ to the kernel matrix.

The proof of Theorem 1.2 calls for some standard algebraic and probabilistic manipulations in RMT, that will be introduced in Chapter 2. For self-consistency, we include the intuitive proof for Theorem 1.2 in Section A.1.1 of the appendix.

Theorem 1.2 only gives the (limiting) eigenvalue distribution of the kernel matrix $\mathbf{K}$ in (1.13) under a null model which, from a machine learning perspective, is not sufficient to decide for example how to choose f with respect to the data/task at hand. Compared to [CBG16], an important piece of the mixture models is still missing, which leads to our following investigation.

(C2) **Zhenyu Liao** and Romain Couillet. Inner-product kernels are asymptotically equivalent to binary discrete kernels, 2019.

In this contribution we provide, as in Theorem 1.1, a more accessible random matrix $\tilde{\mathbf{K}}$ that is asymptotically equivalent to $\mathbf{K}$, under a two-class multivariate mixture model (as in Definition 1) for data with covariance $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$ that satisfies the non-trivial classification condition of Assumption 2.

Interestingly, $\tilde{\mathbf{K}}$ is (again) nothing but the sum of two matrices: i) the null model $\mathbf{K}_N$ characterized in Theorem 1.2 that depends on f via the two coefficients a_1 and ν and ii) a low rank and informative matrix $\tilde{\mathbf{K}}_I$ that only depends on a_1 and a_2. This means that,

instead of a local behavior of f as in [EK10b, CBG16], the classification performance of the properly scaled dot product kernel $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ is related to f in a more "global", yet still simple, way only via the *three* parameters a_1, a_2 and ν.

On the downside, it must be pointed out that, the studies performed in [CS13, DV13] (as well as in [LC19a] above) only cover the case of identity covariance $\mathbf{C} = \mathbf{I}_p$. By introducing an arbitrary covariance $\mathbf{C}$ and considering $\mathbf{x}_i = \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i$ for $\mathbf{z}_i$ having i.i.d. entries, the limiting spectral measure of $\mathbf{K}$ becomes technically more challenging since it breaks most of the orthogonality properties of the orthogonal polynomial approach in the proofs. But it is a needed extension in pursuit of the general multivariate mixture model, and to compare the performance between different scaling strategies, which is of significant interest for future investigation.

Random feature approximation of kernels

From a practical standpoint, the computation of the kernel matrix $\mathbf{K}$ in (1.9) requires one to evaluate f for all pairs of $(\mathbf{x}_i, \mathbf{x}_j)$. As such, the computational complexity of $\mathbf{K}$ grows quadratically with respect to the number of data points n and can be intense for n large. In this regard, kernel machines are inappropriate to deal with large scale problems. Also, they appear to be less efficient in both online applications [VVLGS12] and applications with privacy concerns [WZWD13].

To cope with this limitation, a large scale kernel approximation technique was originally introduced in [RR08], where the authors proposed to approximate the shift-invariant kernels of the type $k(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i - \mathbf{x}_j)$ with "random Fourier features". This idea was then extended to cover other classes of commonly used kernels such as the additive homogeneous kernels [VZ12] and dot-product kernels [KK12]. The core idea of these random feature approximation techniques is based, in a general manner, on the following remark

$$\mathbf{K}_{ij} = \phi^\mathsf{T}(\mathbf{x}_i)\phi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}_\mathbf{w}[\varphi_\mathbf{w}^\mathsf{T}(\mathbf{x}_i)\varphi_\mathbf{w}(\mathbf{x}_j)]$$

where we shall distinguish the *kernel feature map* $\phi : \mathbb{R}^p \mapsto \mathcal{H}$ that *determines* the kernel function k via Mercer's theorem (and possibly maps to an infinite dimensional space) from the *random feature map* $\varphi_\mathbf{w} : \mathbb{R}^p \mapsto \mathbb{R}^d$ that often maps to low dimensional space (with sometimes $d = 1, 2$) and depends on the random $\mathbf{w}$. As such, by independently drawing $\mathbf{w}_i$ from a predefined distribution, one is able to construct an approximation $\hat{\mathbf{K}}$ of $\mathbf{K}$ by replacing the expectation $\mathbb{E}_\mathbf{w}$ with an empirical average as

$$\hat{\mathbf{K}}_{ij} = \frac{1}{N}\sum_{m=1}^{N} \varphi_{\mathbf{w}_m}(\mathbf{x}_i)^\mathsf{T}\varphi_{\mathbf{w}_m}(\mathbf{x}_j). \tag{1.15}$$

It is desired to have $\hat{\mathbf{K}}$ close to $\mathbf{K}$ in some sense, e.g., in their operator norms $\|\mathbf{K} - \hat{\mathbf{K}}\| \to 0$ as $N \to \infty$ for spectrum-based methods.

In the popular example of random Fourier features, one focuses on the family of shift-invariant kernel functions which depends solely on the difference $\mathbf{x}_i - \mathbf{x}_j$. In the particular case of the Gaussian kernel, Bochner's theorem [Rud62] guarantees that

$$\mathbf{K}_{ij} = e^{-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2} = \mathbb{E}_\mathbf{w}[e^{\imath\mathbf{w}^\mathsf{T}\mathbf{x}_i}e^{-\imath\mathbf{w}^\mathsf{T}\mathbf{x}_j}] \simeq \frac{1}{N}\sum_{m=1}^{N} e^{\imath\mathbf{w}_m^\mathsf{T}\mathbf{x}_i}e^{-\imath\mathbf{w}_m^\mathsf{T}\mathbf{x}_j} \equiv \hat{\mathbf{K}}_{ij} \tag{1.16}$$

for the random Fourier feature $\varphi_\mathbf{w}(\mathbf{x}) = e^{\imath\mathbf{w}^\mathsf{T}\mathbf{x}}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. To avoid complex-valued features, in practice one uses only the real part [RR08], i.e., $\Re[\hat{\mathbf{K}}_{ij}]$. Also, by con-

sidering various distributions of $\mathbf{w}$, several other shift-invariant kernels can be approximated, including the Laplacian and Cauchy kernels, within the general random Fourier feature framework.

We close this subsection with the final remark that, by cascading the Gaussian $\mathbf{w}_m$'s as row vectors, we obtain a random Gaussian matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$, the entries of which are independent standard Gaussian random variables, i.e., $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$. Then, according to (1.16) we have

$$\Re[\hat{\mathbf{K}}_{ij}] = \frac{1}{N}\sigma_1(\mathbf{W}\mathbf{x}_i)^{\mathsf{T}}\sigma_1(\mathbf{W}\mathbf{x}_j) + \frac{1}{N}\sigma_2(\mathbf{W}\mathbf{x}_i)^{\mathsf{T}}\sigma_2(\mathbf{W}\mathbf{x}_j) \tag{1.17}$$

for $\sigma_1(t) = \cos(t)$ and $\sigma_2(t) = \sin(t)$, which coincides with the Gram matrix of the single-layer neural network model with random weights $\mathbf{W}$ and cosine+sine activation functions, as detailed in the following subsection.

1.2.2 Random neural networks

A fundamental link exists between random feature maps (and thus random kernel matrices) and *neural networks with random weights*. Neural networks (NNs) with random weights have their roots in the pioneering works on the perceptron [Ros58] and have then been successively revisited and analyzed in a number of works, both in the feed-forward [SKD92, ASD96] and the recurrent case [Gel93]. More recently, some of these randomized NN models (the so-called extreme learning machine [HZDZ12] as an example for the feed-forward case and the echo state network [Jae01] for the recurrent one) have been shown to achieve satisfactory performances in some problems (see examples in the next paragraph), with a relatively short training time and low model complexity. We refer the readers to [SW17] for a complete overview of randomness in NN models.

Random neural networks are more than a "cheap" way to solve less efficiently a difficult problem, given the limited computational resources at hand. On the contrary, they work sufficiently well in many practical scenarios [HZDZ12] and can even sometimes achieve remarkable accuracies that are comparable to deep and finely tuned structures, in a large number of applications [LML+14]. This seemingly counterintuitive fact is perhaps the key to a deeper comprehension of the intrinsic properties of different NN architectures. Taking the example of convolutional neural networks (CNNs), it was observed in [JKL09] that, for small datasets, completely random filters are sufficient to propagate useful information through a rectified linear (ReLU) network. This observation was used to argue for the significance of ReLU nonlinearities and the (convolutional) pooling operation in image classification problems, as also pointed out in [PDDC09, SKC+11, CP11b, GSB16], and in many other tasks such as image reconstruction [HWH16], denoising and super-resolution [UVL18].

Random NNs have the advantage of being mathematically more tractable, at the same time preserving the "structural" properties of these elaborate learning systems, e.g., the use of appropriate nonlinear functions in DNNs or convolutional filters in CNNs. The study of random NNs can, therefore, shed novel light on the design of more advanced NN models.

In the training procedure of modern NNs, optimization methods such as (stochastic) gradient descent and its variants are always initialized randomly, with the NN weights drawn randomly from a (properly scaled) normal distribution. In this regard, NNs with random weights are essentially the initial states of well-trained NNs, and can provide helpful insight into the understanding of modern DNNs, e.g., to accelerate training, at

least at the very initial stage, by examining the eigenspectrum behavior of the input-output Jacobian matrix [PSG17], or to gain theoretical insights into the geometric properties of DNNs [DLT+18, AZLS18]. In these works, the network weights, or training data, or both of them, are typically assumed to be normally distributed to facilitate theoretical assessment.

In this manuscript, we focus only on feed-forward NNs. For random recurrent NNs, RMT-based analyses have also been established, but are mainly restricted to the linear case [CWSA16].

Let us consider the random weights feed-forward NN with a single hidden layer as in Figure 1.3.

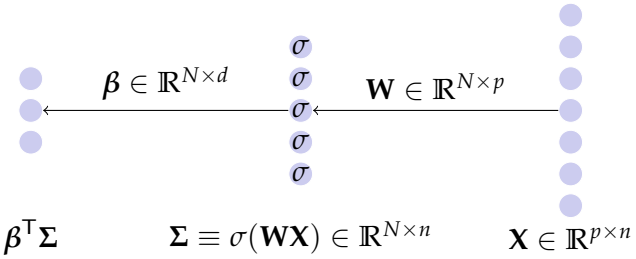


Figure 1.3: Illustration of a single-hidden-layer random NN

For a data matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \in \mathbb{R}^{p\times n}$, we denote $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times n}$ the output of the middle layer comprising of N neurons in total (also referred to as the *random feature matrix*) of $\mathbf{X}$ by premultiplying some random weight matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$ with i.i.d. standard Gaussian entries and then applying entry-wise some nonlinear activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ so that $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{N\times n}$, the column vectors $\sigma(\mathbf{W}\mathbf{x}_i)$ of which are the associated random feature of $\mathbf{x}_i$. With $\boldsymbol{\Sigma}$, the weights of the second layer $\boldsymbol{\beta} \in \mathbb{R}^{N\times d}$ are usually "learned" to adapt the feature matrix $\boldsymbol{\Sigma}$ to the associated target $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$, for example by minimizing $\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\|_F^2$.

In the case where $\boldsymbol{\beta}$ is designed to minimize the regularized mean squared error (MSE) $L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_i)\|^2 + \lambda\|\boldsymbol{\beta}\|_F^2$, for some regularization factor $\lambda \geq 0$, we obtain the explicit *ridge-regressor*

$$\boldsymbol{\beta} \equiv \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{Y}^\mathsf{T} \tag{1.18}$$

which follows from differentiating L with respect to $\boldsymbol{\beta}$ to obtain $\mathbf{0} = \lambda\boldsymbol{\beta} + \frac{1}{n}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\beta} - \mathbf{Y}^\mathsf{T})$ so that $(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\mathsf{T} + \lambda\mathbf{I}_N)\boldsymbol{\beta} = \frac{1}{n}\boldsymbol{\Sigma}\mathbf{Y}^\mathsf{T}$ which, along with $(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\mathsf{T} + \lambda\mathbf{I}_N)^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda\mathbf{I}_n)^{-1}$, gives the result. Note that, similar to the kernel ridge regression investigated in [LC19b], the resolvent $\left(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}$ also plays a central role in the performance analysis of this network.

The single-hidden-layer random NN model presented above, with the first layer fixed at random and the second layer performing a ridge regression, is sometimes referred to as "extreme learning machines" in the literature [HZDZ12]. Note from (1.18) that,

$$\frac{1}{n}[\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}]_{ij} = \frac{1}{n}\sigma(\mathbf{W}\mathbf{x}_i)^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_j)$$

which coincides with the approximated kernel matrix from random feature-based techniques in (1.17). From this perspective, the random weights NN model in Figure 1.3 is in

essence equivalent to a random feature-based kernel ridge regression, i.e., by replacing the kernel matrix $\mathbf{K}$ with its random feature approximation $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ and taking $b = 0$ in the kernel ridge regression formula in (1.12).

Exploiting the close link between random feature maps (and thus random kernel matrices) and random weights NN models, we evaluate the asymptotic performance of the single-hidden-layer random NN model in Figure 1.3, in the regime where $n, p, N \to \infty$ with $p/n \to \bar{c}_1 \in (0, \infty)$ and $N/n \to \bar{c}_2 \in (0, \infty)$, in the following contribution.

(J2) Cosme Louart, **Zhenyu Liao**, Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

By focusing on the randomness of the network weights $\mathbf{W}$ and considering the data-target pair $(\mathbf{X}, \mathbf{Y})$ deterministic, we show that the asymptotic training MSE defined by

$$E_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_i)\|^2 = \frac{1}{n}\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\mathbf{\Sigma}\|_F^2 \tag{1.19}$$

has an asymptotically deterministic behavior. A conjecture is also provided for the test MSE on an independent test set $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ of size $\hat{n}$: $E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{Y}} - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\hat{\mathbf{X}})\|_F^2$, which is later proved in [LC18b] under additional assumptions on $\mathbf{X}$ and $\hat{\mathbf{X}}$. Both the training and test performances of the network depend on the dimensionality of the problem, and on the data as well as the activation function $\sigma(\cdot)$ through the "equivalent" kernel matrix

$$\mathbf{K} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})] \tag{1.20}$$

for $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, which can be explicitly computed via an integration trick [Wil97] for most commonly used σ. This again confirms the strong connection between random weights NNs/random feature models and kernel methods.

Moreover, by (additionally) leveraging the stochastic nature of the data, we are able to dive deeper into the comprehension of (the nonlinearity for instance) in random NN models, with the help of our previous knowledge on random kernel matrices discussed Section 1.1.2. This consideration led to the following contribution.

(C3) **Zhenyu Liao** and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3063–3071. PMLR, 2018.

In a nutshell, we consider data independently drawn from a GMM (see Definition 1), and focus on the interplay between the nonlinear activation $\sigma(\cdot)$ and the data statistics $(\boldsymbol{\mu}_a, \mathbf{C}_a)$. Since only the first and second order statistics are considered here, commonly used activations σ are naturally divided into the following three classes:

1. the *mean-oriented* activations, for which the information in covariance $\mathbf{C}_a$ (asymptotically) does not appear in the equivalent kernel $\mathbf{K}$ in (1.20);

2. the *covariance-oriented* activations, which exploit only the information in $\mathbf{C}_a$ with $\boldsymbol{\mu}_a$ discarded;

3. the "*balanced*" activations, that utilize both first and second order statistics of the data.

A huge gain in classification performance can be observed by applying the nonlinear activation "adapted" to the data/task. For instance, EEG time series data often contain richer information in their covariance structures than the handwritten digits of MNIST, for which the differences in means between classes are more dominant.

In the case of the random feature-based kernel ridge regression in (1.18), we aim to minimize the square loss between the target and the model output. There exist many other loss functions with solutions significantly differing from one another. The square loss turns out to be a convenient choice since it often leads to closed-form solutions as in (1.18). In many other cases, it is hardly possible to obtain closed-form solutions. Indeed, most commonly used machine learning systems (as simple as logistic regression) arise from generic optimization problems that may only take implicit forms. The performance assessment of such implicit systems is, albeit more challenging, a necessary extension for the proposed RMT-based analysis framework, as shall be discussed in the following subsection.

1.2.3 Beyond the square loss: the empirical risk minimization framework

Other loss functions, rather than square loss, can also be used as optimization metrics. In classification applications, it is commonly considered more appropriate to learn a classifier $\boldsymbol{\beta} \in \mathbb{R}^p$ using the following cross-entropy loss

$$L(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\Big[y_i \log(\sigma(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)) + (1-y_i)\log(1-\sigma(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i))\Big] \tag{1.21}$$

with $\sigma(t) = (1+e^{-t})^{-1}$ the *logistic sigmoid* function, for label $y_i \in \{0,1\}$ and feature vectors $\mathbf{x}_i \in \mathbb{R}^p$. This leads to the logistic regression model that, unfortunately, does not admit an explicit solution. The fact that the resulting classifier $\boldsymbol{\beta}$ arises from an implicit optimization problem makes the performance evaluation technically more challenging, mainly due to the complex dependence between the learned parameter $\boldsymbol{\beta}$ and the training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$: it is thus less direct to assess the statistical behavior of $\boldsymbol{\beta}$, and this gets even worse with the logistic sigmoid nonlinearity σ. Despite all the technical difficulties mentioned above, it is nonetheless possible to pursue a stochastic description of $\boldsymbol{\beta}$ and consequently to evaluate the performance of, not only the logistic regression classifier, but also of any differentiable convex loss L that falls into the general empirical risk minimization framework, as briefly recalled below.

First note that the logistic sigmoid function is symmetric with respect to the labels $\{0,1\}$: $\sigma(t)+\sigma(-t) = 1$. Equation (1.21) can thus be rewritten as the minimization of $\frac{1}{n}\sum_{i=1}^n L(\tilde{y}_i\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$, for $L(t) = \log(1+e^{-t})$ and $\tilde{y}_i = (2y_i - 1) \in \{-1,1\}$. In fact, logistic regression is a special case of the general empirical risk minimization framework formulated as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} L(\tilde{y}_i\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{1.22}$$

for some nonnegative convex loss function $L : \mathbb{R} \mapsto \mathbb{R}$ and regularization factor $\lambda \geq 0$. With the logistic loss $L(t) = \log(1+e^{-t})$ one gets the logistic regression formulation as in (1.21), while the least squares classifier (or ridge regressor) is obtained with the square loss $L(t) = (t-1)^2$. Another popular choice is the exponential loss $L(t) = e^{-t}$ that is widely used in boosting algorithms [FSA99].

The empirical minimization principle was first proposed in [Vap92] based on the following consideration: in a binary classification problem, given a training set of n data $\mathbf{x}_i \in \mathbb{R}^p$ with associated label $\tilde{y}_i \in \{-1, 1\}$, $i = 1, \ldots, n$, one wishes to to "learn" a (soft) linear classifier $\boldsymbol{\beta} \in \mathbb{R}^p$ from the available training set such that the thresholded output $\text{sign}(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$ (the "hard" classifier) matches the corresponding label $\tilde{y}_i$, i.e., $\tilde{y}_i = \text{sign}(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$, for all data-label pair $(\mathbf{x}_i, \tilde{y}_i)$ in the training set as well as for unseen $\mathbf{x}$ following the same distribution.

In pursuit of a good classifier $\boldsymbol{\beta}$, a first selection for the loss function is the classification error (also known as the $0-1$ loss, in red below) on the given training set. Despite being the most natural choice, the minimization of the non-convex and non-smooth $0-1$ loss is unfortunately proved to be an NP-hard combinatorial optimization problem in the worst case [BDEL03]. In this regard, many convex surrogate losses are proposed, such as the (non-smooth) hinge loss $L(t) = \max(1-t, 0)$ used in SVMs (in black), the logistic loss (in green) $L(t) = \log(1 + e^{-t})$ used in logistic regression and the square loss $L(t) = (t-1)^2$ (in blue) for linear regression. See Figure 1.4 for an illustration of these different losses for classification.

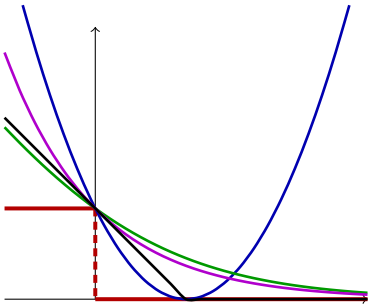


Figure 1.4: Different loss functions for classification: $0-1$ loss (red), logistic loss (green), exponential loss (purple), square loss (blue) and hinge loss (black).

Within the empirical risk framework, the comparison between different designs of loss functions has been long discussed in the statistical learning literature [Vap92, RVC+04, MSV09], mostly in the setting where the number of training data n largely exceeds their dimension p. It was shown in [RVC+04] that, in the limit of large n with p fixed, *all* convex loss functions mentioned above yield the *same* Bayes optimal solution $\boldsymbol{\beta}_*$ that minimizes the $0-1$ loss. Meanwhile, they differ in their rates of convergence (with respect to n) and in this respect, the non-smooth hinge loss is as efficient as the logistic loss, which is much better than the square loss.

The case where n is not much larger than p was not taken into consideration until very recently, mainly due to the fact that, instead of converging to its expectation as when $n \gg p$, the learned parameter $\boldsymbol{\beta}$ remains *random* in the (comparably) large n, p limit and makes its statistical properties less tractable. Relying on ideas from random matrix theory, in [EKBB+13, DM16] the authors managed to capture the stochastic behavior of the M-estimators popularly used in robust statistics that have a similar formulation as in (1.22), but with a loss of the type $L(y_i - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)$.

In a more recent line of works [SC18, CS18], the authors considered the logistic regression model for i.i.d. Gaussian feature $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and showed that, in the regime of $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, the learned $\boldsymbol{\beta}$ is not only biased, but also has a greater variability than classically predicted [SC18]. Indeed, it was shown in [CS18] that, without regularization ($\lambda = 0$ in (1.22)), there exists a sharp "phase transition" threshold for the ratio p/n above which the logistic regression solution almost surely does not

exist.[4]

The results presented above demonstrate the counterintuitive behaviors of not only logistic regression but more generally of the family of empirical risk minimization classifiers in (1.22), where n, p are both large and comparable. As such, a random matrix-based refinement is needed, to gain a better understanding of these statistical learning methods for large dimensional problems, to correct the "bias" induced by the fact that $p \sim n$ and to determine the (problem-dependent) optimal choice of the loss function.

We go beyond the logistic regression setting and consider any twice differentiable and convex loss function L in the following contributions.

(C4) Xiaoyi Mai, **Zhenyu Liao**, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. 2019.

(J3) Xiaoyi Mai and **Zhenyu Liao**. High dimensional classification via empirical risk minimization: Improvements and optimality. (Submitted to) *IEEE Transactions on Signal Processing*, 2019. (Collaborative work not involving the Ph.D. advisors)

Built upon a symmetric two-class GMM (i.e., $-\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$) for the features, we provide, as in [EKBB+13, DM16], an asymptotic stochastic description of the learned parameter $\boldsymbol{\beta}$ (that minimizes (1.22) with $\lambda = 0$)

$$\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\| \to 0, \quad \tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\frac{\eta}{2\tau}\boldsymbol{\beta}_*, \frac{\gamma}{n\tau}\mathbf{I}_p\right)$$

with $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$ the optimal Bayes solution in this setting and constant parameters $(\eta, \gamma, \tau) \in \mathbb{R}^3$ determined by the loss L and the statistics $\boldsymbol{\mu}, \mathbf{C}$, via a system of fixed point equations. In particular, we show that in separating two large dimensional Gaussians with the same covariance $\mathbf{C}$ and opposite means $\pm\boldsymbol{\mu}$, the optimal loss (in the formulation of (1.22) with $\lambda = 0$) is the square loss, which *always* outperforms the maximum likelihood solution (logistic regression in this case), in the regime $n, p \to \infty$ and $p/n \to \bar{c} \in (0, \infty)$. This result is achieved by combining RMT techniques with a heuristic "leave-one-out" approach and will be discussed at length in Section 3.3.

1.2.4 Summary of Section 1.2

In this section, under the high dimensional setting $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$ and a multivariate mixture model (Definition 1) for the input data, we discussed three closely related objects: the kernel matrices built from pairwise comparisons of data points, the random feature maps designed to approximate kernel matrices and the resulting random NN models. We examined the performances of these three models in minimizing the square loss and particularly focus on the role played by the nonlinearity therein

[4] To make this clear, let us consider the classical example of *completely separable* training data $(\mathbf{x}_i, \tilde{y}_i)$ of size n, namely, there exists a linear decision boundary $\mathbf{b} \in \mathbb{R}^p$ such that for $i = 1, \ldots, n$, one has $\tilde{y}_i \mathbf{b}^\mathsf{T} \mathbf{x}_i > 0$. In this case, taking $\lambda = 0$ and $L(t) = \log(1 + e^{-t})$, the minimization of (1.22) with iterative methods such as gradient descent leads to the solution $\boldsymbol{\beta} = \alpha\mathbf{b}$ with $\alpha \to \infty$, *for any given* n, p. Hence, the logistic regressor exists only when the data points "overlap". In the large n, p regime, not only the overlap between data points, but also a sufficient large n/p is indispensable for the existence of a logistic regression solution.

(kernel and activation functions). When more general losses are considered, the (extra) nonlinearity from the loss makes these learning systems less tractable and calls for more advanced tools.

We reviewed some recent advances in the understanding of these large dimensional learning systems and discussed very briefly the main technical approaches on which these theoretical results rely. As has been discussed respectively in Section 1.1.1 and 1.1.2, due to the high dimensional nature of the problem when n, p are comparably large, classical asymptotic statistics fail to predict the performance of neither sample covariance matrices, nor random kernel matrices.

Since random feature maps do no more than approximating kernel matrices, and simple random NNs are indeed random feature-based ridge regressions, all these objects of interest are not working in the expected way they were originally designed for. Consequently, we need a careful refinement of these learning methods, if we are not working with $n \gg 100p$ (see again our argument in Remark 1.2), which is almost *always* the case in practical machine learning applications.

In the topics addressed in this section, we hardly specified the underlying *optimization* method used to obtain the desired solution. As mentioned in Section 1.2.3, with the cross-entropy loss in (1.21), optimization methods such as gradient descent is needed to minimize the objective function. In the following section, we consider the "optimization" aspect of these learning systems, with a particular emphasis on NN models trained with gradient-based methods.

1.3 Training Neural Networks with Gradient Descent

Despite their rapidly growing list of successful applications in as diverse fields as computer vision [KSH12], speech recognition [MDH11] and natural language processing [CW08], our theoretical comprehension of DNNs, is developing at a much more modest pace.

Two salient features of these large neural networks are: i) despite being non-convex optimization problems, training DNN models with simple algorithms such as (stochastic) gradient descent seems always to be able to achieve minimal error and ii) even though they often have far more model parameters than the number of data they are trained on, some models still exhibit remarkably good generalization performance, while others generalize poorly in exactly the same setting [ZBH$^+$16]. A much-expected explanation of these phenomena would be the key to more powerful and reliable network structures.

In this section, we examine the training of NN models with gradient descent algorithms from a "dynamical" standpoint and consider the temporal evolution of both the network weights and the resulting performance. Starting from the toy model of a single-layer linear and convex network (with a unique minimum), in Section 1.3.1 we focus on the gradient descent dynamics (GDDs) and provide a precise characterization of both the training and test performances, as a function of the training time, via a random matrix-based analysis.

When deeper networks (with the number of hidden layers $H \geq 1$) are considered, the optimization of such networks becomes non-convex and more challenging. In Section 1.3.2 we review some basic notions of optimization (with an emphasis on non-convex problems), and discuss some particularly interesting situations in which the underlying non-convex optimization becomes tractable and for which one can *always* find an optimal solution with simple gradient-based methods, sometimes even in linear time.

1.3.1 A random matrix viewpoint

In [SMG13] the authors proposed to study the "dynamics" of gradient descent algorithms in linear DNN models. Despite the linearity of their input-output map, linear networks have highly nonlinear training error behaviors as a function of the time when trained with gradient descent methods. This "learning dynamic" of linear NNs was shown to be closely connected to the eigenvalues of the input-output covariance/correlation matrix. As further justification for their study, empirical evidence was provided showing that the learning dynamics of nonlinear NN, despite being mathematically less tractable due to the (entry-wise) nonlinear activations, exhibit similar behavior to the linear case.

More recently in [AS17], under a teacher-student network setting where a student network tries to "learn" from the noisy samples generated by a predefined teacher network, the authors investigated both the training and test errors, as a function of the gradient descent training time. More precisely, assume a training data matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \in \mathbb{R}^{p\times n}$ with associated target/label $\mathbf{y}^\mathsf{T} = \left[y_1, \ldots, y_n\right] \in \mathbb{R}^n$. With the training pair $(\mathbf{X}, \mathbf{y})$, a weight vector $\mathbf{w} \in \mathbb{R}^p$ is learned using "full-batch" gradient descent to minimize the regularized square loss

$$L(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{1.23}$$

The gradient of L with respect to $\mathbf{w}$ is thus given by $\nabla_\mathbf{w} L(\mathbf{w}) = -\frac{1}{n}\mathbf{X}(\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{w}) + \lambda\mathbf{w}$ so that with a small gradient descent step (or, learning rate) α, we obtain, by performing a continuous-time approximation, the following differential equation

$$\frac{d\mathbf{w}(t)}{dt} = -\alpha\nabla_\mathbf{w} L(\mathbf{w}) = \frac{\alpha}{n}\mathbf{X}(\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{w}) - \alpha\lambda\mathbf{w}$$

the solution of which is explicitly given as

$$\mathbf{w}(t) = e^{-\alpha t(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} + \lambda\mathbf{I}_p)}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\alpha t(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} + \lambda\mathbf{I}_p)}\right)\mathbf{w}_{LS} \tag{1.24}$$

where we denote $\mathbf{w}_{LS} = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} + \lambda\mathbf{I}_p)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}$ the classical ridge regression solution with regularization parameter $\lambda \geq 0$ and $\mathbf{w}_0 = \mathbf{w}(t=0)$ the initialization of gradient descent. We recall the definition of the exponential of a symmetric matrix $\mathbf{A}$

$$e^\mathbf{A} = \sum_{k=0}^{\infty}\frac{1}{k!}\mathbf{A}^k = \mathbf{V}_\mathbf{A} e^{\mathbf{\Lambda}_\mathbf{A}}\mathbf{V}_\mathbf{A}^\mathsf{T}$$

with the eigendecomposition of $\mathbf{A} = \mathbf{V}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathsf{T}$ and $e^{\mathbf{\Lambda}_\mathbf{A}}$ the diagonal matrix with elements equal to the exponential of the diagonal elements of $\mathbf{\Lambda}_\mathbf{A}$.

Equation (1.24) tells us that the temporal evolution of both the weight vector $\mathbf{w}(t)$ and the network performance[5] are in fact functionals of the (regularized) sample covariance matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} + \lambda\mathbf{I}_p$, the evaluation of which is made possible with our proposed resolvent-based RMT techniques and led to the following contribution.

(C5) **Zhenyu Liao** and Romain Couillet. The dynamics of learning: a random matrix approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3072–3081. PMLR, 2018.

[5] e.g., the prediction risk $E_{\text{test}}(t) = \mathbb{E}_{(\hat{\mathbf{x}},\hat{y})}\|\hat{y} - \mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}}\|^2$ under a regression setting on a test set $(\hat{\mathbf{x}}, \hat{y})$ or the correct classification rate $P(\text{sign}(\hat{\mathbf{x}}^\mathsf{T}\mathbf{w}(t)) = \hat{y})$ for an unseen new datum $\hat{\mathbf{x}}$ (of underlying label $\hat{y}$).

Different from all aforementioned works [SMG13, AS17] where no structure is considered within the data (i.e., data indeed come from the *same* distribution class), in this contribution, we instead study the GDD of a linear regression model in separating a two-class GMM (Definition 1) with opposite means $\pm\boldsymbol{\mu}$ and identity covariance $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ that are assigned labels -1 and 1, respectively. We propose a general RMT-based framework to characterize the learning dynamics of this simple model, in the regime where $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and extend the analysis in [AS17] to a classification setting for more natural "structural" data. Similar objectives cannot be achieved within the framework presented in [AS17], which conveys more practical interest to our results and the proposed analysis framework.

As an important special case, by taking the training time $t \to \infty$ and $\lambda = 0$ in (1.24), one obtains the "over-trained" model which is the least-squares solution. In this case, the classification error rate on a new datum $\hat{\mathbf{x}}$ can be consequently shown to be

$$P(\text{sign}(\hat{\mathbf{x}}^\mathsf{T}\mathbf{w}_{LS}) \neq \hat{y}) = Q\left(\frac{\|\boldsymbol{\mu}\|^2\sqrt{1-\min(c,c^{-1})}}{\sqrt{\|\boldsymbol{\mu}\|^2+c}}\right)$$

with $Q(x)$ the Gaussian tail function and admits a singularity at $c = 1$; see Figure 1.5 with $\|\boldsymbol{\mu}\|^2 = 5$. This sharp drop in performance around $c = 1$ can be alleviated, if appropriate regularization techniques such as early stopping are adopted.

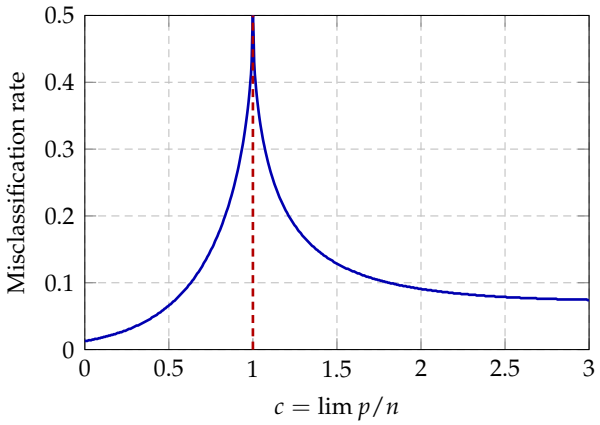


Figure 1.5: Classification error rate as a function of c, $\|\boldsymbol{\mu}\|^2 = 5$.

The fact that the prediction risk blows up around $c \equiv \lim p/n = 1$ in Figure 1.5 indicates the existence of a surprising "double-descent" phenomenon, where one observes, for a given dimension p that the performance of $\mathbf{w}_{LS}$ first gets *worse* as the sample size n grows *large* (when $n < p$), which totally collapses to random guess at the point $n = p$ and starts to increases rapidly when $n > p$. Moreover, in this simple model, the number of model parameters N coincides with the data dimension p, so that for a fixed sample size n, having a more complicated system (with larger N) can lead to a lower prediction risk in the "over-parameterized" regime where $N > n$. This suggests that one must reconsider the golden rule of the *bias-variance tradeoff* in classical statistical learning theory [FHT01].

1.3.2 A geometric approach

From an optimization viewpoint, understanding DNN models is challenging mainly due to the cascading layers with entry-wise nonlinearity between them. This construction

naturally gives rise to non-convex optimization problems that are seemingly intractable. In general, finding the global minimum of a non-convex function is, in the worst case, an NP-complete problem [MK87] and it is unfortunately the case for NN models [BR89]. Yet, many non-convex optimization problems including tensor decomposition, matrix completion, dictionary learning, matrix sensing, and phase retrieval are known to have well-behaved optimization landscapes: under some technical conditions, all local minima are also global [GJZ17, CLC18]. This interesting *global* geometric property, together with some *local* considerations, e.g., any saddle point has a negative directional curvature (i.e., has at least one strictly negative eigenvalue in its Hessian, which makes it possible to continue to decrease the objective function locally), allows for solving efficiently these problems with basic optimization algorithms such as (stochastic) gradient descent. We will discuss this geometric aspect of optimizing NN models in this subsection.

More generally, consider the following minimization problem (which is a more general form of (1.22) with the possible regularization term included in L)

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \frac{1}{n}\sum_{i=1}^{n} L(h(\mathbf{x}_i,\boldsymbol{\theta}),\mathbf{y}_i) \tag{1.25}$$

for a given training set $\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^n$ of size n. The loss function L is often convex (with respect to both $h(\mathbf{x}_i,\boldsymbol{\theta})$ and $\mathbf{y}_i$), for example variants of those in Figure 1.4. Yet, depending on the application of interest, the system model $h(\mathbf{x}_i,\boldsymbol{\theta})$ can be highly nonlinear, non-convex (with respect to the model parameter $\boldsymbol{\theta}$) and even non-smooth, which is more challenging from an optimization standpoint.

Stationary points: a local image. In the case of large-scale problems ($N \gg n \gg 1$), gradient-based methods are often among the most efficient, if not the only, ways to solve (1.25), despite the existence of more advanced optimization methods in the general differentiable programming context. For instance, some second-order (or Hessian-based) optimization algorithms with Hessian information (such as the Newton's method) is capable of finding second-order stationary points (see definition below) [NP06], but they typically require to compute at *every* iteration the inverse or spectral decomposition of the full Hessian matrix of dimension $N \times N$, which is computationally unaffordable for N large.

Solving non-convex problems without higher order information (e.g., Hessian) is challenging, since gradient-based methods stop moving forward as long as the gradient norm $\|\nabla_{\boldsymbol{\theta}} L(\cdot)\|$ is small, which can be far away from the desired *global* minimizer of the non-convex (with respect to the model parameter $\boldsymbol{\theta}$) objective function $L(\cdot)$ in (1.25). To make this more clear, let us introduce the following definitions.

Definition 2 (Stationary points)**.** *Denote $\nabla L(\cdot)$ and $\nabla L^2(\cdot)$ respectively the gradient and the Hessian of the twice differentiable objective function L with respect to the model parameter $\boldsymbol{\theta}$. We say $\boldsymbol{\theta}^*$ is a* first-order stationary point *(or* critical point*) of L if*

$$\nabla L(\boldsymbol{\theta}^*) = \mathbf{0} \Leftrightarrow \|\nabla L(\boldsymbol{\theta}^*)\| = 0$$

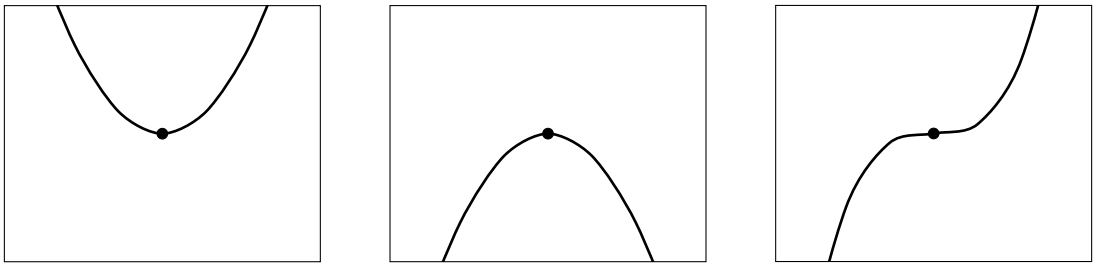
*In practice, the notion of first-order stationarity can be extended to include some numeric tolerance ϵ: we say $\boldsymbol{\theta}^*$ is an ϵ-*first-order stationary point *if $\|\nabla L(\boldsymbol{\theta}^*)\| \leq \epsilon$ for some $\epsilon > 0$.*

First-order stationary points can be divided into the following three categories (see Figure 1.6 for examples of each type), depending on the local *behavior of L in the neighborhood of $\boldsymbol{\theta}^*$:*

1. local minima*: we say $\boldsymbol{\theta}^*$ is a local minimum if there exists $\epsilon > 0$ such that for all $\boldsymbol{\theta}$ in the ϵ-neighborhood of $\boldsymbol{\theta}^*$ (e.g., $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \epsilon$) we have $L(\boldsymbol{\theta}^*) \leq L(\boldsymbol{\theta})$, as in Figure 1.6a;*

2. local maxima*: we say $\boldsymbol{\theta}^*$ is a local maximum if there exists $\epsilon > 0$ such that for all $\boldsymbol{\theta}$ in the ϵ-neighborhood of $\boldsymbol{\theta}^*$ we have $L(\boldsymbol{\theta}^*) \geq L(\boldsymbol{\theta})$, as in Figure 1.6b;*

3. saddle points*: first-order stationary points that are neither minima nor maxima, as in Figure 1.6c.*



(a) Local minimum (b) Local maximum (c) Saddle point

Figure 1.6: Illustration of three types of stationary points in one dimension.

Whether a given first-order stationary point is a local minimum, a local maximum or a saddle point is closely related to the associated Hessian matrix $\nabla L^2(\boldsymbol{\theta}^*) \in \mathbb{R}^{N\times N}$, more precisely

1. *If $\boldsymbol{\theta}^*$ is a local minimum, the Hessian must be* positive semidefinite*, i.e., $\lambda_{\min}(\nabla L^2(\boldsymbol{\theta}^*)) \geq 0$. Conversely, if the Hessian is* positive definite*, then $\boldsymbol{\theta}^*$ must be a local minimum.*

2. *If $\boldsymbol{\theta}^*$ is a local maximum, the Hessian must be* negative semidefinite*, i.e., $\lambda_{\max}(\nabla L^2(\boldsymbol{\theta}^*)) \leq 0$. Conversely, if the Hessian is* negative definite*, then $\boldsymbol{\theta}^*$ must be a local maximum.*

3. *Consequently, if the associated Hessian is neither positive nor negative semidefinite, or equivalently, has both positive and negative eigenvalues, $\boldsymbol{\theta}^*$ must be a saddle point.*

As such, by examining the smallest or largest eigenvalue of the associated Hessian matrix, one can sometimes determine weather a first-order stationary point is a local minimum or maximum: this is referred to as the "second partial derivative test". It is worth pointing out that, if the associated Hessian is degenerate*, i.e., $\det(\nabla L^2(\boldsymbol{\theta}^*)) = 0$ and thus admits (at least one)* zero *eigenvalue, the second partial derivative is inconclusive: in the case of the first and second item listed above, having a zero eigenvalue rules out the possibility of both $\lambda_{\min}(\nabla L^2(\boldsymbol{\theta}^*)) > 0$ and $\lambda_{\max}(\nabla L^2(\boldsymbol{\theta}^*)) < 0$, but $\boldsymbol{\theta}^*$ can still be a local extremum or a saddle point.*

Aiming at minimizing the objective function L, we say a first-order stationary point $\boldsymbol{\theta}^*$ is a second-order stationary point *if*

$$\lambda_{\min}(\nabla L^2(\boldsymbol{\theta}^*)) \geq 0$$

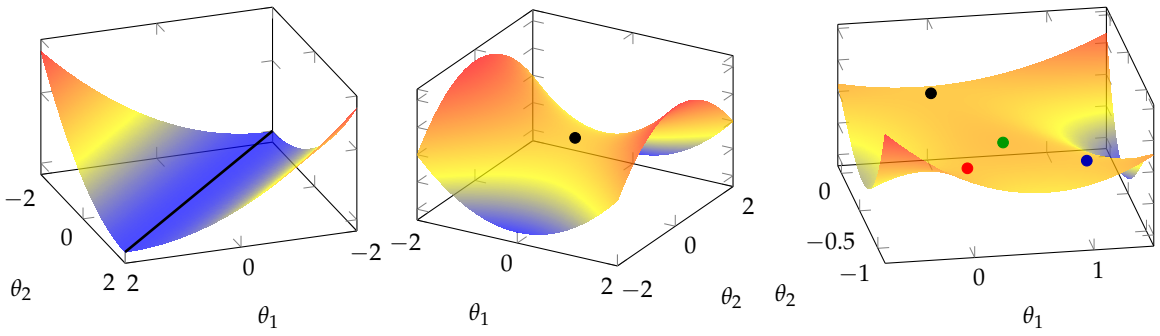
which is indeed a necessary but not sufficient *condition for a local minimum since it does not rule out the possibility of being a saddle point with degenerate Hessian (such that $\lambda_{\min}(\nabla L^2(\boldsymbol{\theta}^*)) = 0$). For all objects defined above, their ϵ-tolerance versions can be similarly defined by replacing zero with some $\epsilon > 0$.*

More generally, a stationary point with degenerate Hessian often demands higher ($\geq$ 3) order information to decide to which type it belongs. If *all* eigenvalues of the Hessian has relatively small amplitude, the local "landscape" of the objective function L is very "flat" and taking a small step in *no matter which* direction results in a very slight change in the value of L. Therefore, it often takes a long time, if still possible, to escape from this kind of regions.

Let us consider some concrete examples in 2D, i.e., $\boldsymbol{\theta} \in \mathbb{R}^2$. In Figure 1.7a we plot the function $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, the set of first-order stationary points of which is the line $\theta_1 = \theta_2$, with the corresponding Hessian given by

$$\nabla^2 L(\theta_1, \theta_2) = \begin{bmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$

that has eigenvalues (constantly) equal to $\lambda_1 = 0$ and $\lambda_2 = 4$. In this example, albeit the degeneration of the Hessian (and thus the invariance of the objective function when moving in the direction of the eigenvector[6] associated to the zero eigenvalue), all stationary points are indeed local minima. To institute a comparison, we plot in Figure 1.7b a more "classical" saddle point (at $(0,0)$) of the objective $L(\theta_1, \theta_2) = \theta_1^2 - \theta_2^2$, the Hessian of which admits two eigenvalues $\lambda_1 = -2$ and $\lambda_2 = 2$, so that the saddle point is "instable" and there is a way "out" locally (to continue to decrease the value of L) due to the existence of the negative eigenvalue $\lambda_1 = -2$.



(a) $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$. (b) $L(\theta_1, \theta_2) = \theta_1^2 - \theta_2^2$. (c) $L = (\theta_1 + \theta_2)(\theta_1\theta_2 + \theta_1\theta_2^2)$.

Figure 1.7: Examples of stationary points in two-dimensional case.

Things become more complicated in the example of Figure 1.7c with $L(\theta_1, \theta_2) = (\theta_1 + \theta_2)(\theta_1\theta_2 + \theta_1\theta_2^2)$ that admits the following four first-order stationary points

$$z_1 : (0,0); \quad z_2 : (0,-1); \quad z_3 : (1,-1); \quad z_4 : (3/8, -3/4)$$

with Hessian eigenvalues respectively given as

$$z_1 : \lambda_1 = \lambda_2 = 0; \quad z_2 : \lambda_1 = -1, \lambda_2 = 1;$$
$$z_3 : \lambda_1 = -1292/305, \lambda_2 = 305/1292; \quad z_4 : \lambda_1 = -1743/1436, \lambda_2 = -450/1351.$$

Therefore, z_2 and z_3 are saddle points with both positive and negative eigenvalues, z_4 is a local maximum; while for z_1 the second derivative test is not enough and one must use higher order test to examine the behavior of L at this point.

To avoid the difficulty in the case of degenerate Hessians, in [GHJY15] the authors introduced the following "strict saddle" property.

Definition 3 (Strict saddle, Definition 4 in [GHJY15])**.** *A twice differentiable function is said to have* strict saddles*, if, apart from the local minima, all its first-order stationary points have at least one negative eigenvalue for the associated Hessian, i.e.,* $\lambda_{\min}(\nabla^2 L(\boldsymbol{\theta}^*)) < 0$*. A more numerically robust version can be similarly defined as* $\lambda_{\min}(\nabla^2 L(\boldsymbol{\theta}^*)) < -\epsilon$ *for some* $\epsilon > 0$*.*

Having "strict saddles" makes it possible to make some progress to (continue to) decrease the objective function, at least locally.

[6] Which happens to be again the line $\theta_1 = \theta_2$ in Figure 1.7a.

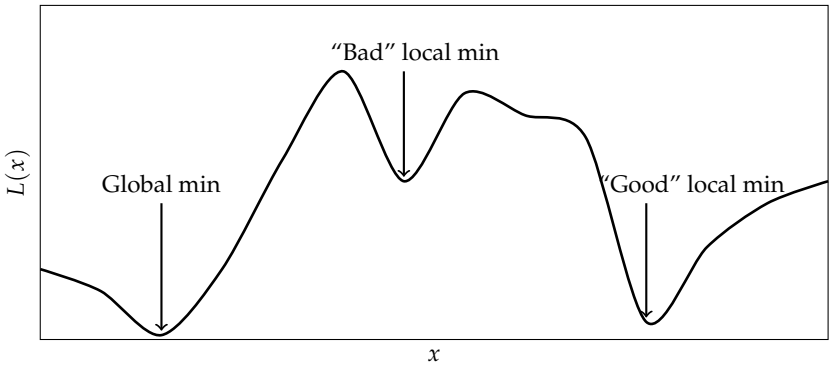


Figure 1.8: Examples of different local minima.

Global guarantee of non-convex optimization. Until now, we have been focusing on the *local* image of the optimization landscape, which is unfortunately far from our initial goal of achieving a *global* minimum. Indeed, if the "strict saddle" property is satisfied, it is possible for gradient-based algorithm to (locally) escape from saddle points and continue to decrease the loss function. Nonetheless, it may still get stuck in a local minimum. Different from the convex case in Figure 1.7a, where the convex (but not strictly convex) objective function $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$ has a *connected* region of local (and global) minima: $\theta_1 = \theta_2$, a non-convex loss function L may have many "isolated" local minimum regions that are separated by higher values of the loss L. Among these regions, some are desired global minima that reach the smallest possible loss $L_{\min}$ and some are "bad" local minima that have much greater loss values than $L_{\min}$. There may also exist "good" local minima with a relatively small loss but still slightly larger than $L_{\min}$ and can be considered satisfying solutions of the minimization, as illustrated in Figure 1.8.

The isolation between local minimum regions creates many "local valleys" of the loss landscape. Very interestingly, there are both empirical evidence [DVSH18, GIP+18] and theoretical arguments [FB16] stating that in modern DNNs, most of these "local valleys" are essentially "connected": i.e., it is possible to construct *continuous* paths that connect *all* these local minimum regions, along which the loss L remains very close to $L_{\min}$, as shown in Figure 1.9 (which could be considered a possible 2D representation of Figure 1.8).

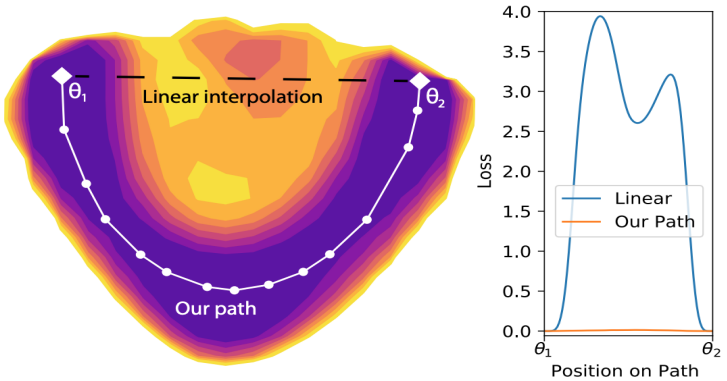


Figure 1.9: Minimum energy path (left) and associated loss value (right) in [DVSH18, Figure 1].

Consider now the (more ideal) case that *all* local minima (connected or isolated) have exactly the *same* loss value. This means that for the optimization problem under study,

all local minima are global. This rules out the presence of "bad" local minima which, together with the strict saddle points property in Definition 3, makes it possible for simple gradient-based algorithms (such as GD or SGD) to achieve a *global* convergence with a guaranteed rate.

To sum up, the key to handle many non-convex optimization problems is to have both

1. a nice local property: for example the "strict saddle" property in Definition 3;
2. some global guarantee: for instance the "all local minima are global" property.

Although seemingly restrictive, both properties are actually satisfied (under some dimensionality condition and at least with high probability) by many non-convex optimization problems of considerable practical interest: many low rank problems including matrix sensing, phase retrieval, matrix completion, robust PCA and many others [GJZ17, CLC18].

In the particular case of neural networks, it was shown in [BH89] that for single-hidden-layer linear NNs with square loss and under some dimensionality condition, all local minima are essentially global and there is no local maximum (the rest are all saddle points). This result was then extended to deep linear NNs in [Kaw16] in which the author showed that the same conclusion holds for DNNs with arbitrary $H \geq 1$ hidden layers. Perhaps more importantly, a necessary condition (on the rank on the weight matrix product) was given in [Kaw16] to ensure the associated Hessian have at least one negative eigenvalue [Kaw16, Theorem 2.3]. This implies the existence of saddle points with positive semidefinite Hessian, for which higher order derivatives will be indispensable for a thorough understanding of these (first-order) stationary points. More precisely, single-hidden-layer ($H = 1$) linear NN models are proven to have strict saddles, which is used for instance in [LSJR16] to show almost sure (with respect to the initialization in a Lebesgue measure sense) convergence to global minima in this case; while for $H \geq 2$ a counterexample was provided in [Kaw16].

We focus in this manuscript on the gradient flow, for linear NN models in the following contribution.

(J4) Yacine Chitour, **Zhenyu Liao**, and Romain Couillet. A geometric approach of gradient descent algorithms in neural networks. (Submitted to) Journal of Differential Equations, 2019.

Based on a cornerstone "invariance" in the parameter/network weights space induced by the network cascading structure, we prove the existence for all time of trajectories associated with gradient descent algorithms in linear networks with an arbitrary number of layers and the convergence of these trajectories to first order stationary points of the loss function. We also prove the exponential/linear convergence of trajectories under an extra condition on their initializations. We propose alternative proof of the "almost sure convergence to global minima" fact in single-hidden-layer linear NN model, which holds true without the technical and unrealistic assumptions demanded in [LSJR16].

1.3.3 Summary of Section 1.3

In this section, we discussed the optimization perspective of large dimensional machine learning problems, and in particular, NN models trained with gradient descent methods. In Section 1.3.1, inspired by recent advances in [SMG13, AS17], we introduced a

RMT-based analysis framework to investigate the GDDs in a linear convex NN model under a GMM (see Definition 1) for the input feature. We found a huge gap between the training and test performances of the least-square solution when the number of parameters gets close to the number of training data, that can be efficiently reduced with the early stopping strategy. This suggests that the use of gradient descent plays an important role in *regularizing* the least-square solution, and sheds new light on the training of large dimensional learning systems.

In Section 1.3.2 we recalled the preliminary local descriptions of (first-order) stationary points in the general context of differentiable programming, with a particular focus on non-convex optimization. We saw that, if both conditions of "strict saddles" and "all local are global" are met, simple gradient-based algorithms such as GD or SGD can benefit from a global convergence to global minima, from almost all initializations. It is moreover possible to speed up the convergence, if the underlying non-convex problem is known to have some nice (statistical) structure, by properly initializing (often with spectral methods) the descent algorithm within some locally convex valley: we will come back to this point in Chapter 5.

1.4 Outline and Contributions

As stated in the previous sections, this manuscript covers both a "structural" viewpoint of large random kernel matrices and neural networks by considering their eigenspectrum behavior, as well as the "dynamical" optimization aspect of learning in neural networks by means of the corresponding gradient descent dynamics. As a first major contribution, we study the eigenspectrum properties of these large nonlinear learning systems, under a multivariate mixture model (see Definition 1) for the input feature, so as to examine the subtle yet crucial interplay between the **nonlinearity** (of the kernel function f, the nonlinear activation σ and the loss function L), the **feature statistics** (limited to first and second here but is envisioned to go well beyond, see our arguments in Remark 1.3 and 2.5 later in Section 2.2.2) as well as the problem **dimensionality**. The second main contribution is the characterization of the gradient descent dynamics (**GDD**s) in learning both zero- and single-hidden-layer linear NN models, which opens the door to a more general **RMT-based** consideration of non-convex **optimization** problems.

The remainder of the manuscript is organized as follows. In Chapter 2 we introduce the mathematical framework of random matrix theory, with helpful preliminaries on linear algebra, probability theory, and complex analysis. In Chapter 3, we start the discussion of large random kernel matrices with the concrete example of kernel ridge regression, or equivalently least squares support vector machines (LS-SVMs), which serves as the cornerstone of subsequent models such as random feature-based methods and random NNs. We will also briefly mention the additional tool needed in handling implicit (convex) optimization problems (e.g., logistic regression) that are technically more involved. Then in Chapter 4, random matrix techniques are combined with ideas from dynamical systems to study the gradient descent dynamics in training simple NNs, first from a statistical standpoint without hidden layer (i.e., linear regression model) and then with a "geometric" approach to deal with the extra hidden layers in linear networks. In Chapter 5 we provide a quick summary of this manuscript, and perhaps more importantly, discuss some possible limitations of the presented results and eventually some future perspectives within the general context of large dimensional machine learning.

We start with the analysis of the classification performance of kernel ridge regression

(or LS-SVM) in Section 3.1.1, which helps understand the counterintuitive behavior of kernel matrices for large dimensional data discussed in Section 1.1.2. Based on a GMM for the input data (see again Definition 1), we show that, for any locally twice differentiable function f, the soft decision function of the kernel ridge regression is asymptotically normally distributed, the mean and variance of which depend on the statistics of mixture model $(\boldsymbol{\mu}_a, \mathbf{C}_a)$, the dimensionality and the function f in a local manner. This theoretical result leads to a novel understanding of the negative impact of imbalanced data in LS-SVM classification and shows an unexpectedly close match when applied on popular image classification datasets. This finding was first reported in

(C1) **Zhenyu Liao** and Romain Couillet. Random matrices meet machine learning: a large dimensional analysis of LS-SVM. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2397–2401. 2017.

and then with more discussions and detailed proofs in

(J1) **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.

The theoretical analysis of LS-SVM presented above are essentially built upon a local expansion of the nonlinearity, which is only possible when the similarity measures under study ($\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$) establish a "concentration" around a *single* point of the nonlinear function f. With a more natural $\sqrt{p}$ scaling, one results in the inner-product kernel that takes the form $\mathbf{K}_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$. Very interestingly, the eigenspectrum of $\mathbf{K}$ is shown to only depend on three scalar parameters of the possibly non-differentiable f. This finding is discussed at length in Section 3.1.2 and reported in

(C2) **Zhenyu Liao** and Romain Couillet. Inner-product kernels are asymptotically equivalent to binary discrete kernels, 2019.

Section 3.2 begins with the discussion on single-hidden-layer random weights NN, or equivalently the random feature-based ridge regression in Figure 1.3. For comparably large n, p, N, we show that for all dimension-free Lipschitz nonlinear activation $\sigma(\cdot)$, the random Gram matrix $\mathbf{G} \equiv \frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ has an asymptotically tractable behavior, that depends solely on σ via the associated underlying kernel $\mathbf{K} \equiv \mathbb{E}_\mathbf{w}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$ for $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. As a consequence, the training MSE can be shown to have an asymptotically (data-dependent) deterministic behavior, with a conjecture provided for the test MSE (and confirmed later in [LC18b]). This work, discussed in more detail in Section 3.2.1, led to the following publication.

(J2) Cosme Louart, **Zhenyu Liao**, Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

In pursuit of a more explicit recognition of the mechanism of different nonlinear activations, we examine the eigenspectrum of the underlying kernel $\mathbf{K}$, by additionally leveraging the stochastic nature of the data (that are assumed to follow a GMM). We show that, depending on whether the first order $(\boldsymbol{\mu}_a)$, or second order $(\mathbf{C}_a)$ information or both, are conserved after the application of both the random weights $\mathbf{W}$ and the entry-wise nonlinearity $\sigma(\cdot)$. As briefly discussed in Section 1.2.2, commonly used activations are divided into three categories of mean-oriented, covariance-oriented and balanced. This result, discussed in Section 3.2.2, led to the following contribution.

(C3) **Zhenyu Liao** and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3063–3071. PMLR, 2018.

In Section 3.3 we consider the more generic (and more involved) convex losses (rather than the squares loss treated in Section 3.2.1 that conducts to explicit solutions) in a mixture classification context and focus mainly on the technical ingredients needed to handle implicit and highly dependent learning systems. This work was reported in ICASSP 2019 with a journal version recently submitted to IEEE Transactions on Signal Processing.

(C4) Xiaoyi Mai, **Zhenyu Liao**, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

(J3) Xiaoyi Mai and **Zhenyu Liao**. High dimensional classification via empirical risk minimization: Improvements and optimality. 2019.

As regards gradient decent dynamics (GDDs) in NN models, we begin in Section 4.1 with the discussion on a statistical approach of evaluating the GDD in simple neural networks. Based on a toy mixture model for the input feature, we retrieve, in the high dimensional regime, the (asymptotic) training and test risks of a linear regression model, as a function of the GD training time. From a methodological standpoint, we introduce a RMT-based framework (Cauchy's integral formula + deterministic equivalent) to examine GDDs of simple NN models, which can be reexpressed as some functional of the feature sample covariance matrix. As a byproduct, we prove for the first time the existence of a "phase-transition" singularity in Figure 1.5, in a high dimensional classification context. This result was first announced in

(C5) **Zhenyu Liao** and Romain Couillet. The dynamics of learning: a random matrix approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3072–3081. PMLR, 2018.

In Section 4.2 we consider a geometric aspect of GDDs in single-hidden-layer linear NN models. By precisely evaluating the (union of) basin of attraction of *all* saddle points, we show that almost all initializations lead to an (equivalent) global minimum for the simplest gradient descent algorithm. Most importantly, this is true without additional assumptions on these stationary points (e.g., on the gradient norms during descent or their Hessians). Indeed, this "global convergence to global minima" conclusion essentially holds due to a key invariant structure (in the network weight space) during the gradient descent procedure imposed by the network cascading structure. As a practical outcome, we also derive critical initialization schemes with exponential convergence rate. These results are summarized in

(J4) Yacine Chitour, **Zhenyu Liao**, and Romain Couillet. A geometric approach of gradient descent algorithms in neural networks. 2019.

now under review for Journal of Differentiable Equation, which concludes the Chapter 4.

In Chapter 5 we draw a speedy conclusion of the manuscript and mention some possible limitations of the presented theoretical results, which are admittedly below the performance of state-of-the-art models in modern machine learning problems today. We will also discuss some possible future research directions, in the continuation of our preliminary findings presented here.

Chapter 2

Mathematical Background: Random Matrix Theory

Random matrix theory, at its inception, primarily dealt with the eigenvalue distribution (also referred to as spectral measure) of large random matrices. One of the key technical tools to study these measures is the Stieltjes transform, often presented as the central object of the theory [BS10].

But signal processing and machine learning alike are more fundamentally interested in subspaces and eigenvectors (which often carry the structural data information) than in eigenvalues of random matrices. Subspace or spectral methods, such as principal component analysis (PCA) [WEG87], spectral clustering [NJW02] and some semi-supervised learning techniques [Zhu05] are built directly upon the eigenspace spanned by the several top eigenvectors.

Consequently, beyond the Stieltjes transform, a more general object, the *resolvent* of large random matrices will constitute the cornerstone of this manuscript. The resolvent of a matrix gives access to its spectral measure, to the location of its isolated eigenvalues, to the statistical behavior of their associated eigenvectors when random, and consequently provides an entry-door to the performance analysis of numerous learning methods.

2.1 Fundamental Objects

2.1.1 The resolvent

We first introduce the resolvent of a matrix.

Definition 4 (Resolvent)**.** *For a symmetric matrix* $\mathbf{M} \in \mathbb{R}^{n \times n}$*, the resolvent* $\mathbf{Q}_{\mathbf{M}}(z)$ *of* $\mathbf{M}$ *is defined, for* $z \in \mathbb{C}$ *not an eigenvalue of* $\mathbf{M}$*, as*

$$\mathbf{Q}_{\mathbf{M}}(z) \equiv (\mathbf{M} - z\mathbf{I}_n)^{-1} \tag{2.1}$$

which is also denoted $\mathbf{Q}$ *when there is not ambiguity.*

Although our focus here will exclusively be on resolvents of (random) matrices, it must be noted that the resolvent operator is in fact a very classical tool in the analysis of linear operators in general Hilbert spaces [AG13] as well as in monotone operator theory of importance to modern convex optimization theory [BC11].

2.1.2 Spectral measure and the Stieltjes transform

The first use of the resolvent $\mathbf{Q}_{\mathbf{M}}$ is in its relation to the *empirical spectral measure* $\mu_{\mathbf{M}}$ of a square matrix $\mathbf{M}$, through the *Stieltjes transform* $m_{\mu_{\mathbf{M}}}$, which we all define next.

Definition 5 (Empirical spectral measure)**.** *For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, the* spectral measure *or* empirical spectral measure *or* empirical spectral distribution $\mu_{\mathbf{M}}$ *of* $\mathbf{M}$ *is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M})$ of $\mathbf{M}$,*

$$\mu_{\mathbf{M}}(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})}(x). \tag{2.2}$$

Since $\mu_{\mathbf{M}}(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{\mathbb{R}} \mu_{\mathbf{M}}(x)dx = 1$, the spectral measure $\mu_{\mathbf{M}}$ of a matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ (random or not) is a probability measure. For (probability) measures, we can define their associated Stieltjes transforms as follows.

Definition 6 (Stieltjes transform)**.** *For a real probability measure μ with support* $\mathrm{supp}(\mu)$*, the* Stieltjes transform *$m_\mu(z)$ is defined, for all $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$, as*

$$m_\mu(z) \equiv \int_{\mathbb{R}} \frac{1}{t-z} d\mu(t). \tag{2.3}$$

The Stieltjes transform m_μ has numerous interesting properties: it is complex analytic on its domain of definition $\mathbb{C} \setminus \mathrm{supp}(\mu)$, it is bounded by $|m_\mu(z)|\,\mathrm{dist}(z, \mathrm{supp}(\mu)) \leq 1$, it satisfies $\Im[z] > 0 \Rightarrow \Im[m(z)] > 0$, and it is an increasing function on all connected components of its restriction to $\mathbb{R} \setminus \mathrm{supp}(\mu)$ (since $m'_\mu(x) = \int_{\mathbb{R}} (t-x)^{-2}dt > 0$) with $\lim_{x\to\pm\infty} m_\mu(x) = 0$ if $\mathrm{supp}(\mu)$ is bounded.

As a transform, m_μ admits an inverse formula to recover μ, as per the following result.

Theorem 2.1 (Inverse Stieltjes transform)**.** *For a, b continuity points of the probability measure μ, we have*

$$\mu([a,b]) = \frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im\left[m_\mu(x+\imath y)\right] dx. \tag{2.4}$$

Besides, if μ admits a density f at x, i.e., $\mu(x)$ is differentiable in a neighborhood of x and

$$\lim_{\epsilon\to 0}(2\epsilon)^{-1}\mu([x-\epsilon, x+\epsilon]) = f(x)$$

then

$$f(x) = \frac{1}{\pi}\lim_{y\downarrow 0}\Im\left[m_\mu(x+\imath y)\right]. \tag{2.5}$$

Finally, if μ has an isolated mass at x, then

$$\mu(\{x\}) = -\frac{1}{\pi}\lim_{y\downarrow 0}\imath y m_\mu(x+\imath y). \tag{2.6}$$

Proof. Since $|\frac{y}{(t-x)^2+y^2}| \leq \frac{1}{y}$ for $y > 0$, by Fubini's theorem,

$$\begin{aligned}
&\frac{1}{\pi}\int_a^b \Im\left[m_\mu(x+\imath y)\right] dx = \frac{1}{\pi}\int_a^b \left[\int_{\mathbb{R}} \frac{y}{(t-x)^2+y^2} d\mu(t)\right] dx \\
=&\frac{1}{\pi}\int_{\mathbb{R}} \left[\int_a^b \frac{y}{(t-x)^2+y^2} dx\right] d\mu(t) \\
=&\frac{1}{\pi}\int_{\mathbb{R}} \left[\arctan\left(\frac{b-t}{y}\right) - \arctan\left(\frac{a-t}{y}\right)\right] d\mu(t).
\end{aligned}$$

As $y \downarrow 0$, the difference in brackets converges either to $\pm\pi$ or 0 depending on the relative position of a, b, t. By the dominated convergence theorem, limits and integrals can be exchanged, and the limit, as $y \downarrow 0$, is $\int_{\mathbb{R}} 1_{[a,b]} d\mu(t) = \mu([a,b])$. □

The important relation between the spectral measure of $\mathbf{M} \in \mathbb{R}^{n\times n}$, the Stieltjes transform $m_{\mu_\mathbf{M}}(z)$ and the resolvent $\mathbf{Q}_\mathbf{M}$ lies in the fact that

$$m_{\mu_\mathbf{M}}(z) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t-z} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i(\mathbf{M})-z} = \frac{1}{n}\operatorname{tr}(\mathbf{Q}_\mathbf{M}). \tag{2.7}$$

Combining inverse Stieltjes transform and the relation above thus provides a link between $\mathbf{Q}_\mathbf{M}$ and the eigenvalue distribution of $\mathbf{M}$. While seemingly contorted, this link is in general the only efficient way to study the spectral measure of *large dimensional random matrices* $\mathbf{M}$.

Remark 2.1 (Resolvent as a matrix-valued Stieltjes transform)**.** *As proposed in [HLN07], it can be convenient to extrapolate Definition 6 of Stieltjes transforms to $n \times n$ matrix-valued positive measures $\mathbf{M}(dt)$,*[1] *in which case Equation* (2.7) *can be generalized as*

$$\mathbf{Q}_\mathbf{M}(z) = \int_{\mathbb{R}} \frac{\mathbf{M}(dt)}{t-z} = \mathbf{U}\operatorname{diag}\left\{\frac{1}{\lambda_i(\mathbf{M})-z}\right\}_{i=1}^{n}\mathbf{U}^\mathsf{T}$$

where we used the spectral decomposition $\mathbf{M} = \mathbf{U}\operatorname{diag}\{\lambda_i(\mathbf{M})\}_{i=1}^n\mathbf{U}^\mathsf{T}$. In particular, $\mathbf{Q}_\mathbf{M}(z)$ enjoys similar properties as Stieltjes transforms of real-valued measures: $\|\mathbf{Q}_\mathbf{M}(z)\| \leq \operatorname{dist}(z, \operatorname{supp}(\mu_\mathbf{M}))^{-1}$, and $x \mapsto \mathbf{Q}_\mathbf{M}(x)$ for $x \in \mathbb{R} \setminus \operatorname{supp}(\mu)$ is an increasing matrix-valued function with respect to symmetric matrix partial ordering (i.e., $\mathbf{A} \succeq \mathbf{B}$ whenever $\mathbf{z}^\mathsf{T}(\mathbf{A}-\mathbf{B})\mathbf{z} \geq 0$ for all $\mathbf{z}$).

2.1.3 Cauchy's integral, linear eigenvalue functionals, and eigenspaces

Being complex analytic, the resolvent $\mathbf{Q}_\mathbf{M}$ can be manipulated using advanced tools from complex analysis. Of particular interest to spectrum-based machine learning methods is the relation between the resolvent and Cauchy's integral theorem.

Theorem 2.2 (Cauchy's integral formula)**.** *For $\Gamma \in \mathbb{C}$ a positively (i.e., counterclockwise) oriented simple closed curve and a complex function $f(z)$ analytic in a region containing Γ and its interior, then*

1) *if $z_0 \in \mathbb{C}$ is enclosed by Γ, $f(z_0) = -\frac{1}{2\pi\imath}\oint_\Gamma \frac{f(z)}{z_0-z}dz$;*

2) *if not, $\frac{1}{2\pi\imath}\oint_\Gamma \frac{f(z)}{z-z_0}dz = 0$.*

This result provides an immediate link between the *linear functionals of the eigenvalues of* $\mathbf{M}$ and the Stieltjes transform through

$$\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i(\mathbf{M})) = -\frac{1}{2\pi\imath n}\oint_\Gamma f(z)\operatorname{tr}(\mathbf{Q}_\mathbf{M}(z))dz = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)m_{\mu_\mathbf{M}}(z)dz$$

for all f complex analytic in a compact neighborhood of $\operatorname{supp}(\mu_\mathbf{M})$, by choosing the contour Γ to enclose $\operatorname{supp}(\mu_\mathbf{M})$ (i.e., all the $\lambda_i(\mathbf{M})$'s). More generally,

$$\frac{1}{n}\sum_{\lambda_i(\mathbf{M})\in\Gamma^\circ} f(\lambda_i(\mathbf{M})) = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)m_{\mu_\mathbf{M}}(z)dz$$

[1] Defined by the fact that $\mu(dt;\mathbf{z}) = \mathbf{z}^\mathsf{T}\mathbf{M}(dt)\mathbf{z} = \sum_{ij} z_i z_j \mathbf{M}_{ij}(dt)$ is a positive real-valued measure for all $\mathbf{z}$. See [RRR67] for an introduction.

for Γ° the interior of contour Γ.

Another quantity of interest relates to eigenvectors and eigenspaces. Decomposing the symmetric matrix $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ in its spectral decomposition with $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M}))$, we have

$$\mathbf{Q}_\mathbf{M}(z) = \sum_{i=1}^{n} \frac{\mathbf{u}_i \mathbf{u}_i^\mathsf{T}}{\lambda_i(\mathbf{M}) - z}$$

and thus the direct access to the i-th eigenvector of $\mathbf{M}$ through

$$\mathbf{u}_i \mathbf{u}_i^\mathsf{T} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{M})}} \mathbf{Q}_\mathbf{M}(z) dz$$

for $\Gamma_{\lambda_i(\mathbf{M})}$ a contour circling around $\lambda_i(\mathbf{M})$ only (if $\lambda_i(\mathbf{M})$ is of unit multiplicity). More generally,

$$\mathbf{U} f(\mathbf{\Lambda}; \Gamma) \mathbf{U}^\mathsf{T} = -\frac{1}{2\pi\imath} \oint_{\Gamma} f(z) \mathbf{Q}_\mathbf{M}(z) dz$$

for f analytic in a neighborhood of Γ and its interior and $f(\mathbf{\Lambda}; \Gamma) = \operatorname{diag}(\{f(\lambda_i(\mathbf{M})) 1_{\lambda_i(\mathbf{M}) \in \Gamma^\circ}\}_{i=1}^n)$.

Of interest in this manuscript will be the projection of the individual eigenvectors $\mathbf{u}_i$ of $\mathbf{M}$ onto a deterministic eigenvector $\mathbf{v}$. In particular, from the above,

$$|\mathbf{v}^\mathsf{T} \mathbf{u}_i|^2 = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{M})}} \mathbf{v}^\mathsf{T} \mathbf{Q}_\mathbf{M}(z) \mathbf{v} dz.$$

It is important to note that the resolvent provides access to *scalar observations* of the eigenstructure of $\mathbf{M}$ through *linear functionals of the resolvent* $\mathbf{M}$, i.e., the scalar observations $\frac{1}{n}\sum_i f(\lambda_i(\mathbf{M}))$ and $|\mathbf{v}^\mathsf{T}\mathbf{u}_i|$ accessible from $\operatorname{tr} \mathbf{Q}_\mathbf{M}$ and $\mathbf{v}^\mathsf{T}\mathbf{Q}_\mathbf{M}\mathbf{v}$, respectively.

2.1.4 Deterministic and random equivalents

This manuscript is concerned with the situation where $\mathbf{M}$ is a *large dimensional random matrix*, the eigenvalues and eigenvectors of which need be related to the statistical nature of the model design of $\mathbf{M}$.

In the early days of random matrix theory, the main focus was on the *limiting spectral measure* of $\mathbf{M}$, that is the characterization of a certain "limit" to the spectral measure $\mu_\mathbf{M}$ of $\mathbf{M}$ as the size of $\mathbf{M}$ increases. To this purpose, the natural approach is to study the *random* Stieltjes transform $m_{\mu_\mathbf{M}}(z)$ and to show that is admits a deterministic limit (in probability or almost surely) $m(z)$. However, this method shows strong limitations today: i) it supposes that such a limit does exist, therefore restricting the study to very isotropic models for $\mathbf{M}$ and ii) it only quantifies $\operatorname{tr} \mathbf{Q}_\mathbf{M}$ (through the Stieltjes transform), thereby discarding all subspace information about $\mathbf{M}$ carried in $\mathbf{Q}_\mathbf{M}$ (as a consequence, a further study of the eigenvectors of $\mathbf{M}$ requires a complete rework).

To avoid these limitations, modern random matrix theory uses the notion of *deterministic equivalents* which are *non-asymptotic* deterministic matrices having (in probability or almost surely) asymptotically the same *scalar observations* as the random ones.

Definition 7 (Deterministic Equivalent)**.** *We say that $\bar{\mathbf{Q}} \in \mathbb{R}^{n\times n}$ is a deterministic equivalent for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ if, for (sequences of) deterministic matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (operator and Euclidean, respectively), we have, as $n \to \infty$,*

$$\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \bar{\mathbf{Q}}) \to 0, \quad \mathbf{a}^\mathsf{T}(\mathbf{Q} - \bar{\mathbf{Q}})\mathbf{b} \to 0$$

where the convergence is either in probability or almost surely.

This definition has the advantage to bring forth the two key elements giving access to spectral information about a random matrix $\mathbf{M}$: traces and bilinear forms (of its resolvent $\mathbf{Q}_{\mathbf{M}}(z)$ for some z). Deterministic equivalents of resolvents $\mathbf{Q}_{\mathbf{M}}$ then encode most of the information necessary to statistically quantify a random matrix $\mathbf{M}$.

A practical use of deterministic equivalents is to establish that, for a random matrix $\mathbf{M}$, $\frac{1}{n}\operatorname{tr}(\mathbf{Q}_{\mathbf{M}}(z) - \bar{\mathbf{Q}}(z)) \to 0$, say almost surely, for all $z \in \mathcal{C}$ with $\mathcal{C} \subset \mathbb{C}$ some region of $\mathbb{C}$. Denoting $\bar{m}_n(z) = \frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}(z)$, this convergence implies that the Stieltjes transform of $\mu_{\mathbf{M}}$ "converges" in the sense that $m_{\mu_{\mathbf{M}}}(z) - \bar{m}_n(z) \to 0$. As we will see, this will indicate that $\mu_{\mathbf{M}}$ gets increasingly well approximated by a probability measure $\bar{\mu}_n$ having Stieltjes transform $\bar{m}_n(z)$. Identifying $\bar{m}_n(z)$, which uniquely defines $\bar{\mu}_n$, will often be as far as *the Stieltjes transform method* will lead us. But in some rare cases (such as with the Marčenko-Pastur and the semi-circle laws), $\bar{\mu}_n$ will be explicitly identifiable.

In the remainder of the manuscript, we will often characterize the large dimensional behavior of random matrix models $\mathbf{M}$ through an approximation by deterministic equivalents $\bar{\mathbf{Q}}(z)$ of their associated resolvents $\mathbf{Q}_{\mathbf{M}}(z)$, as this offers access not only to their asymptotic spectral measure but also to their eigenspaces. We shall therefore often extrapolate some of the traditional results, such as the Marčenko–Pastur law [MP67], the sample covariance matrix model [SB95], etc., under this more general form.

Remark 2.2 ($\bar{\mathbf{Q}}$ versus $\mathbb{E}\mathbf{Q}$)**.** *For $\bar{\mathbf{Q}}$ a deterministic equivalent for $\mathbf{Q}$, the probabilistic convergences $\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q} - \bar{\mathbf{Q}}) \to 0$ and $\mathbf{a}^{\mathsf{T}}(\mathbf{Q} - \bar{\mathbf{Q}})\mathbf{b} \to 0$ will in general unfold from the fact that*

$$\|\mathbb{E}\mathbf{Q} - \bar{\mathbf{Q}}\| \to 0$$

and from a control of the variance of $\frac{1}{n}\operatorname{tr}(\mathbf{AQ})$ and $\mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{b}$; this will often be the strategy followed in our proofs. But note importantly that, if the above relation is met, then $\mathbb{E}\mathbf{Q}$ itself is a deterministic equivalent for $\mathbf{Q}$ by Definition 7. However, $\mathbb{E}\mathbf{Q}$ is often not convenient to work with and a "truly" deterministic matrix $\bar{\mathbf{Q}}$ involving no integration over probability spaces will be systematically preferred.

In some situations, deterministic equivalents may either not exist or, as will often be the case, will only be reachable through the access to an intermediary random matrix. To simplify the readability of the main results and proofs in the remainder of the manuscript, which involve both deterministic and random equivalents, we introduce the following shortcut notation.

Notation 1 (Deterministic and Random Equivalents)**.** *For $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ two random or deterministic matrices, we write*

$$\mathbf{X} \leftrightarrow \mathbf{Y}$$

if, for all $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (respectively, operator and Euclidean), we have the simultaneous results

$$\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{X} - \mathbf{Y}) \xrightarrow{a.s.} 0, \quad \mathbf{a}^{\mathsf{T}}(\mathbf{X} - \mathbf{Y})\mathbf{b} \xrightarrow{a.s.} 0, \quad \|\mathbb{E}[\mathbf{X} - \mathbf{Y}]\| \to 0.$$

2.2 Foundational Random Matrix Results

In this section we introduce the main historical results of random matrix theory (appropriately updated under a deterministic equivalent form), which will serve as supporting

models to most applications to machine learning. For readability and accessibility to the readers new to random matrix theory, we mostly stick to intuitive and short sketches of proofs. Yet, for the readers to have a glimpse on the technical details and modern tools of the field, some of the proof sketches will be appended by a complete exhaustive proof.

Both sketches and detailed proofs rely on a set of elementary lemmas and identities will be introduced below in Section 2.2.1. The main difference between sketches and detailed proofs then relies on additional technical *probability* theory arguments to prove various convergence results. These arguments strongly depend on the underlying random matrix model hypotheses (Gaussian independent, i.i.d., concentrated random vectors, etc.); for readability, we will focus in our proofs on one specific line of proof (that we claim to be the "historical" one) and will introduce some side remarks concerning alternative approaches.

2.2.1 Key lemmas and identities

Resolvent identities

Most results discussed in this section consist in approximating random resolvents $\mathbf{Q}(z)$ via deterministic resolvents $\bar{\mathbf{Q}}(z)$, which are both inverse of matrices. The following first identity provides a comparison of inverse matrices.

Lemma 2.1 (Resolvent identity)**.** *For invertible matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, we have*

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}.$$

Proof. This can be easily checked by multiplying both sides on the left by $\mathbf{A}$ and on the right by $\mathbf{B}$. □

Another useful lemma that helps directly connect the resolvent of $\mathbf{BA}$ to that of $\mathbf{AB}$, is given as follows.

Lemma 2.2. *For* $\mathbf{A} \in \mathbb{R}^{p\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times p}$ *, we have*

$$\mathbf{A}(\mathbf{BA} - z\mathbf{I}_n)^{-1} = (\mathbf{AB} - z\mathbf{I}_p)^{-1}\mathbf{A}$$

for $z \in \mathbb{C}$ *distinct from* 0 *and from the eigenvalues of* $\mathbf{AB}$*.*

For $\mathbf{AB}$ and $\mathbf{BA}$ symmetric, Lemma 2.2 is a special case of the more general relation $\mathbf{A}f(\mathbf{BA}) = f(\mathbf{AB})\mathbf{A}$, with $f(\mathbf{M}) \equiv \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^\mathsf{T}$ under the spectral decomposition $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ and f complex analytic. Since f is analytic, $f(\mathbf{BA}) = \sum_{i=0}^{\infty} c_i(\mathbf{BA})^i$ for some sequence $\{c_i\}_{i=0}^{\infty}$ and thus $\mathbf{A}f(\mathbf{BA}) = \sum_{i=0}^{\infty} c_i(\mathbf{AB})^i\mathbf{A} = f(\mathbf{AB})\mathbf{A}$.

The next lemma, known as *Sylvester's identity*, similarly relates the resolvents of $\mathbf{AB}$ and $\mathbf{BA}$ through their determinant.

Lemma 2.3 (Sylvester's identity)**.** *For* $\mathbf{A} \in \mathbb{R}^{p\times n}$*,* $\mathbf{B} \in \mathbb{R}^{n\times p}$ *and* $z \in \mathbb{C}$*,*

$$\det\left(\mathbf{AB} - z\mathbf{I}_p\right) = \det\left(\mathbf{BA} - z\mathbf{I}_n\right)(-z)^{p-n}.$$

An immediate consequence of Sylvester's identity is that $\mathbf{AB}$ and $\mathbf{BA}$ have the same *non-zero* eigenvalues (those non-zero z's for which both left- and right-hand sides vanish). Thus, say $n \geq p$, $\mathbf{AB} \in \mathbb{R}^{p\times p}$ and $\mathbf{BA} \in \mathbb{R}^{n\times n}$ have the same spectrum, except for the additional $n - p$ zero eigenvalues of $\mathbf{AB}$. This remark implies the next identity.

Lemma 2.4 (Trace of resolvent and co-resolvent). *Let* $\mathbf{A} \in \mathbb{R}^{p\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, *and* $z \in \mathbb{C}$ *not an eigenvalue of* $\mathbf{AB}$ *nor zero. Then*

$$\operatorname{tr} \mathbf{Q}_{\mathbf{AB}}(z) = \operatorname{tr} \mathbf{Q}_{\mathbf{BA}}(z) + \frac{n-p}{z}.$$

In particular, if $\mathbf{AB}$ *and* $\mathbf{BA}$ *are symmetric,*

$$m_{\mu_{\mathbf{AB}}}(z) = \frac{n}{p} m_{\mu_{\mathbf{BA}}}(z) + \frac{n-p}{pz}.$$

Perturbation identities

Quantifying the asymptotic global (e.g., spectral distribution) or local (e.g., isolated eigenvalues or projection on eigenvector) behavior of random matrices $\mathbf{M}$ will systematically involve a *perturbation approach*. The idea often lies in comparing the behavior of the resolvent $\mathbf{Q} = \mathbf{Q}_{\mathbf{M}}$ to the resolvent $\mathbf{Q}_{-i}$ of $\mathbf{M}_{-i}$, with $\mathbf{M}_{-i}$ defined as $\mathbf{M}$ with either row and column i, or some i-th contribution (e.g., $\mathbf{M}_{-i} = \sum_{j\neq i} \mathbf{x}_j \mathbf{x}_j^\mathsf{T}$ if $\mathbf{M} = \sum_j \mathbf{x}_j \mathbf{x}_j^\mathsf{T}$), discarded. A number of so-called *perturbation* identities are then needed.

The first one involves the segmentation of $\mathbf{M}$ under the form of sub-blocks, in general consisting of one large block and three small sub-matrices. The resolvent $\mathbf{Q}_{\mathbf{M}}$ can correspondingly be segmented in sub-blocks according to the following block inversion lemma.

Lemma 2.5 (Block matrix inversion). *For* $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times n}$, $\mathbf{C} \in \mathbb{R}^{n\times p}$ *and* $\mathbf{D} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{D}$ *invertible, we have*

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{S}^{-1} & -\mathbf{S}_{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{S}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

where $\mathbf{S} \equiv \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ *is the* Schur complement *(for the block* $\mathbf{D}$*) of* $\left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right)$.[2]

As a consequence of Lemma 2.5, we have the following explicit form for all diagonal entries of an invertible matrix $\mathbf{A}$.

Lemma 2.6 (Diagonal entries of matrix inverse). *For invertible* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *and* $\mathbf{A}_{-i} \in \mathbb{R}^{(p-1)\times(p-1)}$, *the matrix obtained by removing the i-th row and column from* $\mathbf{A}$, $i = 1, \ldots, p$, *we have*

$$(\mathbf{A}^{-1})_{ii} = \frac{1}{\mathbf{A}_{ii} - \boldsymbol{\alpha}_i^\mathsf{T} (\mathbf{A}_{-i})^{-1} \boldsymbol{\beta}_i}$$

for $\boldsymbol{\alpha}_i^\mathsf{T}, \boldsymbol{\beta}_i \in \mathbb{R}^{p-1}$ *the i-th row and column of* $\mathbf{A}$ *with i-th entries removed, respectively.*

The result is a direct consequence of the fact that $\mathbf{A}^{-1} = \frac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})}$, with $\operatorname{adj}(\mathbf{A})$ the adjugate matrix of $\mathbf{A}$, together with the block determinant formula in Lemma 2.5.

Perturbations by addition or subtraction of low-rank matrices to $\mathbf{M}$ induce modifications in the resolvent $\mathbf{Q}_{\mathbf{M}}$ that involve Woodbury's identity as follows.

Lemma 2.7 (Woodbury). *For* $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p\times n}$, *such that both* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T}$ *are invertible, we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{V}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_n + \mathbf{V}^\mathsf{T}\mathbf{A}\mathbf{U})^{-1}\mathbf{V}^\mathsf{T}\mathbf{A}^{-1}.$$

[2]The Schur complement $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ is particularly known for its providing the block determinant formula $\det\left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right) = \det(\mathbf{D}) \det(\mathbf{S})$.

Note importantly that, while $(\mathbf{A}+\mathbf{U}\mathbf{V}^\mathsf{T})^{-1}$ is of size $p \times p$, $(\mathbf{I}_n + \mathbf{V}\mathbf{A}\mathbf{U})^{-1}$ is of size $n \times n$. This will turn out useful to relate resolvents of large dimensional matrices to resolvents of more elementary fixed small size matrices. In particular, for $n = 1$, i.e., $\mathbf{U}\mathbf{V}^\mathsf{T} = \mathbf{u}\mathbf{v}^\mathsf{T}$ for $\mathbf{U} = \mathbf{u} \in \mathbb{R}^p$ and $\mathbf{V} = \mathbf{v} \in \mathbb{R}^p$, Woodbury's identity specializes to the Sherman–Morrison formula.

Lemma 2.8 (Sherman–Morrison). *For $\mathbf{A} \in \mathbb{R}^{p\times p}$ invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, $\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T}$ is invertible if and only if $1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u} \neq 0$ and*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\mathsf{T}\mathbf{A}^{-1}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}.$$

Besides,

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\mathsf{T})^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}.$$

Letting $\mathbf{A} = \mathbf{M} - z\mathbf{I}_p$, $z \in \mathbb{C}$, and $\mathbf{v} = \tau\mathbf{u}$ for $\tau \in \mathbb{R}$ in the previous lemma leads to the following rank-1 perturbation lemma for the resolvent of $\mathbf{M}$.

Lemma 2.9 (From Lemma 2.6 in [SB95]). *For $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p\times p}$ symmetric, $\mathbf{u} \in \mathbb{R}^p$, $\tau \in \mathbb{R}$ and $z \in \mathbb{C} \setminus \mathbb{R}$,*

$$\left|\operatorname{tr} \mathbf{A}(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T} - z\mathbf{I}_p)^{-1} - \operatorname{tr} \mathbf{A}(\mathbf{M} - z\mathbf{I}_p)^{-1}\right| \leq \frac{\|\mathbf{A}\|}{|\Im(z)|}.$$

Also, for $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p\times p}$ symmetric and nonnegative definite, $\mathbf{u} \in \mathbb{R}^p$, $\tau > 0$ and $z < 0$,

$$\left|\operatorname{tr} \mathbf{A}(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T} - z\mathbf{I}_p)^{-1} - \operatorname{tr} \mathbf{A}(\mathbf{M} - z\mathbf{I}_p)^{-1}\right| \leq \frac{\|\mathbf{A}\|}{|z|}.$$

It is interesting (and possibly counterintuitive at first) to note that $\|\mathbf{u}\|$ does not intervene in this inequality. In particular, irrespective of the amplitude of the rank-1 perturbation, under the conditions of the lemma

$$m_{\mu_{\mathbf{M}+\tau\mathbf{u}\mathbf{u}^\mathsf{T}}}(z) = m_{\mu_\mathbf{M}}(z) + O(p^{-1})$$

and thus, by the link between spectrum and Stieltjes transform, the spectral measure of $\mathbf{M}$ is asymptotically close to that of $\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T}$ for any $\mathbf{u}$, in the large p limit. This result can be understood through the following two arguments: i) for large p, the spectrum of $\mathbf{M}$ (say $\|\mathbf{M}\| = O(1)$ without generality restriction) is only non-trivial if the vast majority of the p eigenvalues of $\mathbf{M}$ are of order $O(1)$: thus, as p eigenvalues use a space of size $O(1)$, they tend to aggregate; ii) by Weyl's interlacing lemma presented next (Lemma 2.10) for symmetric matrices, the eigenvalues of $\mathbf{M}$ and of $\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T}$ are interlaced. Both arguments thus indicate that, in the large p limit, the spectral measures are indeed asymptotically the same.

Unlike non-symmetric matrices, symmetric matrices indeed enjoy the nice property of having stable spectra with respect to rank-1 perturbations. For $\lambda \in \mathbb{R}$ an eigenvalue of $\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T}$ but not of $\mathbf{M}$ with, say $\tau > 0$, we indeed have

$$\begin{aligned} 0 &= \det(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\mathsf{T} - \lambda\mathbf{I}_p) = \det(\mathbf{Q}_\mathbf{M}(\lambda)) \det(\mathbf{I}_p + \tau\mathbf{Q}_\mathbf{M}(\lambda)\mathbf{u}\mathbf{u}^\mathsf{T}) \\ &= \det(\mathbf{Q}_\mathbf{M}(\lambda)) \left(1 + \tau\mathbf{u}^\mathsf{T}\mathbf{Q}_\mathbf{M}(\lambda)\mathbf{u}\right) \end{aligned}$$

where the second equality unfolds from factoring out $\mathbf{M} - \lambda \mathbf{I}_p$ (which is not singular as λ is not an eigenvalue of $\mathbf{M}$) and the third from Sylverster's identity (Lemma 2.3). As a consequence, λ is one of the solutions to

$$-1 = \tau \mathbf{u}^\mathsf{T} \mathbf{Q}_\mathbf{M}(\lambda) \mathbf{u} = \tau \sum_{i=1}^p \frac{|\mathbf{v}_i^\mathsf{T} \mathbf{u}|^2}{\lambda_i(\mathbf{M}) - \lambda}, \quad \left(\mathbf{M} = \sum_{i=1}^p \lambda_i(\mathbf{M}) \mathbf{v}_i \mathbf{v}_i^\mathsf{T} \right)$$

which, seen as a function of λ, has asymptotes at each $\lambda_i(\mathbf{M})$ and is increasing (from $-\infty$ to ∞) on the segments $(\lambda_i(\mathbf{M}), \lambda_{i+1}(\mathbf{M}))$ (eigenvalues being sorted in increasing order). The eigenvalues of $\mathbf{M} + \tau \mathbf{u}\mathbf{u}^\mathsf{T}$ are therefore *interlaced* with those of $\mathbf{M}$. This idea generalizes to finite rank perturbation as follows.

Lemma 2.10 (Weyl, Theorem 4.3.1 in [HJ12])**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$ be symmetric matrices and let the respective eigenvalues of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A} + \mathbf{B}$ arranged in nondecreasing order, i.e., $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_{p-1} \leq \lambda_p$. Then, for all $i \in \{1, \ldots, p\}$,*

$$\begin{aligned} \lambda_i(\mathbf{A} + \mathbf{B}) &\leq \lambda_{i+j}(\mathbf{A}) + \lambda_{p-j}(\mathbf{B}), \quad j = 0, 1, \ldots, p-i, \\ \lambda_{i-j+1}(\mathbf{A}) + \lambda_j(\mathbf{B}) &\leq \lambda_i(\mathbf{A} + \mathbf{B}), \quad j = 1, \ldots, i, \end{aligned}$$

In particular, taking $i = 1$ in the first equation and $i = p$ in the second inequation, together with the fact $\lambda_j(\mathbf{B}) = -\lambda_{p+1-j}(-\mathbf{B})$ for $j = 1, \ldots, p$, implies

$$\max_{1 \leq j \leq p} |\lambda_j(\mathbf{A}) - \lambda_j(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|.$$

Probability identities

The results of the previous sections are algebraic identities allowing for handling the resolvent $\mathbf{Q}_\mathbf{M}$ of the deterministic matrix $\mathbf{M}$. The second ingredient of random matrix analysis lies in asymptotic probability approximations as the dimensions of $\mathbf{M}$ increase. Quite surprisingly, most results essentially revolve around the convergence of a certain quadratic form, which is often nothing more than a mere extension of the *law of large numbers*.

Those quadratic form convergence results come under multiple forms. The historical form, due to Bai and Silverstein, sometimes referred to as the "trace lemma", is as follows.

Lemma 2.11 (Quadratic-form-close-to-the-trace, Lemma B.26 in [BS10])**.** *Let $\mathbf{x} \in \mathbb{R}^p$ have i.i.d. entries of zero mean, unit variance and $\mathbb{E}[|x_i|^L] \leq \nu_L$ for some $L \geq 1$. Then for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $l \geq 1$*

$$\mathbb{E}\left[\left| \mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x} - \operatorname{tr} \mathbf{A} \right|^l \right] \leq K_l \left[\left(\nu_4 \operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{T}) \right)^{l/2} + \nu_{2l} \operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})^{l/2} \right]$$

for some constant $K_l > 0$ independent of p. In particular, if $\|\mathbf{A}\| \leq 1$ and the entries of $\mathbf{x}$ have bounded eighth-order moment,

$$\mathbb{E}\left[\left(\mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x} - \operatorname{tr} \mathbf{A} \right)^4 \right] \leq K p^2$$

for some K independent of p, and consequently, as $p \to \infty$,

$$\frac{1}{p} \mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x} - \frac{1}{p} \operatorname{tr} \mathbf{A} \xrightarrow{a.s.} 0.$$

This last result is rather intuitive. For $\mathbf{A} = \mathbf{I}_p$, this is simply an instance of the law of large numbers. For generic $\mathbf{A}$, first note that, by the independence of the entries of $\mathbf{x}$, $\mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}] = \mathbb{E}[\mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^\mathsf{T})] = \mathrm{tr}\,\mathbf{A}$. Exploiting the fact that $\mathrm{Var}\left[\frac{1}{p}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}\right] = O(p^{-1})$ then ensures that $\frac{1}{p}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \frac{1}{p}\mathrm{tr}\,\mathbf{A} \to 0$, but only in probability; since the variance calculus involves exponentiating the entries x_i of $\mathbf{x}$ to power 4, they need to be of finite fourth power. The almost sure convergence is achieved by showing the faster moment convergence $\mathbb{E}[(\frac{1}{p}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \frac{1}{p}\mathrm{tr}\,\mathbf{A})^4] = O(p^{-2})$ which is the second statement of the lemma and requires 8-th order exponentiation of the x_i's. The request for $\mathbf{A}$ to be of bounded norm with respect to p in this case "stabilizes" the quadratic form $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ by maintaining its random concentration properties.

Recalling that $\|\mathbf{Q}_\mathbf{M}(z)\| \leq (\mathrm{dist}(z, \mathrm{supp}(\mu_\mathbf{M})))^{-1}$, Lemma 2.11 can be exploited for $\mathbf{A} = \mathbf{Q}_\mathbf{M}(z)$ for all z away from the support of $\mu_\mathbf{M}$ and all $\mathbf{x}$ independent of $\mathbf{Q}_\mathbf{M}(z)$. The core of the proofs of the main random matrix results is uniquely based on this last remark.

These identities constitute the main technical ingredients needed to understand the proofs of both historical and recent random matrix results. The next section introduces the most fundamental of those which will be called after over and over in the remainder of the manuscript.

2.2.2 The Marčenko-Pastur law

We start by illustrating how the aforementioned tools are used to prove one of the most popular results in random matrix theory: the Marčenko-Pastur law. Another important result is the Wigner semi-circle law, which, despite being popular in many graph-based problems, is less covered in the works addressed in this manuscript and thus omitted.

To simplify the exposition of the results, we will use the notation for deterministic equivalents introduced in Notation 1. That is, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$, we will denote $\mathbf{X} \leftrightarrow \mathbf{Y}$ if, for all unit norm $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\frac{1}{n}\mathrm{tr}\,\mathbf{A}(\mathbf{X}-\mathbf{Y}) \xrightarrow{a.s.} 0$, $\mathbf{a}^\mathsf{T}(\mathbf{X}-\mathbf{Y})\mathbf{b} \xrightarrow{a.s.} 0$ and $\|\mathbb{E}[\mathbf{X}-\mathbf{Y}]\| \to 0$.

Most of the results involve Stieltjes transforms $m_\mu(z)$ of probability measures with support $\mathrm{supp}(\mu)$. Since Stieltjes transforms are such that $m_\mu(z) > 0$ for $z < \inf \mathrm{supp}(\mu)$, $m_\mu(z) < 0$ for $z > \sup \mathrm{supp}(\mu)$ and $\Im[z]\Im[m_\mu(z)] > 0$ if $z \in \mathbb{C} \setminus \mathbb{R}$, it will be convenient in the following to consider the set

$$\mathcal{Z}(\mathcal{A}) = \left\{(z, m) \in \mathcal{A}^2,\ (\Im[z]\Im[m] > 0 \text{ if } \Im[z] \neq 0) \text{ or } (zm < 0 \text{ if } \Im[z] = 0)\right\}.$$

We present the Marčenko-Pastur law under the slightly modified form of a deterministic equivalent for the resolvent $\mathbf{Q}(z)$.

Theorem 2.3 (From [MP67])**.** *Consider the resolvent $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$, for $\mathbf{X} \in \mathbb{R}^{p\times n}$ having i.i.d. zero mean and unit variance entries. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, we have*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p \tag{2.8}$$

with $(z, m(z))$ the unique solution in $\mathcal{Z}(\mathbb{C} \setminus [(1-\sqrt{c})^2, (1+\sqrt{c})^2])$ of

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{2.9}$$

The function $m(z)$ is the Stieltjes transform of the probability measure μ given explicitly by

$$\mu(dx) = (1 - c^{-1})^+\delta(x) + \frac{1}{2\pi cx}\sqrt{(x-a)^+(b-x)^+}dx \tag{2.10}$$

where $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$ *and* $(x)^+ = \max(0, x)$, *which is known as the* Marčenko-Pastur distribution. *In particular, with probability one, the empirical spectral measure* $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ *converges weakly to* μ.

Figure 2.1 depicts the density of the Marčenko-Pastur distribution for different values of c. For a fixed dimension p, the ratio c decreases as the number of samples n grows large, so that the eigenvalues of the sample covariance matrix become more "concentrated" (their spread is given by the length of the support $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$) around the unique population covariance matrix eigenvalue equal to 1.

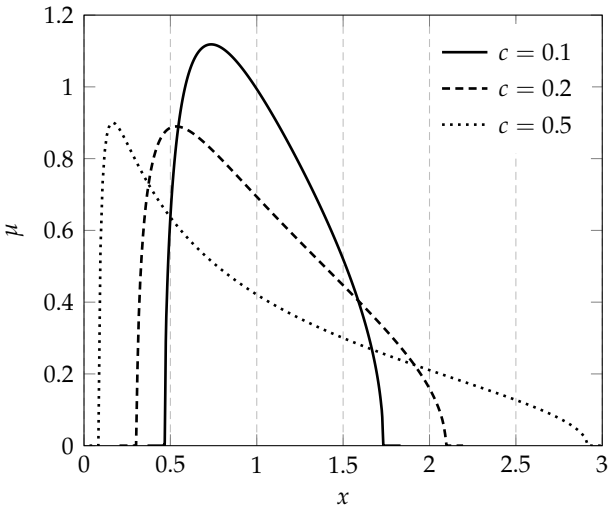


Figure 2.1: Marčenko-Pastur distribution for different c

Proof. Before going into the details of the proof, we first give a few intuitive arguments.

Intuitive idea. A first heuristic derivation consists in iteratively "guessing" the form of $\bar{\mathbf{Q}}(z) = \mathbf{F}(z)^{-1}$ for some matrix $\mathbf{F}(z)^{-1}$. To this end, from Lemma 2.1, it first appears that

$$\begin{aligned}\mathbf{Q}(z) - \bar{\mathbf{Q}}(z) &= \mathbf{Q}(z)\left(\mathbf{F}(z) + z\mathbf{I}_p - \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)\bar{\mathbf{Q}}(z) \\ &= \mathbf{Q}(z)\left(\mathbf{F}(z) + z\mathbf{I}_p - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)\bar{\mathbf{Q}}(z)\end{aligned}$$

For $\bar{\mathbf{Q}}(z)$ to be a deterministic equivalent for $\mathbf{Q}(z)$, we wish in particular that $\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)) \xrightarrow{a.s.} 0$, for $\mathbf{A}$ deterministic with $\|\mathbf{A}\| = 1$. That is

$$\frac{1}{p}\operatorname{tr}(\mathbf{F}(z) + z\mathbf{I}_p)\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z) - \frac{1}{n}\sum_{i=1}^n \frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i \xrightarrow{a.s.} 0. \tag{2.11}$$

We recognize in $\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i$ a quadratic form on which we would like to use Lemma 2.11 to turn it into a trace term independent of $\mathbf{x}_i$. Yet, Lemma 2.11 cannot be used as $\mathbf{Q}(z)$ depends on $\mathbf{x}_i$. To counter the difficulty, we then use Lemma 2.8 to write

$$\mathbf{Q}(z)\mathbf{x}_i = \frac{\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}(z)\mathbf{x}_i}$$

where $\mathbf{Q}_{-i}(z) = (\frac{1}{n}\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j^\mathsf{T} - z\mathbf{I}_p)^{-1}$ is independent of $\mathbf{x}_i$. Now legitimately applying Lemma 2.11, we find that

$$\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i = \frac{\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}(z)\mathbf{x}_i} \simeq \frac{\frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}(z)}. \tag{2.12}$$

From Lemma 2.9, normalized traces involving $\mathbf{Q}_{-i}(z)$ or $\mathbf{Q}(z)$ are asymptotically identical and thus this further reads

$$\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i \simeq \frac{\frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)}.$$

Getting back to (2.11), we thus end up with the approximation

$$\frac{1}{p}\operatorname{tr}(\mathbf{F}(z)+z\mathbf{I}_p)\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z) \simeq \frac{\frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)}.$$

As a consequence, we can now "guess" the form of $\mathbf{F}(z)$. Indeed, if it is to exist, $\mathbf{F}(z)$ must be of the type

$$\mathbf{F}(z) \simeq \left(-z + \frac{1}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)}\right)\mathbf{I}_p$$

for the approximation above to hold. To close the loop, taking $\mathbf{A} = \mathbf{I}_p$, $\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)$ appearing in this display must be well approximated by $m(z) \equiv \frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)$ so that

$$\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) \simeq m(z) = \frac{1}{-z+\frac{1}{1+\frac{p}{n}\frac{1}{p}\operatorname{tr}\mathbf{Q}(z)}} \simeq \frac{1}{-z+\frac{1}{1+\frac{p}{n}m(z)}} \tag{2.13}$$

and we thus have finally

$$\bar{\mathbf{Q}}(z) = \mathbf{F}(z)^{-1} = m(z)\mathbf{I}_p$$

where, in the large n, p limit, $m(z)$ is solution to

$$m(z) = \frac{1}{-z+\frac{1}{1+cm(z)}}$$

or equivalently

$$zcm^2(z) - (1-c-z)m(z) + 1 = 0.$$

This equation has two solutions defined by the two roots of the complex square root function

$$m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{((1+\sqrt{c})^2-z)((1-\sqrt{c})^2-z)}}{2cz}$$

only one of which is such that $\Im[z]\Im[m(z)] > 0$ as imposed by the definition of Stieltjes transforms. Now, from the inverse Stieltjes transform theorem, Theorem 2.1, we find that $m(z)$ is the Stieltjes transform of the measure μ with

$$\mu([a,b]) = \frac{1}{\pi}\lim_{\epsilon\downarrow 0}\int_a^b \Im[m(x+\imath\epsilon)]dx$$

for all continuity points $a, b \in \mathbb{R}$ of μ. The term under the square root in $m(z)$ being negative only in the set $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$, the latter defines the support of the continuous part of the measure μ with density $\frac{\sqrt{((1+\sqrt{c})^2-x)(x-(1-\sqrt{c})^2)}}{2c\pi x}$ at point x in the set. The case $x = 0$ brings a discontinuity in μ with weight equal to

$$\mu(\{0\}) = -\lim_{y \downarrow 0} \imath y m(\imath y) = \frac{c-1}{2c} \pm \frac{c-1}{2c}$$

where the sign is established by a second order development of $zm(z)$ in the neighborhood of zero.

Detailed proof. Having heuristically identified $\bar{\mathbf{Q}}(z)$, we shall now use sound mathematical tools to prove that, indeed, $\bar{\mathbf{Q}}(z)$ is a deterministic equivalent for $\mathbf{Q}(z)$ in the sense of the theorem statement. Let us first show that $\mathbb{E}[\mathbf{Q}(z)] = \bar{\mathbf{Q}}(z) + o_{\|\cdot\|}(1)$, where $o_{\|\cdot\|}(1)$ denotes a matrix term of vanishing operator norm as $n, p \to \infty$.

Convergence in mean. For mathematical convenience, we will take $z < 0$ in what follows. Since $\mathbf{Q}(z)$ and $\bar{\mathbf{Q}}(z)$ from the theorem statement are complex analytic functions for $z \notin \mathbb{R}^+$ (matrix-valued Stieltjes transforms are analytic), obtaining the convergence results on $\mathbb{R}^-$ is equivalent to obtaining the result on all of $\mathbb{C} \setminus \mathbb{R}^+$.

We proceed in two steps by first introducing the intermediate deterministic quantities $\alpha(z) \equiv \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}(z)]$ and $\bar{\tilde{\mathbf{Q}}}(z) \equiv (-z + \frac{1}{1+\alpha(z)})^{-1} \mathbf{I}_p$. From Lemma 2.1, we have (the argument z in $\alpha(z)$, $\mathbf{Q}(z)$ and $\bar{\tilde{\mathbf{Q}}}(z)$ is dropped when confusion is not possible)

$$\begin{aligned}
\mathbb{E}[\mathbf{Q} - \bar{\tilde{\mathbf{Q}}}] &= \mathbb{E}\mathbf{Q}\left(\frac{\mathbf{I}_p}{1+\alpha} - \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right)\bar{\tilde{\mathbf{Q}}} = \frac{\mathbb{E}[\mathbf{Q}]}{1+\alpha}\bar{\tilde{\mathbf{Q}}} - \frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}]\bar{\tilde{\mathbf{Q}}} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+\alpha}\bar{\tilde{\mathbf{Q}}} - \sum_{i=1}^n \frac{1}{n}\mathbb{E}[\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}]\bar{\tilde{\mathbf{Q}}} = \frac{\mathbb{E}[\mathbf{Q}]}{1+\alpha}\bar{\tilde{\mathbf{Q}}} - \sum_{i=1}^n \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right]\bar{\tilde{\mathbf{Q}}}
\end{aligned}$$

where we applied Lemma 2.8 to obtain the last equality and denoted $\mathbf{Q}_{-i} \equiv (\sum_{j \neq i} \frac{1}{n}\mathbf{x}_j\mathbf{x}_j^\mathsf{T} - z\mathbf{I}_p)^{-1}$ as previously.

Since we expect $\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i$ to be close to α (as a consequence of Lemma 2.11), we rewrite

$$\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} = \frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\alpha} - \frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}(\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i - \alpha)}{(1+\alpha)(1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i)}$$

so that

$$\begin{aligned}
\mathbb{E}[\mathbf{Q} - \bar{\tilde{\mathbf{Q}}}] &= \frac{\mathbb{E}[\mathbf{Q}]}{1+\alpha}\bar{\tilde{\mathbf{Q}}} - \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha} + \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}d_i\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha} \\
&= \frac{\mathbb{E}[\mathbf{Q}]}{1+\alpha}\bar{\tilde{\mathbf{Q}}} - \sum_{i=1}^n \frac{\mathbb{E}\left[\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha} + \frac{\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha}
\end{aligned}$$

where we introduced $\mathbf{D} = \operatorname{diag}\{d_i\}_{i=1}^n$ for $d_i = \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i - \alpha$, and used again Lemma 2.8 to write $\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} = \mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ in the first equality. Since $\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbb{E}[\mathbf{Q}_{-i}]$, this further reads

$$\mathbb{E}[\mathbf{Q} - \bar{\tilde{\mathbf{Q}}}] = \frac{1}{n}\sum_{i=1}^n (\mathbb{E}[\mathbf{Q}] - \mathbb{E}[\mathbf{Q}_{-i}])\frac{\bar{\tilde{\mathbf{Q}}}}{1+\alpha} + \frac{\mathbb{E}\left[\frac{1}{n}\mathbf{Q}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha}.$$

Again from Lemma 2.1 and 2.8,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{Q}-\mathbf{Q}_{-i}] &= -\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\right]\\ &= -\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}\left(1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i\right)\right]\\ &= -\frac{1}{n}\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{D}_2\mathbf{X}^\mathsf{T}\mathbf{Q}\right]\end{aligned}$$

where $\mathbf{D}_2 = \operatorname{diag}\left\{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i\right\}_{i=1}^n$ and thus

$$\mathbb{E}[\mathbf{Q}-\bar{\tilde{\mathbf{Q}}}] = -\frac{1}{n}\mathbb{E}\left[\mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{D}_2\mathbf{X}^\mathsf{T}\mathbf{Q}\right]\frac{\bar{\tilde{\mathbf{Q}}}}{1+\alpha} + \frac{\mathbb{E}\left[\frac{1}{n}\mathbf{Q}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\right]\bar{\tilde{\mathbf{Q}}}}{1+\alpha}. \tag{2.14}$$

It remains to show that the right-hand side terms vanish in the large p, n limit.

For the first term, note that

$$0 \preceq \mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{D}_2\mathbf{X}^\mathsf{T}\mathbf{Q} \preceq \mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}\max_{1\le i\le n}[\mathbf{D}_2]_{ii}$$

in the order of symmetric matrices. Since $\mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} = \mathbf{I}_p + z\mathbf{Q}$ which is of bounded operator norm, controlling $\|\mathbb{E}[\mathbf{Q}\frac{1}{n}\mathbf{X}\mathbf{D}_2\mathbf{X}^\mathsf{T}\mathbf{Q}]\|$ boils down to controlling $\mathbb{E}[\max_i[\mathbf{D}_2]_{ii}]$. This can be established in various ways. From the union bound and the i.i.d. nature of the $\mathbf{x}_i$'s,

$$P\left(\max_i[\mathbf{D}_2]_{ii} > t\right) \le nP\left([\mathbf{D}_2]_{11} > t\right).$$

Now, by Markov's inequality $P(X > a) \le \mathbb{E}[X^k]/a^k$ for every k (for $X, a > 0$) and the moment inequality in Lemma 2.11 for, say $l = 4$, $P(\max_i[\mathbf{D}_2]_{ii} > t)$ may be bounded by a function decreasing as t^{-2}, for all $t > 1+\alpha(z)$, and of order n^{-1}. Since $\mathbb{E}[X] = \int_{X>0}P(X > t)dt$, we then find that $\mathbb{E}[\max_i[\mathbf{D}_2]_{ii}]$ is bounded. Alternatively, one may have used a concentration inequality argument to show the same. Consequently, due to the leading $1/n$ factor in front of the first right-hand side term of (2.14), this term vanishes as $n, p \to \infty$.

To handle the second right-hand side term in (2.14), one needs to control the norm of $\frac{1}{n}\mathbf{Q}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\bar{\tilde{\mathbf{Q}}}$. This is not a symmetric matrix, but $\mathbb{E}[\mathbf{Q}-\bar{\tilde{\mathbf{Q}}}]$ is. We may thus rewrite (2.14) as the half-sum of itself and its transpose and we are thus left to controlling the operator norm of $\frac{1}{n}\mathbf{Q}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\bar{\tilde{\mathbf{Q}}} + \frac{1}{n}\bar{\tilde{\mathbf{Q}}}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\mathbf{Q}$. Using the matrix inequalities $\mathbf{A}\mathbf{B}^\mathsf{T}+\mathbf{B}\mathbf{A}^\mathsf{T} \preceq \mathbf{A}\mathbf{A}^\mathsf{T}+\mathbf{B}\mathbf{B}^\mathsf{T}$ (from $(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\mathsf{T} \succeq 0$) and $\mathbf{A}\mathbf{B}^\mathsf{T}+\mathbf{B}\mathbf{A}^\mathsf{T} \succeq -\mathbf{A}\mathbf{A}^\mathsf{T}-\mathbf{B}\mathbf{B}^\mathsf{T}$ (from $(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})^\mathsf{T} \succeq 0$), we are left to bounding the norm of

$$\mathbb{E}\left[\frac{1}{n\sqrt{n}}\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}\right] + \mathbb{E}\left[\frac{1}{\sqrt{n}}\bar{\tilde{\mathbf{Q}}}\mathbf{X}\mathbf{D}^2\mathbf{X}^\mathsf{T}\bar{\tilde{\mathbf{Q}}}\right]$$

where the division of the $1/n^2$ term into $1/(n\sqrt{n})$ and $1/\sqrt{n}$ is essential. The first term above is easily seen to be of order $1/\sqrt{n}$. As for the second, using as above the moment inequality in Lemma 2.11 along with Markov's inequality, it appears to be also of order $1/\sqrt{n}$. This can be anticipated by noticing that $d_i = \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i - \alpha$ fluctuates as $1/\sqrt{n}$ (by a central limit theorem argument) and thus d_i^2 is essentially of order $1/n$.

Gathering the pieces together, we thus conclude that

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\tilde{\mathbf{Q}}}\| \to 0, \quad \text{with } \bar{\tilde{\mathbf{Q}}} = \left(\frac{1}{1+\alpha(z)} - z\right)^{-1}\mathbf{I}_p.$$

Since

$$\alpha(z) = \frac{1}{n}\operatorname{tr}\mathbb{E}[\mathbf{Q}(z)] = \frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}(z) + o(1) = \frac{c}{\frac{1}{1+\alpha(z)} - z} + o(1)$$

by defining $m(z)$ as the unique Stieltjes transform solution with $\Im(z)\Im[m(z)] > 0$ of

$$\frac{1}{m(z)} = \frac{1}{1+cm(z)} - z \Leftrightarrow zcm^2(z) - (1-c-z)m(z) + 1 = 0$$

(the uniqueness is easily shown by solving the quadratic equation), we finally have $cm(z) - \alpha(z) \to 0$, which concludes the proof of (2.8).

Almost sure convergence. To now prove the almost sure convergence $\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}}) \xrightarrow{a.s.} 0$ and $\mathbf{a}^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b} \xrightarrow{a.s.} 0$, it suffices to show

$$\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q}) \xrightarrow{a.s.} 0, \quad \mathbf{a}^\mathsf{T}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\mathbf{b} \xrightarrow{a.s.} 0.$$

We will only show here the leftmost convergence. This follows from either a moment or a concentration argument. The historical approach, due to Bai and Silverstein (see e.g., in [BS10]), exploits the following martingale difference inequality.

Lemma 2.12 (Burkholder inequality, Lemma 2.1 in [BS10])**.** *Let $\{X_i\}_{i=1}^\infty$ be a martingale difference for the increasing σ-field $\{\mathcal{F}_i\}$ and denote $\mathbb{E}_k$ the expectation with respect to $\mathcal{F}_k$. Then, for $k \geq 2$,*

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^k\right] \leq C_k\left(\mathbb{E}\left[\sum_{i=1}^n \mathbb{E}_{k-1}[|X_i|^2]\right]^{k/2} + \sum_{i=1}^n \mathbb{E}[|X_i|^k]\right).$$

Remarking that

$$\begin{aligned}\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q}) &= \sum_{i=1}^n \mathbb{E}_i\left[\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{Q}\right] - \mathbb{E}_{i-1}\left[\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{Q}\right] \\ &= \frac{1}{p}\sum_{i=1}^n (\mathbb{E}_i - \mathbb{E}_{i-1})\left[\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbf{Q}_{-i})\right]\end{aligned}$$

(since $\mathbb{E}_i[\operatorname{tr}\mathbf{A}\mathbf{Q}_{-i}] = \mathbb{E}_{i-1}[\operatorname{tr}\mathbf{A}\mathbf{Q}_{-i}]$) for $\mathcal{F}_i$ the σ-field generating the columns $\mathbf{x}_{i+1}, \ldots, \mathbf{x}_n$ of $\mathbf{X}$ and with the convention $\mathbb{E}_0[f(\mathbf{X})] = f(\mathbf{X})$, which forms a martingale difference sequence, we fall under the scope of Burkholder's lemma. Now, from the identity $\mathbf{Q} = \mathbf{Q}_{-i} - \frac{1}{n}\frac{\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}$ (Lemma 2.8),

$$(\mathbb{E}_i - \mathbb{E}_{i-1})\left[\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbf{Q}_{-i})\right] = -(\mathbb{E}_i - \mathbb{E}_{i-1})\frac{\frac{1}{pn}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{A}\mathbf{Q}_{-i}\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}$$

which is bounded by $1/p$. As a consequence, from Lemma 2.12,

$$\mathbb{E}\left[\left|\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\mathbb{E}\mathbf{Q})\right|^4\right] = O(n^{-2}).$$

From Markov's inequality (i.e., $P(|X| > t) \leq \mathbb{E}[|X|^k]/t^k$) and Borel-Cantelli's lemma (i.e., $P(|X_n| > t) = O(n^{-\ell})$ for some $\ell > 1$ for all $t > 0$ implies $X_n \xrightarrow{a.s.} 0$), we then conclude that

$$\frac{1}{p} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \mathbb{E}\mathbf{Q}) \xrightarrow{a.s.} 0$$

as requested. □

Remark 2.3 (Proof by Stein's lemma and Nash–Poincaré inequality)**.** In [PS11], Pastur and Scherbina propose an alternative proof to Theorem 2.3, based on a two-fold method: i) a proof for Gaussian $\mathbf{X}$ and ii) an interpolation method to non-Gaussian $\mathbf{X}$. The Gaussian case is handled through the powerful Stein's lemma.

Convergence in mean by Stein's lemma.

Lemma 2.13 ([Ste81])**.** *Let $x \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \mapsto \mathbb{R}$ a continuously differentiable function such that $\mathbb{E}[f'(x)] < \infty$. Then,*

$$\mathbb{E}[xf(x)] = \mathbb{E}[f'(x)]. \tag{2.15}$$

In particular, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p \times p}$ and $f : \mathbb{R}^p \mapsto \mathbb{R}$ a continuously differentiable function with derivatives having at most polynomial growth with respect to p,

$$\mathbb{E}[x_i f(\mathbf{x})] = \sum_{j=1}^{p} \mathbf{C}_{ij} \mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial x_j}\right] \tag{2.16}$$

where $\partial/\partial x_i$ indicates differentiation with respect to the i-th entry of $\mathbf{x}$.

The lemma, sometimes referred to as the integration-by-parts formula for Gaussian variables, simply follows from

$$\mathbb{E}[xf(x)] = \int x f(x) e^{-\frac{1}{2}x^2} dx = [-f(x)e^{-\frac{1}{2}x^2}]_{-\infty}^{\infty} + \int f'(x) e^{-\frac{1}{2}x^2} dx = \mathbb{E}[f'(x)]$$

by integration by parts $\int u'v = [uv] - \int uv'$ for $u(x) = -e^{-\frac{1}{2}x^2}$ and $v(x) = f(x)$.

To exploit Lemma 2.13, let us thus assume $\mathbf{X}$ Gaussian, i.e., $\mathbf{X}_{ij} \sim \mathcal{N}(0,1)$. Observe that $\mathbf{Q} = \frac{1}{z}\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q} - \frac{1}{z}\mathbf{I}_p$, so that

$$\mathbb{E}[\mathbf{Q}_{ij}] = \frac{1}{zn} \sum_{k=1}^{n} \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] - \frac{1}{z}\delta_{ij}$$

in which $\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \mathbb{E}[xf(x)]$ for $x = \mathbf{X}_{ik}$ and $f(x) = [\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}$. Therefore, from the lemma and the fact that $\partial\mathbf{Q} = -\frac{1}{n}\mathbf{Q}\partial(\mathbf{X}\mathbf{X}^\mathsf{T})\mathbf{Q}$,

$$\mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \mathbb{E}\left[\frac{\partial[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}}{\partial \mathbf{X}_{ik}}\right] = \mathbb{E}[\mathbf{Q}_{ij}] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{X}]_{kk}\mathbf{Q}_{ij}\right] - \mathbb{E}\left[\frac{1}{n}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ki}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}\right].$$

so that, summing over k,

$$\frac{1}{z}\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[\mathbf{X}_{ik}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{kj}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}_{ij}] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}_{ij}\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T})] - \frac{1}{z}\frac{1}{n^2}\mathbb{E}[\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}]_{ij}.$$

It is not too difficult to see that the rightmost term has vanishing operator norm (of order $O(1/n)$) as $n, p \to \infty$. Also recall that $\operatorname{tr}(\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}) = np + zn \operatorname{tr}\mathbf{Q}$. As a result, matrix-wise, we obtain

$$\mathbb{E}[\mathbf{Q}] + \frac{1}{z}\mathbf{I}_p = \mathbb{E}[\mathbf{X}_{\cdot k}[\mathbf{X}^\mathsf{T}\mathbf{Q}]_{k\cdot}] = \frac{1}{z}\mathbb{E}[\mathbf{Q}] - \frac{1}{z}\frac{1}{n}\mathbb{E}[\mathbf{Q}(p + z\operatorname{tr}\mathbf{Q})] + o_{\|\cdot\|}(1).$$

As $\frac{1}{n}\operatorname{tr}\mathbf{Q}$ is expected to converge to some $m(z)$, it can be taken out of the expectation in the limit so that, gathering all terms proportional to $\mathbb{E}[\mathbf{Q}]$ on the left-hand side, we finally have

$$\mathbb{E}[\mathbf{Q}](1 - p/n - z - p/nzm(z)) = \mathbf{I}_p + o_{\|\cdot\|}(1)$$

which, taking the trace to identify $m(z)$, concludes the proof for the Gaussian case.

Almost sure convergence by Nash–Poicaré inequality. To prove the almost sure convergence of traces and bilinear forms of the resolvent in the case of Gaussian $\mathbf{X}$, one may then use the powerful Nash–Poincaré inequality proposed by Pastur.

Lemma 2.14 (Nash–Poincaré inequality from [Pas05])**.** *For $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \in \mathbb{R}^{p\times p}$ and $f : \mathbb{R}^p \mapsto \mathbb{R}$ continuously differentiable with derivatives having at most polynomial growth with respect to p,*

$$\operatorname{Var}[f(\mathbf{x})] \leq 2 \sum_{1\leq i,j\leq n} \mathbf{C}_{ij}\mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial x_i}\frac{\partial f(\mathbf{x})}{\partial x_j}\right].$$

In the present case, for Gaussian $\mathbf{X}$ with $\mathbf{X}_{ij} \sim \mathcal{N}(0,1)$,

$$\operatorname{Var}\left[\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{Q}\right] \leq \frac{2}{p^2}\sum_{1\leq i,j\leq n}\mathbb{E}\left[\left|\frac{\partial \operatorname{tr}\mathbf{A}\mathbf{Q}}{\partial \mathbf{X}_{ij}}\right|^2\right].$$

Again using $\partial\mathbf{Q} = -\frac{1}{n}\mathbf{Q}\partial(\mathbf{X}\mathbf{X}^\mathsf{T})\mathbf{Q}$, we find

$$\frac{\partial \operatorname{tr}\mathbf{A}\mathbf{Q}}{\partial \mathbf{X}_{ij}} = -\frac{1}{n}[\mathbf{Q}\mathbf{A}\mathbf{Q}\mathbf{X} + \mathbf{Q}\mathbf{A}^\mathsf{T}\mathbf{Q}\mathbf{X}]_{ij}$$

so that, from $(a+b)^2 \leq 2(a^2+b^2)$ and $\|\mathbf{A}\| = 1$,

$$\frac{2}{p^2}\sum_{1\leq i,j\leq n}\mathbb{E}\left[\left|\frac{\partial \operatorname{tr}\mathbf{A}\mathbf{Q}}{\partial \mathbf{X}_{ij}}\right|^2\right] \leq \frac{4}{p^2n^2}\big(\operatorname{tr}(\mathbf{Q}\mathbf{A}\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{A}^\mathsf{T}\mathbf{Q}) + \operatorname{tr}(\mathbf{Q}\mathbf{A}^\mathsf{T}\mathbf{Q}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{A}\mathbf{Q})\big) = O(n^{-2}).$$

By Markov's inequality and the Borel Cantelli lemma, we thus have that $\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q} - \mathbb{E}\mathbf{Q}) \xrightarrow{a.s.} 0$.

When it comes to evaluating the fluctuations of $\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}$ with the same approach, it appears that $\operatorname{Var}[\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}] = O(n^{-1})$ which is enough to ensure convergence in probability (by Markov's inequality) but not almost surely (as the Borel Cantelli lemma cannot be applied). Thus one needs to resort to evaluating a higher moment bound, such as $\mathbb{E}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^4]$. To this end, we may use the fact that

$$\begin{aligned}\mathbb{E}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^4] &= \operatorname{Var}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^2] + \mathbb{E}\left[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^2\right]^2 \\ &= \operatorname{Var}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^2] + \operatorname{Var}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|]^2.\end{aligned}$$

Since we know that the rightmost term is of order $O(n^{-2})$, it remains to show, again through Nash–Poincaré inequality, that $\operatorname{Var}[|\mathbf{a}^\mathsf{T}(\mathbf{Q} - \mathbb{E}\mathbf{Q})\mathbf{b}|^2] = O(n^{-2})$ which is a cumbersome but easily obtained result as well.

Interpolation trick to non-Gaussian $\mathbf{X}$. To "interpolate" these results from Gaussian $\mathbf{X}$ to non-Gaussian $\mathbf{X}$, one may then use a generalized version of Stein's lemma to non-Gaussian distributions, for which we have:

Lemma 2.15 (Interpolation trick, Corollary 3.1 in [LP09])**.** *For $x \in \mathbb{R}$ a random variable with zero mean and unit variance, $y \sim \mathcal{N}(0,1)$, and f a $k+2$-differentiable function with bounded derivatives,*

$$\mathbb{E}[f(x)] - \mathbb{E}[f(y)] = \sum_{l=2}^{k} \frac{\kappa_{l+1}}{2l!} \int_0^1 \mathbb{E}[f^{(l+1)}x(t)]t^{(l-1)/2}dt + \epsilon_k$$

where κ_l is the l^{th} cumulant of x, $x(t) = \sqrt{t}x + (1-\sqrt{t})y$, and $|\epsilon_k| \leq C_k \mathbb{E}[|x|^{k+2}] \sup_t |f^{(k+2)}(t)|$ for some constant C_k only dependent on k.

All Gaussian expectations (means and variance) in the proof above can then be expressed as their non-Gaussian form up to a sum of moment control on the derivatives of f.

Remark 2.4 (On the convergence rates)**.** *In the course of the proofs above, we saw examples of a general concentration trend for linear statistics and quadratic forms of random matrices. We shall indeed typically have for most of the models of random matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ under study that*

- *linear statistics $\frac{1}{n}\sum_{i=1}^n f(\lambda_i(\mathbf{X}))$ for sufficiently well-behaved f (so for instance $\frac{1}{n}\operatorname{tr}\mathbf{Q}_{\mathbf{X}}(z) = \frac{1}{n}\sum_i(\lambda_i(\mathbf{X})-z)^{-1}$) converge at a speed $O(1/n)$ (their variance scales as $O(1/n^2)$. From a central-limit theorem viewpoint, this is as fast as it can get. Indeed, $\mathbf{X}$ is maximally composed of $p \times n = O(n^2)$ "degrees of freedom" and thus, by the central limit theorem, fluctuations are at most at speed $O(1/\sqrt{n^2}) = O(1/n)$.*
- *quadratic forms $\mathbf{a}^\mathsf{T} f(\mathbf{X})\mathbf{b}$ where $f(\mathbf{X}) = \mathbf{U}\operatorname{diag}(f(\lambda_i(\mathbf{X}))\mathbf{U}^\mathsf{T}$ (in the spectral decomposition of $\mathbf{X}$) typically converge at a slower $O(1/\sqrt{n})$ speed.*

This remark is particularly interesting as it indicates, from a statistics viewpoint that, for $\mathbf{X} \in \mathbb{R}^{p\times n}$, asymptotic approximations may gain accuracy by doubly exploiting the degrees of freedom in both the sample (n) and feature (p) direction.

Remark 2.5 (On the assumptions on $\mathbf{X}$)**.** *The Marčenko–Pastur law has been widely generalized and several times proved using different techniques. For instance [A+11, O'R12] assume the $\mathbf{X}_{ij}$ are "weakly" dependent in the sense that their correlation or higher order cross-moments vanish at a certain speed as $n, p \to \infty$. Alternatively, the works of Bai and Silverstein (see [BS06]) tend to assume that the entries of $\mathbf{X}$ are not necessarily identically distributed; in this case, an additional condition on the tails $\mathbb{P}(|\mathbf{X}_{ij}| > t)$ of the probability measures of the entries (for instance a uniform bound on some moment higher than 2) is needed. In [EK09], El Karoui provides a first result which assumes the columns $\mathbf{x}_i$ of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ are independent* concentrated *random vectors. (Very) roughly speaking, concentrated random vectors $\mathbf{x} \in \mathbb{R}^p$ can be written as $\mathbf{x} = \varphi(\tilde{\mathbf{x}})$ where $\tilde{\mathbf{x}}$ has standard i.i.d. entries with either a Gaussian law or a bounded support, and $\varphi : \mathbb{R}^p \to \mathbb{R}^p$ is any 1-Lipschitz function: this assumption maintains the p degrees of freedom in $\mathbf{x}$ (arising from $\tilde{\mathbf{x}}$) while allowing for strong nonlinear correlation between the entries of $\mathbf{x}$. In this case, the Marčenko–Pastur law is indeed still valid if $\varphi(\mathbf{x})$ has zero mean and identity covariance.*

2.2.3 Large dimensional sample covariance matrices

The Marčenko–Pastur and semi-circle theorems have long been the gold-standard in both theoretical and applied random matrix theory, in the sense that most mathematical studies and practical results concerned the Wishart and Wigner random matrix models. But

the assumption of data $\mathbf{X}$ with i.i.d., let alone standard Gaussian, entries is often limiting. In statistics where one is interested in the correlation $\mathbf{X}\mathbf{X}^\mathsf{T}$, it is expected that the columns $\mathbf{x}_i \in \mathbb{R}^p$ of $\mathbf{X}$ exhibit a correlation structure and be non-necessarily independent (in particular when they are samples from a time series).

This section introduces generalizations of these results, to a level that is convenient to machine learning applications. In particular, in order to model the existence of classes within the data, $\mathbf{X}$ will often be subdivided into subblocks that can be identified with each class.

Our first result generalizes the Marčenko–Pastur law to sample covariance matrices and is originally due to a long line of works by Bai and Silverstein [SB95].

Theorem 2.4 (Sample covariance matrix, from [SB95])**.** *Let $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$ with $\mathbf{C} \in \mathbb{R}^{p\times p}$, $\mathbf{Z} \in \mathbb{R}^{p\times n}$ having i.i.d. zero mean and unit variance entries. Then, as $n, p \to \infty$ with $p/n \to c > 0$, letting $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ and $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - z\mathbf{I}_n)^{-1}$, we have*

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

where $(z, \tilde{m}_p(z))$ is the unique solution in $\mathcal{Z}(\mathbb{C}\setminus\mathbb{R}^+)$ of

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

In particular, if $\mu_\mathbf{C} \to \nu$ as $p \to \infty$, then $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \xrightarrow{a.s.} \mu$ and $\mu_{\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}} \xrightarrow{a.s.} \tilde{\mu}$ as $p, n \to \infty$ where $\mu, \tilde{\mu}$ are the unique measures having Stieltjes transforms $m(z)$ and $\tilde{m}(z)$ with

$$m(z) = \frac{1}{c}\tilde{m}(z) + \frac{1-c}{cz}, \quad \tilde{m}(z) = \left(-z + c\int\frac{t\nu(dt)}{1+\tilde{m}(z)t}\right)^{-1}.$$

A few remarks are in order to better understand the statement of the theorem.

Remark 2.6 (On the implicit statement)**.** *As opposed to Theorem 2.3, the statement of the theorem is here implicit in the sense that μ is only defined through $m_\mu(z)$, itself implicitly defined as the solution of an implicit equation. The main reason for the explicit nature of Theorem 2.3 is that Equation* (2.13)*, that through a perturbation approach provides the connection between $m(z)$ and a function of itself, boils down to a quadratic equation in $m(z)$ which can be solved and from which Theorem 2.1 can be applied. Due to the presence of $\mathbf{C}$, in the present situation, the equivalent to* (2.13) *will here maintain an implicit form. This will hold true for almost all generalizations of the Marčenko–Pastur theorem to be introduced in this manuscript.*

Remark 2.7 (Numerical evaluation of $m_\mu(z)$)**.** *Due to its implicit nature, determining $m(z)$ for $z \in \mathbb{C}\setminus\mathbb{R}^+$ requires to solve an implicit equation. Using contraction and analyticity arguments, it can be shown that the standard fixed-point algorithm converges, i.e.,*

$$m(z) = \lim_{\ell\to\infty} m^{(\ell)}(z)$$

for $\tilde{m}^{(0)}(z) = 0$ (say) and, for $\ell \geq 0$, $m^{(\ell)}(z) = \frac{1}{c}\tilde{m}^{(\ell)}(z) + \frac{1-c}{cz}$, $\tilde{m}^{(\ell+1)}(z) = (-z + c\int\frac{t\nu(dt)}{1+\tilde{m}^{(\ell)}(z)t})^{-1}$.

One must be careful here that, since $m(z)$ is not formally defined for $z \in \operatorname{supp}(\mu)$, the above argument does not *hold in this set. In practice, trying to solve for $m(z)$ with $z \in \operatorname{supp}(\mu)$*

numerically leads to a non-converging $m^{(\ell)}(z)$ sequence. This, in passing, can be used to actually determine the support $\mathrm{supp}(\mu)$ as the set of z's for which the above algorithm does not converge.

Numerically, when evaluating $m(z)$ close to the real axis (say for $z = x + \imath\epsilon$, $|\epsilon| \ll 1$), the convergence can appear to be quite slow for $x \in \mathrm{supp}(\mu)$. A convenient workaround is to sequentially evaluate $m(z)$ for all z's of the form $x + \imath\epsilon$, starting from some $z_0 = x_0 + \imath\epsilon$ with $x_0 \notin \mathrm{supp}(\mu)$ until reaching the desired value while systematically starting the fixed-point iterations from the value $m(z)$ obtained for the previous z.

Remark 2.8 (Drawing μ)**.** *As shown in [SC95], the limiting measure μ in Theorem 2.4 admits a density. From the inverse Stieltjes transform formula of Theorem 2.1 and Remark 2.7 above, this can be approximated by solving $m(z)$ for $z \in \imath\epsilon + \mathbb{R}$ for some $\epsilon > 0$ small (say $\epsilon = 10^{-6}$) and retrieving the density at x as $\frac{1}{\pi}\Im[m(x + \imath\epsilon)]$.*

This procedure however only allows for a numerical (rather than theoretical) evaluation of μ and of its support (the latter being approximately the set of x's such that $|\frac{1}{\pi}\Im[m(x + \imath\epsilon)]| \sim \epsilon$).

Figure 2.2 depicts the empirical versus limiting measure of $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ for $\mathbf{C}$ having three distinct and evenly numerous eigenvalues. In this particular setting, the limiting spectrum is akin to Marčenko–Pastur shaped connected components. For sufficiently distinct eigenvalues of $\mathbf{C}$, these components are disjoint while for close eigenvalues they tend to merge.

Remark 2.9 (Deterministic equivalent for $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$)**.** *The convergence result $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \xrightarrow{a.s.} \mu$ in Theorem 2.4 imposes that there exists a limit ν to $\mu_\mathbf{C}$, which may not be practically meaningful. In generalized versions of Theorem 2.4 (see e.g., Theorem 2.5 below), even if the spectral measure of the covariance matrices are to converge, $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ may not even have a limit.*

One may instead consider the deterministic equivalent *μ_p for $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ which is a sequence of probability measures, such that $\mathrm{dist}(\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}, \mu_p) \xrightarrow{a.s.} 0$ for some distance between distributions (for instance, such that $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} - \mu_p \xrightarrow{a.s.} 0$ vaguely).*

Practically speaking, since the data dimension p is in general a fixed quantity and $\mathbf{C}$ a given covariance matrix (rather than specific values in a growing sequence of p's and $\mathbf{C}$'s), one will always consider that the "effective" limiting measure ν coincides with $\mu_\mathbf{C} = \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(\mathbf{C})}$. So for instance, if $p = q$ for some fixed q in the data at hand, one may still theoretically consider that $n, p \to \infty$ and apply Theorem 2.4 with $\nu = \mu_\mathbf{C}$ for $\mathbf{C} \in \mathbb{R}^{q\times q}$, the actual data covariance matrix of the finite-dimensional model.

Sketch of Proof of Theorem 2.4. The proof of Theorem 2.4 generally follows the same line of arguments as that of Theorem 2.3. The main difference is that (2.12) here becomes

$$\frac{1}{n}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}\mathbf{A}\mathbf{Q}\mathbf{x}_i = \frac{\frac{1}{n}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}\mathbf{A}\mathbf{Q}_{-i}\mathbf{x}_i}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} \simeq \frac{\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}\mathbf{A}\mathbf{Q}_{-i}\mathbf{C}}{1 + \frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}\mathbf{C}}$$

where we used the fact that, denoting $\mathbf{x}_i = \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i$ for $\mathbf{z}_i$ the i-th column of $\mathbf{Z}$ having i.i.d. zero mean and unit variance entries, by Lemma 2.11,

$$\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i = \frac{1}{n}\mathbf{z}_i^\mathsf{T}\mathbf{C}^{\frac{1}{2}}\mathbf{Q}_{-i}\mathbf{C}^{\frac{1}{2}}\mathbf{z}_i \simeq \frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}\mathbf{C}.$$

Again with Lemma 2.9 and the fact that $\frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}\mathbf{C} \leq \|\mathbf{C}\|\frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}$, we obtain the approximation

$$\frac{1}{n}\operatorname{tr}(\mathbf{F} + z\mathbf{I}_p)\bar{\mathbf{Q}}\mathbf{A}\mathbf{Q} \simeq \frac{\frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}\mathbf{A}\mathbf{Q}}{1 + \frac{1}{n}\operatorname{tr}\mathbf{Q}\mathbf{C}}$$

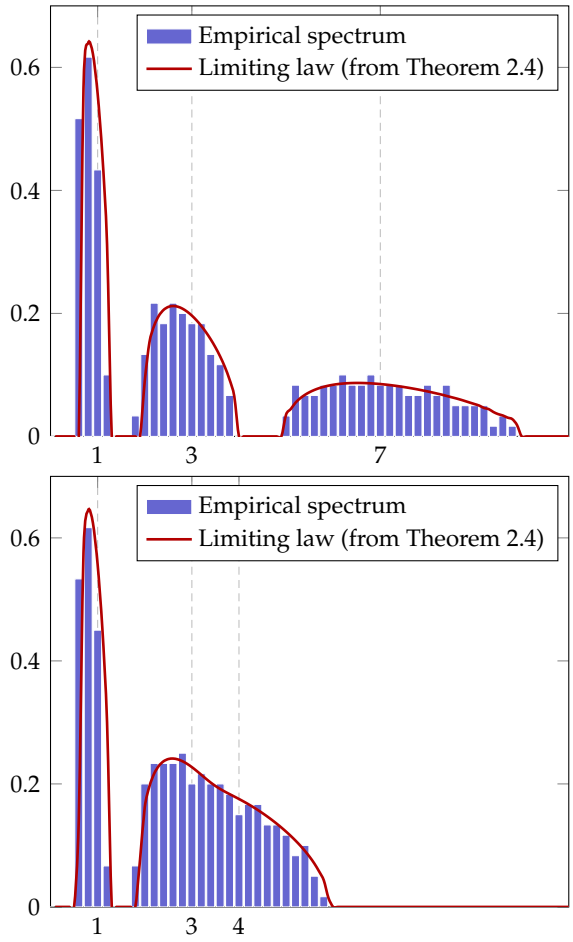


Figure 2.2: Histogram of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, $p = 300$, $n = 3000$, for $\mathbf{C}$ having spectral measure $\mu_\mathbf{C} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_7)$ (top) and $\mu_\mathbf{C} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_4)$ (bottom).

with $\mathbf{F}(z) = \bar{\mathbf{Q}}^{-1}(z)$ the sought-for deterministic equivalent, which then must admit the form

$$\mathbf{F}(z) \simeq \frac{\mathbf{C}}{1 + \frac{1}{n}\operatorname{tr}\mathbf{Q}\mathbf{C}} - z\mathbf{I}_p$$

for the previous approximation to hold. Ultimately, taking $\mathbf{A} = \mathbf{C}$ in $\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q} - \bar{\mathbf{Q}}) \xrightarrow{a.s.} 0$ we deduce

$$\frac{1}{n}\operatorname{tr}\mathbf{C}\mathbf{Q} \simeq \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}} \simeq \frac{1}{n}\operatorname{tr}\mathbf{C}\left(-z\mathbf{I}_p + \frac{\mathbf{C}}{1 + \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}}\right)^{-1} \tag{2.17}$$

or equivalently

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

if we denote $\tilde{m}_p(z) = -\frac{1}{z}\left(1 + \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}(z)\right)^{-1}$, as requested. Note that we implicitly used here the fact that $\|\mathbf{C}\|$ is bounded.

With the deterministic equivalent for $\mathbf{Q}$ in hand, the deterministic equivalent for $\tilde{\mathbf{Q}}$ follows from the direct observation that $\tilde{\mathbf{Q}} = \frac{1}{z}\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{Q}\mathbf{X} - \frac{1}{z}\mathbf{I}_n$ so that

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{Q}}]_{ij} = \frac{1}{z}\frac{1}{n}\mathbb{E}[\mathbf{x}_i^\mathsf{T}\mathbf{Q}\mathbf{x}_j] - \frac{1}{z}\delta_{ij} &= \frac{1}{z}\mathbb{E}\left[\frac{\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_j}{1 + \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right] - \frac{1}{z}\delta_{ij} \\ &\simeq -\frac{1}{z}\left(1 + \frac{1}{n}\operatorname{tr}\mathbf{C}\bar{\mathbf{Q}}\right)^{-1}\delta_{ij} = \tilde{m}_p(z)\delta_{ij}.\end{aligned}$$

□

Kernel methods naturally involve matrices of the type $\mathbf{K} = \{\frac{1}{p}\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\}_{i,j=1}^n = \frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X}$ (inner product kernels) or $\mathbf{K} = \{\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2\}_{i,j=1}^n$ (distance kernels), where the prefactor $1/p$ is necessary under our notation framework to ensure that the spectrum of $\mathbf{K}$ remains of order $O(1)$ as p, n increase. Assuming the vectors $\mathbf{x}_i$ arise from a mixture model, the following generalization of Theorem 2.4 will be of practical relevance.

Theorem 2.5 (Sample covariance of mixture models, from [BGC16]). *Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k] \in \mathbb{R}^{p\times n}$ with $\mathbf{X}_a = [\mathbf{x}_{a1}, \ldots, \mathbf{x}_{an_a}] \in \mathbb{R}^{p\times n_a}$ and $\mathbf{x}_{ai} = \mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_{ai}$ for $\mathbf{z}_{ai}$ a vector with i.i.d. zero mean and unit variance entries. Then, as $n_1, \ldots, n_k, p \to \infty$ in such a way that k is fixed and $n_a/n \to c_a > 0$, $p/n \to c_0 > 0$, letting $\mathbf{Q}(z) = (\frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X} - z\mathbf{I}_n)^{-1}$ and $\tilde{\mathbf{Q}}(z) = (\frac{1}{p}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$, we have*

$$\begin{aligned}\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) &= \operatorname{diag}\{g_a(z)\mathbf{1}_{n_a}\}_{a=1}^k \\ \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) &= -\frac{1}{z}\left(\mathbf{I}_p + \sum_{a=1}^k \frac{c_a}{c_0}g_a(z)\mathbf{C}_a\right)^{-1}\end{aligned}$$

with $(z, g_a(z))$ the unique solutions in $\mathcal{Z}(\mathbb{C}\setminus\mathbb{R}^+)$ of

$$g_a(z) = -\frac{1}{z}(1 + \tilde{g}_a(z))^{-1}, \quad \tilde{g}_a(z) = -\frac{1}{z}\frac{1}{p}\operatorname{tr}\mathbf{C}_a\left(\mathbf{I}_p + \sum_{b=1}^k \frac{c_b}{c_0}g_b(z)\mathbf{C}_b\right)^{-1}.$$

Sketch of proof of Theorem 2.5. Let us begin with the initial guess $\bar{\tilde{\mathbf{Q}}} = \mathbf{F}^{-1}$. By Lemma 2.1,

$$\tilde{\mathbf{Q}} - \bar{\tilde{\mathbf{Q}}} = \tilde{\mathbf{Q}} \left(\mathbf{F} + z\mathbf{I}_p - \frac{1}{p} \sum_{b=1}^{k} \sum_{i=1}^{n_b} \mathbf{x}_{bi} \mathbf{x}_{bi}^\mathsf{T} \right) \bar{\tilde{\mathbf{Q}}}$$

so that, with $\frac{1}{p} \operatorname{tr} \mathbf{A}(\tilde{\mathbf{Q}} - \bar{\tilde{\mathbf{Q}}}) \xrightarrow{a.s.} 0$, it follows

$$\frac{1}{p} \operatorname{tr}(\mathbf{F} + z\mathbf{I}_p) \bar{\tilde{\mathbf{Q}}} \mathbf{A} \tilde{\mathbf{Q}} - \frac{1}{p} \sum_{b=1}^{k} \sum_{i=1}^{n_b} \frac{1}{p} \mathbf{x}_{bi}^\mathsf{T} \bar{\tilde{\mathbf{Q}}} \mathbf{A} \tilde{\mathbf{Q}} \mathbf{x}_{bi} \xrightarrow{a.s.} 0.$$

Applying Lemma 2.8 to remove the dependence in $\tilde{\mathbf{Q}}$ of $\mathbf{x}_{bi}$, together with Lemma 2.9, we deduce

$$\frac{1}{p} \sum_{b=1}^{k} \sum_{i=1}^{n_b} \frac{1}{p} \mathbf{x}_{bi}^\mathsf{T} \bar{\tilde{\mathbf{Q}}} \mathbf{A} \tilde{\mathbf{Q}} \mathbf{x}_{bi} \simeq \sum_{b=1}^{k} \frac{n_b}{p} \frac{\frac{1}{p} \operatorname{tr} \mathbf{C}_b \bar{\tilde{\mathbf{Q}}} \mathbf{A} \tilde{\mathbf{Q}}}{1 + \frac{1}{p} \operatorname{tr} \tilde{\mathbf{Q}} \mathbf{C}_b}$$

so that $\mathbf{F}$ must be written as the following sum over b

$$\mathbf{F} \simeq \sum_{b=1}^{k} \frac{c_b}{c_0} \frac{\mathbf{C}_b}{1 + \frac{1}{p} \operatorname{tr} \tilde{\mathbf{Q}} \mathbf{C}_b} - z\mathbf{I}_p.$$

To eventually close the loop, we take $\mathbf{A} = \mathbf{C}_a$ for $a = 1, \ldots, k$ to establish

$$\begin{aligned} \frac{1}{p} \operatorname{tr} \tilde{\mathbf{C}}_a \mathbf{Q} \simeq \frac{1}{p} \operatorname{tr} \mathbf{C}_a \bar{\tilde{\mathbf{Q}}} \equiv \tilde{g}_a(z) &\simeq \frac{1}{p} \operatorname{tr} \mathbf{C}_a \left(-z\mathbf{I}_p + \sum_{b=1}^{k} \frac{c_b}{c_0} \frac{\mathbf{C}_b}{1 + \frac{1}{p} \operatorname{tr} \bar{\tilde{\mathbf{Q}}} \mathbf{C}_b} \right)^{-1} \\ &\equiv -\frac{1}{z} \frac{1}{p} \operatorname{tr} \mathbf{C}_a \left(\mathbf{I}_p + \sum_{b=1}^{k} \frac{c_b}{c_0} g_b(z) \mathbf{C}_b \right) \end{aligned}$$

where we denote $g_a(z) \equiv -\frac{1}{z} \left(1 + \frac{1}{p} \operatorname{tr} \mathbf{C}_a \bar{\tilde{\mathbf{Q}}} \right)^{-1} = -\frac{1}{z}(1 + \tilde{g}_a(z))^{-1}$, as desired.

To derive the deterministic equivalent for $\mathbf{Q}$ from that for $\tilde{\mathbf{Q}}$, we use again the fact that $\mathbf{Q} = \frac{1}{z} \frac{1}{p} \mathbf{X}^\mathsf{T} \tilde{\mathbf{Q}} \mathbf{X} - \frac{1}{z} \mathbf{I}_n$ and therefore, indexing the set $\{1, \ldots, n\}$ as $\{11, \ldots, 1n_1, \ldots, 1k, \ldots, kn_k\}$,

$$\mathbb{E}[\mathbf{Q}]_{ai,bj} = \frac{1}{z} \frac{1}{p} \mathbb{E}[\mathbf{x}_{ai}^\mathsf{T} \tilde{\mathbf{Q}} \mathbf{x}_{bj}] - \frac{1}{z} \delta_{ai,bj} \simeq -\frac{1}{z} \left(1 + \frac{1}{p} \operatorname{tr} \mathbf{C}_a \bar{\tilde{\mathbf{Q}}} \right)^{-1} \delta_{ai,bj} = g_a(z) \delta_{ai,bj}.$$

□

2.3 Spiked Models

In the last section we discussed the popular sample covariance model, by providing a deterministic equivalent for the associated resolvent, and consequently the associated Stieltjes transform that describes the limiting spectral measure as well as a thorough characterization of linear statistics, quadratic forms and the subspaces of interest. Nonetheless, due to the implicit nature of Theorem 2.4, the aforementioned understanding has only a rather limited practical impact.

In this section, we consider a very special, yet practically far reaching, case of sample covariance matrix models for which the limiting spectral measure coincides with

the Marčenko–Pastur law. Since the Marčenko–Pastur law assumes an explicit well-understood expression (recall Theorem 2.3), the various estimates of interest will be explicit, thus intuitions on their behavior are easily derived.

These special models fundamentally rely on letting the covariance matrix $\mathbf{C}$ be a *low rank* perturbation of the identity matrix $\mathbf{I}_p$, i.e., $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$ for $\mathbf{L} \in \mathbb{R}^{p\times p}$ with $\text{rank}(\mathbf{L}) = k$ fixed with respect to p, n. Such statistical models corresponding to a *low rank update* of a classical random matrix model with well-known behavior are generically called *spiked models*.

2.3.1 Isolated eigenvalues

Let us then consider again the statistical model $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ with $\mathbf{x}_i = \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i$, $\mathbf{z}_i \in \mathbb{R}^p$ with standard i.i.d. entries and where

$$\mathbf{C} = \mathbf{I}_p + \mathbf{L}, \quad \mathbf{L} = \sum_{i=1}^{k} \ell_i \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$$

with k and $\ell_1 \geq \ldots \geq \ell_k > 0$ fixed with respect to n, p and $\|\mathbf{u}_i\| = 1$ for all i.

According to Theorem 2.4, $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ has a limiting measure μ defined through the limiting measure ν of $\mu_\mathbf{C}$. But obviously $\nu = \delta_1$ here since

$$\mu_\mathbf{C} = \frac{p-k}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^{k} \delta_{1+\ell_i} \to \delta_1.$$

As a consequence, while $\mathbf{C}$ is not the identity matrix, the limiting measure μ is the Marčenko–Pastur law introduced in Theorem 2.3. In this case, it is natural to ask, if it is possible to (constantly) observe eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ "jumping out" the support of ν, and if yes, can we further "localize" these isolated eigenvalues.

We will precisely show here that, depending on the values of ℓ_i and $c = \lim p/n$, the i-th largest eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ may indeed *isolate* from $\text{supp}(\mu)$. As such, since most of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ congregate but possibly for a few ones (up to k of them), the latter isolated eigenvalues are seen as "spikes" in the histogram of eigenvalues.

The specific result, here due to Baik (not Bai) and Silverstein, is as follows

Theorem 2.6 (Spiked models, from [BS06])**.** *Under the setting of Theorem 2.4 with* $\mathbb{E}[\mathbf{Z}_{i,j}^4] < \infty$*, let* $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$ *with* $\mathbf{L} = \sum_{i=1}^{k} \ell_i \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$ *in its spectral decomposition, where* k *and* $\ell_1 \geq \ldots \geq \ell_k > 0$ *are fixed with respect to* p, n*. Then, denoting* $\lambda_1 \geq \ldots \geq \lambda_p$ *the eigenvalues of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$*, as* $p, n \to \infty$*,*

$$\lambda_i \xrightarrow{a.s.} \begin{cases} \rho_i = 1 + \ell_i + c\frac{1+\ell_i}{\ell_i} > (1+\sqrt{c})^2 & , \ \ell_i > \sqrt{c} \\ (1+\sqrt{c})^2 & , \ \ell_i \leq \sqrt{c}. \end{cases}$$

The theorem thus identifies an abrupt change in the behavior of the i-th dominant eigenvalue λ_i of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$: if $\ell_i \leq \sqrt{c}$, λ_i converges to the right-edge $(1+\sqrt{c})^2$ of the support of μ and thus does *not* isolate. However, as soon as $\ell_i > \sqrt{c}$, λ_i converges to a limit *beyond* the right-edge of μ and thus *does isolate*, from the Marčenko–Pastur support.

With a statistical physics interpretation, this phenomenon is often referred to as the *phase transition* of the spiked models.

From a statistical viewpoint, the fact that the i-th eigenvalue λ_i of the sample covariance matrix "macroscopically" exceeds or not the other eigenvalues according to whether $\ell_i > \sqrt{c}$ or $\ell_i \leq \sqrt{c}$ can be interpreted as the fact that the "signal strength" ℓ_i of the structured data exceeds the minimal *detectability* threshold $\sqrt{c}$: this is achieved if the signal strength ℓ_i is strong enough, or alternatively if the number of observed independent data n is large enough (so that $c = \lim p/n$ is small), as the intuition would suggest. Indeed, if $\ell_1 < \sqrt{c}$, the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ are all asymptotically compacted in the support $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ and thus it is theoretically (asymptotically) impossible to tell whether $\mathbf{C} = \mathbf{I}_p$ or $\mathbf{C}$ is more structured (e.g., $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$) from the observation of the spectral measure.

Proof of Theorem 2.6. When it comes to assessing the eigenvalues of a given matrix $\mathbf{A}$, it first comes to mind to solve the determinant equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. This approach is not convenient for $\mathbf{X}$ of increasing dimensions and we have seen that the Stieltjes transform method is an appropriate substitute in this case. Here, since the low rank matrix $\mathbf{L}$ only induces a low rank perturbation of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, the use of Sylverster's identity (Lemma 2.3) will turn the large dimensional determinant equation into a small (fixed) dimensional one, and the method is now valid. This is the approach we pursue here.

Specifically, let us seek for the presence of an eigenvalue λ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ which is asymptotically greater than $(1+\sqrt{c})^2$. Our approach is to "isolate" the low rank contribution due to $\mathbf{L}$ from the "whitened" sample covariance matrix model with identity covariance. To this end, we use the following sequence of equivalences

$$\begin{aligned} 0 &= \det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \lambda\mathbf{I}_p\right) \\ &= \det\left(\frac{1}{n}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}} - \lambda\mathbf{I}_p\right) \\ &= \det\left(\mathbf{I}_p+\mathbf{L}\right)\det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda(\mathbf{I}_p+\mathbf{L})^{-1}\right) \end{aligned}$$

where we recall $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z}$. Obviously, $\det(\mathbf{I}_p+\mathbf{L}) \neq 0$ so that the first determinant can be discarded. For the second term, first recall from the resolvent identity (Lemma 2.1) that

$$(\mathbf{I}_p+\mathbf{L})^{-1} = \mathbf{I}_p - (\mathbf{I}_p+\mathbf{L})^{-1}\mathbf{L}$$

so that we can isolate the (now well-understood) resolvent of the "whitened" model. That is, letting $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda\mathbf{I}_p)^{-1}$, we write

$$\begin{aligned} 0 &= \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda\mathbf{I}_p + \lambda(\mathbf{I}_p+\mathbf{L})^{-1}\mathbf{L}\right) \\ &= \det\mathbf{Q}^{-1}(\lambda)\det\left(\mathbf{I}_p + \lambda\mathbf{Q}(\lambda)(\mathbf{I}_p+\mathbf{L})^{-1}\mathbf{L}\right). \end{aligned}$$

Inverting the matrix $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda\mathbf{I}_p$ is (almost surely) licit for all large n, p as we demanded $\lambda > (1+\sqrt{c})^2$. Now, denoting $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$ with $\mathbf{L} = \mathrm{diag}(\ell_1, \ldots, \ell_k)$ and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k] \in \mathbb{R}^{p\times k}$, we further have

$$(\mathbf{I}_p+\mathbf{L})^{-1}\mathbf{L} = (\mathbf{I}_p+\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T})^{-1}\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T} = \mathbf{U}(\mathbf{I}_k+\mathbf{D})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}.$$

Plugging this expansion into the above equation, this is

$$\begin{aligned} 0 &= \det\mathbf{Q}^{-1}(\lambda)\det\left(\mathbf{I}_p + \lambda\mathbf{Q}(\lambda)\mathbf{U}(\mathbf{I}_k+\mathbf{D})^{-1}\mathbf{D}\mathbf{U}^\mathsf{T}\right) \\ &= \det\mathbf{Q}^{-1}(\lambda)\det\left(\mathbf{I}_k + \lambda\mathbf{U}^\mathsf{T}\mathbf{Q}(\lambda)\mathbf{U}(\mathbf{I}_k+\mathbf{D})^{-1}\mathbf{D}\right) \end{aligned}$$

where in the last equality we applied Sylvester's identity. Since the first determinant does not cancel with $\lambda > (1+\sqrt{c})^2$, we finally have for all large n, p,

$$0 = \det\left(\mathbf{I}_k + \lambda \mathbf{U}^\mathsf{T}\mathbf{Q}(\lambda)\mathbf{U}(\mathbf{I}_k + \mathbf{D})^{-1}\mathbf{D}\right).$$

From Theorem 2.3, we now know that

$$\mathbf{U}^\mathsf{T}\mathbf{Q}(\lambda)\mathbf{U} = m(\lambda)\mathbf{I}_k + o_{\|\cdot\|}(1)$$

almost surely, for $m(z)$ the Stieltjes transform of the Marčenko–Pastur law μ (the term $\mathbf{I}_k$ arises from the fact that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_k$). Consequently, by continuity of the determinant (as a polynomial of its entries), we have

$$0 = \det\left(\mathbf{I}_k + \lambda m(\lambda)(\mathbf{I}_k + \mathbf{D})^{-1}\mathbf{D}\right) + o(1)$$

and thus, if such a λ exists, it must satisfy

$$\lambda m(\lambda) = -\frac{1+\ell_i}{\ell_i} + o(1).$$

for some $i \in \{1, \ldots, k\}$. We thus need to understand when this equation has a solution. To this end, observe that the function $\mathbb{R} \setminus \mathrm{supp}(\mu) \to \mathbb{R}$, $x \mapsto xm(x) = \int \frac{x}{t-x}\mu(dt)$ is increasing on its domain of definition and that $m(x) \to 0$ as $x \to \infty$. Solving the expression of $m(z)$ in Theorem 2.3, i.e.,

$$zcm(z)^2 - (1-c-z)m(z) + 1 = 0 \tag{2.18}$$

we further find that

$$\lim_{x \downarrow (1+\sqrt{c})^2} xm(x) = -\frac{1+\sqrt{c}}{\sqrt{c}}.$$

Thus, $m(x)$ increases from $-\frac{1+\sqrt{c}}{\sqrt{c}}$ to 0 on the interval $((1+\sqrt{c})^2, \infty)$. The equation $\lambda m(\lambda) = -\frac{1+\ell_i}{\ell_i}$ thus only has a solution if and only if

$$-\frac{1+\ell_i}{\ell_i} > -\frac{1+\sqrt{c}}{\sqrt{c}}$$

that is, whenever $\ell_i > \sqrt{c}$. Assuming this holds, we may then use again (2.18) to obtain, after multiplication by λ, that

$$c(\lambda m(\lambda))^2 - (1-c-\lambda)(\lambda m(\lambda)) + \lambda = 0$$

which, after replacement of $\lambda m(\lambda)$ by $-\frac{1+\ell_i}{\ell_i}$ finally gives, as expected, that

$$\lambda \to 1 + \ell_i + c\frac{1+\ell_i}{\ell_i}$$

concluding the proof. □

Figure 2.3 depicts the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ versus the Marčenko–Pastur law, in the scenario where $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$ with $\mathbf{L}$ of rank four, for various ratios p/n. As predicted by Theorem 2.6, the number of visible "spikes" outside the limiting support of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ varies with p/n: as the ratio decreases, less spikes are visible.

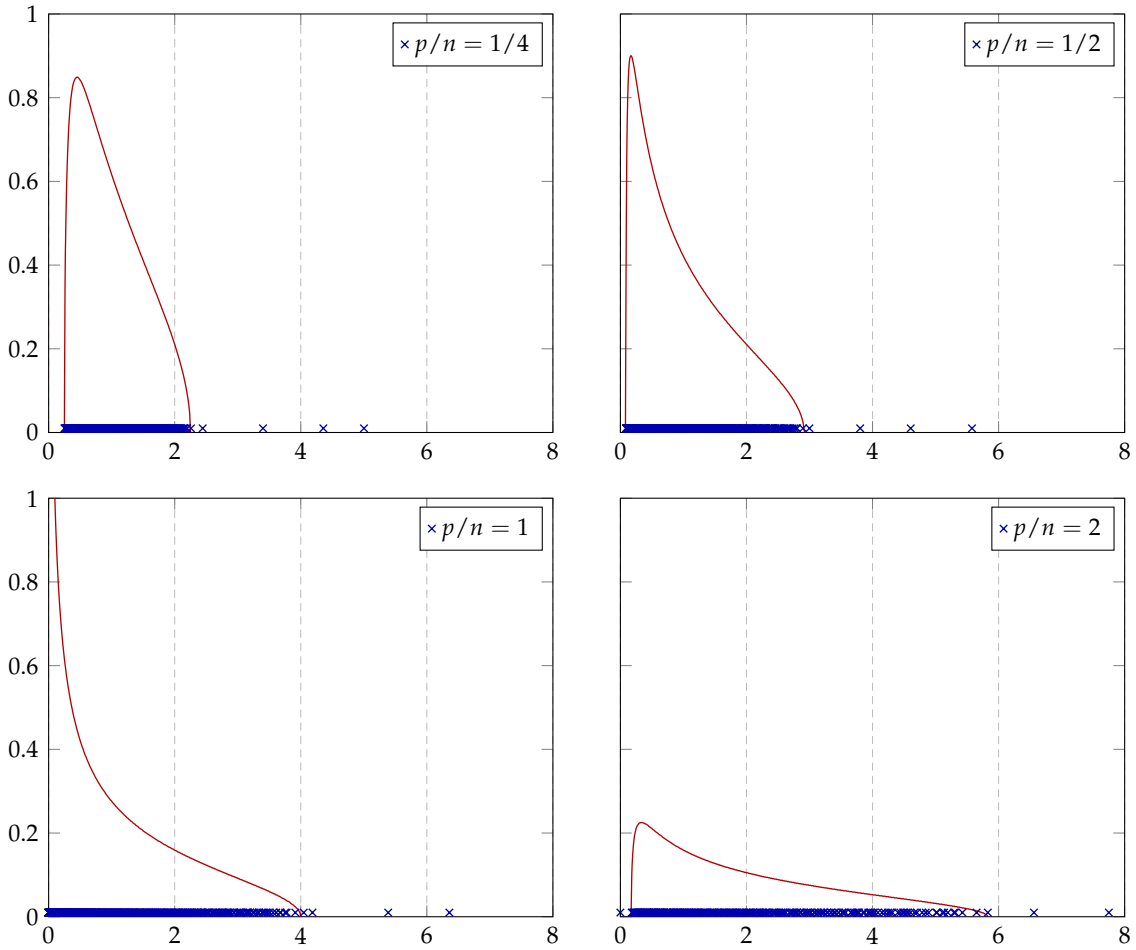


Figure 2.3: Eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ (BLUE crosses) and the Marčenko–Pastur law (red line) for $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\tilde{\mathbf{X}}$, $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$ with $\mu_\mathbf{L} = \frac{p-4}{p}\delta_0 + \frac{1}{p}(\delta_1 + \delta_2 + \delta_3 + \delta_4)$, for $p = 500$ and different values of n.

2.3.2 Isolated eigenvectors

From a practical standpoint, we have seen that the presence of isolated eigenvalues in the spectrum of the sample covariance $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ reveals the presence of some "structure" in the population covariance $\mathbf{C}$ in the sense that $\mathbf{C} \neq \mathbf{I}_p$. We have however also seen that the converse is not true: assuming a spiked model for $\mathbf{C}$, the absence of isolated eigenvalue does not imply $\mathbf{C} = \mathbf{I}_p$.

More interestingly, whether this "structure" is detected or not, one may wonder whether it can be estimated at all. More specifically, if $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$ with $\mathbf{L} = \sum_{i=1}^k \ell_i \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$, are the eigenvectors $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_k$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ associated to the k largest eigenvalues $\lambda_1, \ldots, \lambda_k$ good estimators for $\mathbf{u}_1, \ldots, \mathbf{u}_k$?

Not surprisingly, the answer is here again twofold: i) if $\ell_i \leq \sqrt{c}$ then $\hat{\mathbf{u}}_i$ tends to be orthogonal to $\mathbf{u}_i$, while ii) if $\ell_i > \sqrt{c}$, $\hat{\mathbf{u}}_i$ is to some extent aligned to $\mathbf{u}_i$. The following theorem, originally due to Paul, quantifies this "to some extent".

Theorem 2.7 (Eigenvector alignment, from [Pau07])**.** *Under the setting of Theorem 2.6, let* $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_k$ *be the eigenvectors associated with the largest* k *eigenvalues* $\lambda_1 > \ldots > \lambda_k$ *of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$*. Further assume that* $\ell_1 > \ldots > \ell_k > 0$ *are all distinct. Then, for* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^k$ *unit norm*

deterministic vectors

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} - \mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}\frac{1-c\ell_i^{-2}}{1+c\ell_i^{-1}}1_{\ell_i>\sqrt{c}} \xrightarrow{a.s.} 0.$$

In particular,

$$|\mathbf{u}_i^\mathsf{T}\hat{\mathbf{u}}_i|^2 \xrightarrow{a.s.} \frac{1-c\ell_i^{-2}}{1+c\ell_i^{-1}}1_{\ell_i>\sqrt{c}}.$$

Proof. We may first write that, for all large n, p almost surely,

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} = -\frac{1}{2\pi\imath}\oint_{\Gamma_{\rho_i}} \mathbf{a}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{b}dz$$

for Γ_{ρ_i} a small contour enclosing the almost sure limit ρ_i of the eigenvalue λ_i of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ only.

Isolating $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ from $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ as in the previous proof, we have from Woodbury's identity, Lemma 2.7,

$$\begin{aligned}
&\mathbf{a}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{b} = \mathbf{a}^\mathsf{T}\left(\frac{1}{n}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}} - z\mathbf{I}_p\right)^{-1}\mathbf{b}\\
&= \mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p - z\mathbf{L}\right)^{-1}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b}\\
&= \mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{Q}(z)(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b} + z\mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{Q}(z)\mathbf{U}\mathbf{L}(\mathbf{I}_k - z\mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{U}\mathbf{L})^{-1}\mathbf{U}^\mathsf{T}\mathbf{Q}(z)(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b}\\
&= \mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{Q}(z)(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b} + zm(z)^2\mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{L}(\mathbf{I}_k - zm(z)\mathbf{L})^{-1}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b} + o(1)
\end{aligned}$$

where $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ and for the last equality we used $\mathbf{U}^\mathsf{T}\mathbf{Q}(z)\mathbf{U} = m(z)\mathbf{I}_k + o(1)$ as per Theorem 2.3. The complex integration of $\mathbf{Q}(z)$ on the contour Γ_{ρ_i} only brings a positive residue for the second right-hand side term owing to the inverse $(\mathbf{I}_k - zm(z)\mathbf{L})^{-1}$ which is singular for $z = \rho_i$. We thus finally have

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} = -\frac{1}{2\pi\imath}\oint_{\Gamma_{\rho_i}} zm(z)^2\mathbf{a}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{L}(\mathbf{I}_k - zm(z)\mathbf{L})^{-1}(\mathbf{I}_p+\mathbf{L})^{-\frac{1}{2}}\mathbf{b}dz + o(1).$$

Since $\lim_{z\to\rho_i}(z-\rho_i)(\mathbf{I}_k - zm(z)\mathbf{L})^{-1} = -(\rho_i m'(\rho_i) + m(\rho_i))^{-1}\ell_i^{-1}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}$ with $m(\rho_i) = -1/(c+\ell_i)$ and $m'(\rho_i) = \ell_i^2(c+\ell_i)^{-2}(-c+\ell_i)^{-1}$ (these are obtained from the elements of proof of Theorem 2.6 and from (2.13)), we get that the residue associated to $(\mathbf{I}_k - zm(z)\mathbf{L})^{-1}$ is

$$(-\rho_i m'(\rho_i) - m(\rho_i))^{-1}\ell_i^{-1}\mathbf{u}_i\mathbf{u}_i^\mathsf{T} = (\ell_i^2 - c)\ell_i^{-1}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}.$$

Thus, we finally get after elementary algebra,

$$\mathbf{a}^\mathsf{T}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^\mathsf{T}\mathbf{b} = -\mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}\frac{\rho_i m^2(\rho_i)\ell_i}{1+\ell_i}(\ell_i^2 - c)\ell_i^{-1} = \mathbf{a}^\mathsf{T}\mathbf{u}_i\mathbf{u}_i^\mathsf{T}\mathbf{b}\frac{1-c\ell_i^{-2}}{1+c\ell_i^{-1}}.$$

□

Figure 2.4 compares, in a single-spike scenario, the theoretical limit of $|\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1|^2$ versus its empirical value for different ℓ_1 and different p, n with constant ratio p/n. It is important to note that the theoretical *asymptotic* phase transition phenomenon at $\ell_1 = \sqrt{c}$ corresponds to a sharp non-differentiable change in the function $\ell_1 \mapsto |\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1|^2$; on real data, this sharp transition is only observed for extremely large values of n, p. This in particular means that, in practice, residual information is present below the phase transition.

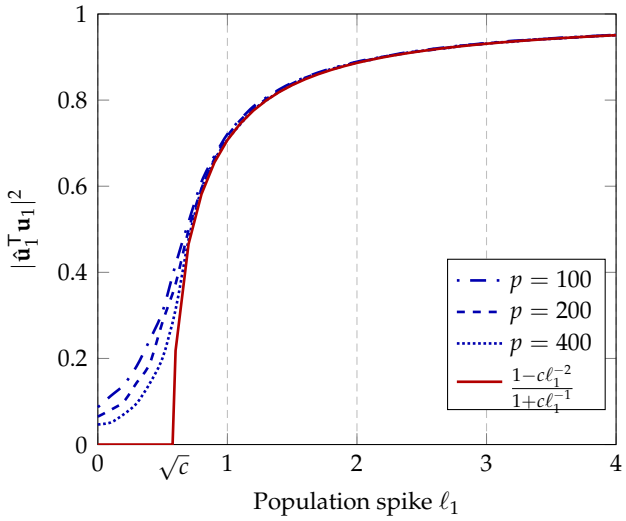


Figure 2.4: Simulated versus limiting $|\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1|^2$ for $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z}$, $\mathbf{C} = \mathbf{I}_p + \ell_1\mathbf{u}_1\mathbf{u}_1^\mathsf{T}$, $p/n = 1/3$, for varying ℓ_1.

2.3.3 Further discussions and other spiked models

As briefly discussed above, the "spiked model" terminology goes beyond sample covariance matrix models with $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$, and $\mathbf{L}$ a low rank matrix. In the literature, spiked models loosely are referred to as "low rank perturbative" models in the following sense: there exists an underlying random matrix model $\mathbf{X}$, the spectral measure of which converges to a well-defined measure with compact support *and* having eigenvalues converging to the support (i.e., no single eigenvalue isolates) which is modified in some way by a low rank matrix $\mathbf{L}$; the resulting matrix has the same limiting spectral measure as that of $\mathbf{X}$ but with possibly some spurious eigenvalues.

Baik and Silverstein in [BS06] were the first to study spiked models, but their approach relied on applying the results on sample covariance matrix models (Theorem 2.4) to the specific case where $\mathbf{C} = \mathbf{I}_p + \mathbf{L}$. This approach requires to have a full understanding of a "more complex" statistical model before particularizing it to a low rank perturbation. The proof of Theorem 2.6 that we proposed follows a second wave of advances in spiked models, mostly triggered by the work of Benaych and Rao [BGN12] (with a free probability approach), which is rather based on relating the perturbation matrix model to the underlying simple (non perturbed) matrix.

Among the popular spiked models, we have the following cases:

- the *information-plus-noise* model of the type

$$\frac{1}{n}(\mathbf{X} + \mathbf{L})(\mathbf{X} + \mathbf{L})^\mathsf{T}$$

  with $\mathbf{X} \in \mathbb{R}^{p\times n}$ having i.i.d. standard entries (zero mean, unit variance and finite fourth order moment) and $\mathbf{L} \in \mathbb{R}^{p\times n}$ deterministic (or at least independent of $\mathbf{X}$) of fixed rank k.

- the additive model of the type

$$\mathbf{Y} + \mathbf{L}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times n}$ is either of the type $\mathbf{Y} = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$, $\mathbf{X} \in \mathbb{R}^{p\times n}$ with standard i.i.d. entries, of $\mathbf{Y} = \frac{1}{\sqrt{n}}\mathbf{X}$ with $\mathbf{X}$ symmetric with standard i.i.d. entries above and on the diagonal and $\mathbf{L} \in \mathbb{R}^{n\times n}$ deterministic.

Each of these models has its own phase transition threshold (i.e., the value that eigenvalues of $\mathbf{L}$ must exceed for a spike to be observed), dominant eigenvalue limits, and eigenvector projections. These can all be determined with the aforementioned proof approach.

However, we will see later that, in practice, we will be confronted with more general forms of low rank perturbation models that do not fit this conventional "random matrix $\mathbf{X}$ and deterministic perturbation $\mathbf{L}$" assumption.

In particular, $\mathbf{L}$ will often be a (possibly elaborate) function of $\mathbf{X}$. Also, $\mathbf{X}$ itself, which will often stand for the "noisy" part of the data model (while $\mathbf{L}$ will in general comprises both the relevant information and some extra noise), may induce its own isolated eigenvalues. For instance, we shall see in Remark 3.7 in Section 3.1.2 that, depending on the ratios p/n and $\operatorname{tr}\mathbf{C}^4/(\operatorname{tr}\mathbf{C}^2)^2$, the random matrix $\{[\mathbf{X}\mathbf{X}^\mathsf{T}]_{ij}^2\}_{i\neq j}^p$ where $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z}$ and $\mathbf{Z}$ with i.i.d. standard entries, may have two isolated eigenvalues although all the eigenvalues of $\mathbf{C}$ remain in their limiting support.

Yet, despite these technical differences, the proof approaches of Theorem 2.6 and Theorem 2.7 remain essentially valid. We thus propose here to generalize the notion of "spiked models" to models of the type $\mathbf{X} + \mathbf{L}$ where $\mathbf{X}$ is some reference, well understood, random matrix model (possibly inducing its own spikes) and $\mathbf{L}$ is a low rank matrix, possibly depending on $\mathbf{X}$.

With this definition, the aforementioned *sample covariance*, *information-plus-noise* and *additive* models are in fact all equivalent to an additive model. Indeed, we may write

$$\frac{1}{n}(\mathbf{X}+\mathbf{L})(\mathbf{X}+\mathbf{L})^\mathsf{T} = \mathbf{Y} + \mathbf{L}'$$
$$\mathbf{Y} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}, \quad \mathbf{L}' = \frac{1}{n}\mathbf{X}\mathbf{L}^\mathsf{T} + \frac{1}{n}\mathbf{L}\mathbf{X}^\mathsf{T} + \frac{1}{n}\mathbf{L}\mathbf{L}^\mathsf{T}$$

and

$$\frac{1}{n}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}}\mathbf{X}\mathbf{X}^\mathsf{T}(\mathbf{I}_p+\mathbf{L})^{\frac{1}{2}} = \mathbf{Y} + \mathbf{L}'$$
$$\mathbf{Y} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}, \quad \mathbf{L}' = \frac{1}{n}\mathbf{X}\mathbf{L}''^\mathsf{T} + \frac{1}{n}\mathbf{L}''\mathbf{X} + \frac{1}{n}\mathbf{L}''\mathbf{L}''^\mathsf{T}$$

where in the second equation, letting $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$, we denoted $\mathbf{L}'' = \mathbf{U}((\mathbf{I}_k+\mathbf{D})^{\frac{1}{2}} - \mathbf{I}_k)\mathbf{U}^\mathsf{T}$. In the remainder of the manuscript, we shall systematically exploit this generic modeling approach to spiked models.

Chapter 3

Spectral Behavior of Large Kernels Matrices and Neural Nets

3.1 Random Kernel Matrices

3.1.1 Kernel ridge regression

Kernel ridge regression, or least squares support vector machine (LS-SVM), as mentioned in Section 1.2.1, is a modification of the standard SVM [SV99] to overcome the drawbacks of SVM related to computational efficiency. In this subsection, we will focus on a two-class GMM (see Definition 1) classification using LS-SVM as described below.

Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ of size n, where data $\mathbf{x}_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, 1\}$, the objective of LS-SVM is to devise a decision function $h(\mathbf{x})$ that ideally maps all $\mathbf{x}_i$ in the training set to y_i and subsequently all unknown data $\mathbf{x}$ to their corresponding y value. Here we denote $\mathbf{x}_i \in \mathcal{C}_1$ if $y_i = -1$ and $\mathbf{x}_i \in \mathcal{C}_2$ if $y_i = 1$ and shall say that $\mathbf{x}_i$ belongs to class $\mathcal{C}_1$ or class $\mathcal{C}_2$, respectively. Due to the often nonlinear separability of the data in their input space $\mathbb{R}^p$, in most cases, one associates the training data $\mathbf{x}_i$ to some feature space $\mathcal{H}$ through a nonlinear mapping $\phi : \mathbf{x}_i \mapsto \phi(\mathbf{x}_i) \in \mathcal{H}$. Optimization methods are then used to define a separating hyperplane in $\mathcal{H}$ with direction vector $\boldsymbol{\alpha}$ and a function $h(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\phi(\mathbf{x}) + b$ that minimizes the training errors $e_i = y_i - (\boldsymbol{\alpha}^\mathsf{T}\phi(\mathbf{x}_i) + b)$ that yields good generalization performance by minimizing the norm of $\boldsymbol{\alpha}$ [SS04]. More specifically, the LS-SVM approach consists in minimizing the squared errors e_i^2, thus resulting in

$$\begin{aligned} \min_{\boldsymbol{\alpha}, b} \quad & L(\boldsymbol{\alpha}, e) = \|\boldsymbol{\alpha}\|^2 + \frac{\lambda}{n}\sum_{i=1}^n e_i^2 \\ \text{s.t.} \quad & y_i = \boldsymbol{\alpha}^\mathsf{T}\phi(\mathbf{x}_i) + b + e_i,\ i = 1, \ldots, n \end{aligned} \tag{3.1}$$

where $\lambda > 0$ is a penalty factor that weights the structural risk $\|\boldsymbol{\alpha}\|^2$ against the empirical one $\frac{1}{n}\sum_{i=1}^n e_i^2$.

The problem can be solved by introducing Lagrange multipliers $\beta_i, i = 1, \ldots, n$ with solution $\boldsymbol{\alpha} = \sum_{i=1}^n \beta_i \phi(\mathbf{x}_i)$, where, letting $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}, \boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]^\mathsf{T}$, we obtain

$$\begin{cases} \boldsymbol{\beta} &= \mathbf{Q}\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n}\right)\mathbf{y} = \mathbf{Q}\left(\mathbf{y} - b\mathbf{1}_n\right) \\ b &= \frac{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{y}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n} \end{cases} \tag{3.2}$$

with $\mathbf{Q} = \left(\mathbf{K} + \frac{n}{\lambda}\mathbf{I}_n\right)^{-1}$ and $\mathbf{K} \equiv \left\{\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j)\right\}_{i,j=1}^n$ referred to as the kernel matrix [SV99].

Given $\boldsymbol{\alpha}$ and b, a new datum $\mathbf{x}$ is then classified into class $\mathcal{C}_1$ or $\mathcal{C}_2$ depending on the value of the decision function

$$g(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b \tag{3.3}$$

where $\mathbf{k}(\mathbf{x}) = \left\{\phi(\mathbf{x})^\mathsf{T}\phi(\mathbf{x}_j)\right\}_{j=1}^n \in \mathbb{R}^n$. More precisely, $\mathbf{x}$ is associated to class $\mathcal{C}_1$ if $g(\mathbf{x})$ takes a small value (below a certain threshold ξ) and to class $\mathcal{C}_2$ otherwise.[1]

With the "kernel trick" [SS02], as shown in (3.2) and (3.3), both in the training and testing steps, one only needs to evaluate the inner product $\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j)$ or $\phi(\mathbf{x})^\mathsf{T}\phi(\mathbf{x}_j)$, and never needs to know explicitly the mapping $\phi(\cdot)$. We assume that the kernel is *shift-invariant* and focus on kernel functions $f : \mathbb{R} \to \mathbb{R}$ that satisfy $\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ and shall redefine $\mathbf{K}$ and $\mathbf{k}(\mathbf{x})$ for data point $\mathbf{x}$ as

$$\mathbf{K} = \left\{f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)\right\}_{i,j=1}^n, \quad \mathbf{k}(\mathbf{x}) = \left\{f\left(\|\mathbf{x} - \mathbf{x}_j\|^2/p\right)\right\}_{j=1}^n \tag{3.4}$$

We will focus on the performance of LS-SVM, in the large n, p regime, by studying the asymptotic behavior of the decision function $g(\mathbf{x})$ defined in (3.3). We assume that all $\mathbf{x}_i$'s are extracted from a two-class GMM (as in Definition 1), which allows for a thorough theoretical analysis. For notational convenience, we rewrite the GMM in Definition 1 as follows:

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a) \Leftrightarrow \mathbf{x}_i = \boldsymbol{\mu}_a + \sqrt{p}\boldsymbol{\omega}_i \tag{3.5}$$

such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_a/p)$ for $a \in \{1, 2\}$. In particular, the non-trivial classification condition in Assumption 2 can be adapted to the binary classification setting as follows.

Assumption 4 (Non-trivial binary classification)**.** *As $n \to \infty$, we have for $a \in \{1, 2\}$ that*

1. *$p/n = c \to \bar{c} \in (0, \infty)$ and $n_a/n = c_a \to \bar{c}_a \in (0, 1)$;*
2. *$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = O(1)$ and $\max\{\|\mathbf{C}_a\|, \|\mathbf{C}_a^{-1}\|\} = O(1)$ with $|\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)| = O(\sqrt{p})$, $\|\mathbf{C}_1 - \mathbf{C}_2\|_F^2 = O(p)$;*
3. *for $\mathbf{C}^\circ = \frac{n_i}{n}\mathbf{C}_i$, $\frac{2}{p}\operatorname{tr}\mathbf{C}^\circ \to \tau > 0$ as $n, p \to \infty$.*

A key observation, also made in [CBG16], is that, as a consequence of Assumption 4, for all pairs $i \neq j$,

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \to \tau \tag{3.6}$$

almost surely as $n, p \to \infty$ and the convergence is even uniform across all $i \neq j$. This remark is the crux of all subsequent results and can be seen as a manifestation of the "curse of dimensionality" with respect to the Euclidean distance in high-dimensional space, as discussed in Section 1.1.2.

Our objective here is to assess the performance of LS-SVM, for all kernel function f that is three-times differentiable in a neighborhood of τ, under the setting of Assumptions 4, by studying the asymptotic behavior of the decision function $g(\mathbf{x})$ defined in (3.3). Following [EK10b, CBG16] and our introductory demonstrations performed in Section 1.1.2, under our basic settings, the convergence in (3.6) makes it possible to linearize the kernel matrix $\mathbf{K}$ around the matrix $f(\tau)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, and thus the intractable nonlinear kernel matrix $\mathbf{K}$ can be asymptotically linearized in the large n, p regime. As such, since the decision function $g(\mathbf{x})$ is explicitly defined as a function of $\mathbf{K}$ (through $\boldsymbol{\alpha}$ and b as defined in (3.2)), one can work out an asymptotic linearization of $g(\mathbf{x})$ as a function of the kernel

[1] Since data from $\mathcal{C}_1$ are labeled -1 while data from $\mathcal{C}_2$ are labeled 1.

function f and the statistics of the data. This analysis, presented in detail in Section A.1.2 of the Appendix, allows one to reveal the relationship between the performance of LS-SVM and the kernel function f as well as the given learning task, for Gaussian input data as $n, p \to \infty$.

Mains results. Before going into our main results, a few notations need to be introduced. In the remainder of this section, we shall use the following deterministic and random elements notations:

$$\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T} \in \mathbb{R}^{n\times n}, \quad \boldsymbol{\Omega} \equiv [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n] \in \mathbb{R}^{p\times n}$$
$$\boldsymbol{\psi} \equiv \left\{ \|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_i\|^2\right] \right\}_{i=1}^n \in \mathbb{R}^n.$$

Under Assumptions 4, following up [CBG16] and Section 1.1.2, for three-times differentiable f, one can approximate the kernel matrix $\mathbf{K}$ by $\tilde{\mathbf{K}}$ in such a way that

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{a.s.} 0, \quad \tilde{\mathbf{K}} = -2f'(\tau)(\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P} + \mathbf{A}) + \left(f(0) - f(\tau) + \tau f'(\tau)\right)\mathbf{I}_n$$

where $\tilde{\mathbf{K}}$ follows a spiked model and consists of i) $\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P}$, a standard Gram/covariance-like random matrix model (of operator norm $O(1)$) and ii) $\mathbf{A}$ a small rank (at most eight here) matrix that is of operator norm $O(n)$, which depends both on $\mathbf{P}, \boldsymbol{\Omega}, \boldsymbol{\psi}$ and on the class statistics $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\mathbf{C}_1, \mathbf{C}_2$. The same analysis is applied to the vector $\mathbf{k}(\mathbf{x})$ by similarly defining the following random variables for a new datum $\mathbf{x} \in \mathcal{C}_a$, $a \in \{1, 2\}$:

$$\boldsymbol{\omega}_\mathbf{x} \equiv (\mathbf{x} - \boldsymbol{\mu}_a)/\sqrt{p} \in \mathbb{R}^p, \quad \psi_\mathbf{x} \equiv \|\boldsymbol{\omega}_\mathbf{x}\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_\mathbf{x}\|^2\right] \in \mathbb{R}.$$

Based on the (operator norm) approximation $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$, a Taylor expansion is then performed on $\mathbf{Q} = (\mathbf{K} + \frac{n}{\lambda}\mathbf{I})^{-1}$ to obtain an asymptotic approximation of $\mathbf{Q}$, and subsequently on $\boldsymbol{\beta}$ and b which depend explicitly on $\mathbf{Q}$. At last, plugging these results into (3.3), one obtains the following theorem that characterizes the asymptotic behavior of the decision function $g(\mathbf{x})$.

Theorem 3.1 (Asymptotic behavior of $g(\mathbf{x})$)**.** *Let Assumption 4 hold and $g(\mathbf{x})$ be defined by* (3.3)*. Then, for a kernel function f that is three-times differentiable in a neighborhood of τ, we have, as $n, p \to \infty$,*

$$n\left(g(\mathbf{x}) - \tilde{g}(\mathbf{x})\right) \xrightarrow{a.s.} 0$$

where

$$\tilde{g}(\mathbf{x}) = \begin{cases} c_2 - c_1 + \lambda\left(\mathfrak{R} - 2c_1c_2^2\mathfrak{D}\right), & \text{if } \mathbf{x} \in \mathcal{C}_1 \\ c_2 - c_1 + \lambda\left(\mathfrak{R} + 2c_1^2c_2\mathfrak{D}\right), & \text{if } \mathbf{x} \in \mathcal{C}_2 \end{cases} \tag{3.7}$$

with

$$\mathfrak{R} = -\frac{2f'(\tau)}{n}\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\omega}_\mathbf{x} - \frac{4c_1c_2f'(\tau)}{\sqrt{p}}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^\mathsf{T}\boldsymbol{\omega}_\mathbf{x} + 2c_1c_2f''(\tau)\frac{1}{p}\operatorname{tr}(\mathbf{C}_2 - \mathbf{C}_1)\psi_\mathbf{x} \tag{3.8}$$

$$\mathfrak{D} = -\frac{2f'(\tau)}{p}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 + \frac{f''(\tau)}{p^2}\operatorname{tr}^2(\mathbf{C}_2 - \mathbf{C}_1) + \frac{2f''(\tau)}{p^2}\|\mathbf{C}_2 - \mathbf{C}_1\|_F^2. \tag{3.9}$$

Leaving the proof to Section A.1.2 in the Appendix, Theorem 3.1 tells us that the decision function $g(\mathbf{x})$ has an asymptotic equivalent $\tilde{g}(\mathbf{x})$ that consists of three parts:

1. the deterministic term $c_2 - c_1$ of order $O(1)$ that depends on the number of instances in each class of the training set, which essentially comes from the term $\mathbf{1}_n^\mathsf{T}\mathbf{y}/n$ in b;

2. the "noisy" term $\mathfrak{R}$ of order $O(p^{-1})$ which is a function of the zero mean random variables $\mathbf{\Omega}$, $\boldsymbol{\omega}_{\mathbf{x}}$ and $\psi_{\mathbf{x}}$, thus in particular $\mathbb{E}[\mathfrak{R}] = 0$;

3. the "informative" term containing $\mathfrak{D}$, also of order $O(p^{-1})$, which features the deterministic differences between the two classes.

From Theorem 3.1, under the settings of Assumption 4, for Gaussian data $\mathbf{x} \in \mathcal{C}_a$, $a \in \{1, 2\}$, we can show that $\tilde{g}(\mathbf{x})$ (and therefore $g(\mathbf{x})$) converges to a random Gaussian variable the mean and variance of which are given in the following theorem. The proof is deferred to Section A.1.3 in the Appendix.

Theorem 3.2 (Asymptotic Gaussian behavior)**.** *Under the setting of Theorem 3.1, we have, as* $n, p \to \infty$,

$$n(g(\mathbf{x}) - G_a) \to 0$$

in distribution, where

$$G_a \sim \mathcal{N}(\bar{G}_a, V_{G_a})$$

with

$$\bar{G}_a = \begin{cases} c_2 - c_1 - 2c_2 \cdot c_1 c_2 \lambda \mathfrak{D}, & a = 1 \\ c_2 - c_1 + 2c_1 \cdot c_1 c_2 \lambda \mathfrak{D}, & a = 2 \end{cases}$$
$$V_{G_a} = 8\lambda^2 c_1^2 c_2^2 \left(\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a\right)$$

and

$$\mathcal{V}_1^a = \frac{\left(f''(\tau)\right)^2}{p^4} \operatorname{tr}^2(\mathbf{C}_2 - \mathbf{C}_1) \|\mathbf{C}_a\|_F^2$$
$$\mathcal{V}_2^a = \frac{2\left(f'(\tau)\right)^2}{p^2} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\mathsf{T} \mathbf{C}_a (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$\mathcal{V}_3^a = \frac{2\left(f'(\tau)\right)^2}{np^2} \left(\frac{\operatorname{tr} \mathbf{C}_1 \mathbf{C}_a}{c_1} + \frac{\operatorname{tr} \mathbf{C}_2 \mathbf{C}_a}{c_2} \right).$$

Theorem 3.2 is our main practical result as it allows one to evaluate the large n, p performance of LS-SVM for Gaussian data. Focusing on the implications of Theorem 3.1–3.2, several remarks and discussions are in order.

Remark 3.1 (Dominant bias)**.** *From Theorem 3.1, under the key Assumption 4, both the random noise* $\mathfrak{R}$ *and the deterministic "informative" term* $\mathfrak{D}$ *are of order* $O(n^{-1})$*, which means that* $g(\mathbf{x}) = c_2 - c_1 + O(n^{-1})$*. This result somehow contradicts the classical decision criterion proposed in [SV99], based on the sign of* $g(\mathbf{x})$*, i.e.,* $\mathbf{x}$ *is associated to class* $\mathcal{C}_1$ *if* $g(\mathbf{x}) < 0$ *and to class* $\mathcal{C}_2$ *otherwise. When* $c_1 \neq c_2$*, this would lead to an asymptotic classification of all new data* $\mathbf{x}$*'s in the same class as* $n \to \infty$*. Practically speaking, this means for* n, p *large that the decision function* $g(\mathbf{x})$ *of a new datum* $\mathbf{x}$ *lies (sufficiently) away from* 0 *(*0 *being the classically considered threshold), so that the sign of* $g(\mathbf{x})$ *is constantly positive (in the case of* $\bar{c}_2 > \bar{c}_1$*) or negative (in the case of* $\bar{c}_2 < \bar{c}_1$*). As such, all new data will be trivially classified into the same class. Instead, a first result of Theorem 3.1 is that the decision threshold* ξ *should be taken as* $\xi = \xi_n = c_2 - c_1 + O(n^{-1})$ *for imbalanced classification problems.*

The conclusion of Remark 3.1 was in fact already known since the work of [GSL+02] who reached the same conclusion through a Bayesian inference analysis, *for all finite* n, p. From their Bayesian perspective, the term $c_2 - c_1$ appears in the "bias term" b under the

form of prior class probabilities $P(y = -1)$, $P(y = 1)$ and allows for adjusting classification problems with different prior class probabilities in the training and test sets. This idea of a (static) bias term correction has also been applied in [EPPP00] in order to improve the validation set performance. We visually confirm the problem of imbalanced datasets in Remark 3.1 by Figure 3.1 with $c_1 = 1/4$ and $c_2 = 3/4$, where the histograms of $g(\mathbf{x})$ for $\mathbf{x} \in \mathcal{C}_1$ and $\mathcal{C}_2$ center somewhere close to $c_2 - c_1 = 0.5$, thus resulting in a trivial classification by assigning all new data to $\mathcal{C}_2$ if one takes $\xi = 0$ because $P(g(\mathbf{x}) < \xi \mid \mathbf{x} \in \mathcal{C}_1) \to 0$ and $P(g(\mathbf{x}) > \xi \mid \mathbf{x} \in \mathcal{C}_2) \to 1$ as $n, p \to \infty$ (the convergence being in fact an equality for finite n, p in this particular figure).

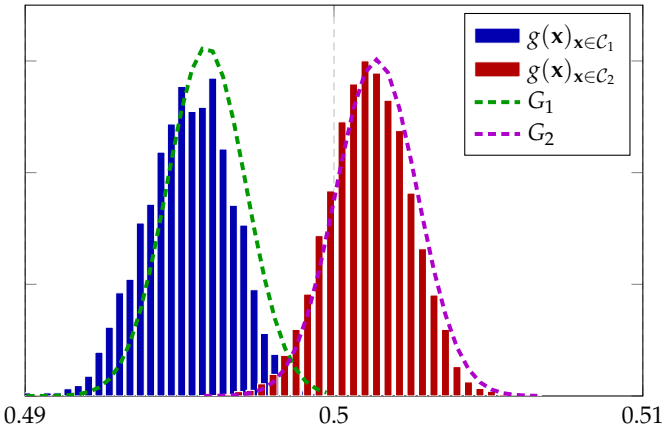


Figure 3.1: Gaussian approximation of $g(\mathbf{x})$ $n = 256, p = 512, c_1 = 1/4, c_2 = 3/4, \lambda = 1$, Gaussian kernel with $\sigma^2 = 1$, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$ with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 3; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = 0.4^{|i-j|}(1 + 5/\sqrt{p})$.

An alternative to alleviate this imbalance issue is to normalize the label vector $\mathbf{y}$. From the proof of Theorem 3.1 in Section A.1.2 we see the term $c_2 - c_1$ is due to the fact that in b one has $\mathbf{1}_n^\mathsf{T}\mathbf{y}/n = c_2 - c_1 \neq 0$. Thus, one may normalize the labels y_i as $y_i^* = -1/c_1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $y_i^* = 1/c_2$ if $\mathbf{x}_i \in \mathcal{C}_2$, so that $\mathbf{1}_n^\mathsf{T}\mathbf{y}^* = 0$. This formulation is also referred to as the *Fisher's targets*: $\{-n/n_1, n/n_2\}$ in the context of kernel Fisher discriminant analysis [BA00, MRW+99]. With these normalized labels $\mathbf{y}^*$, we have the following lemma that reveals the connection between the corresponding decision function $g^*(\mathbf{x})$ and $g(\mathbf{x})$.

Lemma 3.1. *Let $g(\mathbf{x})$ be defined by (3.3) and $g^*(\mathbf{x})$ be defined as $g^*(\mathbf{x}) = (\boldsymbol{\beta}^*)^\mathsf{T}\mathbf{k}(\mathbf{x}) + b^*$, with $(\boldsymbol{\beta}^*, b^*)$ given by (3.2) for $\mathbf{y}^*$ in the place of $\mathbf{y}$, where $y_i^* = -1/c_1$ if $\mathbf{x}_i \in \mathcal{C}_1$ and $y_i^* = 1/c_2$ if $\mathbf{x}_i \in \mathcal{C}_2$. Then,*

$$g(\mathbf{x}) - (c_2 - c_1) = 2c_1c_2 g^*(\mathbf{x}).$$

Proof. From (3.2) and (3.3) we get

$$g(\mathbf{x}) = \mathbf{y}^\mathsf{T}\left(\mathbf{Q} - \frac{\mathbf{Q}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n}\right)\mathbf{k}(\mathbf{x}) + \frac{\mathbf{y}^\mathsf{T}\mathbf{Q}\mathbf{1}_n}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n} = \mathbf{y}^\mathsf{T}\mathbf{v}$$

with $\mathbf{v} = \left(\mathbf{Q} - \frac{\mathbf{Q}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n}\right)\mathbf{k}(\mathbf{x}) + \frac{\mathbf{Q}\mathbf{1}_n}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n}$. Besides, note that $\mathbf{1}_n^\mathsf{T}\mathbf{v} = 1$. We thus have

$$\begin{aligned} g(\mathbf{x}) - (c_2 - c_1) &= \mathbf{y}^\mathsf{T}\mathbf{v} - (c_2 - c_1)\mathbf{1}_n^\mathsf{T}\mathbf{v} \\ &= 2c_1c_2\left(\frac{\mathbf{y} - (c_2 - c_1)\mathbf{1}_n}{2c_1c_2}\right)^\mathsf{T}\mathbf{v} \\ &= 2c_1c_2(\mathbf{y}^*)^\mathsf{T}\mathbf{v} = 2c_1c_2 g^*(\mathbf{x}) \end{aligned}$$

which concludes the proof. □

As a consequence of Lemma 3.1, instead of Theorem 3.2 for standard labels $y_i \in \{-1, 1\}$, one would have the following corollary for the corresponding Gaussian approximation of $g^*(\mathbf{x})$ when normalized labels $y_i^* \in \{-c_1^{-1}, c_2^{-1}\}$ are used.

Corollary 3.1 (Gaussian approximation of $g^*(\mathbf{x})$)**.** *Under the setting of Theorem 3.1, and with $g^*(\mathbf{x})$ defined in Lemma 3.1, $n(g^*(\mathbf{x}) - G_a^*) \to 0$ in distribution as $n, p \to \infty$, where*

$$G_a^* \sim \mathcal{N}(\bar{G}_a^*, V_{G_a^*})$$

with

$$\bar{G}_a^* = \begin{cases} -c_2\lambda\mathfrak{D}, & a = 1 \\ +c_1\lambda\mathfrak{D}, & a = 2 \end{cases}$$
$$V_{G_a^*} = 2\lambda^2\left(\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a\right)$$

and $\mathfrak{D}$ is defined by (3.9)*, $\mathcal{V}_1^a, \mathcal{V}_2^a$ and $\mathcal{V}_3^a$ as in Theorem 3.2.*

Figure 3.2 illustrates this result in the same settings as Figure 3.1. Compared to Figure 3.1, one can observe that both histograms are now centered close to 0 (at distance $O(n^{-1})$ from zero) instead of $c_2 - c_1 = 1/2$. Still, even in the case where normalized labels $\mathbf{y}^*$ are used as observed in Figure 3.2 (where the histograms cross at about $-0.004 \approx 1/n$), taking $\xi = 0$ as a decision threshold may not be an appropriate choice, as $\bar{G}_1^* \neq -\bar{G}_2^*$.

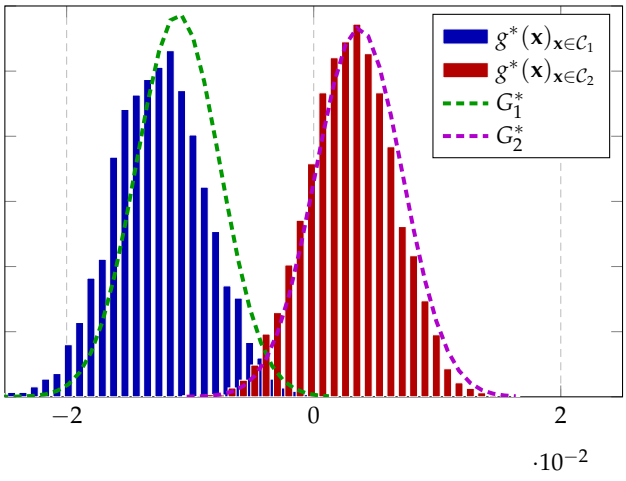


Figure 3.2: Gaussian approximation of $g^*(\mathbf{x})$, under the same setting as Figure 3.1.

Remark 3.2 (Insignificance of λ)**.** *As a direct result of Theorem 3.1 and Remark 3.1, note in* (3.7) *that $\tilde{g}(\mathbf{x}) - (c_2 - c_1)$ is proportional to the hyperparameter λ, which indicates that, rather surprisingly, the tuning of λ is (asymptotically) of no importance when $n, p \to \infty$ since it does not alter the classification statistics when one uses the sign of $g(\mathbf{x}) - (c_2 - c_1)$ for the decision.*

Remark 3.2 is only valid under Assumption 4 and $\lambda = O(1)$, i.e., λ is considered to remain constant as $n, p \to \infty$. Recall that this is in sharp contrast with [CDV07] where $\lambda = O(\sqrt{n})$ (or $O(n)$, depending on the problem) is claimed optimal in the large n only regime. From (1.6), we see here that $\lambda = O(1)$ is rate-optimal under the present large n, p setting; yet we believe that more elaborate kernels (such as those explored in [CS13] and

Section 3.1.2) may allow for improved performances (not in the rate but in the constants), possibly for different scales of λ.

Denote $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-t^2/2\right) dt$ the Q-function with respect to the standard Gaussian distribution. From Theorem 3.2 and Corollary 3.1, we now have the following immediate corollary for the (asymptotic) classification error rate.

Corollary 3.2 (Asymptotic error rate)**.** *Under the setting of Theorem 3.1, for a threshold ξ_n possibly depending on n, as $n, p \to \infty$,*

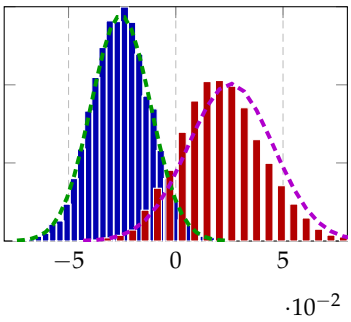
$$P(g(\mathbf{x}) > \xi_n \mid \mathbf{x} \in \mathcal{C}_1) - Q\left(\frac{\xi_n - \bar{G}_1}{\sqrt{V_{G_1}}}\right) \to 0 \tag{3.10}$$

$$P(g(\mathbf{x}) < \xi_n \mid \mathbf{x} \in \mathcal{C}_2) - Q\left(\frac{\bar{G}_2 - \xi_n}{\sqrt{V_{G_2}}}\right) \to 0 \tag{3.11}$$

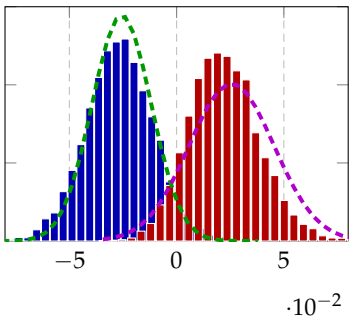
with $\bar{G}_a$ and V_{G_a} given in Theorem 3.2.

Obviously, Corollary 3.2 is only meaningful when $\xi_n = c_2 - c_1 + O(n^{-1})$ as recalled earlier. Besides, it is clear from Lemma 3.1 and Corollary 3.1 that $P(g(\mathbf{x}) > \xi_n \mid \mathbf{x} \in \mathcal{C}_a) = P(2c_1c_2g^*(\mathbf{x}) > \xi_n - (c_2 - c_1) \mid \mathbf{x} \in \mathcal{C}_a)$, so that Corollary 3.2 extends naturally to $g^*(\mathbf{x})$ when normalized labels $\mathbf{y}^*$ are applied.

Corollary 3.2 allows one to compute the asymptotic misclassification rate as a function of $\bar{G}_a$ and V_{G_a} and the threshold ξ_n. Combined with Theorem 3.1, one may note the significance of a proper choice of the kernel function f. For instance, if $f'(\tau) = 0$, the term $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ vanishes from the mean and variance of G_a, meaning that the classification of LS-SVM will not rely (at least asymptotically and under Assumption 4) on the differences in means of the two classes. Figure 3.3 corroborates this finding with the same theoretical Gaussian approximations G_1 and G_2 in subfigures (a) and (b). When $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2$ varies from 0 in (a) to 18 in (b), the distribution of $g(\mathbf{x})$, and in particular, the overlap between two classes, remain almost the same.



(a) $\Delta\mu = 0$



(b) $\Delta\mu = 3$

Figure 3.3: Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 512$, $c_1 = c_2 = 1/2, \lambda = 1$, polynomial kernel with $f(\tau) = 4, f'(\tau) = 0$, and $f''(\tau) = 2$. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = \left[\mathbf{0}_{a-1}; \Delta\mu; \mathbf{0}_{p-a}\right]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 5/\sqrt{p})$.

In terms of the kernel function f, if $f'(\tau) = 0$, the information about the statistical means of the two different classes is lost and will not help perform the classification. Nonetheless, we find that, rather surprisingly, if one further assumes $\operatorname{tr}\mathbf{C}_1 = \operatorname{tr}\mathbf{C}_2 + o(\sqrt{p})$ (which is beyond the minimum "distance" rate in Assumption 4), using a kernel

f that satisfies $f'(\tau) = 0$ results in $V_{G_a} = 0$ while $\bar{G}_a$ may remain non-zero, thereby ensuring a vanishing misclassification rate (as long as $f''(\tau) \neq 0$). Intuitively speaking, the kernels with $f'(\tau) = 0$ play an important role in extracting the covariance "shape" information of both classes, making the classification extremely accurate even in cases that are deemed impossible to classify according to (1.6). This phenomenon was also remarked in [CBG16] and deeply investigated in [CK16]. Figure 3.4 substantiates this finding for $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}$, for which $\operatorname{tr}\mathbf{C}_1 = \operatorname{tr}\mathbf{C}_2 = p$. We observe a rapid drop of the classification error as $f'(\tau)$ gets close to 0.

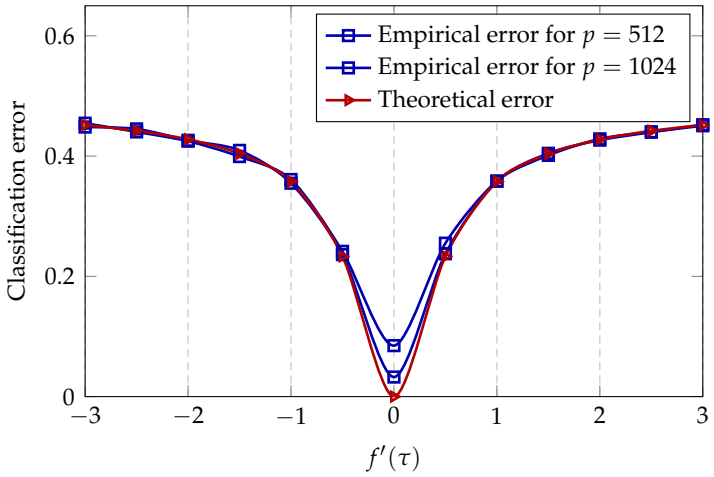


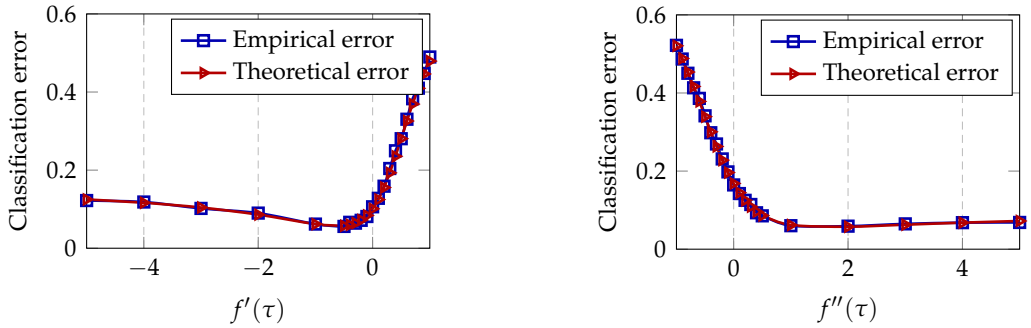
Figure 3.4: Performance of LS-SVM, $c_0 = 1/4$, $c_1 = c_2 = 1/2, \lambda = 1$, polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}$.

Remark 3.3 (Conditions on the kernel function f)**.** *From Theorem 3.2 and Corollary 3.1, one observes that $|\bar{G}_1 - \bar{G}_2|$ is always proportional to the "informative" term $\mathfrak{D}$ and should, for fixed V_{G_a}, be made as large as possible to avoid the overlap of $g(\mathbf{x})$ for $\mathbf{x}$ from different classes. Since V_{G_a} does not depend on the signs of $f'(\tau)$ and $f''(\tau)$, it is easily deduced that, to achieve optimal classification performance, one needs to choose the kernel function f such that $f(\tau) > 0, f'(\tau) < 0$ and $f''(\tau) > 0$.*

Incidentally, the condition in Remark 3.3 is naturally satisfied for the Gaussian kernel $f(x) = \exp\left(-x/(2\sigma^2)\right)$ for any σ, meaning that, even without specific tuning of the kernel parameter σ through cross validation or other techniques, LS-SVM is expected to perform rather well with a Gaussian kernel, which is not always the case for polynomial kernels. This especially entails, for a second-order polynomial kernel given by $f(x) = a_2x^2 + a_1x + a_0$, that attention should be paid to meeting the aforementioned condition when tuning the kernel parameters a_2, a_1 and a_0. Figure 3.5 attests of this remark with Gaussian input data. A rapid increase in classification error rate can be observed both in theory and in practice as soon as the condition $f'(\tau) < 0, f''(\tau) > 0$ is no longer satisfied.

In Figure 3.6 we provide a direct visualization of (the "local" behavior of) different kernel functions f with the same or opposite $f(\tau), f'(\tau), f''(\tau)$. The Gaussian kernel $f(x) = \exp(-x/2)$ (in red) and the quadratic kernel (in BLUE) having the same values of $f(\tau), f'(\tau)$ and $f''(\tau)$ both yield satisfying performance, while by inverting the sign of either $f'(\tau)$ or $f''(\tau)$, the misclassification rate increases rapidly.

Clearly, for practical use, one needs to know in advance the value of τ before training



(a) $f(\tau) = 4,\ f''(\tau) = 1$ (b) $f(\tau) = 4,\ f'(\tau) = -1$

Figure 3.5: Performance of LS-SVM, $n = 256,\ p = 512,\ c_1 = c_2 = 1/2, \lambda = 1$, polynomial kernel. $\mathbf{x} \in \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

so that the kernel f can be properly chosen during the training step. The estimation of τ is possible, in the large n, p regime, with the following lemma.

Lemma 3.2. *Under Assumption 4, as $n \to \infty$,*

$$\frac{2}{n}\sum_{i=1}^{n}\frac{1}{p}\left\|\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right\|^2 \xrightarrow{a.s.} \tau. \tag{3.12}$$

Proof. First observe that

$$\frac{2}{n}\sum_{i=1}^{n}\frac{1}{p}\left\|\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right\|^2 = \frac{2c_1c_2\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2}{p} + \frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2 + \kappa$$

with $\kappa = \frac{4}{n\sqrt{p}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\mathsf{T}\left(-c_2\sum_{\mathbf{x}_i\in\mathcal{C}_1}\boldsymbol{\omega}_i + c_1\sum_{\mathbf{x}_j\in\mathcal{C}_2}\boldsymbol{\omega}_j\right)$ and $\bar{\boldsymbol{\omega}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\omega}_i$.

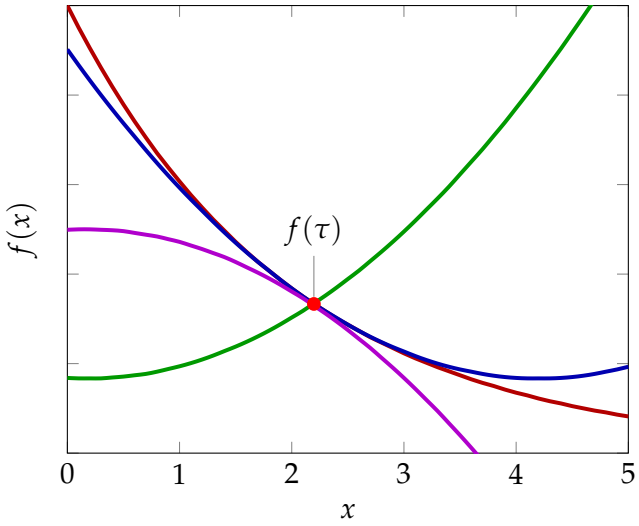
According to Assumption 4 we have $\frac{2c_1c_2}{p}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 = O(n^{-1})$. The term κ is a linear combination of independent zero-mean Gaussian variables and thus $\kappa \sim \mathcal{N}(0, \mathrm{Var}[\kappa])$ with $\mathrm{Var}[\kappa] = \frac{16c_1c_2}{np^2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\mathsf{T}(c_2\mathbf{C}_1 + c_1\mathbf{C}_2)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = O(n^{-3})$. We thus deduce from Chebyshev's inequality and the Borel-Cantelli lemma that $\kappa \xrightarrow{a.s.} 0$.

We then work on the last term $\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2$ as

$$\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i - \bar{\boldsymbol{\omega}}\|^2 = \frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i\|^2 - 2\|\bar{\boldsymbol{\omega}}\|^2.$$

Since $\bar{\boldsymbol{\omega}} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^\circ/np)$, we deduce that $\|\bar{\boldsymbol{\omega}}\|^2 \xrightarrow{a.s.} 0$. Ultimately by the strong law of large numbers, we have $\frac{2}{n}\sum_{i=1}^{n}\|\boldsymbol{\omega}_i\|^2 \xrightarrow{a.s.} \tau$, which concludes the proof. □

Remark 3.4 (Special case: means-dominant). *When the difference in means $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2$ is largely dominant over $(\mathrm{tr}(\mathbf{C}_2 - \mathbf{C}_1)/p)^2$ and $\|\mathbf{C}_2 - \mathbf{C}_1\|_F^2/p$, from Theorem 3.2, both $\bar{G}_a - (c_2 - c_1)$ and $\sqrt{V_{G_a}}$ are (approximately) proportional to $f'(\tau)$, which eventually makes the choice of the kernel irrelevant (as long as $f'(\tau) \neq 0$). This result also holds true for G_a^* and $\sqrt{V_{G_a^*}}$ when normalized labels $\mathbf{y}^*$ are applied, as a result of Lemma 3.1.*



	kernel function	error
	Gaussian kernel $f(x) = \exp(-x/2)$	8.6%
	quadratic kernel with same $f(\tau), f'(\tau)$ and $f''(\tau)$	8.8%
	quadratic kernel with same $f(\tau), f''(\tau)$, while $f'(\tau)$ of opposite sign	66.4%
	quadratic kernel with same $f(\tau), f'(\tau)$ while $f''(\tau)$ of opposite sign	32.9%

Figure 3.6: Classification error rate for different kernel functions f, $n = 256$, $p = 512$, $c_1 = c_2 = 1/2$ and $\lambda = 1$, with $\boldsymbol{\mu}_a = [\mathbf{0}_{a-1}; 2; \mathbf{0}_{p-a}]$, $\mathbf{C}_1 = \mathbf{I}_p$ and $\{\mathbf{C}_2\}_{i,j} = .4^{|i-j|}(1 + 4/\sqrt{p})$.

Practical consequences.[2] When the classification performance of real-world datasets is concerned, our theory may be limited by: i) the fact that it is based on an asymptotic results and allows for an estimation error only of order $O(p^{-1/2})$ between theory and practice and ii) the strong Gaussian assumption for the input data.

However, when applied to real-world datasets, here to the popular MNIST [LBBH98] and Fashion-MNIST [XRV17] datasets, our asymptotic results, theoretically only applicable for Gaussian data, show an unexpectedly good predictive behavior. Here we consider a two-class classification problem with a training set of $n = 256$ vectorized images of size $p = 784$ randomly selected from the MNIST and Fashion-MNIST datasets (numbers 8 and 9 in both cases as an example). Then a test set of $n_{\text{test}} = 256$ is used to evaluate the classification performance. Means and covariances are empirically obtained from the full set of 11 800 MNIST images (5 851 images of number 8 and 5 949 of number 9) and of 11 800 Fashion-MNIST images (5 851 images of number 8 and 5 949 of number 9), respectively. Despite the obvious non-Gaussianity as well as the clearly different nature of the input data (from the two datasets), the distribution of $g(\mathbf{x})$ is still surprisingly close to its Gaussian approximation computed from Theorem 3.2, as shown in Figure 3.7 for MNIST (left) and Fashion-MNIST (right), respectively.

In Figure 3.8 we plot the misclassification rate as a function of the decision threshold ξ for MNIST and Fashion-MNIST data (number 8 and 9). We observe that the conclusion from Remark 3.1, Lemma 3.1 and Corollary 3.1 that the decision threshold should approximately be $c_2 - c_1$ rather than 0 approximately holds true in both cases.

In Figure 3.9 we evaluate the performance of LS-SVM on the MNIST and Fashion-MNIST datasets as a function of the kernel parameter σ of Gaussian kernel $f(x) =$

[2] *Reproducibility*: visit https://github.com/Zhenyu-LIAO/RMT4LSSVM for the Python 3 codes to reproduce the results in this subsection.

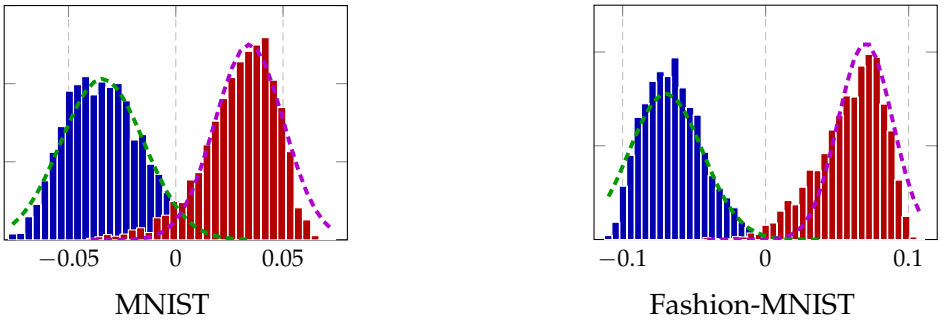


Figure 3.7: Gaussian approximation of $g(\mathbf{x})$, $n = 256$, $p = 784$, $c_1 = c_2 = 1/2, \lambda = 1$, Gaussian kernel with $\sigma = 1$, with MNIST (left) and Fashion-MNIST data (right), numbers 8 and 9.

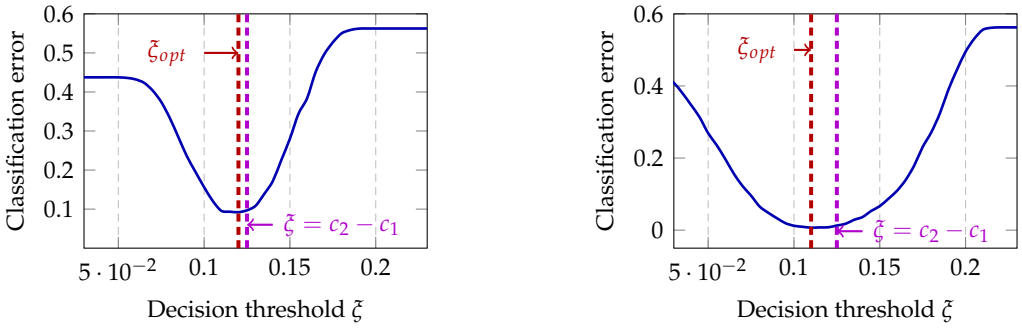


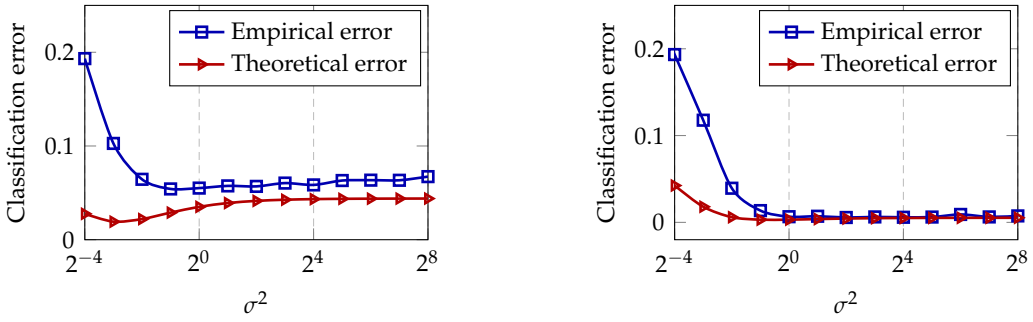
Figure 3.8: LS-SVM classification error rate for $n = 512$, $p = 784$, $c_2 - c_1 = 0.125, \lambda = 1$, Gaussian kernel with $\sigma = 1$ for MNIST (left) and Fashion-MNIST data (right). With optimal decision threshold $\xi_{opt} = 0.12$ (left) and 0.11 (right) in red.

$\exp(-x/2\sigma^2)$. Surprisingly, we face here the situation where there is little difference in the performance of LS-SVM as soon as σ^2 is away from 0, which likely comes from the fact that the difference in means $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$ is so large that it becomes predominant over the influence of covariances as mentioned in Remark 3.4. This argument is numerically sustained by Table 3.1. The gap between theory and practice observed as $\sigma^2 \to 0$ is likely a result of the finite n, p rather than of the Gaussian assumption of the input data, since we continue to observe a similar behavior when one artificiality adds Gaussian white noise to the image data.

Table 3.1: Empirical estimation of differences in means and covariances of MNIST and Fashion-MNIST data (numbers 8 and 9)

	MNIST data	Fashion-MNIST data
$\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2\|^2$	251	483
$\mathrm{tr}^2(\mathbf{C}_2 - \mathbf{C}_1)/p$	19	89
$\|\mathbf{C}_2 - \mathbf{C}_1)^2\|_F^2/p$	30	86

Conclusion. In this section, we investigated the asymptotic performance of kernel ridge regression (or LS-SVM) in separating a two-class Gaussian mixture. We saw that, due to



MNIST Fashion-MNIST

Figure 3.9: Classification performance of LS-SVM, $n = 256$, $p = 784$, $c_1 = c_2 = 1/2, \lambda = 1$, Gaussian kernel, MNIST and Fashion-MNIST data (numbers 8 and 9).

a "concentration" phenomenon of the normalized Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$, the classification performance depends on a local behavior of the kernel function f and the nonlinear kernel matrix $\mathbf{K}$ is asymptotically close to the linear Gram/covariance matrix model with several additive spikes. Although derived from an unrealistic GMM, our theoretical results showed an unexpected close match to simulations on popular real-world datasets.

3.1.2 Inner-product kernels

The results from the previous section are in essence built upon a local expansion of the nonlinear function f, which we recall follows from the "concentration" of the similarity measures $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$ around a *single* value of the smooth domain of f, therefore disregarding most of the domain of f. In this section, following [CS13, DV13], we study the inner product kernel $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$ which avoids the concentration effects with the more natural $\sqrt{p}$ normalization. With the flexible tool of orthogonal polynomials, we are able to prove universal results which solely depend on the first two order moments of the data distribution and allow for nonlinear functions f that need not even be differentiable. As a practical outcome of our theoretical results, we propose an extremely simple piecewise constant function which is spectrally equivalent and thus performs equally well as arbitrarily complex functions f, while inducing enormous gains in both storage and computational complexity.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n feature vectors drawn independently from the following two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) mixture model, which is a special case of the general multivariate mixture model in Definition 1 with $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$, $a = 1, 2$.

$$\begin{cases} \mathcal{C}_1 : & \mathbf{x} = \boldsymbol{\mu}_1 + (\mathbf{I}_p + \mathbf{E}_1)^{\frac{1}{2}}\mathbf{z} \\ \mathcal{C}_2 : & \mathbf{x} = \boldsymbol{\mu}_2 + (\mathbf{I}_p + \mathbf{E}_2)^{\frac{1}{2}}\mathbf{z} \end{cases} \tag{3.13}$$

each having cardinality $n/2$ (see for more discussions in Remark 3.6 below), for some deterministic $\boldsymbol{\mu}_a \in \mathbb{R}^p$, $\mathbf{E}_a \in \mathbb{R}^{p\times p}$, $a = 1, 2$ and random vector $\mathbf{z} \in \mathbb{R}^p$ having i.i.d. entries of zero mean, unit variance and bounded moments. As has been discussed at length in Section 1.1.2, to ensure that the information of $\boldsymbol{\mu}_a, \mathbf{E}_a$ is neither (asymptotically) too simple nor impossible to be extracted from the noisy features, we adapt the two-class non-trivial classification condition in Assumption 4 to the covariance setting $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$ as follows.

Assumption 5 (Non-trivial classification). *As $n \to \infty$, we have for $a \in \{1, 2\}$*

1. $p/n = c \to \bar{c} \in (0, \infty)$,
2. $\|\boldsymbol{\mu}_a\| = O(1)$, $\|\mathbf{E}_a\| = O(p^{-1/4})$, $|\operatorname{tr}(\mathbf{E}_a)| = O(\sqrt{p})$ *and* $\|\mathbf{E}_a\|_F^2 = O(\sqrt{p})$.

Remark 3.5 (On the advantage of proper scaling)**.** *It is worth noting that, compared to Assumption 4 where* $\|\mathbf{C}_1 - \mathbf{C}_2\|_F^2 = O(p)$*, here with Assumption 5 the mixture separation is performed with a higher precision with* $\|\mathbf{E}_1 - \mathbf{E}_2\|_F^2 = O(\sqrt{p})$*. In this respect, the "properly scaled" inner-product kernels investigated here is more powerful than the "improperly scaled" ones of form* $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/p)$*, in separating the covariance "shapes" of Gaussian mixtures.*

Following [EK10a, CS13] we consider the following random inner-product kernel matrix

$$\mathbf{K} = \left\{\delta_{i\neq j} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}\right\}_{i,j=1}^n \tag{3.14}$$

for $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying Assumption 3. As in [EK10a, CS13], the diagonal elements $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p})$ have been discarded. Indeed, under Assumption 7, $\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p} = O(\sqrt{p})$ which is an "improper scaling" for the evaluation by f (unlike $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ which properly scales as $O(1)$ for all $i \neq j$).

In the absence of discriminative information (null model), i.e., if $\boldsymbol{\mu}_a = \mathbf{0}$ and $\mathbf{E}_a = \mathbf{0}$ for $a = 1, 2$, we write $\mathbf{K} = \mathbf{K}_N$ with

$$[\mathbf{K}_N]_{ij} = \delta_{i\neq j} f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})/\sqrt{p}. \tag{3.15}$$

Letting $\xi_p \equiv \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p}$, by the central limit theorem, $\xi_p \xrightarrow{d} \mathcal{N}(0,1)$ as $p \to \infty$. As such, the $[\mathbf{K}_N]_{ij}$, $1 \leq i \neq j \leq n$, asymptotically behave like a family of *dependent* standard Gaussian variables to which f is applied. In order to analyze the joint behavior of this family, we exploit some useful concepts of the theory of orthogonal polynomials and, in particular, of the class of Hermite polynomials defined with respect to the standard Gaussian distribution [AS65, AAR00].

For a real probability measure μ, we denote the set of orthogonal polynomials with respect to the scalar product $\langle f, g\rangle = \int f g d\mu$ as $\{P_l(x), l = 0, 1, \ldots\}$, obtained from the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$ such that $P_0(x) = 1$, P_l is of degree l and $\langle P_{l_1}, P_{l_2}\rangle = \delta_{l_1 - l_2}$. By the Riesz-Fischer theorem [Rud64, Theorem 11.43], for any function $f \in L^2(\mu)$, the set of squared integrable functions with respect to $\langle\cdot,\cdot\rangle$, one can formally expand f as

$$f(x) \sim \sum_{l=0}^{\infty} a_l P_l(x), \quad a_l = \int_{\mathbb{R}} f(x) P_l(x) d\mu(x) \tag{3.16}$$

where "$f \sim \sum_{l=0}^{\infty} P_l$" indicates that $\|f - \sum_{l=0}^{N} P_l\| \to 0$ as $N \to \infty$ (and $\|f\|^2 = \langle f, f\rangle$).

To investigate the asymptotic behavior of $\mathbf{K}$ and $\mathbf{K}_N$ as $n, p \to \infty$, we position ourselves, as [CS13, DV13] under Assumption 3, which roughly says that the polynomial approximation of the kernel function f is accurate, for p large, when we truncate the above expansion at a large but finite degree L.

The limiting parameters a_0, a_1, a_2 and ν appearing in Assumption 3 are simply (generalized) moments of the standard Gaussian measure involving f. Precisely,

$$a_0 = \mathbb{E}[f(\xi)], a_1 = \mathbb{E}[\xi f(\xi)], \sqrt{2}a_2 = \mathbb{E}[(\xi^2 - 1)f(\xi)] = \mathbb{E}[\xi^2 f(\xi)], \nu = \operatorname{Var}[f(\xi)] \geq a_1^2 + a_2^2$$

for $\xi \sim \mathcal{N}(0, 1)$. These parameters are of crucial significance in determining the eigenspectrum behavior of $\mathbf{K}$. Note that a_0 will not affect the classification performance, as described below.

Remark 3.6 (On a_0)**.** *In the present case of balanced mixtures (equal cardinalities for $\mathcal{C}_1$ and $\mathcal{C}_2$), a_0 contributes to the polynomial expansion of $\mathbf{K}_N$ (and $\mathbf{K}$) as a non-informative perturbation of the form $a_0(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} - \mathbf{I}_n)/\sqrt{p}$. Since $\mathbf{1}_n$ is orthogonal to the "class-information vector" $[\mathbf{1}_{n/2}, -\mathbf{1}_{n/2}]$, its presence does not impact the classification performance. If mixtures are unbalanced, the vector $\mathbf{1}_n$ may tend to "pull" eigenvectors aligned to $[\mathbf{1}_{n_1}, -\mathbf{1}_{n_2}]$, with n_i the cardinality in $\mathcal{C}_i$, so away from purely noisy eigenvectors and thereby impacting the classification performance. See [CBG16] for similar considerations.*

It was shown in [CS13, DV13] that, for independent $\mathbf{z}_i$'s with independent entries, the *empirical spectral measure* (see Definition 5) of the null model $\mathbf{K}_N$ has an asymptotically deterministic behavior as $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, which depends on the nonlinear f solely via the *two* parameters (a_1, ν), as recalled below.

Theorem 3.3 (Theorem 3.4 in [CS13], Theorem 3 in [DV13])**.** *Under some regularity condition for the kernel function f (see Assumption 3 below), the empirical spectral measure of $\mathbf{K}$ converges weakly and almost surely, as $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, to a probability measure μ. The latter is uniquely defined through its Stieltjes transform $m : \mathbb{C}^+ \to \mathbb{C}^+$, $z \mapsto \int (t - z)^{-1}\mu(dt)$ (see also Definition 6 in Section 2.2), given as the unique solution in $\mathbb{C}^+$ of the (cubic) equation*[3]

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z)$$

with $a_1 = \mathbb{E}[\xi f(\xi)]$ and $\nu = \mathrm{Var}[f(\xi)] \geq a_1^2$ for standard Gaussian $\xi \sim \mathcal{N}(0, 1)$.

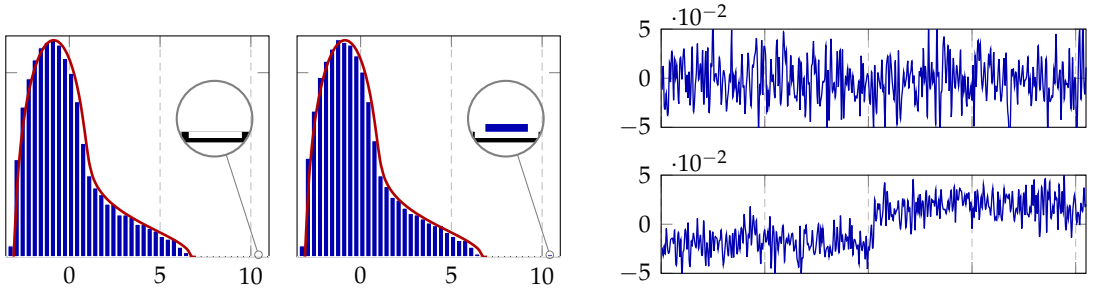
Moreover, Theorem 3.3 is "universal" with respect to the law of the (independent) entries of $\mathbf{z}_i$. While universality is classical in random matrix results, with mostly first and second order statistics involved, the present universality result is much less obvious since i) the nonlinear application $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ depends in an intricate manner on all moments of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ and ii) the entries of $\mathbf{K}_N$ are strongly dependent. In essence, universality still holds here because the convergence speed to Gaussian of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ is sufficiently fast to compensate the residual impact of higher order moments in the spectrum of $\mathbf{K}_N$.

Remark 3.7 (Non-informative spikes)**.** *Despite its "universality", Theorem 3.3 only characterizes the limiting spectral measure μ of the kernel matrix $\mathbf{K}$, and thus allows for a "limited" number of isolated eigenvalues to escape from the support of μ. Indeed, as discussed at the end of Section 2.3, depending on the value of p/n, it was shown in [KC17, FM19] that, for continuously differentiable kernel function f, if $a_2 \neq 0$, then $\mathbf{K}$ has at most two spikes outside* $\mathrm{supp}(\mu)$*, while for $a_2 = 0$ all eigenvalues of $\mathbf{K}$ lie in* $\mathrm{supp}(\mu)$ *almost surely.*

As an illustration, Figure 3.10a compares the empirical spectral measure of $\mathbf{K}_N$ to the limiting measure obtained from Theorem 3.3.

From a technical viewpoint, our objective here is to go beyond the null model described in Theorem 3.3 by providing a tractable *random matrix equivalent* $\tilde{\mathbf{K}}$ for the kernel matrix $\mathbf{K}$, in the sense that $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$ almost surely in operator norm, as $n, p \to \infty$. This convergence allows one to identify the eigenvalues and isolated eigenvectors (that can be used for spectral clustering purpose) of $\mathbf{K}$ by means of those of $\tilde{\mathbf{K}}$ via, for instance Lemma 2.10. More importantly, while not visible from the expression of $\mathbf{K}$, the impact of

[3] $\mathbb{C}^+ \equiv \{z \in \mathbb{C},\ \Im[z] > 0\}$. We also recall that, for $m(z)$ the Stieltjes transform of a measure μ, μ can be obtained from $m(z)$ via $\mu([a, b]) = \lim_{\epsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m(x + \imath\epsilon)]dx$ for all $a < b$ continuity points of μ. See Section 2.1.2 for more details.



(a) Eigenvalues of $\mathbf{K}_N$ (b) Eigenvalues of $\mathbf{K}$ (c) Top eigenvectors of $\mathbf{K}_N$ (top) and $\mathbf{K}$ (bottom)

Figure 3.10: Eigenvalue distribution and top eigenvector of $\mathbf{K}_N$ and $\mathbf{K}$, together with the limiting spectral measure from Theorem 3.3 (in red); $f(x) = \operatorname{sign}(x)$, Gaussian $\mathbf{z}_i$, $n = 2\,048$, $p = 512$, $\boldsymbol{\mu}_1 = -[3/2;\, \mathbf{0}_{p-1}] = -\boldsymbol{\mu}_2$ and $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{0}$. $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{x}_1, \dots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \dots, \mathbf{x}_n \in \mathcal{C}_2$.

the mixture model $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{E}_1, \mathbf{E}_2)$ on $\mathbf{K}$ is readily accessed from $\tilde{\mathbf{K}}$ and easily related to the Hermite coefficients $(\, a_1, a_2, \nu\,)$ of f. This allows us to further investigate how the choice of f impacts the asymptotic feasibility and efficiency of spectral clustering from the top eigenvectors of $\mathbf{K}$.

Main results. The main idea for the asymptotic analysis of $\mathbf{K}$ comes in two steps: first, by an expansion of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ as a function of $\mathbf{z}_i, \mathbf{z}_j$ and the statistical mixture model parameters $\boldsymbol{\mu}, \mathbf{E}$, we decompose $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ (under Assumption 7) into successive orders of magnitudes with respect to p; this, as we will show, further allows for a Taylor expansion of $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ for at least twice differentiable functions f around its dominant term $f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})$. Then, we rely on the orthogonal polynomial approach of [CS13] to "linearize" the resulting matrix terms $\{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$, $\{f'(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ and $\{f''(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ (all terms corresponding to higher order derivatives asymptotically vanish) and use Assumption 3 to extend the result to arbitrary square-summable f.

Our main conclusion is that $\mathbf{K}$ asymptotically behaves like a matrix $\tilde{\mathbf{K}}$ following a so-called "spiked random matrix model" in the sense that $\tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$ is the sum of the full-rank "noise" matrix $\mathbf{K}_N$ having compact limiting spectrum and a low-rank "information" matrix $\tilde{\mathbf{K}}_I$ [BAP05, BGN11].

We first show that $\mathbf{K}$ can be asymptotically approximated as $\mathbf{K}_N + \mathbf{K}_I$ with $\mathbf{K}_N$ defined in (3.15) and $\mathbf{K}_I$ an additional (so far full-rank) term containing the statistical information of the mixture model.

As announced, we start by decomposing $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ into a sequence of terms of successive orders of magnitude using Assumption 7 and $\mathbf{x}_i = \boldsymbol{\mu}_a + (\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}\mathbf{z}_i$, $\mathbf{x}_j = \boldsymbol{\mu}_b + (\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j$ for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$. We have precisely, for $i \neq j$,

$$\begin{aligned}\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\sqrt{p}} &= \frac{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b}{\sqrt{p}} + \frac{1}{\sqrt{p}}(\boldsymbol{\mu}_a^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j + \boldsymbol{\mu}_b^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}\mathbf{z}_i) + \frac{1}{\sqrt{p}}\mathbf{z}_i^\mathsf{T}(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}(\mathbf{I}_p + \mathbf{E}_b)^{\frac{1}{2}}\mathbf{z}_j \\ &= \underbrace{\frac{\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}{\sqrt{p}}}_{O(1)} + \underbrace{\frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a + \mathbf{E}_b)\mathbf{z}_j}{2\sqrt{p}}}_{\equiv \mathbf{A}_{ij} = O(p^{-1/4})} + \underbrace{\frac{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \boldsymbol{\mu}_a^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_b^\mathsf{T}\mathbf{z}_i}{\sqrt{p}} - \frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a - \mathbf{E}_b)^2\mathbf{z}_j}{8\sqrt{p}}}_{\equiv \mathbf{B}_{ij} = O(p^{-1/2})} + o(p^{-1/2}) \end{aligned} \tag{3.17}$$

where in particular we performed a Taylor expansion of $(\mathbf{I}_p + \mathbf{E}_a)^{\frac{1}{2}}$ (since $\|\mathbf{E}_a\| = O(p^{-\frac{1}{4}})$) around $\mathbf{I}_p$, and used the fact that with high probability $\mathbf{z}_i^\mathsf{T}\mathbf{E}_a\mathbf{z}_j = O(p^{1/4})$ and $\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a - \mathbf{E}_b)^2\mathbf{z}_j = O(1)$.

As a consequence of this expansion, for at least twice differentiable $f \in L^2(\mu_p)$, we have

$$\mathbf{K}_{ij} = \frac{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})}{\sqrt{p}} = \frac{f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{\sqrt{p}} + \frac{f'(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{\sqrt{p}}(\mathbf{A}_{ij} + \mathbf{B}_{ij}) + \frac{f''(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})}{2\sqrt{p}}\mathbf{A}_{ij}^2 + o(p^{-1})$$

where $o(p^{-1})$ is understood with high probability and uniformly over $i, j \in \{1, \ldots, n\}$. Recall that, this *entry-wise* expansion up to order $o(p^{-1})$ is sufficient since, *matrix-wise*, if $\mathbf{A}_{ij} = o(p^{-1})$ uniformly on i, j, from $\|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty = p \max_{i,j} |\mathbf{A}_{ij}|$, we have $\|\mathbf{A}\| = o(1)$ as $n, p \to \infty$.

In the particular case where f is a monomial of degree $k \geq 2$, this implies the following result.

Proposition 3.1 (Monomial f)**.** *Under Assumptions 3 and 5, let $f(x) = x^k$, $k \geq 2$. Then, as $n, p \to \infty$,*

$$\|\mathbf{K} - (\mathbf{K}_N + \mathbf{K}_I)\| \to 0 \tag{3.18}$$

almost surely, with $\mathbf{K}_N$ defined in (3.15) *and*

$$\mathbf{K}_I = \frac{k}{\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ (\mathbf{A} + \mathbf{B}) + \frac{k(k-1)}{2\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-2)} \circ (\mathbf{A})^{\circ 2} \tag{3.19}$$

for $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ defined in (3.17) *with $\mathbf{A}_{ii} = \mathbf{B}_{ii} = 0$. Here $\mathbf{X} \circ \mathbf{Y}$ denotes the Hadamard product between $\mathbf{X}, \mathbf{Y}$ and $\mathbf{X}^{\circ k}$ the k-th Hadamard power, i.e., $[\mathbf{X}^{\circ k}]_{ij} = (\mathbf{X}_{ij})^k$.*

Since $f \in L^2(\mu)$ can be decomposed into its Hermite polynomials, Proposition 3.1 along with Theorem 3.3 allows for an asymptotic quantification of $\mathbf{K}$. However, the expression of $\mathbf{K}_I$ in (3.19) does not so far allow for a thorough understanding of the spectrum of $\mathbf{K}$, due to 2) the delicate Hadamard products between purely random $(\mathbf{Z}^\mathsf{T}\mathbf{Z})$ and informative matrices $(\mathbf{A}, \mathbf{B})$ and ii) the fact that $\mathbf{K}_I$ is full rank (so that the resulting spectral properties of $\mathbf{K}_N + \mathbf{K}_I$ remains intractable). We next show that, as $n, p \to \infty$, $\mathbf{K}_I$ admits a tractable low-rank approximation $\tilde{\mathbf{K}}_I$, thereby leading to a spiked-model approximation for $\mathbf{K}$.

Let us consider $\mathbf{K}_I$ defined in (3.19), the (i, j) entry of which can be written as the sum of terms containing $\boldsymbol{\mu}_a, \boldsymbol{\mu}_b$ (treated separately) and random variables of the type

$$\phi = \frac{C}{\sqrt{p}}(\mathbf{x}^\mathsf{T}\mathbf{y}/\sqrt{p})^\alpha(\mathbf{x}^\mathsf{T}\mathbf{F}\mathbf{y})^\beta$$

for independent random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ with i.i.d. zero mean, unit variance and finite moments (uniformly on p) entries, deterministic $\mathbf{F} \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}$, $\alpha \in \mathbb{N}$ and $\beta \in \{1, 2\}$.

For Gaussian $\mathbf{x}, \mathbf{y}$, the expectation of ϕ can be explicitly computed via an integral trick [Wil97, LLC18]. For more generic $\mathbf{x}, \mathbf{y}$ with i.i.d. bounded moment entries, a combinatorial argument controls the higher order moments of the expansion which asymptotically result in (matrix-wise) vanishing terms. See Sections A.1.4 in the appendix. This leads to the following result.

Proposition 3.2 (Low rank asymptotics of $\mathbf{K}_I$)**.** *Under Assumptions 3 and 5, for* $f(x) = x^k$, $k \geq 2$,

$$\|\mathbf{K}_I - \tilde{\mathbf{K}}_I\| \to 0$$

almost surely as $n, p \to \infty$, *for* $\mathbf{K}_I$ *defined in* (3.19) *and*

$$\tilde{\mathbf{K}}_I = \begin{cases} \frac{k!!}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}), & \text{for } k \text{ odd} \\ \frac{k(k-1)!!}{2p}\mathbf{J}\left(\mathbf{T} + \mathbf{S}\right)\mathbf{J}^\mathsf{T}, & \text{for } k \text{ even} \end{cases} \tag{3.20}$$

where[4]

$$\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2] \in \mathbb{R}^{p\times 2},\ \mathbf{T} = \{\mathrm{tr}(\mathbf{E}_a + \mathbf{E}_b)/\sqrt{p}\}_{a,b=1}^2,\ \mathbf{S} = \{\mathrm{tr}(\mathbf{E}_a\mathbf{E}_b)/\sqrt{p}\}_{a,b=1}^2 \in \mathbb{R}^{2\times 2}$$

and $\mathbf{J} = [\mathbf{j}_1, \mathbf{j}_2] \in \mathbb{R}^{n\times 2}$ *with* $\mathbf{j}_a \in \mathbb{R}^n$ *the canonical vector of class* $\mathcal{C}_a$, *i.e.,* $[\mathbf{j}_a]_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$.

We refer the readers to Section A.1.4 of the Appendix for a detailed exposition of the proof.

Proposition 3.2 states that $\mathbf{K}_I$ is asymptotically equivalent to $\tilde{\mathbf{K}}_I$ that is of rank at most two.[5] Note that the eigenvectors of $\tilde{\mathbf{K}}_I$ are linear combinations of the vectors $\mathbf{j}_1, \mathbf{j}_2$ and thus provide the data classes.

From the expression of $\tilde{\mathbf{K}}_I$, quite surprisingly, it appears that for $f(x) = x^k$, depending on whether k is odd or even, either only the information in means ($\mathbf{M}$) or only in covariance ($\mathbf{T}$ and $\mathbf{S}$) can be (asymptotically) preserved.

By merely combining the results of Propositions 3.1–3.2, the latter can be easily extended to polynomial f. Then, by considering $f(x) = P_\kappa(x)$, the Hermite polynomial of degree κ, it can be shown that, quite surprisingly, one has $\tilde{\mathbf{K}}_I = \mathbf{0}$ if $\kappa > 2$. As such, using the Hermite polynomial expansion $P_0, P_1, \ldots$ of an arbitrary $f \in L^2(\mu)$ satisfying Assumption 3 leads to a very simple expression of our main result.

Theorem 3.4 (Spiked-model approximation of $\mathbf{K}$)**.** *For an arbitrary* $f \in L^2(\mu)$ *with* $f \sim \sum_{l=0}^\infty a_l P_l(x)$, *under Assumptions3 and 5,*

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0, \quad \tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$$

with $\mathbf{K}_N$ *defined in* (3.15) *and*

$$\tilde{\mathbf{K}}_I = \frac{a_1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}) + \frac{a_2}{p}\mathbf{J}(\mathbf{T} + \mathbf{S})\mathbf{J}^\mathsf{T}. \tag{3.21}$$

Proof. The proof of Theorem 3.4 follows from the fact that the individual coefficients of the Hermite polynomials $P_\kappa(x) = \sum_{l=0}^\kappa c_{\kappa,l} x^l$ satisfy the following recurrent relation [AS65]

$$c_{\kappa+1,l} = \begin{cases} -\kappa c_{\kappa-1,l} & l = 0; \\ c_{\kappa,l-1} - \kappa c_{\kappa-1,l} & l \geq 1; \end{cases} \tag{3.22}$$

[4]For mental reminder, $\mathbf{M}$ stands for *means*, $\mathbf{T}$ accounts for the difference in *traces* of covariance matrices and $\mathbf{S}$ for the "shapes" of the covariances.

[5]Note that, as defined, $\tilde{\mathbf{K}}_I$ has non-zero diagonal elements, while $[\mathbf{K}_I]_{ii} = 0$. This is not contradictory as the diagonal matrix $\mathrm{diag}(\tilde{\mathbf{K}}_I)$ has vanishing norm and can thus be added without altering the approximation $\|\mathbf{K}_I - \tilde{\mathbf{K}}_I\| \to 0$; it however appears convenient as it ensures that $\tilde{\mathbf{K}}_I$ is low rank (while without its diagonal, $\tilde{\mathbf{K}}_I$ is full rank).

with $c_{0,0} = 1$, $c_{1,0} = 0$ and $c_{1,1} = 1$. As a consequence, by indexing the informative matrix in Proposition 3.2 of the monomial $f(x) = x^l$ as $\tilde{\mathbf{K}}_{I,l}$, we have for odd $\kappa \geq 3$,

$$\tilde{\mathbf{K}}_I = \sum_{l=1,3,\dots}^{\kappa} c_{\kappa,l} \tilde{\mathbf{K}}_{I,l} = \sum_{l=1,3,\dots}^{\kappa} c_{\kappa,l} l!! (\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T})/p - \operatorname{diag}(\cdot) = \mathbf{0}$$

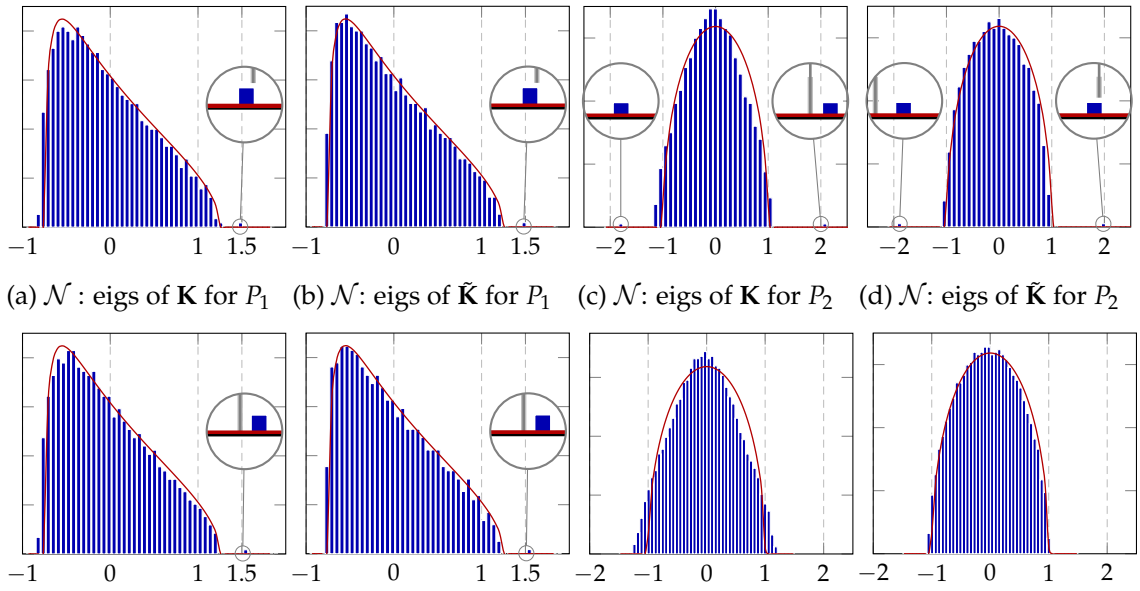
with $[\mathbf{X} - \operatorname{diag}(\cdot)]_{ij} = \mathbf{X}_{ij}\delta_{i\neq j}$. This follows from the fact that, for $\kappa \geq 3$, we have both $\sum_{l=1,3,\dots}^{\kappa} c_{\kappa,l} l!! = 0$ and $\sum_{l=0,2,\dots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$. The latter is proved by induction on κ: first, for $\kappa = 3$, we have $c_{3,1} + 3c_{3,3} = c_{4,0} + 3c_{4,2} + 15c_{4,4} = 0$; then, assuming κ odd, we have $\sum_{l=1,3,\dots}^{\kappa} c_{\kappa,l} l!! = \sum_{l=0,2,\dots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$ so that, together with (3.22)

$$\sum_{l=1,3,\dots}^{\kappa+2} c_{\kappa+2,l} l!! = \sum_{l=1,\dots}^{\kappa+2} (c_{\kappa+1,l-1} - (\kappa+1)c_{\kappa,l}) l!! = \sum_{l=1,\dots}^{\kappa+2} c_{\kappa+1,l-1} l!! = \sum_{l=0,2,\dots}^{\kappa+1} c_{\kappa+1,l}(l+1)!! = 0$$

as well as

$$\sum_{l=0,2,\dots}^{\kappa+3} c_{\kappa+3,l}(l+1)!! = -(\kappa+2)c_{\kappa+1,0} + \sum_{l=2,4,\dots}^{\kappa+3} (c_{\kappa+2,l-1} - (\kappa+2)c_{\kappa+1,l})(l+1)!! = 0$$

where we used $c_{\kappa,l} = 0$ for $l \geq \kappa+1$. Similar arguments hold for the case of κ even, which concludes the proof. □



(a) $\mathcal{N}$: eigs of $\mathbf{K}$ for P_1 (b) $\mathcal{N}$: eigs of $\tilde{\mathbf{K}}$ for P_1 (c) $\mathcal{N}$: eigs of $\mathbf{K}$ for P_2 (d) $\mathcal{N}$: eigs of $\tilde{\mathbf{K}}$ for P_2

(e) Stud: eigs of $\mathbf{K}$ for P_1 (f) Stud: eigs of $\tilde{\mathbf{K}}$ for P_1 (g) Stud: eigs of $\mathbf{K}$ for P_3 (h) Stud: eigs of $\tilde{\mathbf{K}}$ for P_3

Figure 3.11: Eigenvalue distributions of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ from Theorem 3.4 (blue) and $\mathcal{L}$ from Theorem 3.3 (red), for $\mathbf{z}_i$ with Gaussian (top) or Student-t with degree of freedom 7 (bottom) entries; functions $f(x) = P_1(x) = x$, $f(x) = P_2(x) = (x^2-1)\sqrt{2}$, $f(x) = P_3(x) = (x^3-3x)/\sqrt{6}$; $n = 2\,048$, $p = 8\,192$, $\boldsymbol{\mu}_1 = -[2; \mathbf{0}_{p-1}] = -\boldsymbol{\mu}_2$ and $\mathbf{E}_1 = -10\mathbf{I}_p/\sqrt{p} = -\mathbf{E}_2$.

Figure 3.11 compares the spectra of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ for random vectors with independent Gaussian or Student-t entries, for the first three (normalized) Hermite polynomials $P_1(x)$, $P_2(x)$ and $P_3(x)$. These numerical evidences validate Theorem 3.4: only for $P_1(x)$ and $P_2(x)$ is an isolated eigenvalue observed. Besides, as shown in the bottom display of Figure 3.10c, the corresponding eigenvector is, as expected, a noisy version of linear combinations of $\mathbf{j}_1, \mathbf{j}_2$.

Table 3.2: Storage size and top eigenvector running time of $\mathbf{K}$ for piecewise constant and cubic f, in the setting of Figure 3.11 and 3.13.

f	Size (Mb)	Running time (s)
Piecewise	4.15	0.2390
Cubic	16.75	0.4244

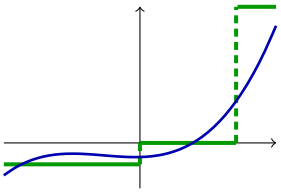


Figure 3.12: Piecewise constant (green) versus cubic (BLUE) function with equal (a_1, a_2, ν).

Remark 3.8 (Even and odd f). *While* $\operatorname{rank}(\tilde{\mathbf{K}}_I) \leq 4$ *(as the sum of two rank-two terms), in Figure 3.11* no more than *two isolated eigenvalues are observed (for $f = P_1$ only one on the right side, for $f = P_2$ one on each side). This follows from $a_2 = 0$ when $f = P_1$ and $a_1 = 0$ for $f = P_2$. More generally, for f* odd *($f(-x) = -f(x)$), $a_2 = 0$ and the statistical information on covariances (through* **E***) asymptotically vanishes in* **K***; for f* even *($f(-x) = f(x)$), $a_1 = 0$ and the information about the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ vanishes. Thus, only f neither odd nor even can preserve both first and second order discriminating statistics (e.g., the popular ReLU function $f(x) = \max(0, x)$). This was previously remarked in [LC18a] based on a local expansion of smooth f in a similar setting.*

Practical consequences. As a direct consequence of Theorem 3.4, the performance of spectral clustering for large dimensional mixture models of the type (3.13) only depends on the *three* parameters of the nonlinear function f: $a_1 = \mathbb{E}[\xi f(\xi)]$, $a_2 = \mathbb{E}[\xi^2 f(\xi)]/\sqrt{2}$ and $\nu = \mathbb{E}[f^2(\xi)]$. The parameters a_1, ν determine the limiting spectral measure $\mathcal{L}$ of $\mathbf{K}$ (since $\mathbf{K}$ and $\mathbf{K}_N$ asymptotically differ by a rank-4 matrix, they share the same limiting spectral measure) while a_2, a_2 determine the low rank structure within $\tilde{\mathbf{K}}_I$.

As an immediate consequence, arbitrary (square-summable) kernel functions f (with $a_0 = 0$) are asymptotically *equivalent* to the simple cubic function $\tilde{f}(x) = c_3 x^3 + c_2 x^2 + c_1 x - c_2$ having the *same* Hermite polynomial coefficients a_1, a_2, ν, due to the following relation

$$a_1 = 3c_3 + c_1, \quad a_2 = \sqrt{2} c_2, \quad \nu = (3c_3 + c_1)^2 + 6c_3^2 + 2c_2^2.$$

The idea here to design a prototypical family $\mathcal{F}$ of functions f having i) universal properties with respect to (a_1, a_2, ν), i.e., for each (a_1, a_2, ν) there exists $f \in \mathcal{F}$ with these Hermite coefficients and ii) having numerically advantageous properties. Thus, any arbitrary kernel function f can be mapped, through (a_1, a_2, ν), to a function in $\mathcal{F}$ with good numerical properties.

One such prototypical family $\mathcal{F}$ can be the set of f, parametrized by (t, s_-, s_+), and defined as

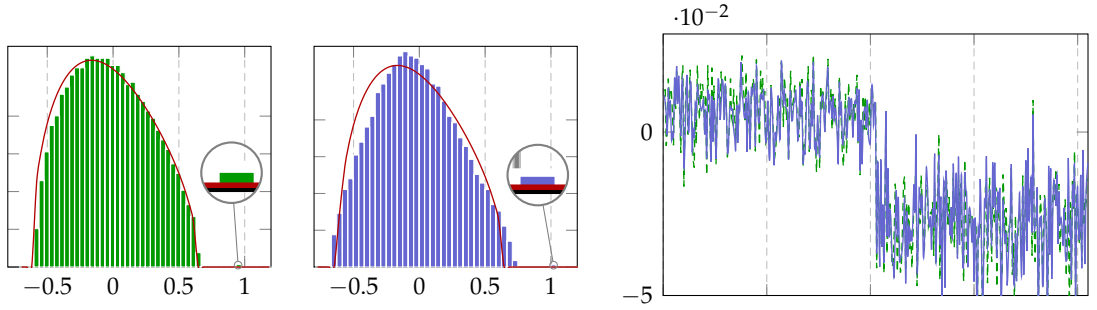
$$f(x) = \begin{cases} -rt & x \leq \sqrt{2} s_- \\ 0 & \sqrt{2} s_- < x \leq \sqrt{2} s_+ \\ t & x > \sqrt{2} s_+ \end{cases} \text{ with } \quad \begin{cases} a_1 = \frac{t}{\sqrt{2\pi}} (e^{-s_+^2} + r e^{-s_-^2}) \\ a_2 = \frac{t}{\sqrt{2\pi}} (s_+ e^{-s_+^2} + r s_- e^{-s_-^2}) \\ \nu = \frac{t^2}{2} (1 - \operatorname{erf}(s_+)) (1 + r) \end{cases} \tag{3.23}$$

where $r \equiv \frac{1 - \operatorname{erf}(s_+)}{1 + \operatorname{erf}(s_-)}$. Figure 3.12 displays f given in (3.23) together with the cubic function $c_3 x^3 + c_2(x^2 - 1) + c_1 x$ sharing the same Hermite coefficients (a_1, a_2, ν).

The class of equivalence of kernel functions induced by this mapping is quite unlike that raised in [EK10b] or [CBG16] in the "improper" scaling $f(\mathbf{x}_i^\mathsf{T} \mathbf{x}_j / p)$ regime. While in

the latter, functions $f(x)$ of the same class of equivalence are those having common $f'(0)$ and $f''(0)$ values, in the present case, these functions may have no similar local behavior (as shown in the example of Figure 3.12).

For the piecewise constant function defined in (3.23) and the associated cubic function having the same (a_1, a_2, ν), a close match is observed for both eigenvalues and top eigenvectors of $\mathbf{K}$ in Figure 3.13, with gains in both storage size and computational time displayed in Table 3.2.



(a) Piecewise constant (b) Cubic polynomial (c) Cubic (blue) versus piecewise (green) function

Figure 3.13: Eigenvalue distribution and top eigenvectors of $\mathbf{K}$ for the piecewise constant function (in green) and the associated cubic function (in blue) with the same (a_1, a_2, ν), performed on Bernoulli distribution with zero mean and unit variance, in the setting of Figure 3.11.

Conclusion. In this last section, we evaluated the eigenspectrum of the "properly-scaled" inner-product kernel matrix $\mathbf{K} = f(\mathbf{X}^\mathsf{T}\mathbf{X}/\sqrt{p})/\sqrt{p}$. Built upon the flexible tool of orthogonal polynomials, we showed that the spectrum of $\mathbf{K}$ depends on the kernel function f via three "global" key parameters (a_1, a_2, ν), and this holds true for a large range of non-linear functions f, including even some non-smooth functions. In this vein, all kernel functions under study are equivalent, in an eigenspectrum sense, to a cubic function, as well as to a piecewise function that takes only three values, as long as they share the same (a_1, a_2, ν).

To close this section on random kernel matrices, according to the discussion in Remark 3.5, it is of future interest to extend the current analyses on $f(\mathbf{X}^\mathsf{T}\mathbf{X}/\sqrt{p})/\sqrt{p}$ to cover data with an arbitrary covariance matrix $\mathbf{C}$ (that is not limited to $\mathbf{C} = \mathbf{I} + \mathbf{E}$ with $\|\mathbf{E}\| = o(1)$). However, this is technically more involved, since it breaks most of the orthogonality properties of the orthogonal polynomial approach of the proofs, but is a needed extension of the result.

It would also be of interest to properly scale not only the inner-product kernels, but also other popular choice of kernels such as the shift-invariant kernels studied in Section 3.1.1. From this perspective, the kernel matrices of the type $f(\|\mathbf{x}_i\|^2/p - 2\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p} + \|\mathbf{x}_j\|^2/p)$ could be a good starting point.

3.2 Random Neural Networks

As discussed in Section 1.2.2, the fundamental connection between random kernel matrices and random weights NN models can be built upon random feature maps. In this sub-

section, we demonstrate how random kernel matrices appear in the performance analysis of single-hidden-layer random NN models, and how this analysis allows us to evaluate the impact on performance of the activation function.

3.2.1 Large neural networks with random weights

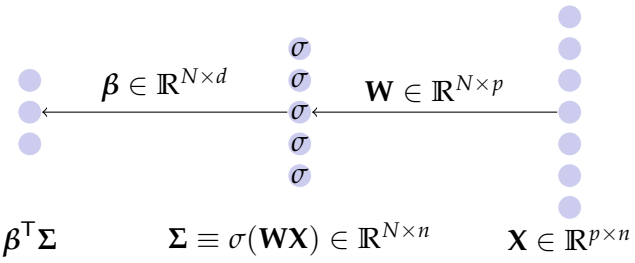


Figure 3.14: Illustration of a single-hidden-layer random NN

In this section, we consider the single-hidden-layer random NN model (or, random feature-based kernel ridge regression) illustrated above (same as Figure 1.3).

For a data matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \in \mathbb{R}^{p\times n}$ with associated targets $\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_n\right] \in \mathbb{R}^{d\times n}$, we denote $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times n}$ the output of the middle layer comprising in total N neurons of $\mathbf{X}$ by premultiplying some (random) weight matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$ with i.i.d. standard Gaussian entries and then passing through some nonlinear activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ to obtain $\boldsymbol{\Sigma} \equiv \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{N\times n}$. In particular, we focus on the case where the second layer weight matrix $\boldsymbol{\beta} \in \mathbb{R}^{N\times d}$ is designed to minimize the regularized MSE: $L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_i)\|^2 + \lambda\|\boldsymbol{\beta}\|_F^2$ on the given deterministic training set $(\mathbf{X}, \mathbf{Y})$, for some regularization factor $\lambda > 0$. This givens the following explicit form (as in (1.18)) for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \equiv \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{Y}^\mathsf{T}. \tag{3.24}$$

We are interested in the asymptotic behavior of the training and test MSE of the network (as in (1.19)) defined by

$$E_{\text{train}} = \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i - \boldsymbol{\beta}^\mathsf{T}\sigma(\mathbf{W}\mathbf{x}_i)\|^2 = \frac{1}{n}\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\|_F^2, \quad E_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{Y}} - \boldsymbol{\beta}^\mathsf{T}\hat{\boldsymbol{\Sigma}}\|_F^2 \tag{3.25}$$

on a (deterministic) test set $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ of size $\hat{n}$, i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{p\times\hat{n}}$, $\hat{\mathbf{Y}} \in \mathbb{R}^{d\times\hat{n}}$, where we similarly denote $\hat{\boldsymbol{\Sigma}} \equiv \sigma(\mathbf{W}\hat{\mathbf{X}}) \in \mathbb{R}^{N\times\hat{n}}$.

Let us start with the training error E_{train}. Note that, by defining the resolvent (see Definition 4) of the Gram matrix $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ as

$$\mathbf{Q}(z = -\lambda) = \mathbf{Q} \equiv \left(\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma} + \lambda\mathbf{I}_n\right)^{-1} \in \mathbb{R}^{n\times n}$$

the training MSE of interest can be rewritten as

$$E_{\text{train}} = \frac{1}{n}\|\mathbf{Y} - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\|_F^2 = \frac{\lambda^2}{n}\operatorname{tr}(\mathbf{Y}\mathbf{Q}^2\mathbf{Y}^\mathsf{T}) = -\frac{\lambda^2}{n}\frac{\partial\operatorname{tr}(\mathbf{Y}\mathbf{Q}\mathbf{Y}^\mathsf{T})}{\partial\lambda} \tag{3.26}$$

and is therefore wished to establish an asymptotically deterministic behavior as the sum of $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ for deterministic vector $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of bounded Euclidean norm (with the normalization n^{-1}). if we could find a deterministic equivalent (see Definition 7) for $\mathbf{Q}$.

To properly state our main results, the following assumptions are needed.

Assumption 6 (Lipschitz σ)**.** *The function σ is Lipschitz continuous with parameter independent of n, p.*

Assumption 7 (Growth rate in random NN)**.** *As $n \to \infty$,*

1. *$p/n \to \bar{c}_1 \in (0, \infty)$ and $N/n \to \bar{c}_2 \in (0, \infty)$;*
2. *$\mathbf{X}$ has a bounded operator norm, i.e., $\|\mathbf{X}\| = O(1)$;*
3. *$\mathbf{Y}$ has bounded entries, i.e., $\|\mathbf{Y}\|_\infty = O(1)$.*

Note that, for the nonlinear Gram matrix $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$, with $\sigma(t) = t$ one merely gets the sample covariance/Gram matrix $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{X}$, the spectrum of which is known to have an asymptotic deterministic behavior (recall Section 2.2.3) with $\|\mathbf{X}\| = O(1)$. Intuitively speaking, Assumption 6 ensures that the nonlinear activation σ "varies" in a "controlled" manner.

Also, by demanding $\mathbf{Y}$ to have bounded entries as $n, p, N \to \infty$, we indeed ask the rows of $\mathbf{Y}$, together with the $n^{-1/2}$ normalization, to have bounded Euclidean norm, so as to apply our deterministic equivalent approach.

Under Assumptions 6 and 7, with a we are now ready to prove the concentration of the resolvent $\mathbf{Q}$ (in the sense that $\|\mathbf{Q} - \mathbb{E}_\mathbf{W}[\mathbf{Q}]\| \to 0$) as $n, p, N \to \infty$, and consequently we can provide an exact characterization of the training and test MSE in (1.19).

Main results. Before going into our main results, let us first re-introduce the following "equivalent" kernel matrix (as in (1.20))

$$\mathbf{K} = \mathbb{E}_\mathbf{w}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})] \tag{3.27}$$

for $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Again, the appearance of kernel matrix $\mathbf{K}$ here is rather natural since the Gram matrix $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ at the heart of our analysis is indeed an approximation of $\mathbf{K}$ under the random feature framework, for instance with $\psi_\mathbf{w}(\mathbf{x}) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x})$ in (1.15).

To show the concentration of the resolvent $\mathbf{Q}$ (as well as the eigenspectrum) of a Gram/sample covariance-like random matrix $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, a concentration of quadratic forms based on the row vectors of $\mathbf{Z}$ is necessary. For instance, recall from Section 2.2.2 that, to show the (almost sure) convergence of $\frac{1}{n}\operatorname{tr}\mathbf{A}\left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n\right)^{-1}$ for $\mathbf{Z} \in \mathbb{R}^{p\times n}$ having i.i.d. standard Gaussian entries, we exploited the concentration of quadratic form result $\frac{1}{n}\mathbf{z}_i^\mathsf{T}\mathbf{A}\mathbf{z}_i - \frac{1}{n}\operatorname{tr}\mathbf{A} \xrightarrow{a.s.} 0$ in Lemma 2.11, to bound the difference between $\left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n\right)^{-1}$ and $\left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - \frac{1}{n}\mathbf{z}_i\mathbf{z}_i^\mathsf{T} - z\mathbf{I}_n\right)^{-1}$, for $\mathbf{z} \in \mathbb{R}^n$ the i-th row of $\mathbf{Z}$ so that $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. In general, such concentration of quadratic form results are obtained by exploiting the independence (or linear dependence) in the vector entries (e.g., $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ in the Marčenko-Pastur case in Section 2.2.2). For the nonlinear model under consideration

$$\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} = \frac{1}{n}\sigma(\mathbf{W}\mathbf{X})^\mathsf{T}\sigma(\mathbf{W}\mathbf{X})$$

the desired independence unfortunately does not hold, since the entries of the vector $\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})$ are in general not independent. To overcome this technical difficulty, we resort to a concentration of measure approach, as advocated in [Kar09]. The following lemma provides this concentration result.

Lemma 3.3 (Concentration of quadratic forms)**.** *Let Assumptions 6 and 7 hold and* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *such that* $\|\mathbf{A}\| \leq 1$*. For* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$*, denote the random vector* $\boldsymbol{\sigma} \equiv \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}) \in \mathbb{R}^n$*. Then,*

$$P\left(\left|\frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma} - \frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{A})\right| > t\right) \leq Ce^{-cn\min(t,t^2)}$$

for some $C, c > 0$ *and* $\mathbf{K}$ *defined in* (3.27)*.*

Lemma 3.3 states that, with exponentially high probability, the quadratic form $\frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}$ is asymptotically close to its expectation (with respect to the random $\mathbf{w}$)

$$\frac{1}{n}\mathbb{E}_\mathbf{w}[\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}] = \frac{1}{n}\operatorname{tr}(\mathbb{E}_\mathbf{w}[\boldsymbol{\sigma}\boldsymbol{\sigma}^\mathsf{T}]\mathbf{A}) = \frac{1}{n}\operatorname{tr}\left(\mathbb{E}_\mathbf{w}\left[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})\right]\mathbf{A}\right) = \frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{A}).$$

This can be seen as an extension of the crucial Lemma 2.11 for random vectors of linearly correlated entries to the nonlinear setting under investigation. Roughly speaking, with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ we have the following *normal concentration* [Led05, Proposition 1.10] result for all Lipschitz applications,

$$\mathbb{E}\left[\left|f - \int f d\mu\right|^k\right] \leq (C\lambda_f)^k$$

holds for some $C > 0$ and all $k \geq 1$, $f : \mathbb{R}^p \mapsto \mathbb{R}$ a λ_f-Lipschitz function and $d\mu(\mathbf{z}) = (2\pi)^{-\frac{p}{2}}e^{-\frac{1}{2}\|\mathbf{z}\|^2}$ the standard Gaussian measure on $\mathbb{R}^p$. Since $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz (Assumption 6), the normal concentration of $\mathbf{w}$ transfers to $\boldsymbol{\sigma}$ and $\boldsymbol{\Sigma}$, which further implying that all Lipschitz *functionals* of $\boldsymbol{\sigma}$ or $\boldsymbol{\Sigma}$ are also concentrated. Nonetheless, note that the quadratic form $\frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}$ is "quadratic" with respect to $\mathbf{w}$ and thus not Lipschitz. AS such, to prove Lemma 3.3 we choose to first provide an exponentially high probability $O(1)$ bound on $n^{-1/2}\|\boldsymbol{\sigma}\|$ such that, conditioned on this event, the mapping $\mathbf{w} \mapsto \frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\boldsymbol{\sigma}$ can be shown to be $O(1)$-Lipschitz.

With the above result in place, we can follow the same idea as in Section 2.2.2 and apply Lemma 3.3 instead of Lemma 2.11, to show subsequently the concentrations of $\frac{1}{n}\boldsymbol{\sigma}^\mathsf{T}\mathbf{A}\mathbf{Q}_{-i}\mathbf{B}\boldsymbol{\sigma}$ as well as $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ for independent $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of bounded norm. This eventually leads to the following result on the deterministic equivalent (see Definition 7) for the resolvent $\mathbf{Q}$ of interest, the proof of which is deferred to Section A.1.5 of the Appendix.

Theorem 3.5 (Asymptotic equivalent for $\mathbb{E}[\mathbf{Q}]$)**.** *Let Assumptions 6 and 7 hold and define* $\bar{\mathbf{Q}}$ *as*

$$\bar{\mathbf{Q}} \equiv \left(\check{\mathbf{K}} - z\mathbf{I}_n\right)^{-1}, \quad \check{\mathbf{K}} \equiv \frac{N}{n}\frac{\mathbf{K}}{1+\delta}$$

where δ *is implicitly defined as the unique solution of* $\delta = \frac{1}{n}\operatorname{tr}(\mathbf{K}\bar{\mathbf{Q}})$*. Then, as* $n, p, N \to \infty$ *one has*

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0.$$

With Theorem 3.5 at hand, we now present the following corollary on the training and test MSE, which follows directly from our observation in (3.26).

Theorem 3.6 (Asymptotic training and test MSE)**.** *Let Assumptions 6 and 7 hold. For* $E_{\text{train}}, E_{\text{test}}$ *given by* (3.25)*,* $\bar{\mathbf{Q}}$ *defined as in Theorem 3.5, denote*

$$\bar{E}_{\text{train}} = \frac{\lambda^2}{n}\operatorname{tr}\mathbf{Y}\bar{\mathbf{Q}}\left[\frac{\frac{1}{N}\operatorname{tr}\left(\bar{\mathbf{Q}}\check{\mathbf{K}}\bar{\mathbf{Q}}\right)}{1-\frac{1}{N}\operatorname{tr}\left(\check{\mathbf{K}}\bar{\mathbf{Q}}\check{\mathbf{K}}\bar{\mathbf{Q}}\right)}\check{\mathbf{K}} + \mathbf{I}_n\right]\bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T}$$

$$\bar{E}_{\text{test}} = \frac{1}{\hat{n}}\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}\bar{\mathbf{Q}}\check{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}\|_F^2 + \frac{\frac{1}{N}\operatorname{tr}\mathbf{Y}\bar{\mathbf{Q}}\check{\mathbf{K}}\bar{\mathbf{Q}}\mathbf{Y}^\mathsf{T}}{1-\frac{1}{N}\operatorname{tr}\left(\check{\mathbf{K}}\bar{\mathbf{Q}}\check{\mathbf{K}}\bar{\mathbf{Q}}\right)}\left[\frac{1}{\hat{n}}\operatorname{tr}\check{\mathbf{K}}_{\hat{\mathbf{X}}\hat{\mathbf{X}}} - \frac{1}{\hat{n}}\operatorname{tr}\left(\mathbf{I}_n + \frac{N}{n}\lambda\bar{\mathbf{Q}}\right)\check{\mathbf{K}}_{\mathbf{X}\hat{\mathbf{X}}}\check{\mathbf{K}}_{\hat{\mathbf{X}}\mathbf{X}}\bar{\mathbf{Q}}\right].$$

Then, as $n, p, N \to \infty$

$$E_{\text{train}} - \bar{E}_{\text{train}} \to 0, \quad E_{\text{test}} - \bar{E}_{\text{test}} \to 0$$

almost surely, with

$$\mathbf{K_{AB}} \equiv \mathbb{E}_{\mathbf{w}}\left[\sigma(\mathbf{w}^\mathsf{T}\mathbf{A})^\mathsf{T}\sigma(\mathbf{w}^\mathsf{T}\mathbf{B})\right], \quad \check{\mathbf{K}}_{\mathbf{AB}} \equiv \frac{N}{n}\frac{\mathbf{K_{AB}}}{1+\delta}, \quad \mathbf{K} \equiv \mathbf{K_{XX}}, \quad \check{\mathbf{K}} \equiv \check{\mathbf{K}}_{\mathbf{XX}}.$$

We also mention here that, while the above finding on the (asymptotic) training MSE always holds true under Assumptions 6 and 7, the test MSE presented is only a conjecture which not be proved under a general situation, for example with only $\|\hat{\mathbf{X}}\| = O(1)$ and $\|\hat{\mathbf{Y}}\|_\infty = O(1)$. Under some statistical assumptions for $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ (e.g., GMM as in Definition 1) the conjecture is valid [LC18b]. Further empirical evidences on real-world datasets also show an extremely close match. This leads to the fundamental question on the minimal (statistical) assumption on the data to establish an asymptotically deterministic behavior for neural networks performances. A first promising answer is given in [LC18b] by assuming the data to be *concentrated* random vectors, e.g., Lipschitz maps of standard Gaussian random vector or of vectors with i.i.d. entries. Intuitively speaking, this means that the high dimensional data are random, but in a "concentrated" manner such that, instead of "spreading" all over their large ambient space, the data live within a rather low dimensional layer. As a consequence, each scalar observation of these data, even with complicated function such as regressor or classifier, tends to have an almost deterministic and predictable behavior, that depends only on the first several order of statistics of the data.

Practical consequences. Theorem 3.6 allows to assess the training and test performance of the network, via the equivalent kernel matrix $\mathbf{K}$ (and $\hat{\mathbf{K}}$) built from the (deterministic) data. It thus remains to compute this kernel matrix $\mathbf{K}$, which is given under the form of an expectation/integral with respect to the standard Gaussian distribution (of dimension p). For most commonly used nonlinear activations $\sigma(\cdot)$, the generic form $\mathbf{K}(\mathbf{a}, \mathbf{b}) = \mathbf{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})]$ can be computed explicitly via an integral trick [Wil97], for arbitrary vector $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. We list the results for commonly used functions in Table 3.3, with computational details provided in Section A.1.6 of the Appendix.

Table 3.3: $\mathbf{K}(\mathbf{a}, \mathbf{b})$ for different $\sigma(\cdot)$, $\angle(\mathbf{a}, \mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

$\sigma(t)$	$\mathbf{K}(\mathbf{a}, \mathbf{b})$
t	$\mathbf{a}^\mathsf{T}\mathbf{b}$
$\max(t, 0)$	$\frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\|\left(\angle(\mathbf{a},\mathbf{b})\arccos\left(-\angle(\mathbf{a},\mathbf{b})\right) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2}\right)$
$\lvert t\rvert$	$\frac{2}{\pi}\|\mathbf{a}\|\|\mathbf{b}\|\left(\angle(\mathbf{a},\mathbf{b})\arcsin\left(\angle(\mathbf{a},\mathbf{b})\right) + \sqrt{1-\angle(\mathbf{a},\mathbf{b})^2}\right)$
$1_{t>0}$	$\frac{1}{2} - \frac{1}{2\pi}\arccos\left(\angle(\mathbf{a},\mathbf{b})\right)$
$\text{sign}(t)$	$\frac{2}{\pi}\arcsin\left(\angle(\mathbf{a},\mathbf{b})\right)$
$\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$	$\varsigma_2^2\left(2\left(\mathbf{a}^\mathsf{T}\mathbf{b}\right)^2 + \|\mathbf{a}\|^2\|\mathbf{b}\|^2\right) + \varsigma_1^2\mathbf{a}^\mathsf{T}\mathbf{b} + \varsigma_2\varsigma_0\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right) + \varsigma_0^2$
$\cos(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right)\cosh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\sin(t)$	$\exp\left(-\frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2\right)\right)\sinh(\mathbf{a}^\mathsf{T}\mathbf{b})$
$\text{erf}(t)$	$\frac{2}{\pi}\arcsin\left(\frac{2\mathbf{a}^\mathsf{T}\mathbf{b}}{\sqrt{(1+2\|\mathbf{a}\|^2)(1+2\|\mathbf{b}\|^2)}}\right)$
$\exp(-t^2/2)$	$\frac{1}{\sqrt{(1+\|\mathbf{a}\|^2)(1+\|\mathbf{b}\|^2)-(\mathbf{a}^\mathsf{T}\mathbf{b})^2}}$

To corroborate the findings in Theorem 3.6 we consider the task of classifying the MNIST database [LBBH98] with a single-hidden-layer random weight neural network composed of $N = 512$ hidden-units. Similar to Section 3.1.1 we represent each image as a $p = 784$-size vector; 1 024 images of sevens and 1 024 images of nines were extracted from the database and were evenly split in 512 training and test images, respectively, so that $n = \hat{n} = 1\,024$. The database images were jointly centered and scaled so to fall close to the setting of Assumption 7 on $\mathbf{X}$ and $\hat{\mathbf{X}}$. The columns of the output values $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ were taken as unidimensional ($d = 1$) with $\mathbf{Y}_i, \hat{\mathbf{Y}}_i \in \{-1, 1\}$ depending on the image class. Figure 3.15 displays the simulated (averaged over 100 realizations of $\mathbf{W}$) versus theoretical values of E_{train} and E_{test} for three choices of Lipschitz continuous functions $\sigma(\cdot)$, as a function of the regularization factor λ. We observe an almost perfect match between the simulations and our theoretical results from Theorem 3.6 for not so large n, p, N.

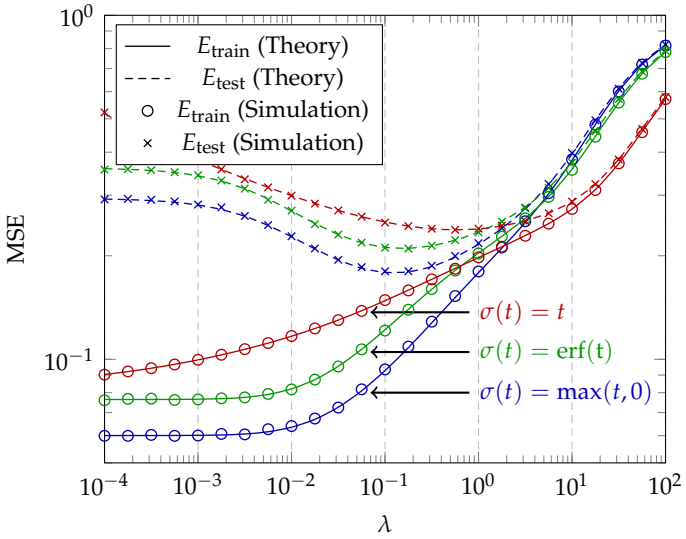


Figure 3.15: Performance of the network for Lipschitz σ, as a function of λ, for MNIST data (number 7 and 9), $N = 512$, $n = \hat{n} = 1024$, $p = 784$.

Conclusion. In this section, we characterized the resolvent $\mathbf{Q}$ of $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ with $\mathbf{\Sigma} = \sigma(\mathbf{WX})$ the output of a random NN with input $\mathbf{X}$ and showed it establishes an asymptotically deterministic behavior for Lipschitz activations $\sigma(\cdot)$. As a direct consequence, we gave the training and test MSE of the random NN model in Figure 1.3 that depends on the data and the activation function $\sigma(\cdot)$ via the key "equivalent" kernel matrix $\mathbf{K} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$. For most commonly used σ, the expression of $\mathbf{K}$ is given in Table 3.3.

Nonetheless, the expressions of $\mathbf{K}$ in Table 3.3 are still in very barely interpretable and provide little understanding of the role played by the nonlinear activations. For instance, we are not able to say which $\sigma(\cdot)$ should be used for a given task. To further exploit the interplay between the data and the activation function, we next investigate in the following section the eigenspectrum of $\mathbf{K}$ by (additionally) considering a GMM for the input data.

3.2.2 Random feature maps and the equivalent kernels

We saw in the previous section that the single-layer random NN performance depends on the deterministic data $\mathbf{X}$ and the nonlinear activation $\sigma(\cdot)$ via the following "equivalent" kernel matrix (as in (1.20))

$$\mathbf{K} \equiv \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$$

the expression of which is explicitly given in Table 3.3 for most commonly used σ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. In this section we develop a deeper understanding of, the eigenspectrum of this kernel matrix $\mathbf{K}$, for structural statistical models for $\mathbf{X}$. In particular, we focus on the interplay between σ and the data statistics $\boldsymbol{\mu}, \mathbf{C}$ of $\mathbf{X}$ under a GMM as detailed in Definition 1.

In the context of random feature maps, $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ is indeed the Gram matrix of the random features $\sigma(\mathbf{W}\mathbf{x}_i)$, that approximates an underlying kernel matrix as discussed at the end of Section 1.2.1. From Theorem 3.5, we know that the resolvent of $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ and thus its limiting spectral measure, if exists, are both closely related to $\mathbf{K}$. In this vein, the kernel matrix $\mathbf{K}$ above is again at the heart of all random feature-based spectral algorithms.

Consider $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independent data vectors, each belonging to one of K distribution classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$. Class $\mathcal{C}_a$ has cardinality n_a, for all $a \in \{1, \ldots, K\}$. The data vector $\mathbf{x}_i$ that belongs to $\mathcal{C}_a$ is assumed to be drawn from the following GMM

$$\mathbf{x}_i = \frac{1}{\sqrt{p}}\boldsymbol{\mu}_a + \boldsymbol{\omega}_i$$

with $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_a/p)$ for some mean $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and covariance $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ that satisfy the non-trivial classification in Assumption 2. Note that we normalize the data by $1/\sqrt{p}$, together with Assumption 2 to ensure $\|\mathbf{x}_i\| = O(1)$ with high probability, which is in consistent with Assumption 7 in the presented random NN model.

Taking $\mathbf{a} = \mathbf{x}_i$ and $\mathbf{b} = \mathbf{x}_j$, we see that the expressions of $\mathbf{K}$ in Table 3.3 are indeed nonlinear functions of the "concentrated" measures of the data, e.g., $\|\mathbf{x}_i\|$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$, as investigated in Section 3.1.1 under the binary setting $K = 2$. Therefore, the spectral analysis of $\mathbf{K}$ for GMM under Assumption 2 follows exactly the same line of arguments as in [CBG16] and discussed at length in Section 3.1.1, that we briefly recall next.

From a random feature map standpoint, the Gram matrix $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ describes the correlation of data in the *feature space*. It is thus natural to recenter $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$, hence $\mathbf{K}$, by pre- and post-multiplying with the projection matrix $\mathbf{P} \equiv \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$. In the case of $\mathbf{K}$, we get

$$\mathbf{K}_c \equiv \mathbf{PKP}.$$

This centering operation technically eliminates, from a technical standpoint, the non-informative isolated eigenvalues that may go to infinity as $n, p \to \infty$, with corresponding non-informative eigen-direction $\mathbf{1}_n$.

Main results. Let us now introduce the key steps of our present analysis. Under Assumption 2, observe that for $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b$, $i \neq j$,

$$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j = \underbrace{\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b/p + \boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j/\sqrt{p} + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}$$

which allows one to perform a Taylor expansion around 0 as $p, T \to \infty$, to give a reasonable approximation of nonlinear functions of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$, such as those appearing in $\mathbf{K}_{ij}$ (see again Table 3.3). For $i = j$, one has instead

$$\|\mathbf{x}_i\|^2 = \underbrace{\|\boldsymbol{\omega}_i\|^2}_{O(1)} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}.$$

From $\mathbb{E}_{\boldsymbol{\omega}_i}[\|\boldsymbol{\omega}_i\|^2] = \operatorname{tr}(\mathbf{C}_a)/p$ it is convenient to further write

$$\|\boldsymbol{\omega}_i\|^2 = \operatorname{tr}(\mathbf{C}_a)/p + \left(\|\omega_i\|^2 - \operatorname{tr}(\mathbf{C}_a)/p\right)$$

where $\operatorname{tr}(\mathbf{C}_a)/p = O(1)$ and $\|\omega_i\|^2 - \operatorname{tr}(\mathbf{C}_a)/p = O(n^{-1/2})$. By definition[6] $\tau \equiv \operatorname{tr}(\mathbf{C}^\circ)/p$ that is assumed to converge in $(0, \infty)$ as in the two-class case in Assumption 4. Exploiting again Assumption 2, one results in,

$$\|\mathbf{x}_i\|^2 = \underbrace{\tau}_{O(1)} + \underbrace{\operatorname{tr}(\mathbf{C}_a^\circ)/p + \|\boldsymbol{\omega}_i\|^2 - \operatorname{tr}(\mathbf{C}_a)/p}_{O(p^{-1/2})} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}}_{O(p^{-1})}$$

which allows for a Taylor expansion of nonlinear functions of $\|\mathbf{x}_i\|^2$ around τ, as done for $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$.

From Table 3.3, it appears that, for every listed $\sigma(\cdot)$, $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ is a smooth function of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ and $\|\mathbf{x}_i\|$, $\|\mathbf{x}_j\|$, despite their possible discontinuities (for example, the ReLU function and $\sigma(t) = |t|$). The above results therefore allow for an entry-wise Taylor expansion of the matrix $\mathbf{K}$ in the large n, p limit.

A critical aspect of the analysis where random matrix theory comes into play now consists in developing $\mathbf{K}$ as a sum of matrices arising from the Taylor expansion and ignoring terms that give rise to a vanishing operator norm, so as to find an asymptotic equivalent matrix $\tilde{\mathbf{K}}$ such that $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$ as $n, p \to \infty$, as described in detail below. This analysis provides a simplified asymptotically equivalent expression for $\mathbf{K}$ with all nonlinearities removed.

To present our main theoretical result, we define, similarly to Section 3.1.1 and 3.1.2 the following notations for random elements

$$\boldsymbol{\Omega} \equiv \left[\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\right] \in \mathbb{R}^{p\times n}, \quad \boldsymbol{\phi} \equiv \left\{\|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\left[\|\boldsymbol{\omega}_i\|^2\right]\right\}_{i=1}^n \in \mathbb{R}^n$$

as well as for deterministic elements[7],

$$\mathbf{M} \equiv \left[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\right] \in \mathbb{R}^{p\times K}, \quad \mathbf{J} \equiv \left[\mathbf{j}_1, \ldots, \mathbf{j}_K\right] \in \mathbb{R}^{n\times K}$$
$$\mathbf{t} \equiv \left\{\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{C}_a^\circ\right\}_{a=1}^K \in \mathbb{R}^K, \quad \mathbf{S} \equiv \left\{\frac{1}{p}\operatorname{tr}(\mathbf{C}_a\mathbf{C}_b)\right\}_{a,b=1}^K \in \mathbb{R}^{K\times K}$$

where $\mathbf{j}_a \in \mathbb{R}^n$ denotes, as usual, the canonical vector of class $\mathcal{C}_a$ such that $(\mathbf{j}_a)_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$.

Theorem 3.7 (Asymptotic equivalent of $\mathbf{K}_c$)**.** *Let Assumption 2 hold and $\mathbf{K}_c$ be defined as $\mathbf{K}_c \equiv \mathbf{PKP}$ for $\mathbf{K}$ given in* (1.20)*. Then, as $n \to \infty$, for all $\sigma(\cdot)$ given in Table 3.3,*

$$\|\mathbf{K}_c - \tilde{\mathbf{K}}_c\| \to 0$$

[6]Note that the notation τ defined here differs from that in Section 1.1.2 and 3.1.1 by a factor 2.

[7]Similarly to Section 3.1.2, $\mathbf{M}$ stands for *means*, $\mathbf{t}$ accounts for (difference in) *traces* while $\mathbf{S}$ for the "*shapes*" of covariances. Note in particular that the definition of $\mathbf{S}$ is different from that in Section 3.1.2, since only the special case $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$ is considered in Section 3.1.2.

almost surely, with $\tilde{\mathbf{K}}_c = \mathbf{P}\tilde{\mathbf{K}}\mathbf{P}$ *and*

$$\tilde{\mathbf{K}} \equiv d_1 \left(\boldsymbol{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}}\right)^\mathsf{T}\left(\boldsymbol{\Omega} + \mathbf{M}\frac{\mathbf{J}^\mathsf{T}}{\sqrt{p}}\right) + d_2\mathbf{U}\mathbf{B}\mathbf{U}^\mathsf{T} + d_0\mathbf{I}_n \equiv \mathbf{K}_N + \tilde{\mathbf{K}}_I + d_0\mathbf{I}_n$$

where we recall that $\mathbf{P} \equiv \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ *and*

$$\mathbf{U} \equiv \left[\tfrac{\mathbf{J}}{\sqrt{p}}, \boldsymbol{\phi}\right], \quad \mathbf{B} \equiv \begin{bmatrix}\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S} & \mathbf{t} \\ \mathbf{t}^\mathsf{T} & 1\end{bmatrix}, \quad \mathbf{K}_N \equiv d_1\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega} + d_2\boldsymbol{\phi}\boldsymbol{\phi}^\mathsf{T}$$

$$\tilde{\mathbf{K}}_I \equiv \frac{d_1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \sqrt{p}\boldsymbol{\Omega}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \sqrt{p}\mathbf{J}\mathbf{M}^\mathsf{T}\boldsymbol{\Omega}) + \frac{d_2}{p}\left(\mathbf{J}(\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S})\mathbf{J}^\mathsf{T} + \sqrt{p}\boldsymbol{\phi}\mathbf{t}^\mathsf{T}\mathbf{J}^\mathsf{T} + \sqrt{p}\mathbf{J}\mathbf{t}\boldsymbol{\phi}^\mathsf{T}\right)$$

with the coefficients d_0, d_1, d_2 *given in Table 3.4.*

Table 3.4: Coefficients d_i in $\tilde{\mathbf{K}}_c$ for different $\sigma(\cdot)$.

$\sigma(t)$	d_0	d_1	d_2
t	0	1	0
$\mathrm{ReLU}(t) \equiv \max(t, 0)$	$\left(\frac{1}{4} - \frac{1}{2\pi}\right)\tau$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\lvert t\rvert$	$\left(1 - \frac{2}{\pi}\right)\tau$	0	$\frac{1}{2\pi\tau}$
$\varsigma_+\max(t,0) + \varsigma_-\max(-t,0)$ $\equiv \mathrm{LReLU}(t)$	$\frac{\pi-2}{4\pi}(\varsigma_+ + \varsigma_-)^2\tau$	$\frac{1}{4}(\varsigma_+ - \varsigma_-)^2$	$\frac{1}{8\tau\pi}(\varsigma_+ + \varsigma_-)^2$
$1_{t>0}$	$\frac{1}{4} - \frac{1}{2\pi}$	$\frac{1}{2\pi\tau}$	0
$\mathrm{sign}(t)$	$1 - \frac{2}{\pi}$	$\frac{2}{\pi\tau}$	0
$\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$	$2\tau^2\varsigma_2^2$	ς_1^2	ς_2^2
$\cos(t)$	$\frac{1}{2} + \frac{e^{-2\tau}}{2} - e^{-\tau}$	0	$\frac{e^{-\tau}}{4}$
$\sin(t)$	$\frac{1}{2} - \frac{e^{-2\tau}}{2} - \tau e^{-\tau}$	$e^{-\tau}$	0
$\mathrm{erf}(t)$	$\frac{2}{\pi}\left(\arccos\left(\frac{2\tau}{2\tau+1}\right) - \frac{2\tau}{2\tau+1}\right)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-t^2/2)$	$\frac{1}{\sqrt{2\tau+1}} - \frac{1}{\tau+1}$	0	$\frac{1}{4(\tau+1)^3}$

Theorem 3.7 tells us as a corollary (from for example Lemma 2.10) that the maximal difference between the eigenvalues of $\mathbf{K}_c$ and $\tilde{\mathbf{K}}_c$ vanishes asymptotically as $n, p \to \infty$. Similarly the distance between the "isolated eigenvectors" also vanishes. This is of tremendous importance as the determination of the leading eigenvalues and eigenvectors of $\mathbf{K}_c$ (that contain crucial information for clustering, for example) can be studied from the equivalent problem performed on $\tilde{\mathbf{K}}_c$ and becomes mathematically more tractable.

It is also of interest to remark from Theorem 3.7 that, albeit derived from different data models and kernel types, Theorem 3.7 provides the same intuition as Theorem 3.4: the first order information ($\mathbf{M}$) always goes with one coefficient (d_1 and a_1, respectively), while the second order information ($\mathbf{t}, \mathbf{S}$ or $\mathbf{T}, \mathbf{S}$, respectively) is always multiplied by another coefficient (d_2 and a_2, respectively).

To make this clear, the major differences between these two results are summarized as follows:

1. Theorem 3.4 characterizes the spectrum of the "properly scaled" inner-product kernel matrix model $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$, with zero on its diagonal, regardless of the underlying distribution; it holds for any square-summable kernel function f (that satisfies

Assumption 3); the coefficients a_1, a_2 and ν are known to be the (generalized) moments of f (with respect to standard Gaussian measure). Nonetheless, due to the lack of more advanced technical tools, only the special case $\mathbf{C}_a = \mathbf{I}_p + \mathbf{E}_a$ is covered in Theorem 3.4.

2. On the other hand, Theorem 3.7 treats the kernel matrices arising from random feature maps (or random NNs), in the Gaussian case (i.e., for GMM data); these kernels are "improperly scaled" (as those studied in Section 3.1.1) with the form $f(\mathbf{x}_i^\top \mathbf{x}_j / p)$ in Table 3.3, and the analysis is based on a local expansion of f, which imperatively demands f to be locally (at least three-times) differentiable near 0 and τ. Also, Theorem 3.7 holds only for the nonlinear activation listed in Table 3.3 for which we are capable of explicitly computing the expression of $\mathbf{K}$. However, Theorem 3.7 holds for arbitrary covariance $\mathbf{C}_a$.

Practical consequences. On closer inspection of Theorem 3.7, the matrix $\tilde{\mathbf{K}}$ is expressed as the sum of three terms, weighted respectively by the three coefficients d_0, d_1 and d_2, that depend on the nonlinear function $\sigma(\cdot)$ via Table 3.4. Note that the statistical structure of the data $\{\mathbf{x}_i\}_{i=1}^n$ (namely the means in $\mathbf{M}$ and the covariances in $\mathbf{t}$ and $\mathbf{S}$) is perturbed by random fluctuations ($\boldsymbol{\Omega}$ and $\boldsymbol{\phi}$). In this perspective, we see that d_1 and d_2 (asymptotically) control the "expression" of the first ($\mathbf{M}$) and second order ($\mathbf{t}$ and $\mathbf{S}$) statistical information in $\mathbf{K}$, respectively. In particular, there exists a balance between the means and covariances, that provides some instructions in the appropriate choice of the nonlinearity. From Table 3.4, the functions $\sigma(\cdot)$ can be divided into the following three groups:

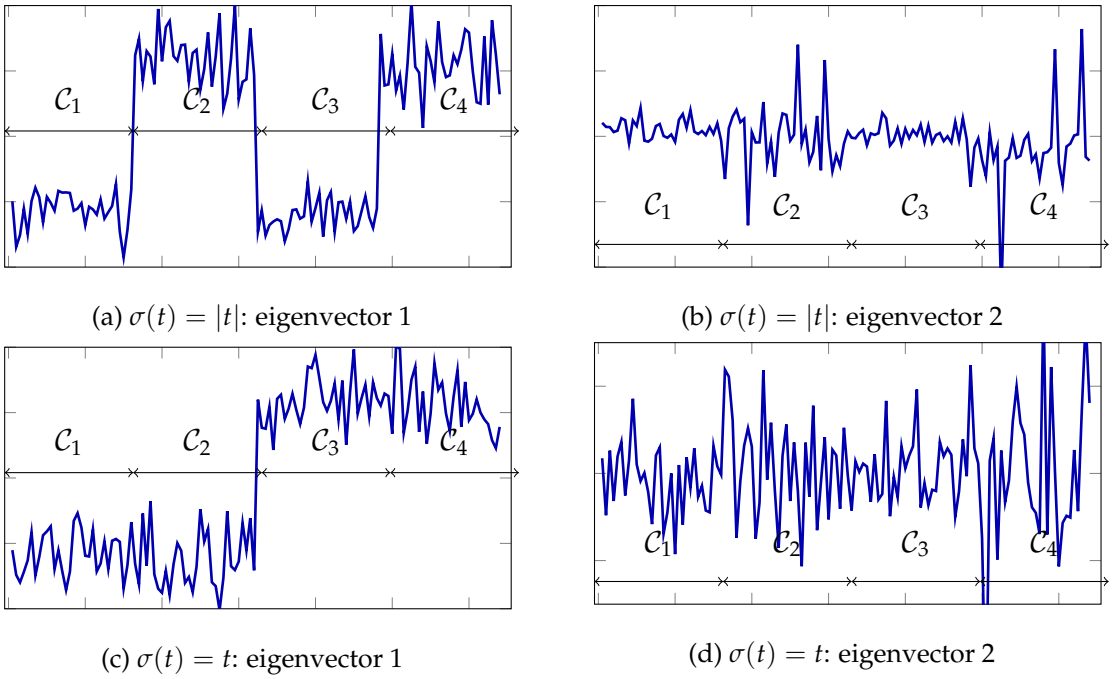
- *mean-oriented*, where $d_1 \neq 0$ while $d_2 = 0$: this is the case of the functions t, $1_{t>0}$, $\mathrm{sign}(t)$, $\sin(t)$ and $\mathrm{erf}(t)$, which asymptotically track only the difference in means (i.e., $\mathbf{t}$ and $\mathbf{S}$ disappear from the expression of $\tilde{\mathbf{K}}_c$);

- *covariance-oriented*, where $d_1 = 0$ while $d_2 \neq 0$: this concerns the functions $|t|$, $\cos(t)$ and $\exp(-t^2/2)$, which asymptotically track only the difference in covariances;

- *balanced*, where both $d_1, d_2 \neq 0$: here for the ReLU function $\max(t, 0)$ and the quadratic function $\varsigma_2 t^2 + \varsigma_1 t + \varsigma_0$.

To corroborate the above classification of different nonlinearity, we perform kernel spectral clustering in Figure 3.16 on four classes of Gaussian data: $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_2)$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$ with the LReLU function that takes different values for ς_+ and ς_-. For $a = 1, 2$, $\boldsymbol{\mu}_a = \left[\mathbf{0}_{a-1}; 5; \mathbf{0}_{p-a}\right]$ and $\mathbf{C}_a = (1 + 15(a-1)/\sqrt{p})\mathbf{I}_p$. By choosing $\varsigma_+ = -\varsigma_- = 1$ (equivalent to $\sigma(t) = |t|$) and $\varsigma_+ = \varsigma_- = 1$ (equivalent to the linear map $\sigma(t) = t$), with the leading two eigenvectors we always recover two classes instead of four, as each setting of parameters only allows for a part of the statistical information of the data to be used for clustering. However, by taking $\varsigma_+ = 1, \varsigma_- = 0$ (the ReLU function) we distinguish all four classes in the leading two eigenvectors, to which the k-means method can then be applied for final classification, as shown in Figure 3.17.

We complete this section by showing that our theoretical results in Theorem 3.7, derived from GMMs, show an unexpected close match in practice when applied to real-world datasets. We consider two different types of classification tasks: one on the MNIST [LBBH98] database (number 6 and 8), and the other on epileptic EEG time series data [ALM+01] (set B and E). These two datasets are typical examples of means-dominant



(a) $\sigma(t) = |t|$: eigenvector 1

(b) $\sigma(t) = |t|$: eigenvector 2

(c) $\sigma(t) = t$: eigenvector 1

(d) $\sigma(t) = t$: eigenvector 2

Figure 3.16: Leading two eigenvectors of $\mathbf{K}_c$ for the LReLU function with $\varsigma_+ = -\varsigma_- = 1$ (top) and $\varsigma_+ = \varsigma_- = 1$ (bottom), performed on four classes Gaussian mixture data with $p = 512$, $n = 256$, $c_a = 1/4$ and $\mathbf{j}_a = [\mathbf{0}_{n_{a-1}}; \mathbf{1}_{n_a}; \mathbf{0}_{n-n_a}]$, for $a = 1, 2, 3, 4$. Expectation estimated by averaging over 500 realizations of $\mathbf{W}$.

Table 3.5: Empirical estimation of (normalized) differences in means and covariances of the MNIST (Figure 3.18) and epileptic EEG (Figure 3.19) datasets.

	$\|\mathbf{M}^\mathsf{T}\mathbf{M}\|$	$\|\mathbf{t}\mathbf{t}^\mathsf{T} + 2\mathbf{S}\|$
MNIST data	172.4	86.0
EEG data	1.2	182.7

(handwritten digits recognition) and covariances-dominant (EEG times series classification) tasks. This is numerically confirmed in Table 3.5 (see also Table 3.1 from Section 3.1.1).

We perform random feature-based spectral clustering on data matrices that consist of $n = 32, 64$ and 128 randomly selected vectorized images of size $p = 784$ from the MNIST dataset. Means and covariances are empirically obtained from the full set of 11 769 MNIST images (5 918 images of number 6 and 5 851 of number 8). Comparing the matrix $\mathbf{K}_c$ built from the data and the theoretically equivalent $\tilde{\mathbf{K}}_c$ obtained as if the data were Gaussian with the (empirically) computed means and covariances, we observe an extremely close fit in the behavior of the eigenvalues and the leading eigenvector in Figure 3.18. The k-means method is then applied to the leading two eigenvectors of the Gram matrix $\frac{1}{n}\boldsymbol{\Sigma}^\mathsf{T}\boldsymbol{\Sigma}$ that consists of $N = 32$ random features to perform unsupervised classification, with resulting accuracies (averaged over 50 runs) reported in Table 3.6. As seen from Table 3.5, the mean-oriented $\sigma(t)$ functions are expected to outperform the covariance-oriented functions in this task, which is consistent with the results in Table 3.6.

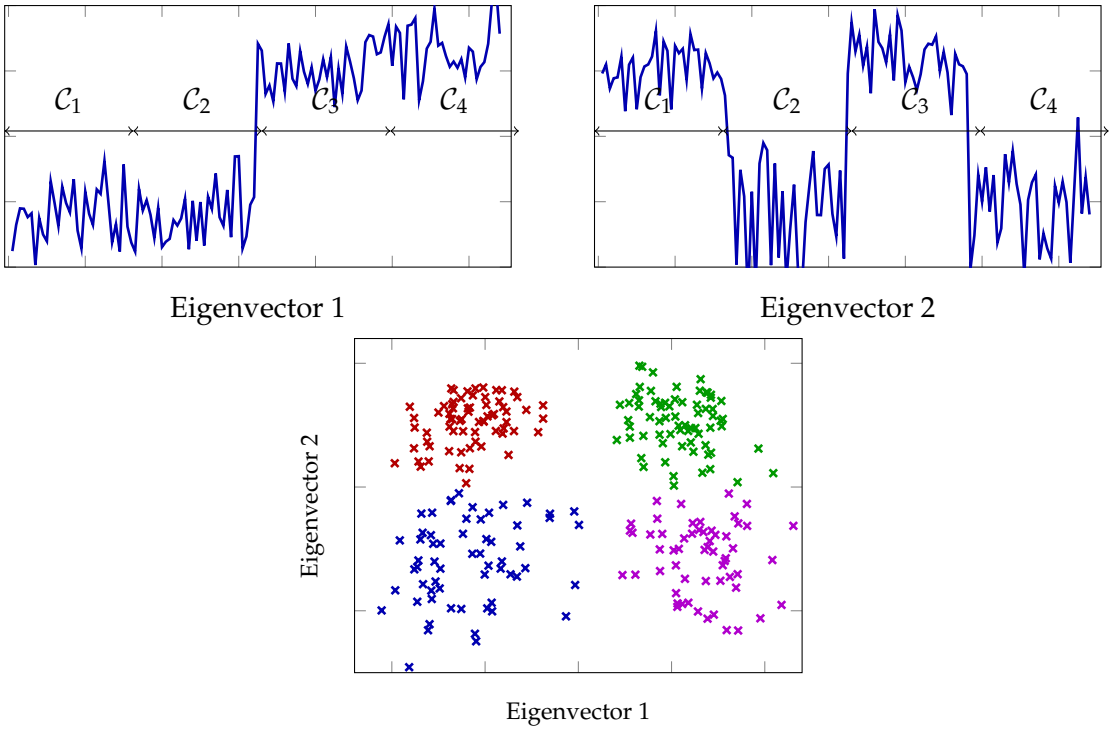


Figure 3.17: Leading two eigenvectors of $\mathbf{K}_c$ (top) for the LReLU function with $\varsigma_+ = 1$, $\varsigma_- = 0$ (equivalent to ReLU(t)) and two dimensional representation of these eigenvectors (bottom), in the same setting as in Figure 3.16.

Table 3.6: Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the MNIST dataset.

	$\sigma(t)$	$n = 32$	$n = 64$	$n = 128$
mean-oriented	t	85.31%	**88.94**%	87.30%
	$1_{t>0}$	86.00%	82.94%	85.56%
	$\text{sign}(t)$	81.94%	83.34%	85.22%
	$\sin(t)$	85.31%	87.81%	**87.50**%
	$\text{erf}(t)$	**86.50**%	87.28%	86.59%
cov-oriented	$\lvert t\rvert$	62.81%	60.41%	57.81%
	$\cos(t)$	62.50%	59.56%	57.72%
	$\exp(-t^2/2)$	64.00%	60.44%	58.67%
balanced	$\text{ReLU}(t)$	82.87%	85.72%	82.27%

The epileptic EEG dataset[8], developed by the University of Bonn, Germany, is described in [ALM+01]. The dataset consists of five subsets (denoted A-E), each containing 100 single-channel EEG segments of 23.6-sec duration. Sets A and B were collected from surface EEG recordings of five healthy volunteers, while sets C, D and E were collected from the EEG records of the pre-surgical diagnosis of five epileptic patients. Here we perform random feature-based spectral clustering on $n = 32, 64$ and 128 randomly picked EEG segments of length $p = 100$ from the dataset. Means and covariances are empir-

[8] `http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html`.

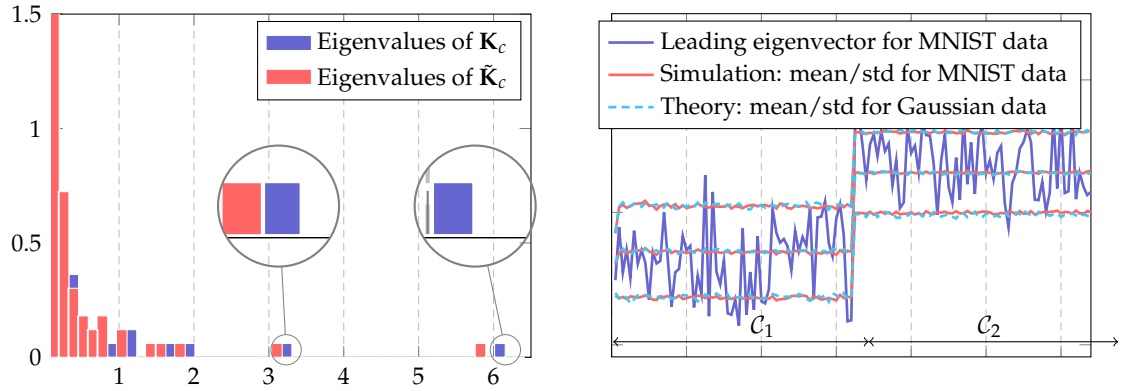


Figure 3.18: Eigenvalue distribution of $\mathbf{K}_c$ and $\tilde{\mathbf{K}}_c$ for the MNIST data (left) and leading eigenvector of $\mathbf{K}_c$ for the MNIST and Gaussian mixture data (right) with a width of ± 1 standard deviations (generated from 500 trials). With the ReLU function, $p = 784, n = 128$ and $c_1 = c_2 = 1/2, \mathbf{j}_1 = [\mathbf{1}_{n_1}; \mathbf{0}_{n_2}]$ and $\mathbf{j}_2 = [\mathbf{0}_{n_1}; \mathbf{1}_{n_2}]$.

ically estimated from the full set (4 097 segments of set B and 4 097 segments of set E). Similar behavior of eigenpairs as for Gaussian mixture models is once more observed in Figure 3.19. After k-means classification on the leading two eigenvectors of the (centered) Gram matrix composed of $N = 32$ random features, the accuracies (averaged over 50 runs) are reported in Table 3.7.

As opposed to the MNIST image recognition task, from Table 3.7 it is easy to check that the covariance-oriented functions (i.e., $\sigma(t) = |t|$, $\cos(t)$ and $\exp(-t^2/2)$) far outperform any other with almost perfect classification accuracies. It is particularly interesting to note that the popular ReLU function is suboptimal in both tasks, but never performs very badly, thereby offering a good risk-performance tradeoff.

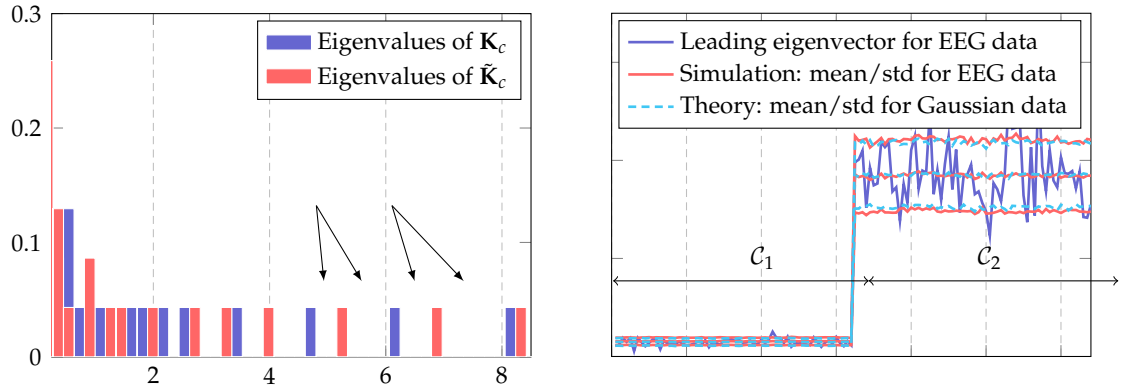


Figure 3.19: Eigenvalue distribution of $\mathbf{K}_c$ and $\tilde{\mathbf{K}}_c$ for the epileptic EEG data (left) and leading eigenvector of $\mathbf{K}_c$ for the EEG and Gaussian mixture data (right) with a width of ± 1 standard deviations (generated from 500 trials). With the ReLU function, $p = 100, n = 128$ and $c_1 = c_2 = 1/2, \mathbf{j}_1 = [\mathbf{1}_{n_1}; \mathbf{0}_{n_2}]$ and $\mathbf{j}_2 = [\mathbf{0}_{n_1}; \mathbf{1}_{n_2}]$.

Conclusion. In this section, by leveraging the randomness of the data, we dived deep into the comprehension of the equivalent kernel matrix $\mathbf{K} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{X}^{\mathsf{T}}\mathbf{w})\sigma(\mathbf{w}^{\mathbf{X}})]$ that appears in the random NN context (in Section 3.2.1) as well as in random feature-based techniques. The same technical approach as in Section 3.1.1 was applied to handle the

Table 3.7: Classification accuracies for random feature-based spectral clustering with different $\sigma(t)$ on the epileptic EEG dataset.

	$\sigma(t)$	$n = 32$	$n = 64$	$n = 128$
mean-oriented	t	71.81%	70.31%	69.58%
	$1_{t>0}$	65.19%	65.87%	63.47%
	$\text{sign}(t)$	67.13%	64.63%	63.03%
	$\sin(t)$	71.94%	70.34%	68.22%
	$\text{erf}(t)$	69.44%	70.59%	67.70%
cov-oriented	$\lvert t\rvert$	99.69%	99.69%	99.50%
	$\cos(t)$	99.00%	99.38%	99.36%
	$\exp(-t^2/2)$	**99.81**%	**99.81**%	**99.77**%
balanced	$\text{ReLU}(t)$	84.50%	87.91%	90.97%

nonlinearity arising from the activation function $\sigma(\cdot)$. Despite the difference in data models and technical approaches, we reached a similar conclusion on the classification of different σ as in Section 3.1.2, i.e., two key parameters (d_1, d_2) control separately the statistical information in means and covariance. Experiments on typical real-world datasets were performed to validate our theoretical arguments.

3.3 Empirical Risk Minimization of Convex Loss

In Section 3.1 and 3.2 we discussed both large random kernel matrices (and kernel ridge regression build on top of it) as well as random NN/feature-based models that are closely connected to one another. However, all aforementioned objects of interest are in these machine learning methods assume closed forms, as they come in essence from the minimization of (regularized) square losses. More generally, almost all machine learning algorithms are given in form of (solutions of) optimization problems, with most of them expressed only in an implicit manner. An important example is the popular logistic regression, where one aims to find an optimal decision vector $\boldsymbol{\beta} \in \mathbb{R}^p$ by minimizing the logistic loss $\frac{1}{n}\sum_{i=1}^n \log(1 + e^{-\tilde{y}_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i})$ over the training set $\{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ with labels $\tilde{y}_i \in \{-1, +1\}$. Different from the linear regression solution that arises from the minimization of square loss, the logistic regression solution $\boldsymbol{\beta}$ takes an implicit form and it is less direct to understand how $\boldsymbol{\beta}$ depends on the data $\mathbf{X}$ so as to investigate its statistical behavior. The technical challenge from implicit optimization problems appears not only in the analysis of logistic regression, but also in many other machine learning algorithms in daily use.

In this regard, it is of crucial importance to adapt the proposed random matrix-based analysis framework to assess the performance of optimization-based learning methods, we evaluate here the classifier obtained by minimizing an arbitrary convex and differentiable loss, with a major emphasis on the "leave-one-out" technical approach to "decouple" the learning system that contains complicated dependences.

In this section, we consider the following (regularized) empirical risk minimization problem (as in (1.22))

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n L(\tilde{y}_i \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{3.28}$$

under a binary symmetric GMM for the input data, i.e.,

$$\begin{cases} \mathcal{C}_1: & \mathbf{x}_i = -\boldsymbol{\mu} + \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i, \quad y_i = -1 \\ \mathcal{C}_2: & \mathbf{x}_i = +\boldsymbol{\mu} + \mathbf{C}^{\frac{1}{2}}\mathbf{z}_i, \quad y_i = +1 \end{cases} \tag{3.29}$$

each with a class prior of $1/2$, for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite $\mathbf{C} \in \mathbb{R}^{p\times p}$.

Let us first consider the case where $\lambda > 0$ and the existence of the minimizer $\boldsymbol{\beta}$ is guaranteed. Similar to the non-trivial classification condition in Assumption 2, we shall position ourselves under the following assumption.

Assumption 8 (Non-trivial classification)**.** *As $n \to \infty$, we have $p/n \to \bar{c} \in (0, \infty)$, $\|\boldsymbol{\mu}\| = O(1)$ and $\max\{\|\mathbf{C}\|, \|\mathbf{C}^{-1}\|\} = O(1)$ with respect to p.*

Note that the GMM in (3.29) satisfies the logistic regression model in the sense that the conditional class probability is given by

$$P(y_i = 1 \mid \mathbf{x}_i \in \mathcal{C}_2) = \frac{1}{1 + \exp(-2\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\mathbf{x}_i)} \equiv \sigma(\boldsymbol{\beta}_*^\mathsf{T}\mathbf{x}_i)$$

with $\sigma(t) = (1 + e^{-t})^{-1}$ the *logistic sigmoid* function and the optimal Bayes solution $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$. With the GMM in (3.29) and labels $y_i \in \{-1, +1\}$, it is convenient to denote the shortcut $\tilde{\mathbf{x}}_i \equiv \tilde{\mathbf{y}}_i\mathbf{x}_i$ so that

$$\tilde{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$

regardless of the class to which $\mathbf{x}_i$ belongs.

To investigate the asymptotic performance of the empirical risk-based classifier in (3.28), it is of crucial importance to understand the statistical properties of $\boldsymbol{\beta}$ that minimizes (3.28). The main technical difficulty of this analysis lies in the fact that $\boldsymbol{\beta}$, as the solution of a convex optimization problem, does not have an explicit form. Nonetheless, by canceling the loss function derivative with respect to $\boldsymbol{\beta}$ we obtain the following implicit relation

$$\lambda\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n} -L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i)\tilde{\mathbf{x}}_i \tag{3.30}$$

where we assume the loss function L is convex and at least three-times differentiable. In this section, in addition to RMT techniques, in order to handle the implicit nature of $\boldsymbol{\beta}$, we mainly focus on the additional "leave-one-out" tool to handle the complex dependences in the optimization problem.

Main results. From (3.30), $\boldsymbol{\beta}$ can be seen as a linear combination of all $\tilde{\mathbf{x}}_i$'s, weighted by the coefficient $-L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i)$. The idea is to understand how $\tilde{\mathbf{x}}_i$ (and its statistical properties) affects the corresponding coefficient $-L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i)$. However, as a solution of (3.30), $\boldsymbol{\beta}$ depends on all $\tilde{\mathbf{x}}_i$'s in an intricate manner. We handle this correlation by establishing a "leave-one-out" version of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}_{-i}$, that is asymptotically close to $\boldsymbol{\beta}$ and independent of $\tilde{\mathbf{x}}_i$, by solving (3.28) for all data $\tilde{\mathbf{x}}_j$ different from $\tilde{\mathbf{x}}_i$, i.e., for $j = 1, \ldots, i-1, i+1, \ldots, n$.

In details, since by definition $\lambda\boldsymbol{\beta}_{-i} = -\frac{1}{n}\sum_{j\neq i} L'(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j)\tilde{\mathbf{x}}_j$, the difference $\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i})$ is given by

$$\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}) = \frac{1}{n}\sum_{j\neq i}\left(L'(\boldsymbol{\beta}_{-i}^\mathsf{T}\tilde{\mathbf{x}}_j) - L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_j)\right)\tilde{\mathbf{x}}_j - \frac{1}{n}L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i)\tilde{\mathbf{x}}_i \tag{3.31}$$

Intuitively speaking, under Assumption 8, the difference $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}\|$ should be small, in the sense that it goes to 0 as $n, p \to \infty$. As such, by a Taylor expansion of $L'(x)$ around $x = \boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_j$, $j \neq i$, we obtain

$$L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_j) = L'(\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_j) + L''(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_j)(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i})^\mathsf{T} \tilde{\mathbf{x}}_j + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}\|^2).$$

This approximation, together with (3.31), leads to the following relation on $\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}$

$$\lambda(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}) \simeq -\frac{1}{n} \sum_{j \neq i} L''(\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_j) \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^\mathsf{T} (\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}) - \frac{1}{n} L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \tilde{\mathbf{x}}_i$$

or in matrix form

$$\left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D}_{-i} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p \right) (\boldsymbol{\beta} - \boldsymbol{\beta}_{-i}) \simeq -\frac{1}{n} L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \tilde{\mathbf{x}}_i$$

with $\mathbf{D}_{-i} \equiv \operatorname{diag}\{L''(\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_j)\}_{j \neq i}^n \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\mathbf{X}_{-i} \in \mathbb{R}^{p \times (n-1)}$ the data matrix without $\mathbf{x}_i$.

Note here that by convexity of L, $L''(t) \geq 0$ for all t, so that with $\lambda > 0$, the matrix $\left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p\right)$ is invertible and we get

$$\boldsymbol{\beta} - \boldsymbol{\beta}_{-i} \simeq -\frac{1}{n} L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D}_{-i} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p \right)^{-1} \tilde{\mathbf{x}}_i$$

so that its projection on $\tilde{\mathbf{x}}_i$ gives

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i})^\mathsf{T} \tilde{\mathbf{x}}_i \simeq -\frac{1}{n} L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \tilde{\mathbf{x}}_i^\mathsf{T} \left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D}_{-i} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p \right)^{-1} \tilde{\mathbf{x}}_i.$$

Note that $\left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p\right)^{-1}$ is of bounded operator norm and independent of $\tilde{\mathbf{x}}_i$. Following the idea of Lemma 2.11 we deduce the following approximation.

Lemma 3.4 (Asymptotic approximation of quadratic forms)**.** *Let Assumption 8 holds. Then as $n, p \to \infty$,*

$$\frac{1}{n} \tilde{\mathbf{x}}_i^\mathsf{T} \left(\frac{1}{n} \tilde{\mathbf{X}}_{-i} \mathbf{D}_{-i} \tilde{\mathbf{X}}_{-i}^\mathsf{T} + \lambda \mathbf{I}_p \right)^{-1} \tilde{\mathbf{x}}_i - \kappa \xrightarrow{a.s.} 0$$

where κ is the unique positive solution of $\kappa = \frac{1}{n} \operatorname{tr}(\bar{\mathbf{Q}} \mathbf{C})$ with

$$\bar{\mathbf{Q}} \equiv \left(\mathbb{E}\left[\frac{L''(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}})}{1 + \kappa L''(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}})} \right] \mathbf{C} + \lambda \mathbf{I}_p \right)^{-1}.$$

As a consequence of Lemma 3.4, one has immediately

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_{-i})^\mathsf{T} \tilde{\mathbf{x}}_i \simeq -L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \kappa$$

and therefore

$$\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i = \boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_i + (\boldsymbol{\beta} - \boldsymbol{\beta}_{-i})^\mathsf{T} \tilde{\mathbf{x}}_i \simeq \boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_i - L'(\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i) \kappa$$

where we observe the quantity $\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i$ on both sides, which is given by the explicit equation

$$\boldsymbol{\beta}^\mathsf{T} \tilde{\mathbf{x}}_i \simeq \operatorname{prox}_{\kappa L}(\boldsymbol{\beta}_{-i}^\mathsf{T} \tilde{\mathbf{x}}_i)$$

with $\text{prox}_{\kappa L}(t)$ the *proximal operator* [BC11] defined as the *unique* solution of the following minimization problem

$$\min_{x\in\mathbb{R}} \kappa L(x) + \frac{1}{2}(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i - x)^2.$$

The uniqueness of the minimization problem above is guaranteed by the convexity of the loss L. Moreover, the proximal operator can be defined for all lower semi-continuous convex functions, and thus covers non-smooth loss $L(t)$ such as the hinge losses $L(t) = \max(1-t, 0)$. In this vein, our results are envisioned to extend to non-smooth losses, for which more refinements (to replace the gradient with the subgradient in the sense of Clarke [Cla90] and to rework on Lemma 3.4) are needed.

Since $\boldsymbol{\beta}_{-i}$ is independent of $\tilde{\mathbf{x}}_i$, we have $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i \sim \mathcal{N}(m, \sigma^2)$ in the large p limit for some unknown deterministic m and σ^2, as summarized in the following lemma.

Lemma 3.5. *Under Assumption 8, there exist two positive constants m, σ^2 such that, as $n, p \to \infty$,*

$$\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i - r \xrightarrow{d} 0, \quad r \sim \mathcal{N}(m, \sigma^2)$$

and

$$L'(\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{x}}_i) - f(r) \xrightarrow{d} 0, \quad f(r) \equiv L'(\text{prox}_{\kappa L}(r)) = \frac{r - \text{prox}_{\kappa L}(r)}{\kappa}$$

with the proximal operator $\text{prox}_{\kappa L}(r)$ defined as the unique *solution of the following minimization problem*

$$\min_{x\in\mathbb{R}} \kappa L(x) + \frac{1}{2}(x-r)^2$$

for some $\kappa > 0$ determined by the fixed point equation in Lemma 3.4.

With Lemma 3.5, to characterize the stochastic behavior of $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ and $\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ (and subsequently $\boldsymbol{\beta}$), it remains to determine the unknown constants m and σ^2. This can be done by taking the expectation on both sides of (3.30) as

$$\lambda\mathbb{E}[\boldsymbol{\beta}] = \frac{1}{n}\sum_{i=1}^{n} -\mathbb{E}\left[f(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)\tilde{\mathbf{x}}_i\right]$$

which helps connect the statistics of $\boldsymbol{\beta}$ to m and σ^2. Nonetheless, the right-hand-side expectation in the above equation is non-trivial, due to the dependence between the random vector $\tilde{\mathbf{x}}_i$ and its projection $\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}_{-i}$ onto the independent $\boldsymbol{\beta}_{-i}$.

Fortunately, in the case of Gaussian $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, this dependence can be separated and treated explicitly. By writing $\tilde{\mathbf{x}}_i = \boldsymbol{\mu} + \sqrt{p}\boldsymbol{\omega}_i$ with $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}/p)$ we can decompose $\boldsymbol{\omega}_i$ as the following sum

$$\boldsymbol{\omega}_i = \frac{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i}{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\mathbf{C}\boldsymbol{\beta}_{-i}}\mathbf{C}\boldsymbol{\beta}_{-i} + \boldsymbol{\omega}_i^{\perp} \simeq \frac{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i}{\text{tr}\,\mathbf{C}\mathbb{E}[\boldsymbol{\beta}_{-i}\boldsymbol{\beta}_{-i}^{\mathsf{T}}]}\mathbf{C}\boldsymbol{\beta}_{-i} + \boldsymbol{\omega}_i^{\perp} \simeq \frac{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i}{\sigma^2}\mathbf{C}\boldsymbol{\beta}_{-i} + \boldsymbol{\omega}_i^{\perp}$$

for σ^2 defined in Lemma 3.4 and such that $\boldsymbol{\omega}_i^{\mathsf{T}}\boldsymbol{\beta}_{-i}$ is independent of $\boldsymbol{\omega}_i^{\perp}$, i.e.,

$$\mathbb{E}_{\boldsymbol{\omega}_i}\left[\boldsymbol{\omega}_i^{\perp}\boldsymbol{\omega}_i^{\mathsf{T}}\boldsymbol{\beta}_{-i}\right] = \mathbb{E}_{\boldsymbol{\omega}_i}\left[\left(\boldsymbol{\omega}_i - \frac{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i}{\boldsymbol{\beta}_{-i}^{\mathsf{T}}\mathbf{C}\boldsymbol{\beta}_{-i}}\mathbf{C}\boldsymbol{\beta}_{-i}\right)\boldsymbol{\omega}_i^{\mathsf{T}}\boldsymbol{\beta}_{-i}\right] = \mathbf{0}$$

and in particular $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i^{\perp} = 0$. As a consequence, (3.30) can be reduced to

$$\lambda\boldsymbol{\beta} \simeq \frac{1}{n}\sum_{i=1}^{n} -f(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)\left(\boldsymbol{\mu} + \frac{\sqrt{p}\boldsymbol{\beta}_{-i}^{\mathsf{T}}\boldsymbol{\omega}_i}{\sigma^2}\mathbf{C}\boldsymbol{\beta}_{-i} + \sqrt{p}\boldsymbol{\omega}_i^{\perp}\right)$$

which, since $\boldsymbol{\beta}_{-i} \simeq \boldsymbol{\beta}$, further leads to

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)(\tilde{\mathbf{x}}_i-\boldsymbol{\mu})}{\sigma^2}\mathbf{C}+\lambda\mathbf{I}_p\right)\boldsymbol{\beta} \simeq \left(\frac{1}{n}\sum_{i=1}^{n}-f(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)\right)\boldsymbol{\mu}+\mathbf{u}$$

for the random vector $\mathbf{u} \equiv \frac{1}{n}\sum_{i=1}^{n} -f(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)\sqrt{p}\boldsymbol{\omega}_i^{\perp}$ such that

$$\mathbb{E}[\mathbf{u}]=\mathbf{0}, \quad \mathbb{E}[\mathbf{u}\mathbf{u}^{\mathsf{T}}] = \frac{\mathbb{E}[f^2(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i)]}{n}\left(\mathbf{C}-\frac{1}{\sigma^2}\mathbf{C}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}}]\mathbf{C}\right).$$

This allows us to conclude that

$$\left(\frac{\mathbb{E}[f(r)(r-m)]}{\sigma^2}\mathbf{C}+\lambda\mathbf{I}_p\right)\boldsymbol{\beta} \simeq \mathbb{E}[-f(r)]\boldsymbol{\mu}+\mathbf{u}$$

or

$$\boldsymbol{\beta} \simeq (\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}+(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{u}$$

where we denote $\eta \equiv \mathbb{E}[-f(r)]$, $\gamma \equiv \mathbb{E}[f^2(r)]$ and $\tau \equiv \mathbb{E}[f(r)(r-m)/\sigma^2]$ for $r \sim \mathcal{N}(m,\sigma^2)$ as defined in Lemma 3.4.

Lastly, with

$$\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}}] \simeq \eta^2(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}+(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbb{E}[\mathbf{u}\mathbf{u}^{\mathsf{T}}](\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}$$

we deduce

$$\sigma^2 = \operatorname{tr}\mathbf{C}\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}}] \simeq \eta^2\boldsymbol{\mu}^{\mathsf{T}}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{C}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}+\frac{\gamma}{n}\|(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{C}\|_F^2$$

and finally reach the following theorem.

Theorem 3.8 (Asymptotic behavior of $\boldsymbol{\beta}$)**.** *Let Assumption 8 hold. Then, under the notations of Lemma 3.4, we have, as $n, p \to \infty$*

$$\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\| \to 0, \quad (\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\tilde{\boldsymbol{\beta}} \sim \mathcal{N}(\eta\boldsymbol{\mu}, \gamma\mathbf{C}/n)$$

with $(\eta,\gamma,\tau) \in \mathbb{R}^3$ the unique solution of

$$\eta \equiv \mathbb{E}[-f(r)], \quad \gamma \equiv \mathbb{E}[f^2(r)], \quad \tau \equiv \mathbb{E}[f(r)(r-m)/\sigma^2] = \mathbb{E}[-f'(r)]$$

for $f(r) \equiv L'(\operatorname{prox}_{\kappa L}(r))$ defined in Lemma 3.5, $r \sim \mathcal{N}(m,\sigma^2)$ with

$$m = \eta\boldsymbol{\mu}^{\mathsf{T}}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}, \quad \sigma^2 = \eta^2\boldsymbol{\mu}^{\mathsf{T}}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{C}(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}+\frac{\gamma}{n}\|(\tau\mathbf{C}+\lambda\mathbf{I}_p)^{-1}\mathbf{C}\|_F^2.$$

As a direct consequence of Theorem 3.8 and Lemma 3.5, we obtain the following corollary on the (asymptotic) training and test performance.

Corollary 3.3. *Under the conditions and notations of Theorem 3.8, the asymptotic test classification error rate is given by*

$$P(\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i < 0) - Q\left(\frac{m}{\sigma}\right) \to 0$$

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

$$P(\boldsymbol{\beta}^{\mathsf{T}}\tilde{\mathbf{x}}_i < 0) - P\left(\operatorname{prox}_{\kappa L}(r) < 0\right) \to 0$$

for $r \sim \mathcal{N}(m,\sigma^2)$ as in Lemma 3.5.

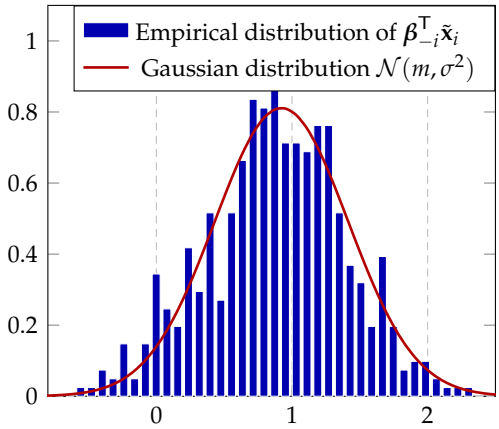


Figure 3.20: Comparison between the histogram of $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ and the Gaussian distribution $\mathcal{N}(m,\sigma^2)$ defined in Theorem 3.8 with $\boldsymbol{\mu} = [2,\ \mathbf{0}_{p-1}]$, $\mathbf{C}_p = \mathbf{I}_p$, for $\lambda = 1$, $p = 256$ and $n = 512$.

Practical consequences. To validate Theorem 3.8 and Lemma 3.5 for n, p of reasonable sizes, we compare in Figure 3.20 the empirical distribution of $\boldsymbol{\beta}_{-i}^{\mathsf{T}}\tilde{\mathbf{x}}_i$ and the Gaussian distribution $\mathcal{N}(m,\sigma^2)$ from the system of fixed point equations in Theorem 3.8. Our theoretical results fit the simulation almost perfectly, already for $p = 256$ and $n = 512$.

To interpret Theorem 3.8 we first restrict ourselves to the unregularized case where $\lambda = 0$ and assume the minimization problem (3.28) admits a solution $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\| = O(1)$, at least with high probability. Note that depending on the problem ($\boldsymbol{\mu}, \mathbf{C}$ and the loss L) and the dimensionality (n, p), such a solution may not exist with $\lambda = 0$ (for instance as pointed out in [CS18] for the logistic regression $L(t) = \log(1 + e^{-t})$).

With $\lambda = 0$, the notations in Theorem 3.8 are reduced to

$$\tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\frac{\eta}{\tau}\mathbf{C}^{-1}\boldsymbol{\mu}, \frac{\gamma}{n\tau}\mathbf{I}_p\right)$$

with $m = \frac{\eta}{\tau}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu}$ and $\sigma^2 = \frac{\eta^2}{\tau^2}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu} + \frac{\gamma}{\tau^2}\frac{p}{n}$. Several remarks are in order:

1. $\mathbb{E}[\tilde{\boldsymbol{\beta}}]$ is aligned to the optimal Bayes solution $\boldsymbol{\beta}_* = 2\mathbf{C}^{-1}\boldsymbol{\mu}$. However, unlike in the regime where $n \to \infty$ and p fixed, $\tilde{\boldsymbol{\beta}}$ contains an additional zero mean Gaussian noise of covariance $\frac{\gamma}{n\tau}\mathbf{I}_p$, which results in a larger variability of the soft output $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}$ than classical asymptotics predict.

2. The scaling of $\tilde{\boldsymbol{\beta}}$ depends on the interplay between the data statistics $(\boldsymbol{\mu}, \mathbf{C})$, the problem dimensionality (n, p) and the loss function L. In general, $\mathbb{E}[\tilde{\boldsymbol{\beta}}]$ differs from the optimal Bayes solution $\boldsymbol{\beta}_*$ by a multiplicative (constant) factor, i.e., $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \frac{\eta}{2\tau}\boldsymbol{\beta}_*$. In Figure 3.21 we examine the empirical mean (so as to estimate the expectation) of $\boldsymbol{\beta}$, the rescaled version $\frac{2\tau}{\eta}\boldsymbol{\beta}$ and the optimal Bayes solution $\boldsymbol{\beta}_*$. We see that by rescaling the obtained $\boldsymbol{\beta}$ by the factor $2\tau/\eta$, one gets (in expectation) the optimal $\boldsymbol{\beta}_*$. Nonetheless, note that for classification applications, one has $\operatorname{sign}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}) = \operatorname{sign}(a\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})$ for $a > 0$, meaning that a positive rescaling of $\boldsymbol{\beta}$ does not affect the classification performance.

3. As a consequence of Corollary 3.3, in pursuit of an optimal design of the loss function L, one must find (η, γ, τ) that maximizes the ratio m^2/σ^2, or equivalently

$$\underset{(\eta,\gamma,\tau)\in\mathbb{R}^3}{\arg\max}\ \frac{m^2}{\sigma^2} = \underset{(\eta,\gamma,\tau)\in\mathbb{R}^3}{\arg\max}\ \frac{(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu})^2}{\boldsymbol{\mu}^{\mathsf{T}}\mathbf{C}^{-1}\boldsymbol{\mu} + \frac{\gamma}{\eta^2}\frac{p}{n}} = \underset{(\eta,\gamma,\tau)\in\mathbb{R}^3}{\arg\max}\ \frac{\eta^2}{\gamma}$$

where we recall $\eta = \mathbb{E}[-f(r)]$ and $\gamma = \mathbb{E}[f^2(r)]$, so that by Cauchy–Schwarz inequality

$$\eta^2 \leq \gamma$$

with equality if and only if $f(r)$ is a constant function. The minimal misclassification error rate is therefore given by $Q(\frac{m}{\sigma}) = Q\left(\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\mathbf{C}^{-1}\boldsymbol{\mu}+p/n}}\right)$ for the model under consideration. On the other hand, it is of interest to note that, by the law of large numbers, the empirical average $\frac{1}{n}\sum_{i=1}^n -L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{x}}_i)$ converges to its expectation η as $n \to \infty$, and similarly for γ. As such, one may alternatively wish to maximize the following empirical version

$$\max_{(\eta,\gamma,\tau)\in\mathbb{R}^3} \frac{L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{X}})\mathbf{1}_n}{\sqrt{L'(\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{X}})L'(\tilde{\mathbf{X}}^\mathsf{T}\boldsymbol{\beta})}}$$

where the function L' is applied entry-wise on the vector $\boldsymbol{\beta}^\mathsf{T}\tilde{\mathbf{X}} \in \mathbb{R}^n$. Under the above form, note from (3.30) that in the unregularized case with $\lambda = 0$ one must has

$$\tilde{\mathbf{X}}L'(\tilde{\mathbf{X}}^\mathsf{T}\boldsymbol{\beta}) = \mathbf{0}$$

so that by considering a singular value decomposition of $\tilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U}\begin{bmatrix}\mathbf{S} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^\mathsf{T} \\ \mathbf{V}_2^\mathsf{T}\end{bmatrix}$ for $\tilde{\mathbf{X}} \in \mathbb{R}^{p\times n}$ with $n > p$ such that $\mathbf{S} \in \mathbb{R}^{p\times p}$, $\mathbf{V}_1 \in \mathbb{R}^{n\times p}$ and $\mathbf{V}_2 \in \mathbb{R}^{n\times(n-p)}$, one must have $\mathbf{V}_1^\mathsf{T}L'(\tilde{\mathbf{X}}^\mathsf{T}\boldsymbol{\beta}) = \mathbf{0}$, or equivalently, $L'(\tilde{\mathbf{X}}^\mathsf{T}\boldsymbol{\beta})$ lies on the subspace spanned by the columns vectors of $\mathbf{V}_2$ such that there exists $\mathbf{a} \in \mathbb{R}^{n-p}$ for which

$$L'(\tilde{\mathbf{X}}^\mathsf{T}\boldsymbol{\beta}) = \mathbf{V}_2\mathbf{a}.$$

With the formulation, one aims to find

$$\arg\max_{(\eta,\gamma,\tau)\in\mathbb{R}^3} \frac{\mathbf{a}^\mathsf{T}\mathbf{V}_2^\mathsf{T}\mathbf{1}_n}{\|\mathbf{a}\|}$$

which attains the maximum if and only if $\mathbf{a}$ is aligned to $\mathbf{V}_2^\mathsf{T}\mathbf{1}_n$ $\mathbf{a} = a\mathbf{V}_2^\mathsf{T}\mathbf{1}_n$ for some $a > 0$. This optimality condition is met for instance with the square loss $L(t) = (t-1)^2$.

Conclusion. In this section, we considered the separation of a symmetric binary GMM with opposite means $\pm\boldsymbol{\mu}$ and identical covariance $\mathbf{C}$, by minimizing the empirical convex and differentiable risk L on a given training set. In this scenario, the statistical of the resulting classifier $\boldsymbol{\beta}$ depends on the chosen loss metric via the generalized moment of a nonlinear function $f(x) \equiv L'(\text{prox}_{\kappa L}(x))$ with respect to the standard Gaussian measure, in an implicit manner via the proximal operator and the constant $\kappa > 0$. As a consequence of the analysis, we saw that the in the regime $n, p \to \infty$ with $p/n \to \bar{c} \in (0, \infty)$, $\boldsymbol{\beta}$ is the sum of a rescaled of $\boldsymbol{\beta}_*$, the optimal Bayes solution, together with an additional homogeneous Gaussian noise.

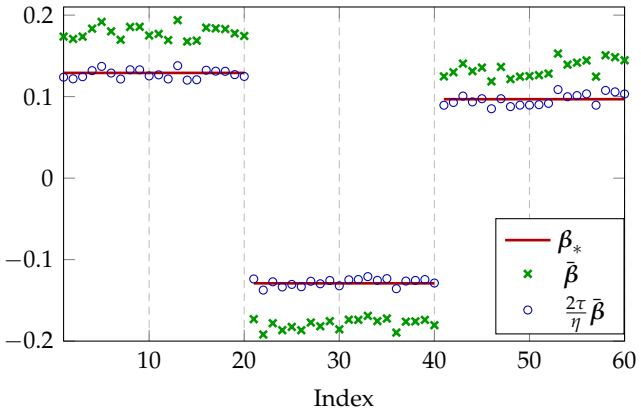


Figure 3.21: Comparison of the empirical mean of $\boldsymbol{\beta}$ (denoted $\bar{\boldsymbol{\beta}}$) the solution of (1.22) for logistic loss $L(t) = \log(1 + e^{-t})$, the optimal Bayes solution $\boldsymbol{\beta}_*$ and the rescaled solution $\frac{2\tau}{\eta}\bar{\boldsymbol{\beta}}$ for (η, τ) derived from Theorem 3.8 with $\boldsymbol{\mu} = [\mathbf{1}_{p/3}, -\mathbf{1}_{p/3}, \frac{3}{4}\mathbf{1}_{p/3}]/\sqrt{p}$, $\mathbf{C} = 2\mathbf{I}_p$, for $p = 60$, $n = 300$. Empirical means obtained by averaging over 500 realizations.

3.4 Summary of Chapter 3

In this chapter, we focused on the eigenspectrum of large dimensional random kernel matrices and neural networks. Built upon a simple Gaussian mixture modeling of the input data, we provided a tractable characterization of the spectral behavior of kernel matrices and nonlinear Gram matrices, from which we further deduced the (asymptotic) performance of various machine learning methods, as a function of the data statistics, the dimensionality, and the hyperparameters of the algorithm. The presented theoretical findings shed novel light on the understanding, as well as a better designing of the aforementioned machine learning methods.

To further address more involved learning algorithms, for example, those (implicitly) defined by convex optimization problems (e.g., logistic regression), we resort to a "leave-one-out" approach and assume that the optimality conditions are met. By working on these conditions, we deduced the (asymptotic) performance reached by the optimal points of the these optimization problems.

In practice, optimization methods such a gradient descent are regularly used to reach these optimal points. The solution of the optimization problem, and therefore its performance, naturally depends on the optimization method applied. The impact of the optimization method is more significant in non-convex problems, where depending on the initialization, totally different solutions can be reached. In the following chapter, we focus precisely on the gradient descent algorithm and study its behavior in convex and non-convex machine learning problems.

Chapter 4

Gradient Descent Dynamics in Neural Networks

In Chapter 3 we discussed either kernel-based methods that are of explicit forms (in Section 3.1.1–3.2.2), or the statistical behavior of the unique minimum from a convex optimization problem (in Section 3.3), without considering by which means can we reach this optimum. In this chapter, we discuss the possibly most widely used gradient descent method, and in particular, the temporal evolution of (the performance of) the learning system when trained with gradient descent, starting from the simple and convex linear regression model in the section below.

4.1 A Random Matrix Approach to GDD

In the section we consider the gradient descent dynamics (GDDs) for the training of a ridge regression, under a binary symmetric GMM (see Definition 1) with identity covariance for the input pattern, i.e.,

$$\begin{cases} \mathcal{C}_1: & \mathbf{x}_i = -\boldsymbol{\mu} + \mathbf{z}_i, \quad y_i = -1 \\ \mathcal{C}_2: & \mathbf{x}_i = +\boldsymbol{\mu} + \mathbf{z}_i, \quad y_i = +1 \end{cases}$$

for $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\|\boldsymbol{\mu}\| = O(1)$ as $n, p \to \infty$ with $p/n = c \to \bar{c} \in (0, \infty)$.

Let us detail our basic settings: for a given training data matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \in \mathbb{R}^{p \times n}$ with associated labels $\mathbf{y} = \left[y_1, \ldots, y_n\right] \in \mathbb{R}^n$, a weight vector $\mathbf{w} \in \mathbb{R}^p$ is learned using gradient descent to minimize the square loss

$$L(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}^\mathsf{T} \mathbf{w}\|^2.$$

The gradient of L with respect to $\mathbf{w}$ is thus given by $\nabla_{\mathbf{w}} L(\mathbf{w}) = -\frac{1}{n} \mathbf{X}(\mathbf{y} - \mathbf{X}^\mathsf{T} \mathbf{w})$ so that with small gradient descent step (or, learning rate) α, we obtain, by performing a continuous-time approximation, the following differential equation

$$\frac{d\mathbf{w}(t)}{dt} = -\alpha \nabla_{\mathbf{w}} L(\mathbf{w}) = \frac{\alpha}{n} \mathbf{X}(\mathbf{y} - \mathbf{X}^\mathsf{T} \mathbf{w})$$

the solution of which is given explicitly by (as in (1.24))

$$\mathbf{w}(t) = e^{-\frac{\alpha t}{n} \mathbf{X}\mathbf{X}^\mathsf{T}} \mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n} \mathbf{X}\mathbf{X}^\mathsf{T}} \right) \mathbf{w}_{LS}$$

where we denote $\mathbf{w}_{LS} = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\frac{1}{n}\mathbf{X}\mathbf{y} \in \mathbb{R}^p$ the classical ridge regression solution with regularization parameter $\lambda \geq 0$ and $\mathbf{w}_0 = \mathbf{w}(t=0)$ the initialization of gradient descent. We recall that the exponential of a symmetric matrix $\mathbf{A}$ is given by $e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k = \mathbf{V}_\mathbf{A} e^{\mathbf{\Lambda}_\mathbf{A}} \mathbf{V}_\mathbf{A}^\mathsf{T}$, with $\mathbf{A} = \mathbf{V}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathsf{T}$ the spectral decomposition of $\mathbf{A}$. We apply, as proposed in [GB10], the following random initialization for gradient descent.

Assumption 9 (Random initialization)**.** *Let $\mathbf{w}_0 \equiv \mathbf{w}(t=0)$ be a random vector with i.i.d. entries of zero mean, variance σ^2/p for some $\sigma > 0$ and finite fourth moment.*

Given the expression of $\mathbf{w}(t)$ above, we are interested in the training and test misclassification error rate, i.e.,

$$P(\mathbf{x}_i^\mathsf{T}\mathbf{w}(t) > 0 \mid y_i = -1), \quad P(\hat{\mathbf{x}}^\mathsf{T}\mathbf{w}(t) > 0 \mid \hat{y} = -1).$$

Due to the symmetry of the GMM under consideration, we denote, similar to Section 3.3, $\tilde{\mathbf{x}}_i \equiv y_i \mathbf{x}_i$ so that $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ and use, with a slight abuse of notations for test data $\hat{\mathbf{x}} = \tilde{\hat{\mathbf{x}}}$. As such, we further write

$$\mathbf{X}\mathbf{y} = \tilde{\mathbf{X}}\mathbf{1}_n, \quad \mathbf{X}\mathbf{X}^\mathsf{T} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}$$

so that the classification error rates become

$$P(\tilde{\mathbf{x}}_i^\mathsf{T}\mathbf{w}(t) < 0), \quad P(\hat{\mathbf{x}}^\mathsf{T}\mathbf{w}(t) < 0)$$

where $\hat{\mathbf{x}} \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p)$ is a new (test) datum independent of $\mathbf{X}$.

From a random matrix perspective, all aforementioned objects are functionals of the eigenvalues of the sample covariance matrix $\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}$ and can thus be evaluated with the resolvent-based technique discussed in Chapter 2.

Main results. Let us start with the test misclassification error rate. Since the new datum $\hat{\mathbf{x}}$ is independent of $\mathbf{w}(t)$, conditioned on $\mathbf{w}(t)$, $\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}}$ is a Gaussian random variable of mean $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}$ and variance $\|\mathbf{w}(t)\|^2$. The above probabilities can therefore be given via the Q-function of Gaussian distribution. We thus resort to the computation of $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}$ as well as $\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t)$ to evaluate the aforementioned classification error.

For $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$, with Cauchy's integral formula we have

$$\begin{aligned}
\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t) &= \boldsymbol{\mu}^\mathsf{T} e^{-\alpha t(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T})}\mathbf{w}_0 + \boldsymbol{\mu}^\mathsf{T}\left(\mathbf{I}_p - e^{-\alpha t(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T})}\right)\mathbf{w}_{LS} \\
&= -\frac{1}{2\pi i}\oint_\gamma f_t(z)\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\, dz - \frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z+\lambda}\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n\, dz
\end{aligned}$$

with $f_t(z) \equiv \exp(-\alpha t z)$ and γ a positive closed path circling around all eigenvalues of $\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}$. Note that the (unified) data matrix $\tilde{\mathbf{X}}$ can be rewritten as

$$\tilde{\mathbf{X}} = \boldsymbol{\mu}\mathbf{1}_n^\mathsf{T} + \mathbf{Z}$$

with $\mathbf{Z} \equiv \left[\mathbf{z}_1, \ldots, \mathbf{z}_n\right] \in \mathbb{R}^{p\times n}$ of i.i.d. $\mathcal{N}(0,1)$ entries. To isolate the deterministic vectors $\boldsymbol{\mu}$ from the random $\mathbf{Z}$ in the expression of $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$, we exploit Woodbury's identity to obtain

$$\begin{aligned}
\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} &= \mathbf{Q}(z) \\
&\quad - \mathbf{Q}(z)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\boldsymbol{\mu} & 1+\frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{1}_n \\ * & -1+\frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Q}(z)\frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z).
\end{aligned}$$

where we denote the resolvent $\mathbf{Q}(z) \equiv \left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}$ of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$, a deterministic equivalent of which is given by

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) \equiv m(z)\mathbf{I}_p$$

with $m(z)$ determined by the popular Marčenko–Pastur equation [MP67]

$$m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2 - 4cz}}{2cz} \tag{4.1}$$

where the branch of the square root is selected in such a way that $\Im(z) \cdot \Im m(z) > 0$, i.e., for a given z there exists a *unique* corresponding $m(z)$.

Substituting $\mathbf{Q}(z)$ by the simple form deterministic equivalent $m(z)\mathbf{I}_p$, we are able to estimate the random variable $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$ with a contour integral of some deterministic quantities as $n, p \to \infty$. Similar arguments also hold for $\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t)$, together leading to the following theorem, the proof of which is deferred to Section A.2.1 in the Appendix.

Theorem 4.1 (Test performance)**.** *Let Assumptions 9 hold. As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one*

$$P(\mathbf{w}(t)^\mathsf{T}\hat{\mathbf{x}} < 0) - Q\left(\frac{E}{\sqrt{V}}\right) \to 0$$

where

$$E \equiv -\frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)\, dz}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1}$$
$$V \equiv \frac{1}{2\pi i}\oint_\gamma \left[\frac{\frac{1}{z^2}\left(1-f_t(z)\right)^2}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1} - \sigma^2 f_t^2(z) m(z)\right] dz$$

with γ a closed positively oriented contour that surrounds the support of Marčenko–Pastur law, the origin $(0,0)$ and the point $(\lambda_s, 0)$ with $\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2}$, $f_t(z) \equiv \exp(-\alpha t z)$ and $m(z)$ given by Equation (4.1)*.*

To compare test versus training performances, we now evaluate the training error by considering

$$\frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{X}}^\mathsf{T}\mathbf{w}(t) = -\frac{1}{2\pi i}\oint_\gamma f_t(z,t)\frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{X}}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\, dz$$
$$-\frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{X}}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n\, dz$$

which yields the following results.

Theorem 4.2 (Training performance)**.** *Under the assumptions and notations of Theorem 4.1, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$,*

$$P(\mathbf{w}(t)^\mathsf{T}\tilde{\mathbf{x}}_i < 0) - Q\left(\frac{E_*}{\sqrt{V_*}}\right) \to 0$$

almost surely, with

$$E_* \equiv \frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\frac{dz}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1}$$
$$V_* \equiv \frac{1}{2\pi i}\oint_\gamma \left[\frac{\frac{1}{z}\left(1-f_t(z)\right)^2}{(\|\boldsymbol{\mu}\|^2 + c)\, m(z) + 1} - \sigma^2 f_t^2(z) z m(z)\right] dz.$$

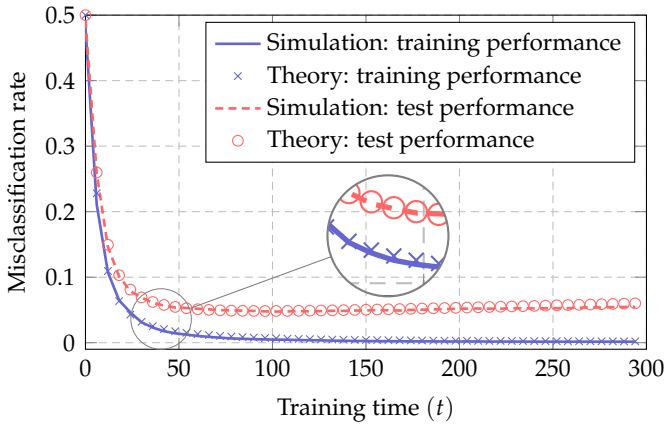


Figure 4.1: Training and test performance for $\boldsymbol{\mu} = [2; \mathbf{0}_{p-1}]$, $p = 256$, $n = 512$, $\sigma^2 = 0.1$, $\alpha = 0.01$ and $c_1 = c_2 = 1/2$. Results obtained by averaging over 50 runs.

The proofs of Theorem 4.1 and 4.2 are provided in Section A.2.1 of the Appendix.

In Figure 4.1 we compare finite dimensional simulations with theoretical results obtained from Theorem 4.1 and 4.2 and observe a very close match, already for not too large n, p. As t grows large, the test error first drops rapidly with the training error, then increases, although slightly, while the training error continues to decrease to zero. This is because the classifier eventually over-fits the training data $\mathbf{X}$ and performs badly on unseen data. To avoid over-fitting, one effectual approach is to apply regularization strategies [Bis07]; for example, to "early stop" (at $t = 100$ for instance in the setting of Figure 4.1) in the training process. This introduces new hyperparameters such as the optimal stopping time t_{opt} that is of crucial importance for the network performance and is often tuned through cross-validation in practice. Theorem 4.1 and 4.2 tell us that the training and test performances, although random themselves, have asymptotically deterministic behaviors described by (E_*, V_*) and (E, V), respectively, which allows for a deeper understanding on the choice of t_{opt}, since E, V are in fact functions of t via $f_t(z) \equiv \exp(-\alpha t(z + \lambda))$.

Nonetheless, the expressions in Theorem 4.1 and 4.2 of contour integrations are not easily analyzable nor interpretable. To gain more insight, we shall rewrite (E, V) and (E_*, V_*) in a more readable way. First, note from Figure 4.2 that the matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ has (possibly) two types of eigenvalues: those inside the *main bulk* (between $\lambda_- \equiv (1 - \sqrt{c})^2$ and $\lambda_+ \equiv (1 + \sqrt{c})^2$) of the Marčenko–Pastur distribution

$$\nu(dx) = \frac{\sqrt{(x - \lambda_-)^+(\lambda_+ - x)^+}}{2\pi c x} dx + (1 - c^{-1})^+ \delta(x) \tag{4.2}$$

and a (possibly) isolated one lying away from $[\lambda_-, \lambda_+]$, that will be treated separately. We rewrite the path γ (that contains all eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$) as the sum of two paths γ_b and γ_s, that circle around the main bulk and the isolated eigenvalue (if any), respectively. To handle the first integral of γ_b, we use the fact that for any nonzero $\lambda \in \mathbb{R}$, the limit $\lim_{z \in \mathbb{C} \to \lambda} m(z) \equiv \check{m}(\lambda)$ exists [SC95] and follow the idea in [BS08] by choosing the contour γ_b to be a rectangle with sides parallel to the axes, intersecting the real axis at 0 and λ_+ and the horizontal sides being at a distance $\varepsilon \to 0$ from the real axis, to split the contour integral into four real line integrals. The second integral circling around γ_s can be computed by residue calculus. This together leads to the expressions of (E, V) and

(E_*, V_*) as follows.

$$E = \int \frac{1 - f_t(x)}{x} \mu(dx) \tag{4.3}$$

$$V = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{(1 - f_t(x))^2 \mu(dx)}{x^2} + \sigma^2 \int f_t^2(x) \nu(dx) \tag{4.4}$$

$$E_* = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{1 - f_t(x)}{x} \mu(dx) \tag{4.5}$$

$$V_* = \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2} \int \frac{(1 - f_t(x))^2 \mu(dx)}{x} + \sigma^2 \int x f_t^2(x) \nu(dx) \tag{4.6}$$

where we recall $f_t(x) = \exp(-\alpha t x)$, $\nu(x)$ given by (4.2) and denote the measure

$$\mu(dx) \equiv \frac{\sqrt{(x - \lambda_-)^+ (\lambda_+ - x)^+}}{2\pi(\lambda_s - x)} dx + \frac{(\|\boldsymbol{\mu}\|^4 - c)^+}{\|\boldsymbol{\mu}\|^2} \delta_{\lambda_s}(x) \tag{4.7}$$

as well as

$$\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2} \geq (\sqrt{c} + 1)^2 \tag{4.8}$$

with equality if and only if $\|\boldsymbol{\mu}\|^2 = \sqrt{c}$.

A first remark on the expressions of (4.3)-(4.6) is that E_* differs from E only by a factor $\frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2}$. Also, both V and V_* are the sum of two parts: the first part that strongly depends on $\boldsymbol{\mu}$ and the second one that is independent of $\boldsymbol{\mu}$. One thus deduces for $\|\boldsymbol{\mu}\| \to 0$ that $E \to 0$ and

$$V \to \int \frac{(1 - f_t(x))^2}{x^2} \rho(dx) + \sigma^2 \int f_t^2(x) \nu(dx) > 0$$

with $\rho(dx) \equiv \frac{\sqrt{(x - \lambda_-)^+ (\lambda_+ - x)^+}}{2\pi(c+1)} dx$. Therefore the test performance goes to $Q(0) = 0.5$. On the other hand, for $\|\boldsymbol{\mu}\| \to \infty$, one has $\frac{E}{\sqrt{V}} \to \infty$ and hence the classifier makes perfect predictions.

In a more general context (i.e., for Gaussian mixture models with generic means and covariances as investigated in [BGC16], and obviously for practical datasets), there may be more than one eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ lying outside the main bulk, which may not be limited to the interval $[\lambda_-, \lambda_+]$. In this case, the expression of $m(z)$, instead of being explicitly given by (4.1), may be determined through more elaborate formulations as in the sample covariance model in Section 2.2.3. While handling more generic models is technically reachable within the present analysis scheme, the results are much less intuitive. Similar objectives cannot be achieved within the framework presented in [AS17]; this conveys more practical interest to our results and the proposed analysis framework.

Practical consequences. With a careful inspection of (4.3) and (4.4), we now discuss several different aspects of their practical implications.

- *On the test performance.* First of all, recall that the test performance is simply given by $Q\left(\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)}{\|\mathbf{w}(t)\|}\right)$, with the term $\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ describing the alignment between $\mathbf{w}(t)$ and $\boldsymbol{\mu}$. Therefore the best possible test performance is simply $Q(\|\boldsymbol{\mu}\|)$. Nonetheless, this "best" performance can never be achieved as long as $p/n \to c > 0$, as described in the following remark.

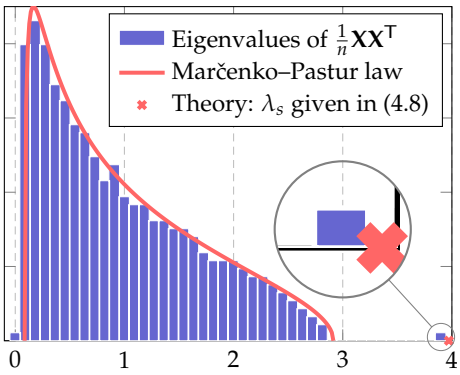


Figure 4.2: Eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\boldsymbol{\mu} = [3/2; \mathbf{0}_{p-1}]$, $p = 512$, $n = 1\,024$ and $c_1 = c_2 = 1/2$.

Remark 4.1 (Optimal test performance)**.** *Note that, with Cauchy–Schwarz inequality and the fact that* $\int \mu(dx) = \|\boldsymbol{\mu}\|^2$ *from* (4.7)*, one has*

$$E^2 \leq \int \frac{(1 - f_t(x))^2}{x^2} d\mu(x) \cdot \int d\mu(x) \leq \frac{\|\boldsymbol{\mu}\|^4}{\|\boldsymbol{\mu}\|^2 + c} V$$

with equality in the right-most inequality if and only if the variance $\sigma^2 = 0$*. One thus concludes that* $E/\sqrt{V} \leq \|\boldsymbol{\mu}\|^2/\sqrt{\|\boldsymbol{\mu}\|^2 + c}$ *and the best test performance (lowest misclassification rate) is* $Q(\|\boldsymbol{\mu}\|^2/\sqrt{\|\boldsymbol{\mu}\|^2 + c})$ *which can be reached only when* $\sigma^2 = 0$*. Note in particular that the optimal (test) classification performance agrees with the theoretical optimum derived under the empirical risk minimization framework in Section 3.3.*

This remark is of particular interest because, for a given task (thus for $p, \boldsymbol{\mu}$ fixed) it gives access to the *minimum* number of training data number n needed to fulfill a certain classification accuracy.

As a side remark, note that in the expression of $E/\sqrt{V}$ the initialization variance σ^2 only appears in V, meaning that random initializations impair the test performance of the network. As such, one should initialize with σ^2 very close, but not equal, to zero, to obtain symmetry breaking between hidden units [GBC16] as well as to mitigate the drop of performance due to large σ^2. In Figure 4.3 we plot the optimal test performance with the corresponding optimal stopping time as a function of σ^2, showing that small initializations help training both in terms of accuracy and efficiency.

- *Approximation for* t *close to* 0. Although the integrals in (4.3) and (4.4) do not have nice closed forms, note that, for t close to 0, with a Taylor expansion of $f_t(x) \equiv \exp(-\alpha t x)$ around $\alpha t x = 0$, one gets more interpretable forms of E and V without integrals, as presented next.

  Taking $t = 0$, one has $f_t(x) = 1$ and therefore $E = 0$, $V = \sigma^2 \int \nu(dx) = \sigma^2$, with $\nu(dx)$ the Marčenko–Pastur distribution given in (4.2). As a consequence, at the beginning stage of training, the test performance is $Q(0) = 0.5$ for $\sigma^2 \neq 0$ and the classifier makes random guesses.

  For t not equal but close to 0, the Taylor expansion of $f_t(x) \equiv \exp(-\alpha t x)$ around $\alpha t x = 0$ gives

$$f_t(x) \equiv \exp(-\alpha t x) \approx 1 - \alpha t x + O(\alpha^2 t^2 x^2).$$

Making the substitution $x = 1 + c - 2\sqrt{c}\cos\theta$ and with the fact that $\int_0^\pi \frac{\sin^2\theta}{p+q\cos\theta}d\theta = \frac{p\pi}{q^2}\left(1 - \sqrt{1-q^2/p^2}\right)$ (see for example 3.644-5 in [GR14]), one gets $E = \tilde{E} + O(\alpha^2 t^2)$ and $V = \tilde{V} + O(\alpha^2 t^2)$, where

$$\tilde{E} \equiv \frac{\alpha t}{2} g(\boldsymbol{\mu}, c) + \frac{(\|\boldsymbol{\mu}\|^4 - c)^+}{\|\boldsymbol{\mu}\|^2}\alpha t = \|\boldsymbol{\mu}\|^2 \alpha t$$

$$\begin{aligned}\tilde{V} \equiv & \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2}\frac{(\|\boldsymbol{\mu}\|^4 - c)^+}{\|\boldsymbol{\mu}\|^2}\alpha^2 t^2 + \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2}\frac{\alpha^2 t^2}{2} g(\boldsymbol{\mu}, c) \\ & + \sigma^2(1+c)\alpha^2 t^2 - 2\sigma^2 \alpha t + \left(1 - \frac{1}{c}\right)^+ \sigma^2 + \frac{\sigma^2}{2c}\left(1 + c - (1+\sqrt{c})|1-\sqrt{c}|\right) \\ = & (\|\boldsymbol{\mu}\|^2 + c + c\sigma^2)\alpha^2 t^2 + \sigma^2(\alpha t - 1)^2\end{aligned}$$

with $g(\boldsymbol{\mu}, c) \equiv \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2} - \left(\|\boldsymbol{\mu}\| + \frac{\sqrt{c}}{\|\boldsymbol{\mu}\|}\right)\left|\|\boldsymbol{\mu}\| - \frac{\sqrt{c}}{\|\boldsymbol{\mu}\|}\right|$ and consequently $\frac{1}{2}g(\boldsymbol{\mu}, c) + \frac{(\|\boldsymbol{\mu}\|^4 - c)^+}{\|\boldsymbol{\mu}\|^2} = \|\boldsymbol{\mu}\|^2$. It is interesting to note from the above calculation that, although E and V seem to have different behaviors for $\|\boldsymbol{\mu}\|^2 > \sqrt{c}$ or $c > 1$, it is in fact not the case as the extra term involving $\|\boldsymbol{\mu}\|^2 > \sqrt{c}$ (or $c > 1$) exactly compensates the singularity of the integral. As such, the test performance of the classifier is a smooth function of both $\|\boldsymbol{\mu}\|^2$ and c, and there is indeed no "phase transition" as far as performance is concerned.

Taking the derivative of $\frac{\tilde{E}}{\sqrt{\tilde{V}}}$ with respect to t, one has

$$\frac{\partial}{\partial t}\frac{\tilde{E}}{\sqrt{\tilde{V}}} = \frac{\alpha(1-\alpha t)\sigma^2}{\tilde{V}^{3/2}}$$

which implies that the maximum of $\frac{\tilde{E}}{\sqrt{\tilde{V}}}$ is $\frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2 + c + c\sigma^2}}$ and can be attained with $t = 1/\alpha$. Moreover, taking $t = 0$ in the above equation one gets $\frac{\partial}{\partial t}\frac{\tilde{E}}{\sqrt{\tilde{V}}}\big|_{t=0} = \frac{\alpha}{\sigma}$. Therefore, large σ is harmful to the training efficiency, which coincides with the conclusion from Remark 4.1.

Yet, the approximation error arising from Taylor expansion can be large for t away from 0, e.g., at $t = 1/\alpha$ the difference $E - \tilde{E}$ is of order $O(1)$ and thus cannot be neglected.

- *As $t \to \infty$: the least-squares solution.* As $t \to \infty$, one has $f_t(x) \to 0$ which results in the least-square solution $\mathbf{w}_{LS} = (\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\mathbf{y}$ and consequently

$$\frac{\boldsymbol{\mu}^\mathsf{T}\mathbf{w}_{LS}}{\|\mathbf{w}_{LS}\|} = \frac{\|\boldsymbol{\mu}\|^2}{\sqrt{\|\boldsymbol{\mu}\|^2 + c}}\sqrt{1 - \min\left(c, \frac{1}{c}\right)}. \tag{4.9}$$

Comparing (4.9) with the expression in Remark 4.1, one observes that when $t \to \infty$ the network becomes "over-trained" and the performance drops by a factor $\sqrt{1 - \min(c, c^{-1})}$. This becomes worse when c gets close to 1, which is consistent with the empirical findings in [AS17]. However, the point $c = 1$ is a singularity for (4.9), but not for $\frac{E}{\sqrt{V}}$ as in (4.3) and (4.4). One may thus expect to have a smooth and reliable behavior of the well-trained network for c close to 1, which is a noticeable advantage of gradient-based training compared to the simple least-squares

method. This coincides with the conclusion of [YRC07] in which the asymptotic behavior in the limit $n \to \infty$ with p fixed is considered.

In Figure 4.4 we plot the test performance from simulation (blue line), the approximation from Taylor expansion of $f_t(x)$ as described above (red dashed line), together with the performance of $\mathbf{w}_{LS}$ (cyan dashed line). One observes a close match between the result from Taylor expansion and the true performance for t small, with the former being optimal at $t = 100$ and the latter slowly approaching the performance of $\mathbf{w}_{LS}$ as t goes to infinity.

In Figure 4.5 we underline the case $c = 1$ by taking $p = n = 512$ with all other parameters unchanged from Figure 4.4. One observes that the simulation curve (blue line) increases much faster compared to Figure 4.4 and is supposed to end up at 0.5, which is the performance of $\mathbf{w}_{LS}$ (cyan dashed line). This confirms a serious degradation of performance for c close to 1 of the classical least-squares solution.

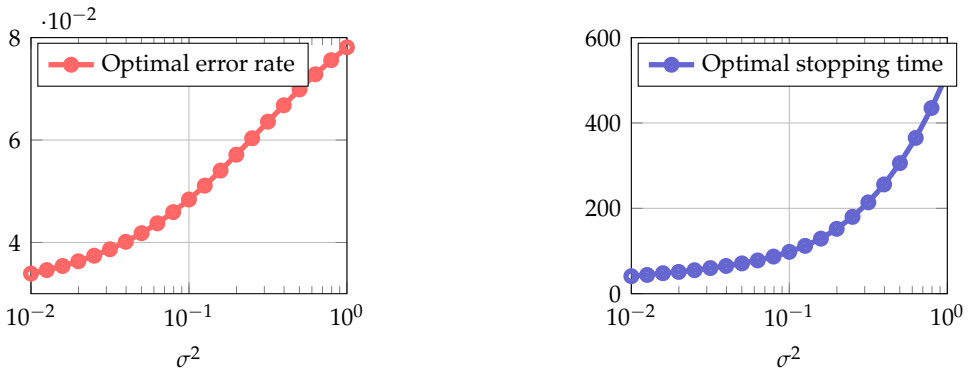


Figure 4.3: Optimal performance and corresponding stopping time as functions of σ^2, with $c = 1/2$, $\|\boldsymbol{\mu}\|^2 = 4$ and $\alpha = 0.01$.

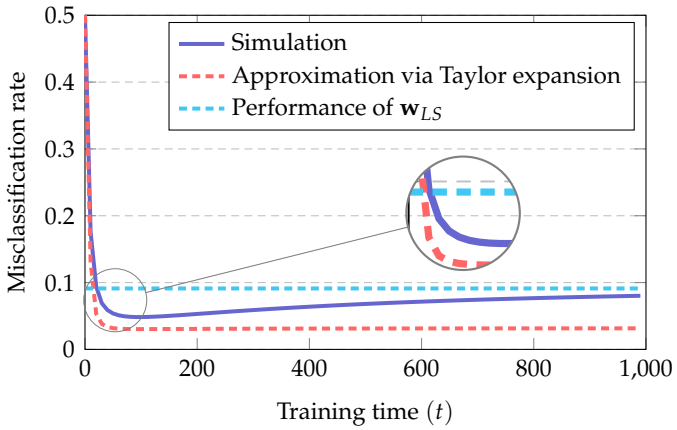


Figure 4.4: Generalization performance for $\boldsymbol{\mu} = [2; \mathbf{0}_{p-1}]$, $p = 256$, $n = 512$, $c_1 = c_2 = 1/2$, $\sigma^2 = 0.1$ and $\alpha = 0.01$. Simulation results obtained by averaging over 50 runs.

Conclusion. In this section we discussed the learning dynamic of a single-layer NN model when trained with gradient descent, under a two-class GMM of opposite means

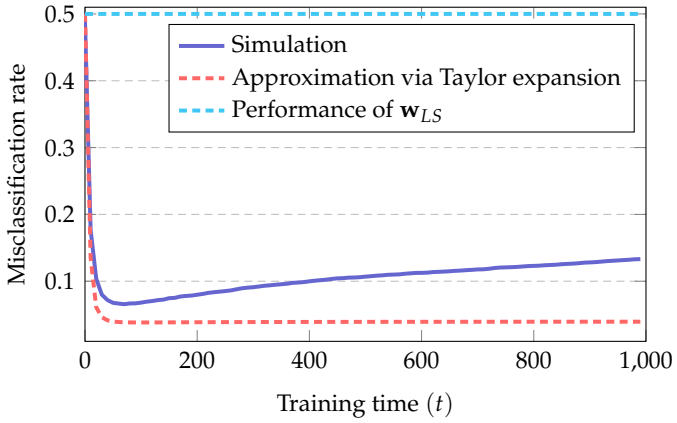


Figure 4.5: Generalization performance for $\boldsymbol{\mu} = [2; \mathbf{0}_{p-1}]$, $p = 512$, $n = 512$, $c_1 = c_2 = 1/2$, $\sigma^2 = 0.1$ and $\alpha = 0.01$. Simulation results obtained by averaging over 50 runs.

$\pm\boldsymbol{\mu}$ and identity covariance $\mathbf{C} = \mathbf{I}_p$. Based on the random matrix-based analysis, both the training and test performance of the network establish asymptotically tractable behaviors as $n, p \to \infty$ with $p/n \to c > 0$. These performances can be expressed generally in the form of contour integrations, which can further be reduced to real integrals for the simple model here with $\mathbf{C} = \mathbf{I}_p$. We obtain several interesting yet counterintuitive conclusions: one should prohibit the use of large random initialization, as it not only harms the network performance but also slows down the convergence rate (i.e., having a larger optimal stopping time); also, different from the initial stage of training, when the network is over-trained as t goes large, the performance drops by a factor $\sqrt{1 - \min(c, c^{-1})}$, and the misclassification error rate blows up when $n \approx p$, i.e., the number of training sample approximates the number of network parameters (and also the data dimension in this case).

4.2 A Geometric Approach to GDD of Linear NNs

In this section, we move forward to consider linear DNN models with $H \geq 1$ hidden layers as in [SMG13, Kaw16] and to evaluate the dynamics of the associated gradient system in a "continuous" manner. We prove that, in the case $H = 1$, for almost every choice of training data-target pair $(\mathbf{X}, \mathbf{Y})$ and almost every initialization for the weight matrices $\mathbf{W}_1, \mathbf{W}_2$ (see network configuration in Figure 4.6 below), the corresponding trajectory of the gradient system exists for all $t \geq 0$ and converges to a global minimizer of the square loss function. Based on a key "invariant" structure in the network weight space induced by the network cascading structure, we further propose a generic framework for the geometric understanding of linear deep neural networks, including a critical initialization scheme that ensures exponential convergence rate, a detailed description of the (first-order) stationary points, as well as the associated Hessians and basins of attraction of these stationary points.

For a deep linear neural network with H hidden layers as illustrated in Figure 4.6, we introduce the following notations and basic settings.

Denote $(\mathbf{X}, \mathbf{Y})$ the training data and associated targets, with $\mathbf{X} \equiv \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right] \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d_y \times n}$, where n denotes the number of instances in the training set and d_x, d_y the dimensions of data and targets, respectively. We denote the weight matrix

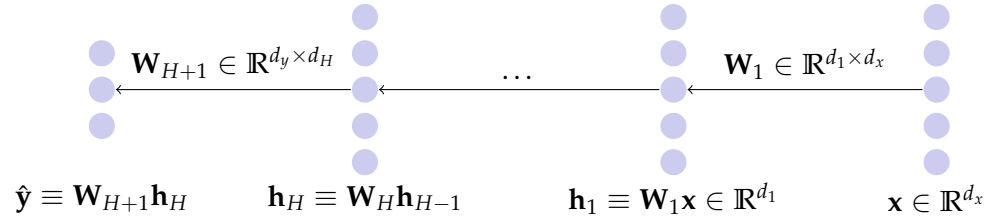


Figure 4.6: Illustration of the H-hidden-layer linear neural network

$\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ that connects $\mathbf{h}_{i-1}$ to $\mathbf{h}_i$ for $i = 1, \dots H+1$ and set $\mathbf{h}_0 = \mathbf{x}$, $\mathbf{h}_{H+1} = \hat{\mathbf{y}}$ as in Figure 4.6. The network output is thus given by $\hat{\mathbf{Y}} = \mathbf{W}_{H+1} \dots \mathbf{W}_1 \mathbf{X}$. We denote $\mathbf{W}$ the $(H+1)$-tuple of $(\mathbf{W}_1, \dots, \mathbf{W}_{H+1})$ for simplicity and work on the square loss $\mathcal{L}(\mathbf{W})$ given by the Frobenius norm below,

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2}\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2 = \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_{H+1} \dots \mathbf{W}_1 \mathbf{X}\|_F^2 . \tag{4.10}$$

We position ourselves under the following assumptions:

Assumption 10 (Dimension condition)**.** *$n \geq d_x \geq \max(d_1, \dots, d_H) \geq \min(d_1, \dots, d_H) > d_y$. In particular in the case $H = 1$ this condition yields $n \geq d_x \geq d_1 > d_y$.*

Assumption 11 (Full rank data and targets)**.** *The matrices $\mathbf{X}$ and $\mathbf{Y}$ are of full (row) rank, i.e., of rank d_x and d_y, respectively, accordingly with Assumption 10.*

Assumption 10 and 11 on the dimension and rank of the training data are realistic and practically easy to satisfy, as discussed in previous works [BH89, Kaw16]. Indeed, Assumption 10 is demanded here for convenience and our results can be extended to handle more elaborate dimension settings. Similarly, when the training data is rank deficient, the learning problem can be reduced to a lower-dimensional one by removing these non-informative data in such a way that Assumption 11 holds.

Under Assumptions 10 and 11, with the singular value decomposition on $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$ with $\mathbf{V}_\mathbf{X} = \left[\begin{array}{c|c} \mathbf{V}_\mathbf{X}^1 & \mathbf{V}_\mathbf{X}^2 \end{array}\right]$, $\mathbf{V}_\mathbf{X}^1 \in \mathbb{R}^{n \times d_x}$ and then on $\mathbf{Y}\mathbf{V}_\mathbf{X}^1 \equiv \tilde{\mathbf{Y}} = \mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^\mathsf{T}$, together with a change of variable, we get $\mathcal{L}(\mathbf{W}) = L(\mathbf{W}) + \frac{1}{2}\|\mathbf{Y}\mathbf{V}_\mathbf{X}^2\|_F^2$ with

$$L(\mathbf{W}) \equiv \frac{1}{2}\|\mathbf{\Sigma}_\mathbf{Y} - \mathbf{W}_{H+1}\mathbf{W}_H \dots \mathbf{W}_2\mathbf{W}_1\|_F^2 \tag{4.11}$$

where $\mathbf{\Sigma}_\mathbf{X} \equiv \left[\begin{array}{c|c} \mathbf{S}_\mathbf{X} & \mathbf{0} \end{array}\right] \in \mathbb{R}^{d_x \times n}$, $\mathbf{\Sigma}_\mathbf{Y} \in \mathbb{R}^{d_y \times d_x}$ and we denote, with a slight abuse of notations that $\mathbf{W}_{H+1} \equiv \mathbf{U}_\mathbf{Y}^\mathsf{T}\mathbf{W}_{H+1} \in \mathbb{R}^{d_y \times d_H}$ and $\mathbf{W}_1 \equiv \mathbf{W}_1\mathbf{U}_\mathbf{X}\mathbf{S}_\mathbf{X}\mathbf{V}_\mathbf{Y} \in \mathbb{R}^{d_1 \times d_x}$. Therefore the state space[1] of $\mathbf{W} \equiv (\mathbf{W}_{H+1}, \dots, \mathbf{W}_1)$ is equal to $\mathcal{X} = \mathbb{R}^{d_y \times d_H} \times \dots \times \mathbb{R}^{d_1 \times d_x}$. In particular, for $H = 1$ we have $d_H = d_1$ and $\mathcal{X}$ has dimension $d_1(d_x + d_y)$.

With the above notations, we demand also the following assumption on the target $\mathbf{Y}$.

Assumption 12 (Distinct singular values)**.** *The target $\mathbf{Y}$ has d_y distinct singular values.*

The objective of this section is to study the gradient descent [BV04] dynamics (GDD) defined as follows.

[1]The network (weight) parameters $\mathbf{W}$ evolve through time and are considered to be *state variables* of the dynamical system, while the pair $(\mathbf{X}, \mathbf{Y})$ is fixed and thus referred as the "*parameters*" of the given system.

Definition 8 (GDD). *The gradient descent dynamic of L is the dynamical system defined on $\mathcal{X}$ by*

$$\frac{d\mathbf{W}}{dt} = -\nabla_{\mathbf{W}} L(\mathbf{W}) \tag{4.12}$$

where $\nabla_{\mathbf{W}} L(\mathbf{W})$ denotes the gradient of the loss function L with respect to $\mathbf{W}$. A point $\mathbf{W} \in \mathcal{X}$ is a (first-order) stationary point of L if and only if $\nabla_{\mathbf{W}} L(\mathbf{W}) = \mathbf{0}$ and we denote $\mathcal{S}(L)$ the set of (first-order) stationary points.

In the following, we work directly on the equivalent equation (4.11) and start by evaluating the gradient of L. With the previous notations, for $\mathbf{w} \equiv (\mathbf{w}_{H+1}, \dots, \mathbf{w}_1)$, we expand the variation of $L(\mathbf{W} + \mathbf{w})$ as

$$L(\mathbf{W} + \mathbf{w}) = L(\mathbf{W}) + D_{\mathbf{W}}(\mathbf{w}) + O\left(\|\mathbf{w}\|^2\right)$$

with $\mathbf{M} \equiv \mathbf{\Sigma_Y} - \mathbf{W}_{H+1} \dots \mathbf{W}_1$ so that $L(\mathbf{W}) = \frac{1}{2}\|\mathbf{M}\|_F^2$ and the differential term given by $D_{\mathbf{W}}(\mathbf{w}) \equiv -\sum_{j=1}^{H+1} \operatorname{tr}\left(\mathbf{M}^\mathsf{T} \mathbf{W}_{H+1} \dots \mathbf{W}_{j+1} \mathbf{w}_j \mathbf{W}_{j-1} \dots \mathbf{W}_1\right)$. We thus derive from Definition 8 the dynamics of L, for $j = 1, \dots, H+1$, as

$$\frac{d\mathbf{W}_j}{dt} \equiv -\nabla_{\mathbf{W}_j} L(\mathbf{W}) = \left(\mathbf{W}_{H+1} \dots \mathbf{W}_{j+1}\right)^\mathsf{T} \mathbf{M} \left(\mathbf{W}_{j-1} \dots \mathbf{W}_1\right)^\mathsf{T}. \tag{4.13}$$

Main results. We start with the global convergence to stationary points of all gradient descent trajectories. While one expects the gradient descent algorithm to converge to stationary points, this may not always be the case. Two possible (undesirable) situations are i) a trajectory is unbounded or ii) it oscillates "around" several stationary points without convergence, i.e., along an ω-limit set made of a continuum of stationary points. The property of an iterative algorithm (like gradient descent) to converge to a stationary point for any initialization is referred to as "global convergence" [Zan69]. However, it is very important to stress the fact that it does not imply (contrary to what the name might suggest) convergence to a global (or good) minimum for all initializations.

To answer the convergence question, we resort to Lojasiewicz's theorem for the convergence of a gradient descent flow as (4.13) with real analytic loss L, as recalled below.

Theorem 4.3 (Lojasiewicz, from [Loj82]). *Let L be a real analytic function and $\mathbf{W}(\cdot)$ be a solution trajectory of the gradient system given by Definition 8 such that $\mathbf{W}(t)$ remains bounded for all $t \geq 0$. Then $\mathbf{W}(\cdot)$ converges to a stationary point of L as $t \to \infty$, with the convergence rate determined by the associated Lojasiewicz component [DK05].*

Since the loss function $L(\mathbf{W})$ is a polynomial of degree $(H+1)^2$ in the component of $\mathbf{W}$, Lojasiewicz's theorem ensures that a bounded trajectory of the gradient descent flow must converge to a stationary point with a guaranteed rate of convergence. In particular, the aforementioned phenomenon of "oscillation" cannot occur and we are left to ensure the absence of unbounded trajectories. The following lemma characterizes the "invariants" along trajectories of GDD, inspired by [SMG13] which essentially considered the case where all dimensions are equal to one. These invariants will be used at several stages: to prove convergence to stationary points, to ensure an exponential convergence rate, as well as to help understand the basin of attractions of the (undesired) saddle points.

Lemma 4.1 (Invariant in GDD). *Consider any trajectory of the gradient system given by* (4.13). *Then, for $j = 1, \dots, H$, the value of $\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1} - \mathbf{W}_j \mathbf{W}_j^\mathsf{T}$ remains constant, i.e.,*

$$\mathbf{W}_{j+1}^\mathsf{T}(t) \mathbf{W}_{j+1}(t) - \mathbf{W}_j(t) \mathbf{W}_j^\mathsf{T}(t) = \mathbf{C}_j^0 \equiv \mathbf{W}_{j+1}(0)^\mathsf{T} \mathbf{W}_{j+1}(0) - \mathbf{W}_j(0) \mathbf{W}_j(0)^\mathsf{T}, \quad \forall t \geq 0.$$

In particular, in the case of $H = 1$ we get $\mathbf{W}_2^\mathsf{T}\mathbf{W}_2 - \mathbf{W}_1\mathbf{W}_1^\mathsf{T} = \mathbf{C}^0 \equiv (\mathbf{W}_2^\mathsf{T}\mathbf{W}_2 - \mathbf{W}_1\mathbf{W}_1^\mathsf{T})\big|_{t=0}$.

Proof. Simply check that $\frac{d}{dt}\left(\mathbf{W}_{j+1}^\mathsf{T}\mathbf{W}_{j+1} - \mathbf{W}_j\mathbf{W}_j^\mathsf{T}\right) = 0$. □

Note in particular that, as a result of Lemma 4.1, there exists constant c_j, $1 \le j \le H$ such that for all $t \ge 0$,

$$\|\mathbf{W}_{j+1}(t)\|_F^2 - \|\mathbf{W}_j(t)\|_F^2 = c_j \tag{4.14}$$

i.e., the difference between the (Frobenius) norms of successive weight matrices $\mathbf{W}_j$ are preserved along with the gradient descent process.

With the key Lemma 4.1 and the remark above, to prove the boundedness of $\mathbf{W}_j(t)$ for all $j = 1, \dots, H+1$, it suffices to bound any single one of them, say for $j = H+1$. To this end, first note that

$$\begin{aligned}\|\mathbf{W}_{H+1}\dots\mathbf{W}_1\|_F^2 &= \operatorname{tr}(\mathbf{W}_{H+1}^\mathsf{T}\mathbf{W}_{H+1}\dots\mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^\mathsf{T}\mathbf{W}_2^\mathsf{T}\dots\mathbf{W}_H^\mathsf{T}) \\ &= \operatorname{tr}(\mathbf{W}_{H+1}^\mathsf{T}\mathbf{W}_{H+1}\dots(\mathbf{W}_2\mathbf{W}_2^\mathsf{T})^2\dots\mathbf{W}_H^\mathsf{T}) - \operatorname{tr}(\mathbf{W}_{H+1}^\mathsf{T}\mathbf{W}_{H+1}\dots\mathbf{W}_2\mathbf{C}_1^0\mathbf{W}_2^\mathsf{T}\dots\mathbf{W}_H^\mathsf{T})\end{aligned}$$

together with the fact that for symmetric positive semi-definite $\mathbf{A}$ we have $\lambda_{\min}(\mathbf{B})\operatorname{tr}(\mathbf{A}) \le \operatorname{tr}(\mathbf{AB}) \le \lambda_{\max}(\mathbf{B})\operatorname{tr}(\mathbf{A})$ so that

$$|\operatorname{tr}(\mathbf{W}_{H+1}^\mathsf{T}\mathbf{W}_{H+1}\dots\mathbf{W}_2\mathbf{C}_1^0\mathbf{W}_2^\mathsf{T}\dots\mathbf{W}_H^\mathsf{T})| \le \|\mathbf{C}_1^0\|\|\mathbf{W}_{H+1}\dots\mathbf{W}_2\|_F^2$$

so that by successively applying Lemma 4.1 one deduces the following bound

$$P_l(\mathbf{W}_{H+1}) \le \|\mathbf{W}_{H+1}\dots\mathbf{W}_1\|_F^2 \le \|P_u(\mathbf{W}_{H+1})$$

for some polynomial P_l and P_u of degree $H+1$ with positive leading coefficients. As a consequence, by considering the time derivative of $\|\mathbf{W}_{H+1}\|_F^2$, one can similarly get

$$\frac{d\|\mathbf{W}_{H+1}\|_F^2}{dt} \le -2C_0\|\mathbf{W}_{H+1}\|_F^{2(H+1)} + C_1(1 + \|\mathbf{W}_{H+1}\|_F^{2H})$$

for some positive constants C_0 and C_1, so that $\frac{d\|\mathbf{W}_{H+1}\|_F^2}{dt} < 0$ for some $\|\mathbf{W}_{H+1}\|_F^2$ large enough and $\mathbf{W}_{H+1}$ remains bounded for $t \ge 0$ and this holds for all $\mathbf{W}_j$, $j = 1, \dots, H+1$ via (4.14). As such, by Lojasiewicz's theorem one achieves the global convergence to stationary points, stated as below.

Proposition 4.1 (Global convergence to stationary points)**.** *Let $(\mathbf{X}, \mathbf{Y})$ be a data-target pair satisfying Assumptions 10 and 11. Then, every GDD trajectory defined in Definition 8 converges to a (first-order) stationary point as $t \to \infty$, at rate at least $t^{-\alpha}$, for some fixed $\alpha > 0$ only depending on the problem.*

As another important byproduct of Lemma 4.1, we have the following corollary that ensures an exponential convergence with a critical initialization scheme, and perhaps more importantly, to convergence to the *global* minima with $L = 0$.

Corollary 4.1 (Exponential convergence with critical initialization)**.** *Let Assumptions 10 and 11 hold. If we have in addition that $d_1 \ge d_2 \ge \dots \ge d_H$ and the initialization*

$$\mathbf{C}_j^0 \equiv \mathbf{W}_{j+1}(0)^\mathsf{T}\mathbf{W}_{j+1}(0) - \mathbf{W}_j(0)\mathbf{W}_j(0)^\mathsf{T} \in \mathbb{R}^{d_j \times d_j}$$

has at least d_{j+1} positive eigenvalues, i.e., $\lambda_{d_{j+1}}(\mathbf{C}_j^0) > 0$ for $j = 1, \dots, H$. Then, every trajectory of the GDD converges to a global minimum at least at the rate of $\exp(-2\alpha t)$, for some $\alpha > 0$.

Proof. Recall the definition $\mathbf{M} = \mathbf{\Sigma_Y} - \mathbf{W}_{H+1} \dots \mathbf{W}_1$ and consider its time derivative as

$$\begin{aligned}\frac{dM}{dt} &= -\sum_{j=1}^{H+1} \mathbf{W}_{H+1} \dots \mathbf{W}_{j+1} \frac{d\mathbf{W}_j}{dt} \mathbf{W}_{j-1} \dots \mathbf{W}_1 \\ &= -\sum_{j=1}^{H+1} \mathbf{W}_{H+1} \dots \mathbf{W}_{j+1} \mathbf{W}_{j+1}^\mathsf{T} \dots \mathbf{W}_{H+1}^\mathsf{T} \mathbf{M} \mathbf{W}_1^\mathsf{T} \dots \mathbf{W}_{j-1}^\mathsf{T} \mathbf{W}_{j-1} \dots \mathbf{W}_1\end{aligned}$$

so that

$$\begin{aligned}\frac{dL}{dt} = \frac{d\|\mathbf{M}\|_F^2}{dt} &= -2 \sum_{j=1}^{H+1} \operatorname{tr}\left(\mathbf{M}^\mathsf{T} \mathbf{W}_{H+1} \dots \mathbf{W}_{j+1} \mathbf{W}_{j+1}^\mathsf{T} \dots \mathbf{W}_{H+1}^\mathsf{T} \mathbf{M} \mathbf{W}_1^\mathsf{T} \dots \mathbf{W}_{j-1}^\mathsf{T} \mathbf{W}_{j-1} \dots \mathbf{W}_1\right) \\ &\le -2 \sum_{j=1}^{H+1} \prod_{k=1}^{j-1} \lambda_{\min}(\mathbf{W}_k^\mathsf{T} \mathbf{W}_k) \operatorname{tr}\left(\mathbf{M}^\mathsf{T} \mathbf{W}_{H+1} \dots \mathbf{W}_{j+1} \mathbf{W}_{j+1}^\mathsf{T} \dots \mathbf{W}_{H+1}^\mathsf{T} \mathbf{M}\right) \\ &\le -2 \sum_{j=1}^{H+1} \prod_{k=1}^{j-1} \lambda_{\min}(\mathbf{W}_k^\mathsf{T} \mathbf{W}_k) \prod_{l=j+1}^{H+1} \lambda_{\min}(\mathbf{W}_l \mathbf{W}_l^\mathsf{T}) \|\mathbf{M}\|_F^2\end{aligned}$$

where we constantly use the fact that for symmetric and positive semi-definite $\mathbf{A}$ we have $|\operatorname{tr}(\mathbf{AB})| \ge \lambda_{\min}(\mathbf{B}) \operatorname{tr}(\mathbf{A})$. Therefore, if there exists at least $1 \le j \le H+1$ such that

$$\prod_{l=j+1}^{H+1} \lambda_{\min}(\mathbf{W}_l \mathbf{W}_l^\mathsf{T}) \prod_{k=1}^{j-1} \lambda_{\min}(\mathbf{W}_k^\mathsf{T} \mathbf{W}_k) > 0 \tag{4.15}$$

then we obtain $\frac{d\|\mathbf{M}\|_F^2}{dt} \le -C\|\mathbf{M}\|_F^2$ for some $C > 0$ and thus the conclusion. To this end, we derive, from Lemma 4.1 and Weyl's inequality (see Lemma 2.10) that, for $j = 1, \dots, H$

$$\lambda_i(\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1}) \ge \lambda_i(\mathbf{C}_j^0) + \lambda_{\min}(\mathbf{W}_j \mathbf{W}_j^\mathsf{T}) \ge \lambda_i(\mathbf{C}_j^0), \quad i = 1, \dots, d_j$$

with $\lambda_i(\mathbf{C}_j^0)$ the i-th eigenvalue of $\mathbf{C}_j^0$ arranged in nondecreasing order so that $\lambda_1(\mathbf{C}_j^0) = \lambda_{\min}(\mathbf{C}_j^0)$. Then since $d_1 \ge \dots \ge d_j \ge d_{j+1} \ge \dots \ge d_H$, the matrix $\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1} \in \mathbb{R}^{d_j \times d_j}$ is of rank maximum d_{j+1} and thus admits at least $d_j - d_{j+1}$ zero eigenvalues so that

$$\lambda_i(\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1}) = 0, \quad \lambda_i(\mathbf{C}_j^0) \le 0$$

for $i = 1, \dots, d_j - d_{j+1}$. Moreover, since for $i = 1, \dots, d_{j+1}$ we also have,

$$\lambda_{i+d_j-d_{j+1}}(\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1}) = \lambda_i(\mathbf{W}_{j+1} \mathbf{W}_{j+1}^\mathsf{T})$$

we further deduce that for $i = 1, \dots, d_{j+1}$,

$$\lambda_i(\mathbf{W}_{j+1} \mathbf{W}_{j+1}^\mathsf{T}) = \lambda_{i+d_j-d_{j+1}}(\mathbf{W}_{j+1}^\mathsf{T} \mathbf{W}_{j+1}) \ge \lambda_{i+d_j-d_{j+1}}(\mathbf{C}_j^0)$$

so that by taking $j = 1$ in (4.15) we result in

$$\prod_{l=1}^{H} \lambda_{\min}(\mathbf{W}_{l+1} \mathbf{W}_{l+1}^\mathsf{T}) \ge \prod_{l=1}^{H} \lambda_{d_l-d_{l+1}+1}(\mathbf{C}_l^0) \equiv \alpha > 0$$

which concludes the proof. □

Remark 4.1 entails that, for a linear network, although in general only polynomial convergence rate can be established, it is possible to wisely initialize the gradient descent algorithm to achieve exponential convergence.

Apart from the critical initialization scheme proposed in Corollary 4.1 that leads to a global convergence to *global* minima, Proposition 4.1 itself does not guarantee the "quality" of stationary point to which the trajectory converges.

In the following, we move on to consider the general situation that we initialize the gradient descent algorithm from a arbitrary point in the network weight space $\mathcal{X}$ and focus on the set of the first-order stationary points $\mathcal{S}$. We consider for simplicity the case $H = 1$ and comment on $H > 1$.

Proposition 4.1 ensures, for all initializations, the convergence of the gradient descent to a stationary point, i.e., a point $\mathbf{W}$ in the state space $\mathcal{X}$ verifying $\nabla_{\mathbf{W}} L(\mathbf{W}) = 0$. Nonetheless, the information on the "quality" (e.g., the loss value L) of the solution achieved is still unknown. To obtain a clearer picture, we now consider the case $H = 1$ and focus on the set of *all* stationary points by further decomposing the loss L with $\mathbf{\Sigma_Y} \equiv \left[\begin{array}{c|c} \mathbf{S_Y} & \mathbf{0} \end{array}\right]$ for diagonal $\mathbf{S_Y} \in \mathbb{R}^{d_y \times d_y}$ with $[\mathbf{S_Y}]_{ii} > 0$ as

$$L(\mathbf{W}_1, \mathbf{W}_2) = \frac{1}{2}\|\mathbf{\Sigma_Y} - \mathbf{W}_2\mathbf{W}_1\|_F^2 = \frac{1}{2}\|\mathbf{S_Y} - \mathbf{CA}\|_F^2 + \frac{1}{2}\|\mathbf{CB}\|_F^2 \tag{4.16}$$

with $\mathbf{C} \equiv \mathbf{W}_2 \in \mathbb{R}^{d_y \times d_1}$, $\mathbf{A} \in \mathbb{R}^{d_1 \times d_y}$ and $\mathbf{B} \in \mathbb{R}^{d_1 \times (d_x - d_y)}$ such that $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \end{array}\right] \equiv \mathbf{W}_1$.

Under the notations above, we further expand $L(\mathbf{W} + \mathbf{w})$ to obtain its higher order variation as

$$L(\mathbf{A} + \mathbf{a}, \mathbf{B} + \mathbf{b}, \mathbf{C} + \mathbf{c}) \equiv L(\mathbf{W} + \mathbf{w}) = L(\mathbf{W}) + D_{\mathbf{W}}(\mathbf{w}) + H_{\mathbf{W}}(\mathbf{w}) + O\left(\|\mathbf{w}\|^3\right)$$

with $\mathbf{M} \equiv \mathbf{S_Y} - \mathbf{CA}$, $L(\mathbf{W}) = \frac{1}{2}\|\mathbf{M}\|_F^2 + \frac{1}{2}\|\mathbf{CB}\|_F^2$ and

$$D_{\mathbf{W}}(\mathbf{w}) \equiv -\operatorname{tr}\left(\mathbf{M}^\mathsf{T}(\mathbf{Ca} + \mathbf{cA})\right) + \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{C}^\mathsf{T}(\mathbf{Cb} + \mathbf{cB})\right) = O(\|\mathbf{w}\|)$$

$$H_{\mathbf{W}}(\mathbf{w}) \equiv -\operatorname{tr}(\mathbf{M}^\mathsf{T}\mathbf{ca}) + \frac{1}{2}\|\mathbf{Ca} + \mathbf{cA}\|_F^2 + \operatorname{tr}(\mathbf{B}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{cb}) + \frac{1}{2}\|\mathbf{Cb} + \mathbf{cB}\|_F^2 = O(\|\mathbf{w}\|^2)$$

that give the differential and the Hessian of L, respectively. Recall that $\mathcal{S}(L) \equiv \{\mathbf{W} \mid D_{\mathbf{W}}(\mathbf{w}) = \mathbf{0}\}$ and denote $\mathbf{M} \equiv \mathbf{S_Y} - \mathbf{CA}$, so that, by Definition 8,

$$\begin{cases} \frac{d\mathbf{A}}{dt} &= \mathbf{C}^\mathsf{T}\mathbf{M} = \mathbf{0} \\ \frac{d\mathbf{B}}{dt} &= -\mathbf{C}^\mathsf{T}\mathbf{CB} = \mathbf{0} \\ \frac{d\mathbf{C}}{dt} &= \mathbf{MA}^\mathsf{T} - \mathbf{CBB}^\mathsf{T} = \mathbf{0} \end{cases} \Leftrightarrow \begin{cases} \mathbf{C}^\mathsf{T}\mathbf{S_Y} = \mathbf{C}^\mathsf{T}\mathbf{CA} \\ \mathbf{CB} = \mathbf{0} \\ \mathbf{AS_Y} = \mathbf{AA}^\mathsf{T}\mathbf{C}^\mathsf{T}. \end{cases} \tag{4.17}$$

Observing the symmetric structure of $\mathbf{A}, \mathbf{C}$ in (4.17) we have the following lemma.

Lemma 4.2 (Same kernel for $\mathbf{A}$ and $\mathbf{C}^\mathsf{T}$)**.** *Let Assumptions 10 and 11 hold. Then for all* $\mathbf{W} \in \mathcal{S}(L)$,

$$\ker \mathbf{A} = \ker \mathbf{C}^\mathsf{T}, \text{ with } \ker \mathbf{A} \equiv \{\mathbf{x}, \mathbf{Ax} = \mathbf{0}\}.$$

Moreover, denote r *the common rank of* $\mathbf{A}$ *and* $\mathbf{C}$ *with* $0 \leq r \leq d_y$*. Then there exists some orthogonal matrix* $\mathbf{U} \in \mathbb{R}^{d_y \times d_y}$ *such that*

$$\begin{cases} \mathbf{AU} = \left[\begin{array}{c|c} \bar{\mathbf{A}} & \mathbf{0}_{d_1 \times (d_y - r)} \end{array}\right] \\ \mathbf{C}^\mathsf{T}\mathbf{U} = \left[\begin{array}{c|c} \bar{\mathbf{C}}^\mathsf{T} & \mathbf{0}_{d_1 \times (d_y - r)} \end{array}\right] \\ \mathbf{U}^{-1}\mathbf{S_Y}\mathbf{U} = \mathbf{S_Y} \end{cases} \tag{4.18}$$

with $\bar{\mathbf{A}}, \bar{\mathbf{C}}^\mathsf{T} \in \mathbb{R}^{d_1 \times r}$. *Moreover, if* $\mathbf{S_Y}$ *has distinct eigenvalues (i.e.,* $\mathbf{Y}$ *has* d_y *distinct singular values, as demanded in Assumption 12), then* $\mathbf{U}$ *is a diagonal matrix made of* 1 *and* -1.

Sketch of proof of Lemma 4.2. It can be shown with basic algebraic manipulations that the eigenvectors of $\mathbf{S_Y^2}$ (thus of $\mathbf{S_Y}$) form a basis of both $\ker \mathbf{A}$ and $\ker \mathbf{C}^\mathsf{T}$. Therefore $\ker \mathbf{A} = \ker \mathbf{C}^\mathsf{T}$ and in particular $\dim \ker \mathbf{A} = \dim \ker \mathbf{C}^\mathsf{T}$. We denote this dimension $d_y - r$ and $\mathbf{A}, \mathbf{C}$ are thus both of rank r. Choose $\mathbf{U}_2$ from $\ker \mathbf{A}$ and $\mathbf{U}_1 \perp \ker \mathbf{A}$; we deduce $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \mid \mathbf{U}_2 \end{bmatrix}$ so that (4.18) holds. □

In the general case $H \geq 1$, similar conclusion as in Lemma 4.2 holds for the matrix product $(\mathbf{W}_{H+1} \dots \mathbf{W}_2)^\mathsf{T}$ and $\mathbf{W}_{H+1} \dots \mathbf{W}_2 \mathbf{A}$ for $\mathbf{W}_1 \equiv \begin{bmatrix} \mathbf{A} \mid \mathbf{B} \end{bmatrix}$.

Remark from (4.18) in Lemma 4.2 that, for arbitrary $\mathbf{S_Y}$, there are infinitely many possibilities on the choice of $\mathbf{U}$ with the risk of occupying too much of the state space $\mathcal{X}$, since, with the change of variable in Lemma 4.2 the state variable now becomes the tuple $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{U})$. Using Assumption 12, $\mathbf{U}$ only takes a finite number of values for a given $\mathbf{W} \in \mathcal{S}_r(L)$, hence the state variable essentially becomes the tuple $(\mathbf{A}, \mathbf{B}, \mathbf{C})$.

For $\mathbf{W} \in \mathcal{S}(L)$ with $\mathbf{A}, \mathbf{C}$ of rank r with $0 \leq r \leq d_y$, rewriting $\mathbf{S_Y}$ in two blocks $\mathbf{S_Y} = \begin{bmatrix} \mathbf{D_Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_Y} \end{bmatrix}$, with $\mathbf{D_Y} \in \mathbb{R}^{r \times r}$ and $\mathbf{E_Y} \in \mathbb{R}^{(d_y - r) \times (d_y - r)}$. With Lemma 4.2, we then simplify (4.17) as

$$\begin{cases} \mathbf{C}\,\mathbf{A} = \mathbf{D_Y} \\ \mathbf{C}\mathbf{B} = \mathbf{0} \end{cases}, \ \mathbf{U}^\mathsf{T}\mathbf{M}\mathbf{U} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_Y} \end{bmatrix} \tag{4.19}$$

with the fact that $\mathbf{C}^\mathsf{T}, \mathbf{A}$ are both of full rank (equal to r). The loss $L(\mathbf{W})$ (at stationary points) can thus be simplified as $L(\mathbf{W}) = \frac{1}{2}\|\mathbf{E_Y}\|_F^2$ where $\mathbf{E_Y}$ measures the "quality" of each stationary points.

For any $\mathbf{W} \in \mathcal{S}(L)$, with Lemma 4.2 we are allowed to "extract" the full rank (sub-)structures of $\mathbf{A}, \mathbf{C}$ with $\mathbf{S_Y}$ unchanged, via a simple change of basis. For $0 \leq r \leq d_y$, let $\mathcal{S}_r(L)$ be the subset of $\mathcal{S}(L)$ such that the rank of $\mathbf{A}$ and of $\mathbf{C}$ is equal to r. Then, one has the following disjoint union

$$\mathcal{S}(L) = \cup_{r=0}^{d_y} \mathcal{S}_r(L).$$

Proposition 4.2 (Landscape of single-hidden-layer linear NN model)**.** Under Assumptions 10–12, the loss function $L(\mathbf{W})$ has the following properties:

1. The set of possible limits of L along the GDD given by (4.12) is equal to the finite set made of the sum of the squares of any subset of the singular values of $\bar{\mathbf{Y}}$.
2. The set $\mathcal{S}_{d_y}(L)$ is in fact the set of local (and global) minima, with $L = 0$ and $\mathbf{M} = \mathbf{0}$.
3. Every first-order stationary point $\mathbf{W} \in \mathcal{S}_r(L)$ with $0 \leq r \leq d_y - 1$ is a saddle point such that the Hessian has at least one negative eigenvalue. In particular, the set of saddle points is an algebraic variety of positive dimension, i.e., (up to a unitary matrix) the zero set of the polynomial functions given in (4.19), with $\mathbf{E_Y} \neq \mathbf{0}$.

For more general DNNs with $H > 1$, similar conclusion holds, except from the negative eigenvalue of the Hessian for all saddle points. As a matter of fact, for $H > 1$, there exists saddle points with positive semidefinite Hessian and we show that a sufficient condition to ensure at least one negative eigenvalue is to have $\text{rank}(\mathbf{W}_H \dots \mathbf{W}_2) > \text{rank}(\mathbf{W}_{H+1} \dots \mathbf{W}_2)$.

The fact that all local minima are equivalently global minima and all critical points that are not global minima are saddle points with at least one negative eigenvalue for the

Hessian is in fact already known for single-hidden-layer linear networks [BH89, Kaw16]. Here, we improve the condition given in [Kaw16], with an alternative and shorter proof based only on the quadratic form involving the Hessian, for deep linear networks.

With Proposition 4.2, and assume in addition that the loss function L admits two by two distinct values over two by two distinct subsets made of (the sum of the squares of) the singular values of the target $\bar{\mathbf{Y}}$, one is able to provide a precise "local" description of all saddle points. Based on this description, one is able to further construct the local stable and unstable manifolds, with the help of the notation of normally hyperbolic [HPS06, Theorem 3.5]. Ultimately, with the fact that the loss value L is decreasing along the gradient descent trajectory and by contradiction, one reaches the convergence of the GDD trajectory to a global minima, from almost all initialization (in the sense of Lebesgue measure).

Conclusion. The essential difference between the case $H = 1$ and deeper NN with $H > 1$ is the existence of saddle points with positive semidefinite Hessian matrices. In this case, one has to evaluate higher order information (≥ 3) to understand more precisely the basins of attraction of these undesired stationary points, or otherwise establish tighter bounds for the dimension of the parameter space $\mathcal{X}$ that is occupied by these bad saddle points and their basins of attraction. Alternatively, one may also wish to provide initialization scheme, in this $H > 1$ case, that can avoid, at least with high probability say, these bad saddle points, so that we can again apply the arguments as in the case $H = 1$.

Chapter 5

Conclusions and Perspectives

5.1 Conclusions

Under a simple GMM for the input data model, we investigated, in Section 3.1.1-3.2.2 the performance of three closely connected objects: large random kernel matrices, large dimensional random feature maps, and large simple neural networks random weights. In the setting where the number of data n and their dimension p are both large and comparable, the MSE of random neural networks, the misclassification error rate of kernel ridge regression, as well as the performance of kernel/random feature-based spectral clustering techniques, all depend heavily on the eigenspectrum or on a functional of a particular random kernel/nonlinear Gram matrix. The major technical challenge to theoretically grasp these kernel matrices, so as to assess the performance of the aforementioned algorithms, lies in the presence of nonlinearities, e.g., the kernel function f and the neural network activation function σ. To handle this difficulty, we placed ourselves under the critical regime of asymptotically non-trivial classification (namely, Assumption 2 in Section 1.2). In this regime, a "concentration" phenomenon emerges, for the similarly measures $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ or $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$, which then helps (asymptotically) "linearize" (or Taylor-expand to first orders) the nonlinear objects of interest. The associated eigenspectrum, therefore, behaves like (a scaled and shifted version of) the classical sample covariance/Gram matrix, perturbed by low rank matrices, and only depends on the local behavior of the nonlinear function at the "concentration" point.

This "concentration" effect, if theoretically convenient, however, hinders the full discriminative power of kernels (such as the discrimination of covariance matrices in different classes). To work this limitation around, a proper scaling as $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p}$ is necessary, which we studied in the second part of the thesis. In this situation, a "concentration" cannot be established so that, in place of Taylor expansion arguments, we exploited the method of orthogonal polynomials to decompose the nonlinear f into the sum of a linear and a purely nonlinear parts, that both contribute to the kernel eigenspectrum, but in a very different way. A surprising outcome is that the kernel matrix eigenspectrum depends on f solely via its first three Hermite coefficients, suggesting that, even without concentration, the space of kernel functions can be mapped to a low dimensional subspace. Similar behavior is also observed for the empirical risk-based classifier in Section 3.3, where the performance only depends on the loss function L through the same three coefficients of the implicit proximal function. However, the performance description in Section 3.3 is more complicated and can only be expressed as the fixed point of a system of nonlinear equations. RMT analyses allow for direct access to the (asymptotic) performance of the aforementioned machine learning methods, so long as their solutions

are explicit or implicitly determined by (convex) optimization problems.

Aiming to study the impact of optimization algorithm in convex or non-convex problems, we considered in Section 4.1 the temporal evolution of a single-layer linear neural network, learned by the gradient descent method, still under a GMM for the input data. Despite being a linear classifier, the training and test performance of this linear regression model is a highly nonlinear function of the training time and involves the sum of a nonlinear function of all the eigenvalues of the input covariance matrix. Based on the proposed RMT analysis, both the training and test performance of the model can be fully characterized in the large n, p limit. This result provides a first precise description of large network performances when trained with gradient descent and sheds new light on many common questions in neural networks, such as overfitting, early stopping and the initialization of training, in the large dimensional regime where $n, p \to \infty$ with $p/n \to c > 0$.

When the number of hidden layers is more than one, the underlying optimization problem to train the neural network becomes non-convex and less tractable. In Section 4.2 we took a geometric and, so far, only deterministic approach to understand the convergence properties of simple gradient descent algorithms in this problem. We improved the existing analysis of the gradient descent dynamics in single-hidden-layer linear neural networks by considering the corresponding dynamical system. Based on a cornerstone "invariance" structure in the network weight space, we provided alternative and simpler proofs and removed some unnecessary assumptions from previous contributions to achieve the "almost sure" convergence to global minima.

5.2 Limitations and Perspectives

Yet, these findings are limited to the simple but seemingly restrictive Gaussian mixture modeling of the data. In the following, we discuss some possible future research directions, in the continuation of our preliminary findings.

Universality of RMT results in machine learning applications. As discussed in the proof of the Marčenko-Pastur law in Section 2.2.2, as well as in the "properly scaled" inner product kernel matrix in Section 3.1.2, by positioning ourselves under the large dimensional setting where n, p are both large and comparable, many random matrix-based analyses, for linear or nonlinear models, lead to *universal* results, in the sense that they hold as long as the (original) random matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ has i.i.d. zero mean and unit variance entries, and are independent of the higher order statistics. Since real data generally do not comply with this stringent i.i.d. assumption, it is of interest to investigate the limitation of this universality phenomenon in machine learning applications. It is worth pointing out that the universality has already been observed to collapse in many nonlinear models, for instance in the "improperly scaled" shift-invariant kernel $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ investigated in [CBG16] where, by considering non-Gaussian $\mathbf{Z}$ with a non-zero excess kurtosis $\kappa = \mathbb{E}[\mathbf{Z}_{ij}^4] - 3 \neq 0$, the random variable $\psi_i \equiv \|\boldsymbol{\omega}_i\|^2 - \mathbb{E}[\|\boldsymbol{\omega}_i\|^2]$ yields a different (statistical) behavior, which nonetheless only moderately impacts the performances of kernel spectral clustering and kernel ridge regression methods. In [LLC18, BP19] where one considers the Gram matrix $\sigma(\mathbf{WX})^\mathsf{T}\sigma(\mathbf{WX})$ in a random neural network context, it has been observed that the higher order moments of the distribution of (the i.i.d. entries of) the weight matrix $\mathbf{W}$ has an important impact on the eigenspectrum of the Gram matrix and consequently on the performance of the resulting neural network model. In this respect, one may wonder on the possibility to identify machine learning methods that

are "universal" and depend only on the first several moments of the data distribution from those that depend on the higher order moments of the data modeling. From a practical standpoint, it is perhaps of more importance to empirically investigate which tasks can be sufficiently well fulfilled by only considering the first and second order information from the data. For non-universal models, a possible idea is to perform preliminary studies on the influence of higher order information, starting from a qualitative study, so as to understand the (systematic) advantage of certain machine learning methods in some specific (e.g., kurtosis-sensitive) applications to provide theoretical guarantees of the existing algorithms as well as new insights in designing more efficient techniques.

RMT-based analyses of optimization problems. The random matrix analyses of optimization problems developed in this thesis are twofold:

1. As in Section 3.3, by considering the optimality condition of convex or non-convex optimization problems (e.g., first- or second-order stationary points condition), RMT analyses provide statistical descriptions of the local or global optima, which further lead to a theoretical understanding of the algorithm performance as well as a better hyperparameter tuning or "unbiasing" for these methods, as discussed at the end of Section 3.3. This scheme mainly works for convex (or strongly convex) optimization problems as in Section 3.3, where the uniqueness of the optimal solution is naturally guaranteed.

2. For non-convex optimization problems as the single-hidden-layer neural networks studied in Section 4.2, there may exist in general a large or even infinite number of first or higher order stationary points. In this case, even if one manages to establish an almost sure global convergence to global minima, the rate of convergence can be extremely slow due to the presence of (possibly numerous) saddle points [DJL$^+$17]. A wisely chosen initialization (as proposed in Corollary 4.1) within a locally convex structure helps accelerate the convergence rate, up to linear/exponential. This is where a RMT analysis may serve practical purposes: by leveraging the random nature of many non-convex optimization problems (e.g., with sensing matrices having i.i.d. standard Gaussian entries [CP11a] or with random measurement of random signals in phase retrieval [CC15]), it is possible to propose the following two-step algorithms:

   - *initialization*, that makes a RMT-based initial guess so as to start the gradient descent *away* from any saddle points and *within* the "local valleys" of local (and global) minima;
   - *iterative refinement* based on the gradient method, without leaving the initial valley.

   In this case, the convergence rate relies heavily on a proper initialization, that can be achieved by means of spectral methods, e.g., to initialize with the top eigenvectors of a carefully designed matrix, as in the case of phase retrieval, matrix sensing/completion and many others [CLC18]. In many cases, by considering a statistical modeling of the problem with repeated and informative patterns, it is possible to construct a data-dependent (random) matrix $\mathbf{M}$ that follows a spiked model as discussed at length in Section 2.3. The RMT analysis can then be applied to capture the statistical behavior of the top eigenvectors (that correspond to the isolated eigenvalues) of $\mathbf{M}$, so as to measure the "distance" between the proposed

(spectral-based) initialization and the desired global minima. In this regard, RMT offers not only a powerful initialization scheme by proposing an optimal design for $\mathbf{M}$, but also a precise description of the conditions under which the proposed initialization works, by establishing "phase transition" conditions as discussed at the end of Section 2.3.

As a closing remark, in the regime where n, p, N are all large and comparable, with N the number of model parameters, many counterintuitive phenomena arise (recall from Figure 1.5 that for $N > n$ the test error decreases as the ratio N/n increases), for which we need a deeper theoretical understanding. Indeed, modern machine learning systems are often highly *over-parameterized* and it is an everyday occurrence in modern DNNs to have $N > \max(n, p)$. As a consequence, random matrix-based approaches are a promising tool to investigate these modern *over-parameterized* learning systems that are of greater importance today.

Appendix A

Mathematical Proofs

A.1 Proofs in Chapter 3

A.1.1 Intuitive calculus for Theorem 1.2

In this section we consider the empirical spectral measure of the following inner-product kernel matrix as in (1.13)

$$\mathbf{K}_{ij} = \begin{cases} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}$$

for independent random vector $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and in particular, the associated Stieltjes transform (see Definition 6) given by

$$m(z) \equiv \frac{1}{n}\operatorname{tr}\mathbf{Q}(z), \quad \mathbf{Q}(z) \equiv (\mathbf{K} - z\mathbf{I}_n)^{-1}.$$

Our objective here is to prove that $m(z)$ is the (unique) solution of the following cubic equation

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z)$$

with $a_1 = \mathbb{E}[\xi f(\xi)]$ and $\nu = \operatorname{Var}[f(\xi)] \geq a_1^2$ for standard Gaussian $\xi \sim \mathcal{N}(0,1)$.

Basic settings and notations. Following the ideas of the sample covariance model in Section 2.2.3, we remove the i-th row and the i-th column of the symmetric matrix $\mathbf{K}$ such that

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{-i} & f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{x}_i/\sqrt{p})/\sqrt{p} \\ f(\mathbf{x}_i^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})/\sqrt{p} & 0 \end{bmatrix}, \quad \mathbf{K}_{-i} = f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})/\sqrt{p} \in \mathbb{R}^{(n-1)\times(n-1)}$$

with zero on the diagonal of $\mathbf{K}_{-i}$ and $\mathbf{X}_{-i} \in \mathbb{R}^{p\times(n-1)}$ the data matrix without $\mathbf{x}_i$. As such, we have that $\mathbf{K}_{-i}$ is independent of $\mathbf{x}_i$. We similarly define the resolvent $\mathbf{Q}_{-i}$ of $\mathbf{K}_{-i}$ as

$$\mathbf{Q}_{-i} = (\mathbf{K}_{-i} - z\mathbf{I}_{n-1})^{-1}$$

so that under the above notations, the (i,i)-th (diagonal) entry of $\mathbf{Q}$ is given by

$$\mathbf{Q}_{ii} = \frac{1}{-z - \frac{1}{p} f(\mathbf{x}_i^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})\mathbf{Q}_{-i} f(\mathbf{X}_{-1}^\mathsf{T}\mathbf{x}_i/\sqrt{p})} \tag{A.1}$$

where we recall that the diagonal of both $\mathbf{K}$ and $\mathbf{K}_{-i}$ contain only zero entries. Since we are interested in the Stieltjes transform $m(z) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_{ii}(z)$, the key object is the (nonlinear) quadratic form $\frac{1}{p}f(\mathbf{x}_i^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})\mathbf{Q}_{-i}f(\mathbf{X}_{-1}^\mathsf{T}\mathbf{x}_i/\sqrt{p})$.

To handle the *nonlinear* random vector $f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{x}_i/\sqrt{p})$, we follow [CS13] and perform the following orthogonal decomposition of $\mathbf{x}_j$: for all $j \neq i$,

$$\mathbf{x}_j = \alpha_j \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} + \mathbf{x}_j^\perp \tag{A.2}$$

where $\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$ is the unit vector in the direction of $\mathbf{x}_i$ and $\mathbf{x}_j^\perp$ lies in the $(p-1)$-dimensional subspace orthogonal to $\mathbf{x}_i$. By orthogonality between $\mathbf{x}_j^\perp$ and $\mathbf{x}_i$ we have

$$\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} = \alpha_j\|\mathbf{x}_i\|/\sqrt{p} \Leftrightarrow \alpha_j = \frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|}$$

for $j \neq i$. Since $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent standard Gaussian vectors, we have, in the large p limit that $\mathbf{x}_j^\mathsf{T}\mathbf{x}_i/\sqrt{p} \sim \mathcal{N}(0,1)$ and $\|\mathbf{x}_i\| \simeq \sqrt{p}$. Moreover, $\alpha_j \sim \mathcal{N}(0,1)$, $\mathbf{x}_j^\perp \sim \mathcal{N}(0, \mathbf{I}_{p-1})$ and both are decorrelated thus independent.

Indeed,

$$\mathbb{E}_{\mathbf{x}_j}[\alpha_i \mathbf{x}_j^\perp] = \mathbb{E}_{\mathbf{x}_j}\left[\frac{\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|}\left(\mathbf{x}_j - \frac{\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{x}_j}{\|\mathbf{x}_i\|^2}\right)\right] = \mathbb{E}_{\mathbf{x}_j}\left[\frac{\mathbf{x}_j\mathbf{x}_j^\mathsf{T}\mathbf{x}_i}{\|\mathbf{x}_i\|}\right] - \mathbb{E}_{\mathbf{x}_j}\left[\frac{\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\mathbf{x}_j^\mathsf{T}\mathbf{x}_i}{\|\mathbf{x}_i\|^3}\right] = 0.$$

As such, we have for $k \neq j, k \neq i$ that

$$\mathbf{x}_j^\mathsf{T}\mathbf{x}_k = \alpha_j\alpha_k + (\mathbf{x}_j^\perp)^\mathsf{T}\mathbf{x}_k^\perp \equiv \alpha_j\alpha_k + \mathbf{\Phi}_{jk}^\perp \tag{A.3}$$

where the cross terms disappear again by orthogonality. Note from (A.2) that with high probability, both $\mathbf{x}_j$ and $\mathbf{x}_j^\perp$ are of (Euclidean) norm $O(\sqrt{p})$ while $\alpha_j = O(1)$, Similarly, in (A.3), both $\mathbf{x}_j^\mathsf{T}\mathbf{x}_k$ and $\mathbf{\Phi}_{jk}^\perp$ are of order $O(\sqrt{p})$, while $\alpha_j\alpha_k = O(1)$. In this sense, $\mathbf{\Phi}_{jk}^\perp$ is asymptotically close to the inner product $\mathbf{x}_j^\mathsf{T}\mathbf{x}_k$, with only the contribution from $\mathbf{x}_i$ excluded and explicitly given by $\alpha_j\alpha_k$.

We further denote $\boldsymbol{\alpha}_{-i} = \left[\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n\right]^\mathsf{T} \in \mathbb{R}^{n-1}$ and $\mathbf{K}_{-i}^\perp \in \mathbb{R}^{(n-1)\times(n-1)}$ with (j,k) entry given by

$$[\mathbf{K}_{-i}^\perp]_{jk} \equiv \delta_{j\neq k} f\left((\mathbf{x}_j^\perp)^\mathsf{T}\mathbf{x}_k^\perp/\sqrt{p}\right)/\sqrt{p} = \delta_{j\neq k} f(\mathbf{\Phi}_{jk}^\perp/\sqrt{p})/\sqrt{p} \tag{A.4}$$

so that $f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{x}_i/\sqrt{p}) \simeq f(\boldsymbol{\alpha}_{-i})$ in the sense that $\|f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{x}_i/\sqrt{p}) - f(\boldsymbol{\alpha}_{-i})\|/\sqrt{p} \to 0$ as $p \to \infty$.

It is worth remarking here that, intuitively speaking, the random vector $\boldsymbol{\alpha}_{-i}$ is merely a standard Gaussian random vector $\boldsymbol{\alpha}_{-i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-1})$ in the large n, p limit in the sense that its each entry is "asymptotically" Gaussian and uncorrelated with each other.

The advantage of introducing $\mathbf{\Phi}^\perp$ (as well as $\mathbf{K}_{-i}^\perp$) is that $\boldsymbol{\alpha}_{-i}$ is "essentially" asymptotically independent of $\mathbf{\Phi}^\perp$ in the sense that the expectation $\mathbb{E}[\mathbf{\Phi}^\perp\boldsymbol{\alpha}_{-i}]$ asymptotically vanishes. Note that this is in particular not the case for $\mathbb{E}[\mathbf{K}_{-i}\boldsymbol{\alpha}_{-i}]$ as an instance.

Since the study of $\mathbf{K}_{-i}$ boils down to that of $\mathbf{K}_{-i}^\perp$, we will need in the remainder its resolvent

$$\mathbf{Q}_{-i}^\perp \equiv \left(\mathbf{K}_{-i}^\perp - z\mathbf{I}_{n-1}\right)^{-1}$$

the resolvent of $\mathbf{K}_{-i}^\perp$ that is therefore also independent of $\boldsymbol{\alpha}_{-i}$.

Nonlinear quadratic forms. We first focus on $\mathbf{K}_{-i} \in \mathbb{R}^{(n-1)\times(n-1)}$. By (A.3), the (k,l)-entry of $\mathbf{K}_{-i} \equiv f(\mathbf{X}_{-i}^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})/\sqrt{p}$ is given by

$$[\mathbf{K}_{-i}]_{jk} = \frac{1}{\sqrt{p}} f\left(\frac{1}{\sqrt{p}}\alpha_j\alpha_k + \frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right)$$

where we recall that $\mathbf{\Phi}_{jk}^\perp/\sqrt{p} = O(1)$, $\alpha_j\alpha_k/\sqrt{p} = O(p^{-1/2})$ and they are independent. As a consequence, with a Taylor expansion of $f\left(\frac{1}{\sqrt{p}}\alpha_j\alpha_k + \frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right)$ around the dominant $\mathbf{\Phi}_{jk}^\perp/\sqrt{p}$ we obtain[1]

$$f\left(\frac{1}{\sqrt{p}}\alpha_j\alpha_k + \frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right) = f\left(\frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right) + f'\left(\frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right)\frac{1}{\sqrt{p}}\alpha_j\alpha_k + O(p^{-1})$$

so that in matrix form

$$\begin{aligned}
[\mathbf{K}_{-i}]_{jk} &= \frac{1}{\sqrt{p}} f\left(\frac{1}{\sqrt{p}}\alpha_j\alpha_k + \frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right) \\
&= \frac{1}{\sqrt{p}} f\left(\frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right) + \frac{1}{p}a_1\alpha_j\alpha_k + \frac{1}{p}g\left(\frac{1}{\sqrt{p}}\mathbf{\Phi}_{jk}^\perp\right)\alpha_j\alpha_k + O(p^{-3/2}) \\
&= [\mathbf{K}_{-i}^\perp]_{jk} + \frac{1}{p}a_1\left(\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T} - \operatorname{diag}(\boldsymbol{\alpha}_{-i}^2)\right)_{jk} + \frac{1}{p}\left(\operatorname{diag}(\boldsymbol{\alpha}_{-i})\mathbf{G}\operatorname{diag}(\boldsymbol{\alpha}_{-i})\right)_{jk} + O(p^{-3/2})
\end{aligned}$$

where we denote the shortcut $g(x) = f'(x) - a_1$, for $\mathbf{K}_{-i}^\perp$ given by (A.4), $[\boldsymbol{\alpha}_{-i}^2]_j = \alpha_j^2$ and $\mathbf{G} \equiv g(\mathbf{\Phi}^\perp/\sqrt{p}) \in \mathbb{R}^{(n-1)\times(n-1)}$.

The linear part $a_1 x$ of the nonlinear function $f(x)$ is treated separately since, intuitively speaking, taking the derivative of $f(x)$ with $a_0 = 0$ (see the last item of Assumption 3), results in $\mathbb{E}[f'(x)] = a_1 \neq 0$. The fact that $f'(x)$ is not centered (with respect to the Gaussian measure), together with $\|\boldsymbol{\alpha}_{-i}\| = O(\sqrt{p})$, implies that $\frac{a_1}{p}\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T}$ that has non-vanishing operator norm as $n, p \to \infty$. By subtracting a_1 from $f'(x)$, one obtains

$$[\mathbf{G}/\sqrt{p}]_{jk} = \delta_{j\neq k}\, g\left((\mathbf{x}_j^\perp)^\mathsf{T}\mathbf{x}_k^\perp/\sqrt{p}\right)/\sqrt{p}.$$

Since the (original) kernel matrix $\mathbf{K}$ is of bounded operator norm for all f with $a_0 = \mathbb{E}[f(\xi)] = 0$ and $\xi \sim \mathcal{N}(0,1)$ (see for a proof in [FM19, Theorem 1.7]), $\mathbf{G}/\sqrt{p}$, which can be seen as another inner-product kernel matrix with centered kernel function g (with $\mathbb{E}[g(\xi)] = 0$), has asymptotically bounded operator norm.

Further noting that $\operatorname{diag}(\boldsymbol{\alpha}_{-i}) = O_{\|\cdot\|}(1)$, $\operatorname{diag}(\boldsymbol{\alpha}_{-i}) = O_{\|\cdot\|}(1)$, we conclude that

$$\mathbf{K}_{-i} = \mathbf{K}_{-i}^\perp + \frac{a_1}{p}\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T} + o_{\|\cdot\|}(1) \tag{A.5}$$

where $o_{\|\cdot\|}(1)$ denotes a matrix with vanishing operator norm as $n, p \to \infty$. Here we use the fact that for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, we have $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ and $\|\mathbf{A}\| \leq n\|\mathbf{A}\|_\infty$.

We now move on to the quadratic form $\frac{1}{p}f(\mathbf{x}_i^\mathsf{T}\mathbf{X}_{-i}/\sqrt{p})\mathbf{Q}_{-i}f(\mathbf{X}_{-1}^\mathsf{T}\mathbf{x}_i/\sqrt{p})$, for $\mathbf{Q}_{-i} \equiv (\mathbf{K}_{-i} - z\mathbf{I}_{n-1})^{-1}$. As a consequence of (A.5) we deduce

$$\begin{aligned}
\mathbf{Q}_{-i} &\simeq \left(\mathbf{K}_{-i}^\perp + \frac{1}{p}a_1\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T} - z\mathbf{I}_{n-1}\right)^{-1} \\
&= \mathbf{Q}_{-i}^\perp - \frac{a_1\mathbf{Q}_{-i}^\perp\frac{1}{p}\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T}\mathbf{Q}_{-i}^\perp}{1 + \frac{a_1}{p}\boldsymbol{\alpha}_{-i}^\mathsf{T}\mathbf{Q}_{-i}^\perp\boldsymbol{\alpha}_{-i}} \simeq \mathbf{Q}_{-i}^\perp - \frac{a_1\mathbf{Q}_{-i}^\perp\frac{1}{p}\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T}\mathbf{Q}_{-i}^\perp}{1 + \frac{a_1}{p}\operatorname{tr}\mathbf{Q}_{-i}^\perp} \simeq \mathbf{Q}_{-i}^\perp - \frac{a_1\mathbf{Q}_{-i}^\perp\frac{1}{p}\boldsymbol{\alpha}_{-i}\boldsymbol{\alpha}_{-i}^\mathsf{T}\mathbf{Q}_{-i}^\perp}{1 + a - 1\frac{n}{p}m(z)}
\end{aligned}$$

[1] Here we consider for the moment f to be a Hermite polynomial, and then extend to square-summable f with Assumption 3.

where we recall $\mathbf{Q}_{-i}^{\perp} \equiv \left(\mathbf{K}_{-i}^{\perp} - z\mathbf{I}_{n-1}\right)^{-1}$ the resolvent of $\mathbf{K}_{-i}^{\perp}$ that is independent of $\boldsymbol{\alpha}_{-i}$. Here we use Lemma 2.8 for the equality and Lemma 2.11 for the approximation in the second line.

With the approximation above and (A.2), the quadratic form of crucial interest can be expanded as

$$\frac{1}{p} f(\mathbf{x}_i^{\mathsf{T}} \mathbf{X}_{-i}/\sqrt{p}) \mathbf{Q}_{-i} f(\mathbf{X}_{-1}^{\mathsf{T}} \mathbf{x}_i/\sqrt{p}) \simeq \frac{1}{p} f(\boldsymbol{\alpha}_{-i})^{\mathsf{T}} \mathbf{Q}_{-i}^{\perp} f(\boldsymbol{\alpha}_{-i}) - a_1 \frac{\left(\frac{1}{p} \boldsymbol{\alpha}_{-i}^{\mathsf{T}} \mathbf{Q}_{-i}^{\perp} f(\boldsymbol{\alpha}_{-i})\right)^2}{1 + a_1 \frac{n}{p} m(z)}$$

$$\simeq \frac{a_1^2}{p} \boldsymbol{\alpha}_{-i}^{\mathsf{T}} \mathbf{Q}_{-i}^{\perp} \boldsymbol{\alpha}_{-i} + \frac{1}{p} f_{>1}(\boldsymbol{\alpha}_{-i}) \mathbf{Q}_{-i}^{\perp} f_{>1}(\boldsymbol{\alpha}_{-i}) - a_1 \frac{\left(\frac{a_1}{p} \boldsymbol{\alpha}_{-i}^{\mathsf{T}} \mathbf{Q}_{-i}^{\perp} \boldsymbol{\alpha}_{-i}\right)^2}{1 + a_1 \frac{n}{p} m(z)}$$

where we write the Hermite polynomial $f(x)$ as the sum of its linear part $a_1 x$ and the purely nonlinear part $f_{>1}(x) = f(x) - a_1 x$ that is *orthogonal* to x in the sense that $\mathbb{E}_{\xi}[\xi f_{>1}(\xi)] = 0$ for $\xi \sim \mathcal{N}(0,1)$. Again by orthogonality, the cross terms of $\boldsymbol{\alpha}_{-i}$ and $f_{>1}(\boldsymbol{\alpha}_{-i})$ vanish. Also, since

$$\frac{1}{p} f_{>1}(\boldsymbol{\alpha}_{-i}) \mathbf{Q}_{-i}^{\perp} f_{>1}(\boldsymbol{\alpha}_{-i}) \simeq (\nu - a_1^2) \frac{n}{p} m(z)$$

one obtains the following approximation for the nonlinear quadratic form of interest

$$\frac{1}{p} f(\mathbf{x}_i^{\mathsf{T}} \mathbf{X}_{-i}/\sqrt{p}) \mathbf{Q}_{-i} f(\mathbf{X}_{-1}^{\mathsf{T}} \mathbf{x}_i/\sqrt{p}) \simeq \frac{a_1^2 \frac{n}{p} m(z)}{1 + a_1 \frac{n}{p} m(z)} + (\nu - a_1^2) \frac{n}{p} m(z).$$

Ultimately with (A.1) one has

$$m(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{-z - \frac{1}{p} f(\mathbf{x}_i^{\mathsf{T}} \mathbf{X}_{-i}/\sqrt{p}) \mathbf{Q}_{-i} f(\mathbf{X}_{-1}^{\mathsf{T}} \mathbf{x}_i/\sqrt{p})} \simeq \frac{1}{-z - \frac{\frac{a_1^2}{c} m(z)}{1 + \frac{a_1}{c} m(z)} - \frac{\nu - a_1^2}{c} m(z).}$$

which concludes the proof of Theorem 1.2.

A.1.2 Proof of Theorem 3.1

Our interest here is on the decision function of LS-SVM: $g(\mathbf{x}) = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{k}(\mathbf{x}) + b$ with $(\boldsymbol{\beta}, b)$ given by

$$\begin{cases} \beta &= \mathbf{Q} \left(\mathbf{I} - \frac{\mathbf{1}_n \mathbf{1}_n^{\mathsf{T}} \mathbf{Q}}{\mathbf{1}_n^{\mathsf{T}} \mathbf{Q} \mathbf{1}_n} \right) \mathbf{y} \\ b &= \frac{\mathbf{1}_n^{\mathsf{T}} \mathbf{Q} \mathbf{y}}{\mathbf{1}_n^{\mathsf{T}} \mathbf{Q} \mathbf{1}_n} \end{cases}$$

and $\mathbf{Q} = \left(\mathbf{K} + \frac{n}{\gamma} \mathbf{I} \right)^{-1}$.

Before going into the detailed proof, as we will frequently deal with *random* variables evolving as n, p grow large, we will use the extension of the $O(\cdot)$ notation introduced in [CBG16]: for a random variable $x \equiv x_n$ and $u_n \geq 0$, we write $x = O(u_n)$ if for any $\eta > 0$ and $D > 0$, we have $n^D \mathrm{P}(x \geq n^{\eta} u_n) \to 0$. Note that under Assumption 4 it is equivalent to use either $O(u_n)$ or $O(u_p)$ since n, p scale linearly. In the following we shall use constantly $O(u_n)$ for simplicity.

When multidimensional objects are concerned, $\mathbf{v} = O(u_n)$ means the maximum entry of a vector (or a diagonal matrix) $\mathbf{v}$ in absolute value is of order $O(u_n)$ and $\mathbf{M} = O(u_n)$ means that the operator norm of $\mathbf{M}$ is of order $O(u_n)$. We refer the reader to [CBG16] for more discussions on these practical definitions.

Under the growth rate settings of Assumption 4, from [CBG16], the approximation of the kernel matrix $\mathbf{K}$ is given by

$$\mathbf{K} = -2f'(\tau)\left(\mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P} + \mathbf{A}\right) + \beta\mathbf{I} + O(n^{-\frac{1}{2}})$$

with $\beta = f(0) - f(\tau) + \tau f'(\tau)$ and $\mathbf{A} = \mathbf{A}_n + \mathbf{A}_{\sqrt{n}} + \mathbf{A}_1$, $\mathbf{A}_n = -\frac{f(\tau)}{2f'(\tau)}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ and $\mathbf{A}_{\sqrt{n}}$, $\mathbf{A}_1$ given by (A.6) and (A.7) at the top of next page, where we denote

$$t_a \equiv \frac{\operatorname{tr}(\mathbf{C}_a - \mathbf{C}^\circ)}{\sqrt{p}} = O(1), \quad (\boldsymbol{\psi})^2 \equiv [(\boldsymbol{\psi}_1)^2, \ldots, (\boldsymbol{\psi}_n)^2]^\mathsf{T}.$$

$$\mathbf{A}_{\sqrt{n}} = -\frac{1}{2}\left[\boldsymbol{\psi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\psi}^\mathsf{T} + \left\{t_a\frac{\mathbf{1}_{n_a}}{\sqrt{p}}\right\}_{a=1}^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\left\{t_b\frac{\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{b=1}^2\right] \tag{A.6}$$

$$\begin{aligned}
\mathbf{A}_1 = {} & -\frac{1}{2}\left\{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{a,b=1}^2 - \left\{\frac{(\boldsymbol{\Omega}\mathbf{P})_a^\mathsf{T}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_a)\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{a,b=1}^2 + \left\{\frac{\mathbf{1}_{n_a}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_a)^\mathsf{T}(\boldsymbol{\Omega}\mathbf{P})_b}{\sqrt{p}}\right\}_{a,b=1}^2 \\
& - \frac{f''(\tau)}{4f'(\tau)}\left[(\boldsymbol{\psi})^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n[(\boldsymbol{\psi})^2]^\mathsf{T} + \left\{t_a^2\frac{\mathbf{1}_{n_a}}{p}\right\}_{a=1}^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\left\{t_b^2\frac{\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{b=1}^2 + 2\left\{t_at_b\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p}\right\}_{a,b=1}^2\right. \\
& + 2\operatorname{diag}\{t_a\mathbf{I}_{n_a}\}_{a=1}^2\boldsymbol{\psi}\frac{\mathbf{1}_n^\mathsf{T}}{\sqrt{p}} + 2\boldsymbol{\psi}\left\{t_b\frac{\mathbf{1}_{n_b}^\mathsf{T}}{\sqrt{p}}\right\}_{b=1}^2 + 2\frac{\mathbf{1}_n}{\sqrt{p}}(\boldsymbol{\psi})^\mathsf{T}\operatorname{diag}\{t_a\mathbf{1}_{n_a}\}_{a=1}^2 \\
& \left. + 2\left\{t_a\frac{\mathbf{1}_{n_a}}{\sqrt{p}}\right\}_{a=1}^2(\boldsymbol{\psi})^\mathsf{T} + 4\left\{\operatorname{tr}(\mathbf{C}_a\mathbf{C}_b)\frac{\mathbf{1}_{n_a}\mathbf{1}_{n_b}^\mathsf{T}}{p^2}\right\}_{a,b=1}^2 + 2\boldsymbol{\psi}(\boldsymbol{\psi})^\mathsf{T}\right]
\end{aligned} \tag{A.7}$$

We start with the resolvent $\mathbf{Q}$. The terms of leading order in $\mathbf{K}$, i.e.,$-2f'(\tau)\mathbf{A}_n$ and $\frac{n}{\gamma}\mathbf{I}$ are both of operator norm $O(n)$. Therefore a Taylor expansion can be performed as

$$\begin{aligned}
\mathbf{Q} &= \left(\mathbf{K} + \frac{n}{\gamma}\mathbf{I}\right)^{-1} = \frac{1}{n}\left[\mathbf{L}^{-1} - \frac{2f'(\tau)}{n}\left(\mathbf{A}_{\sqrt{n}} + \mathbf{A}_1 + \mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P}\right) + \frac{\beta\mathbf{I}}{n} + O(n^{-\frac{3}{2}})\right]^{-1} \\
&= \frac{\mathbf{L}}{n} + \frac{2f'(\tau)}{n^2}\mathbf{L}\mathbf{A}_{\sqrt{n}}\mathbf{L} + \mathbf{L}\left(\mathbf{B} - \frac{\beta}{n^2}\mathbf{I}\right)\mathbf{L} + O(n^{-\frac{5}{2}})
\end{aligned}$$

with $\mathbf{L} = \left(f(\tau)\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{\mathbf{I}}{\gamma}\right)^{-1}$ of order $O(1)$ and $\mathbf{B} = \frac{2f'(\tau)}{n^2}\left(\mathbf{A}_1 + \mathbf{P}\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega}\mathbf{P} + \frac{2f'(\tau)}{n}\mathbf{A}_{\sqrt{n}}\mathbf{L}\mathbf{A}_{\sqrt{n}}\right)$.

With the Sherman-Morrison formula we are able to compute explicitly $\mathbf{L}$ as

$$\begin{aligned}
\mathbf{L} &= \left(f(\tau)\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{\mathbf{I}}{\gamma}\right)^{-1} = \gamma\left(\mathbf{I} - \frac{\gamma f(\tau)}{1+\gamma f(\tau)}\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\right) \\
&= \frac{\gamma}{1+\gamma f(\tau)}\mathbf{I} + \frac{\gamma^2 f(\tau)}{1+\gamma f(\tau)}\mathbf{P} = O(1).
\end{aligned} \tag{A.8}$$

Writing $\mathbf{L}$ as a linear combination of $\mathbf{I}$ and $\mathbf{P}$ is useful when computing $\mathbf{L}\mathbf{1}_n$ or $\mathbf{1}_n^\mathsf{T}\mathbf{L}$, because by the definition of $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, we have $\mathbf{1}_n^\mathsf{T}\mathbf{P} = \mathbf{P}\mathbf{1}_n = \mathbf{0}$.

We shall start with the term $\mathbf{1}_n^\mathsf{T}\mathbf{Q}$, since it is the basis of several other terms appearing in $\boldsymbol{\beta}$ and b,

$$\mathbf{1}_n^\mathsf{T}\mathbf{Q} = \frac{\gamma\mathbf{1}_n^\mathsf{T}}{1+\gamma f(\tau)}\left[\frac{\mathbf{I}}{n} + \frac{2f'(\tau)}{n^2}\mathbf{A}_{\sqrt{n}}\mathbf{L} + \left(\mathbf{B} - \frac{\beta}{n^2}\mathbf{I}\right)\mathbf{L}\right] + O(n^{-\frac{3}{2}})$$

since $\mathbf{1}_n^\mathsf{T}\mathbf{L} = \frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}$.

With $\mathbf{1}_n^\mathsf{T}\mathbf{Q}$ at hand, we next obtain,

$$\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q} = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{n}}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{L}}_{O(n^{-1/2})} + \underbrace{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{B}-\frac{\beta}{n^2}\mathbf{I}\right)\mathbf{L}}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}) \quad \text{(A.9)}$$

$$\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{y} = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{c_2-c_1}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{L}\mathbf{y}}_{O(n^{-1/2})} + \underbrace{\mathbf{1}_n^\mathsf{T}\left(\mathbf{B}-\frac{\beta}{n^2}\mathbf{I}\right)\mathbf{L}\mathbf{y}}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}) \quad \text{(A.10)}$$

$$\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n = \frac{\gamma}{1+\gamma f(\tau)}\Bigg[\underbrace{1}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\frac{\gamma\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{1}_n}{1+\gamma f(\tau)}}_{O(n^{-1/2})} + \underbrace{\frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}\left(\mathbf{B}-\frac{\beta}{n^2}\mathbf{I}\right)\mathbf{1}_n}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}). \quad \text{(A.11)}$$

The inverse of $\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n$ can consequently be computed using a Taylor expansion around its leading order, allowing an error term of $O(n^{-\frac{3}{2}})$ as

$$\frac{1}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n} = \frac{1+\gamma f(\tau)}{\gamma}\Bigg[\underbrace{1}_{O(1)} - \underbrace{\frac{2f'(\tau)}{n^2}\frac{\gamma\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\mathbf{1}_n}{1+\gamma f(\tau)}}_{O(n^{-1/2})} - \underbrace{\frac{\gamma}{1+\gamma f(\tau)}\mathbf{1}_n^\mathsf{T}\left(\mathbf{B}-\frac{\beta}{n^2}\mathbf{I}\right)\mathbf{1}_n}_{O(n^{-1})}\Bigg] + O(n^{-\frac{3}{2}}). \quad \text{(A.12)}$$

Combing (A.9) with (A.12) we deduce

$$\begin{aligned}\frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\mathbf{1}_n} = \underbrace{\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}_{O(1)} + \underbrace{\frac{2f'(\tau)}{n^2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{A}_{\sqrt{n}}\left[\mathbf{L}-\frac{\gamma\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+\gamma f(\tau)}\right]}_{O(n^{-1/2})} \\ + \underbrace{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\left(\mathbf{B}-\frac{\beta}{n^2}\mathbf{I}\right)\left[\mathbf{L}-\frac{\gamma\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+\gamma f(\tau)}\right]}_{O(n^{-1})} + O(n^{-\frac{3}{2}}) \quad \text{(A.13)}\end{aligned}$$

and similarly the following approximation of b as

$$\begin{aligned}b = \underbrace{c_2-c_1}_{O(1)} - \underbrace{\frac{2\gamma}{\sqrt{p}}c_1c_2f'(\tau)(t_2-t_1)}_{O(n^{-1/2})} - \underbrace{\frac{\gamma f'(\tau)}{n}\mathbf{y}^\mathsf{T}\mathbf{P}\boldsymbol{\psi}}_{O(n^{-1})} \\ \underbrace{-\frac{\gamma f''(\tau)}{2n}\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\psi})^2 + \frac{4\gamma c_1c_2}{p}[c_1T_1+(c_2-c_1)D-c_2T_2]}_{O(n^{-1})} + O(n^{-\frac{3}{2}}) \quad \text{(A.14)}\end{aligned}$$

where

$$\begin{aligned}D &= \frac{f'(\tau)}{2}\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|^2 + \frac{f''(\tau)}{4}(t_1+t_2)^2 + f''(\tau)\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_2}{p} \\ T_a &= f''(\tau)t_a^2 + f''(\tau)\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_2}{p}\end{aligned}$$

which gives the asymptotic approximation of b.

Moving to $\boldsymbol{\beta}$, note from (A.8) that $\mathbf{L} - \frac{\gamma}{1+\gamma f(\tau)} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} = \gamma \mathbf{P}$, and we can thus rewrite:

$$\frac{\mathbf{1}_n \mathbf{1}_n^\mathsf{T} \mathbf{Q}}{\mathbf{1}_n^\mathsf{T} \mathbf{Q} \mathbf{1}_n} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} + \frac{2\gamma f'(\tau)}{n^2} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \mathbf{A}_{\sqrt{n}} \mathbf{P} + \gamma \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \left(\mathbf{B} - \frac{\beta}{n^2} \mathbf{I} \right) \mathbf{P} + O(n^{-\frac{3}{2}}).$$

At this point, for $\boldsymbol{\beta} = \mathbf{Q} \left(\mathbf{I} - \frac{\mathbf{1}_n \mathbf{1}_n^\mathsf{T} \mathbf{Q}}{\mathbf{1}_n^\mathsf{T} \mathbf{Q} \mathbf{1}_n} \right) \mathbf{y}$, we have

$$\boldsymbol{\beta} = \mathbf{Q} \left[\mathbf{I} - \frac{2\gamma f'(\tau)}{n^2} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \mathbf{A}_{\sqrt{n}} - \gamma \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \left(\mathbf{B} - \frac{\beta}{n^2} \mathbf{I} \right) \right] \mathbf{P} \mathbf{y} + O(n^{-\frac{5}{2}}).$$

Here again, we use $\mathbf{1}_n^\mathsf{T} \mathbf{L} = \frac{\gamma}{1+\gamma f(\tau)} \mathbf{1}_n^\mathsf{T}$ and $\mathbf{L} - \frac{\gamma}{1+\gamma f(\tau)} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} = \gamma \mathbf{P}$, to eventually get

$$\boldsymbol{\beta} = \underbrace{\frac{\gamma}{n} \mathbf{P} \mathbf{y}}_{O(n^{-1})} + \underbrace{\gamma^2 \mathbf{P} \left(\mathbf{B} - \frac{\beta}{n^2} \mathbf{I} \right) \mathbf{P} \mathbf{y}}_{O(n^{-2})} - \underbrace{\frac{\gamma^2}{1+\gamma f(\tau)} \left(\frac{2f'(\tau)}{n^2} \right)^2 \mathbf{L} \mathbf{A}_{\sqrt{n}} \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \mathbf{A}_{\sqrt{n}} \mathbf{P} \mathbf{y}}_{O(n^{-2})} + O(n^{-\frac{5}{2}}). \tag{A.15}$$

Note here the absence of a term of order $O(n^{-3/2})$ in the expression of $\boldsymbol{\beta}$ since $\mathbf{P} \mathbf{A}_{\sqrt{n}} \mathbf{P} = 0$ from (A.6).

We now work on the vector $\mathbf{k}(\mathbf{x})$ for a new datum $\mathbf{x}$, following the same analysis as in [CBG16] for the kernel matrix $\mathbf{K}$, assuming that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$ and recalling the random variables definitions,

$$\begin{aligned} \boldsymbol{\omega}_\mathbf{x} &\equiv (\mathbf{x} - \boldsymbol{\mu}_a)/\sqrt{p} \\ \psi_\mathbf{x} &\equiv \|\boldsymbol{\omega}_\mathbf{x}\|^2 - \mathbb{E}\|\boldsymbol{\omega}_\mathbf{x}\|^2. \end{aligned}$$

We show that the j-th entry of $\mathbf{k}(\mathbf{x})$ can be written as

$$\begin{aligned} [\mathbf{k}(\mathbf{x})]_j = \underbrace{f(\tau)}_{O(1)} + f'(\tau) \Bigg[\underbrace{\frac{t_a + t_b}{\sqrt{p}} + \psi_x + \psi_j - 2(\boldsymbol{\omega}_\mathbf{x})^\mathsf{T} \boldsymbol{\omega}_j}_{O(n^{-1/2})} + \underbrace{\frac{\|\boldsymbol{\mu}_b - \boldsymbol{\mu}_a\|^2}{p} + \frac{2}{\sqrt{p}} (\boldsymbol{\mu}_b - \boldsymbol{\mu}_a)^\mathsf{T} (\boldsymbol{\omega}_j - \boldsymbol{\omega}_\mathbf{x})}_{O(n^{-1})} \Bigg] \\ + \frac{f''(\tau)}{2} \Bigg[\underbrace{\left(\frac{t_a + t_b}{\sqrt{p}} + \psi_j + \psi_\mathbf{x} \right)^2 + \frac{4}{p^2} \operatorname{tr} \mathbf{C}_a \mathbf{C}_b}_{O(n^{-1})} \Bigg] + O(n^{-\frac{3}{2}}). \end{aligned} \tag{A.16}$$

Combining (A.15) and (A.16), we deduce

$$\boldsymbol{\beta}^\mathsf{T} \mathbf{k}(\mathbf{x}) = \underbrace{\frac{2\gamma}{\sqrt{p}} c_1 c_2 f'(\tau)(t_2 - t_1)}_{O(n^{-1/2})} + \underbrace{\frac{\gamma}{n} \mathbf{y}^\mathsf{T} \mathbf{P} \tilde{\mathbf{k}}(\mathbf{x})}_{O(n^{-1})} + \underbrace{\frac{\gamma f'(\tau)}{n} \mathbf{y}^\mathsf{T} \mathbf{P} (\boldsymbol{\psi} - 2\mathbf{P} \boldsymbol{\Omega}^\mathsf{T} \boldsymbol{\omega}_\mathbf{x})}_{O(n^{-1})} + O(n^{-\frac{3}{2}}) \tag{A.17}$$

with $\tilde{\mathbf{k}}(\mathbf{x})$ given as follows.

$$\tilde{\mathbf{k}}(\mathbf{x}) = f'(\tau)\left\{\frac{\|\boldsymbol{\mu}_b-\boldsymbol{\mu}_a\|^2}{p}\mathbf{1}_{n_b}\right\}_{b=1}^2 - \frac{2f'(\tau)}{\sqrt{p}}\left\{\mathbf{1}_{n_b}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^\mathsf{T}\right\}_{b=1}^2\boldsymbol{\omega}_\mathbf{x}$$
$$+\frac{2f'(\tau)}{\sqrt{p}}\operatorname{diag}\left(\left\{\mathbf{1}_{n_b}(\boldsymbol{\mu}_b-\boldsymbol{\mu}_a)^\mathsf{T}\right\}_{b=1}^2\boldsymbol{\Omega}\right)+\frac{f''(\tau)}{2}\left\{\frac{(t_a+t_b)^2}{p}\mathbf{1}_{n_b}\right\}_{b=1}^2$$
$$+\frac{f''(\tau)}{2}\left[2\operatorname{diag}\left(\left\{\frac{t_a+t_b}{\sqrt{p}}\mathbf{1}_{n_b}\right\}_{b=1}^2\right)\boldsymbol{\psi}+2\left\{\frac{t_a+t_b}{\sqrt{p}}\mathbf{1}_{n_b}\right\}_{b=1}^2\psi_\mathbf{x}+(\boldsymbol{\psi})^2+2\psi_\mathbf{x}\boldsymbol{\psi}+\psi_\mathbf{x}^2\mathbf{1}_n\right.$$
$$\left.+\left\{\frac{4}{p^2}\operatorname{tr}(\mathbf{C}_a\mathbf{C}_b)\mathbf{1}_{n_b}\right\}_{b=1}^2\right] \tag{A.18}$$

At this point, note that the term of order $O(n^{-\frac{1}{2}})$ in the final object $g(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b$ disappears because in both (A.14) and (A.17) the term of order $O(n^{-1/2})$ is $\frac{2\gamma}{\sqrt{p}}c_1c_2f'(\tau)(t_2 - t_1)$ but of opposite signs. Also, we see that the leading term $c_2 - c_1$ in b will remain in $g(\mathbf{x})$ as stated in Remark 3.1.

The development of $\mathbf{y}^\mathsf{T}\mathbf{P}\tilde{\mathbf{k}}(\mathbf{x})$ induces many simplifications, since i) $\mathbf{P}\mathbf{1}_n = \mathbf{0}$ and ii) random variables as $\boldsymbol{\omega}_\mathbf{x}$ and $\boldsymbol{\psi}$ in $\tilde{\mathbf{k}}(\mathbf{x})$, once multiplied by $\mathbf{y}^\mathsf{T}\mathbf{P}$, thanks to probabilistic averaging of independent zero-mean terms, are of smaller order and thus become negligible. We thus get

$$\frac{\gamma}{n}\mathbf{y}^\mathsf{T}\mathbf{P}\tilde{\mathbf{k}}(\mathbf{x}) = 2\gamma c_1c_2f'(\tau)\left[\frac{\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_a\|^2-\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_a\|^2}{p}-2(\boldsymbol{\omega}_\mathbf{x})^\mathsf{T}\frac{\boldsymbol{\mu}_2-\boldsymbol{\mu}_1}{\sqrt{p}}\right]+\frac{\gamma f''(\tau)}{2n}\mathbf{y}^\mathsf{T}\mathbf{P}(\boldsymbol{\psi})^2$$
$$+\gamma c_1c_2f''(\tau)\left[2\left(\frac{t_a}{\sqrt{p}}+\psi_\mathbf{x}\right)\frac{t_2-t_1}{\sqrt{p}}+\frac{t_2^2-t_1^2}{p}+\frac{4}{p^2}\operatorname{tr}\mathbf{C}_a(\mathbf{C}_2-\mathbf{C}_1)\right]+O(n^{-\frac{3}{2}}). \tag{A.19}$$

This result, together with (A.17), completes the analysis of the term $\boldsymbol{\beta}^\mathsf{T}\mathbf{k}(\mathbf{x})$. Combining (A.17)-(A.19) with (A.14) we conclude the proof of Theorem 3.1.

A.1.3 Proof of Theorem 3.2

This section is dedicated to the proof of the central limit theorem for

$$\tilde{g}(\mathbf{x}) = c_2 - c_1 + \gamma\left(\mathfrak{R} + c_\mathbf{x}\mathfrak{D}\right)$$

with the shortcut $c_\mathbf{x} = -2c_1c_2^2$ for $\mathbf{x} \in \mathcal{C}_1$ and $c_\mathbf{x} = 2c_1^2c_2$ for $\mathbf{x} \in \mathcal{C}_2$, and $\mathfrak{R}, \mathfrak{D}$ as defined in (3.8) and (3.9).

Our objective is to show that, for $a \in \{1, 2\}$, $n(\tilde{g}(\mathbf{x}) - G_a) \to 0$ in distribution with

$$G_a \sim \mathcal{N}(\bar{G}_a, V_{G_a})$$

where $\bar{G}_a$ and V_{G_a} are given in Theorem 3.2. We recall that $\mathbf{x} = \boldsymbol{\mu}_a + \sqrt{p}\boldsymbol{\omega}_\mathbf{x}$ with $\boldsymbol{\omega}_\mathbf{x} \sim \mathcal{N}(0, \mathbf{C}_a/p)$.

Letting $\mathbf{z}_\mathbf{x}$ such that $\boldsymbol{\omega}_\mathbf{x} = \mathbf{C}_a^{1/2}\mathbf{z}_\mathbf{x}/\sqrt{p}$, we have $\mathbf{z}_\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and we can rewrite $\tilde{g}(\mathbf{x})$ in the following quadratic form (of $\mathbf{z}_\mathbf{x}$) as

$$\tilde{g}(\mathbf{x}) = \mathbf{z}_\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{z}_\mathbf{x} + \mathbf{z}_\mathbf{x}^\mathsf{T}\mathbf{b} + c$$

with

$$\mathbf{A} = 2\gamma c_1 c_2 f''(\tau)\frac{\operatorname{tr}(\mathbf{C}_2-\mathbf{C}_1)}{p}\frac{\mathbf{C}_a}{p}$$
$$\mathbf{b} = -\frac{2\gamma f'(\tau)}{n}\frac{(\mathbf{C}_a)^{\frac{1}{2}}}{\sqrt{p}}\mathbf{\Omega}\mathbf{P}\mathbf{y} - \frac{4c_1c_2\gamma f'(\tau)}{\sqrt{p}}\frac{(\mathbf{C}_a)^{\frac{1}{2}}}{\sqrt{p}}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)$$
$$c = c_2 - c_1 + \gamma c_{\mathbf{x}}\mathfrak{D} - 2\gamma c_1 c_2 f''(\tau)\frac{\operatorname{tr}(\mathbf{C}_2-\mathbf{C}_1)}{p}\frac{\operatorname{tr}\mathbf{C}_a}{p}.$$

Since $\mathbf{z}_{\mathbf{x}}$ is standard Gaussian and has the same distribution as $\mathbf{U}\mathbf{z}_{\mathbf{x}}$ for any orthogonal matrix $\mathbf{U}$ (i.e., such that $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{I}$), we choose $\mathbf{U}$ that diagonalizes $\mathbf{A}$ such that $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$, with $\mathbf{\Lambda}$ diagonal so that $\tilde{g}(\mathbf{x})$ and $\tilde{g}(\mathbf{x})$ have the same distribution where

$$\tilde{g}(\mathbf{x}) = \mathbf{z}_{\mathbf{x}}^\mathsf{T}\mathbf{\Lambda}\mathbf{z}_{\mathbf{x}} + \mathbf{z}_{\mathbf{x}}^\mathsf{T}\tilde{\mathbf{b}} + c = \sum_{i=1}^n \left(z_i^2\lambda_i + z_i\tilde{b}_i + \frac{c}{n}\right)$$

and $\tilde{\mathbf{b}} = \mathbf{U}^\mathsf{T}\mathbf{b}$, λ_i the diagonal elements of $\mathbf{\Lambda}$ and z_i the elements of $\mathbf{z}_{\mathbf{x}}$.

Conditioning on $\mathbf{\Omega}$, this results in the sum of independent but not identically distributed random variables $r_i = z_i^2\lambda_i + z_i\tilde{b}_i + \frac{c}{n}$. We then resort to the Lyapunov CLT [Bil12, Theorem 27.3].

We begin by estimating the expectation and the variance

$$\mathbb{E}[r_i|\mathbf{\Omega}] = \lambda_i + \frac{c}{n}$$
$$\operatorname{Var}[r_i|\mathbf{\Omega}] = \sigma_i^2 = 2\lambda_i^2 + \tilde{b}_i^2$$

of r_i, so that

$$\sum_{i=1}^n \mathbb{E}[r_i|\mathbf{\Omega}] = c_2 - c_1 + \gamma c_{\mathbf{x}}\mathfrak{D} = \bar{G}_a$$
$$s_n^2 = \sum_{i=1}^n \sigma_i^2 = 2\operatorname{tr}(\mathbf{A}^2) + \mathbf{b}^\mathsf{T}\mathbf{b}$$
$$= 8\gamma^2 c_1^2 c_2^2 \left(f''(\tau)\right)^2 \frac{\left(\operatorname{tr}\left(\mathbf{C}_2-\mathbf{C}_1\right)\right)^2}{p^2}\frac{\operatorname{tr}\mathbf{C}_a^2}{p^2}$$
$$+ 4\gamma^2\left(\frac{f'(\tau)}{n}\right)^2 \mathbf{y}^\mathsf{T}\mathbf{P}\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega}\mathbf{P}\mathbf{y} + \frac{16\gamma^2c_1^2c_2^2(f'(\tau))^2}{p}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\mathsf{T}\frac{\mathbf{C}_a}{p}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + O(n^{-\frac{5}{2}}).$$

We shall rewrite $\mathbf{\Omega}$ into two blocks as:

$$\mathbf{\Omega} = \left[\frac{(\mathbf{C}_1)^{\frac{1}{2}}}{\sqrt{p}}\mathbf{Z}_1, \quad \frac{(\mathbf{C}_2)^{\frac{1}{2}}}{\sqrt{p}}\mathbf{Z}_2\right]$$

where $\mathbf{Z}_1 \in \mathbb{R}^{p\times n_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{p\times n_2}$ with i.i.d. Gaussian entries with zero mean and unit variance. Then

$$\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega} = \frac{1}{p^2}\begin{bmatrix}\mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{Z}_1 & \mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2 \\ \mathbf{Z}_2^\mathsf{T}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{Z}_1 & \mathbf{Z}_2^\mathsf{T}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\end{bmatrix}$$

and with $\mathbf{P}\mathbf{y} = \mathbf{y} - (c_2-c_1)\mathbf{1}_n$, we deduce

$$\mathbf{y}^\mathsf{T}\mathbf{P}\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega}\mathbf{P}\mathbf{y} = \frac{4}{p^2}\Big(c_2^2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_1)^{\frac{1}{2}}bZ_1\mathbf{1}_{n_1}$$
$$-2c_1c_2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_1^\mathsf{T}(\mathbf{C}_1)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\mathbf{1}_{n_2} + c_2^2\mathbf{1}_{n_1}^\mathsf{T}\mathbf{Z}_2^\mathsf{T}(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{C}_a(\mathbf{C}_2)^{\frac{1}{2}}\mathbf{Z}_2\mathbf{1}_{n_2}\Big).$$

Since $\mathbf{Z}_i\mathbf{1}_{n_i} \sim \mathcal{N}(\mathbf{0}, n_i\mathbf{I}_{n_i})$, by applying the trace lemma (Lemma 2.11), we get

$$\mathbf{y}^\mathsf{T}\mathbf{P}\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega}\mathbf{P}\mathbf{y} - \frac{4nc_1^2c_2^2}{p^2}\left(\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\operatorname{tr}\mathbf{C}_2\mathbf{C}_a}{c_2}\right) \xrightarrow{a.s.} 0. \tag{A.20}$$

Consider now the events

$$E = \left\{\left|\mathbf{y}^\mathsf{T}\mathbf{P}\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega}\mathbf{P}\mathbf{y} - \rho\right| < \epsilon\right\}$$
$$\bar{E} = \left\{\left|\mathbf{y}^\mathsf{T}\mathbf{P}\mathbf{\Omega}^\mathsf{T}\frac{\mathbf{C}_a}{p}\mathbf{\Omega}\mathbf{P}\mathbf{y} - \rho\right| > \epsilon\right\}$$

for any fixed ϵ with $\rho = \frac{4nc_1^2c_2^2}{p^2}\left(\frac{\operatorname{tr}\mathbf{C}_1\mathbf{C}_a}{c_1} + \frac{\operatorname{tr}\mathbf{C}_2\mathbf{C}_a}{c_2}\right)$ and write

$$\begin{aligned}\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \bar{G}_a}{s_n}\right)\right] &= \mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \bar{G}_a}{s_n}\right)\middle|E\right]P(E) \\ &\quad + \mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \bar{G}_a}{s_n}\right)\middle|\bar{E}\right]P(\bar{E})\end{aligned} \tag{A.21}$$

We start with the variable $\tilde{g}(\mathbf{x})|E$ and check that Lyapunov's condition for $\bar{r}_i = r_i - \mathbb{E}[r_i]$,

$$\lim_{n\to\infty}\frac{1}{s_n^4}\sum_{i=1}^{n}\mathbb{E}[|\bar{r}_i|^4] = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{60\lambda_i^4 + 12\lambda_i^2\tilde{b}_i^{\,2} + 3\tilde{b}_i^{\,4}}{s_n^4} = 0$$

since both λ_i and $\tilde{b}_i$ are of order $O(n^{-3/2})$.

As a consequence, we have the CLT for the random variable $\tilde{g}(\mathbf{x})|E$. Thus

$$\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \bar{G}_a}{s_n}\right)\middle|E\right] \to \exp(-u^2/2).$$

Next, we see that the second term in (A.21) goes to zero because $\left|\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x})-\bar{G}_a}{s_n}\right)|\bar{E}\right]\right| \leq 1$ and $P(\bar{E}) \to 0$ from (A.20). We eventually deduce

$$\mathbb{E}\left[\exp\left(iun\frac{\tilde{g}(\mathbf{x}) - \bar{G}_a}{s_n}\right)\right] \to \exp(-u^2/2).$$

With the help of Lévy's continuity theorem, we thus prove the CLT of the variable $n\frac{\tilde{g}(\mathbf{x})-\bar{G}_a}{s_n}$. Since $s_n^2 \to V_{G_a}$, with Slutsky's theorem, we have the CLT for $n\frac{\tilde{g}(\mathbf{x})-\bar{G}_a}{\sqrt{V_{G_a}}}$ and eventually for $n\frac{g(\mathbf{x})-\bar{G}_a}{\sqrt{V_{G_a}}}$ by Theorem 3.1 which completes the proof.

A.1.4 Proof of Proposition 3.2

We aim to prove, for $f(x) = x^k$ with $k \geq 2$, the informative kernel matrix $\mathbf{K}_I$ in Proposition 3.2 defined as

$$\mathbf{K}_I = \frac{k}{\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ (\mathbf{A} + \mathbf{B}) + \frac{k(k-1)}{2\sqrt{p}}(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-2)} \circ (\mathbf{A})^{\circ 2} \tag{A.22}$$

for $\mathbf{A}_{ij} \equiv \delta_{i\neq j}\frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a+\mathbf{E}_b)\mathbf{z}_j}{2\sqrt{p}}$ and $\mathbf{B}_{ij} \equiv \delta_{i\neq j}\frac{\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b+\boldsymbol{\mu}_a^\mathsf{T}\mathbf{z}_j+\boldsymbol{\mu}_b^\mathsf{T}\mathbf{z}_i}{\sqrt{p}} - \frac{\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a-\mathbf{E}_b)^2\mathbf{z}_j}{8\sqrt{p}}$, has a tractable low-rank approximation $\tilde{\mathbf{K}}$ given by

$$\tilde{\mathbf{K}}_I = \begin{cases} \frac{k!!}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}), & \text{for } k \text{ odd;} \\ \frac{k(k-1)!!}{2p}\mathbf{J}\left(\mathbf{T}+\mathbf{S}\right)\mathbf{J}^\mathsf{T}, & \text{for } k \text{ even.} \end{cases} \tag{A.23}$$

Define by $\mathbf{L}$ the matrix with $\mathbf{L}_{ij} \equiv [\frac{1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T})]_{ij}$ for $i \neq j$ and $\mathbf{L}_{ii} = 0$. Then $\mathbf{K}_I$ can be written as

$$\begin{aligned} \mathbf{K}_I &= k(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ \mathbf{L} + \boldsymbol{\Phi}, \\ \boldsymbol{\Phi}_{ij} &\equiv \frac{k}{p}(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}\mathbf{z}_i^\mathsf{T}\left(\frac{1}{2}(\mathbf{E}_a+\mathbf{E}_b) - \frac{1}{8}(\mathbf{E}_a-\mathbf{E}_b)^2\right)\mathbf{z}_j \\ &\quad + \frac{k(k-1)}{8p}(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-2}\frac{1}{\sqrt{p}}(\mathbf{z}_i^\mathsf{T}(\mathbf{E}_a+\mathbf{E}_b)\mathbf{z}_j)^2 \end{aligned}$$

for $i \neq j$ and $\boldsymbol{\Phi}_{ii} = 0$. With this expression, the proof of Proposition 3.2 can be divided into three steps.

I. Concentration of $\boldsymbol{\Phi}$. We first show that $\|\boldsymbol{\Phi} - \mathbb{E}[\boldsymbol{\Phi}]\| \to 0$ almost surely, as $n, p \to \infty$. This follows from the observation that $\boldsymbol{\Phi}$ is a $p^{-1/4}$ rescaling (since $\|\mathbf{E}_a\| = O(p^{-1/4})$) of the null model $\mathbf{K}_N$, which concentrates around its expectation in the sense that $\|\mathbf{K}_N - \mathbb{E}[\mathbf{K}_N]\| = O(1)$ for $\mathbb{E}[\mathbf{K}_N] = O(\sqrt{p})$ if $a_0 \neq 0$ (see Remark 3.6). Indeed, it is shown in [FM19, Theorem 1.7] that, the leading eigenvalue of order $O(\sqrt{p})$ discarded (arising from $\mathbb{E}[\mathbf{K}_N]$), $\mathbf{K}_N$ is of bounded operator norm for all large n, p with probability one; this, together with the fact that $\|\mathbb{E}[\boldsymbol{\Phi}]\| = O(1)$ that will be shown subsequently (and independently), allows us to conclude that $\|\boldsymbol{\Phi} - \mathbb{E}[\boldsymbol{\Phi}]\| \to 0$ as $n, p \to \infty$.

II.I Computation of $\mathbb{E}[\boldsymbol{\Phi}]$: Gaussian case. Recall that the entries of $\boldsymbol{\Phi}$ are the sum of random variables of the type

$$\phi = \frac{C}{\sqrt{p}}(\mathbf{x}^\mathsf{T}\mathbf{y}/\sqrt{p})^\alpha(\mathbf{x}^\mathsf{T}\mathbf{F}\mathbf{y})^\beta$$

for independent random vectors $\mathbf{x}, \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ in the Gaussian case. As such, we resort to computing, as in [Wil97, LLC18] (or Section A.1.6), the integral

$$\begin{aligned} &\mathbb{E}_\mathbf{z}[(\mathbf{z}^\mathsf{T}\mathbf{a})^{k_1}(\mathbf{z}^\mathsf{T}\mathbf{b})^{k_2}] = (2\pi)^{-p/2}\int_{\mathbb{R}^p}(\mathbf{z}^\mathsf{T}\mathbf{a})^{k_1}(\mathbf{z}^\mathsf{T}\mathbf{b})^{k_2}e^{-\|\mathbf{z}\|^2/2}d\mathbf{z} \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^2}(\tilde{z}_1\tilde{a}_1)^{k_1}(\tilde{z}_1\tilde{b}_1 + \tilde{z}_2\tilde{b}_2)^{k_2}e^{-(\tilde{z}_1^2+\tilde{z}_2^2)/2}d\tilde{z}_1 d\tilde{z}_2 = \frac{1}{2\pi}\int_{\mathbb{R}^2}(\tilde{\mathbf{z}}^\mathsf{T}\tilde{\mathbf{a}})^{k_1}(\tilde{\mathbf{z}}^\mathsf{T}\tilde{\mathbf{b}})^{k_2}e^{-\|\tilde{\mathbf{z}}\|^2/2}d\tilde{\mathbf{z}} \end{aligned}$$

where we apply the Gram-Schmidt procedure to project $\mathbf{z}$ onto the two-dimensional space[2] spanned by $\mathbf{a}, \mathbf{b}$ with $\tilde{a}_1 = \|\mathbf{a}\|$, $\tilde{b}_1 = \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|}$, $\tilde{b}_2 = \sqrt{\|\mathbf{b}\|^2 - \frac{(\mathbf{a}^\mathsf{T}\mathbf{b})^2}{\|\mathbf{a}\|^2}}$ and denote

[2] By assuming first that $\mathbf{a}, \mathbf{b}$ are linearly independent before extending by continuity to $\mathbf{a}, \mathbf{b}$ proportional.

$\tilde{\mathbf{z}} = [\tilde{z}_1; \tilde{z}_2]$, $\tilde{\mathbf{a}} = [\tilde{a}_1; 0]$ and $\tilde{\mathbf{b}} = [\tilde{b}_1; \tilde{b}_2]$. As a consequence, we obtain, for k even,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k\right] = \mathbb{E}[\xi^k] = (k-1)!!;\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k(\mathbf{z}_i^\mathsf{T}\mathbf{b})\right] = \mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}(\mathbf{z}_i^\mathsf{T}\mathbf{b})^2\right] = 0;\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^{k-1}(\mathbf{z}_i^\mathsf{T}\mathbf{b})\right] = (k-1)!!(\|\mathbf{z}_j\|/\sqrt{p})^{k-2}(\mathbf{z}_j^\mathsf{T}\mathbf{b})/\sqrt{p};\\
&\mathbb{E}_{\mathbf{z}_i}\left[(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})^k(\mathbf{z}_i^\mathsf{T}\mathbf{b})^2\right] = (k-1)!!\left(k(\|\mathbf{z}_j\|/\sqrt{p})^{k-2}(\mathbf{z}_j^\mathsf{T}\mathbf{b}/\sqrt{p})^2 + (\|\mathbf{z}_j\|/\sqrt{p})^k\|\mathbf{b}\|^2\right);
\end{aligned}$$

where $k!!$ is the double factorial of an integral k defined by $k!! = k(k-2)(k-4)\cdots$. This further leads, in the Gaussian case, to the expression of $\tilde{\mathbf{K}}_I$ in Proposition 3.2.

II.II Computation of $\mathbb{E}[\mathbf{\Phi}]$: beyond the Gaussian case. We then show that, for random vectors $\mathbf{z}$ with zero mean, unit variance and bounded moments entries, the expression of $\mathbb{E}[\mathbf{\Phi}]$ coincides with the Gaussian case. To this end, recall that the entries of $\mathbf{\Phi}$ are the sum of random variables of the type

$$\phi = \frac{C}{\sqrt{p}}(\mathbf{x}^\mathsf{T}\mathbf{y}/\sqrt{p})^\alpha(\mathbf{x}^\mathsf{T}\mathbf{F}\mathbf{y})^\beta$$

for independent random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ with i.i.d. zero mean, unit variance and finite moments (uniformly on p) entries, deterministic $\mathbf{F} \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}$, $\alpha \in \mathbb{N}$ and $\beta \in \{1, 2\}$. Let us start with the case $\beta = 1$ and expand ϕ as

$$\phi = \frac{C}{\sqrt{p}}\left(\frac{1}{\sqrt{p}}\sum_{i_1=1}^{p} x_{i_1}y_{i_1}\right)\cdots\left(\frac{1}{\sqrt{p}}\sum_{i_\alpha=1}^{p} x_{i_\alpha}y_{i_\alpha}\right)\left(\sum_{j_1,j_2=1}^{p} F_{j_1,j_2}x_{j_1}y_{j_2}\right) \tag{A.24}$$

with x_i and y_i the i-th entry of $\mathbf{x}$ and $\mathbf{y}$, respectively, so that i) x_i is independent of y_j for all i, j and ii) x_i is independent of x_j for $i \neq j$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i^2] = 1$ and $\mathbb{E}[|x_i|^k] \leq C_k$ for some C_k independent of p (and similarly for $\mathbf{y}$).

At this point, note that to ensure $\mathbb{E}[\mathbf{K}_I]$ has non-vanishing operator norm as $n, p \to \infty$, we need $\mathbb{E}[\phi] \geq O(p^{-1})$ since $\|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$. Also, note that (as $\beta = 1$), all terms in the sum $\sum_{j_1,j_2=1}^{p} F_{j_1,j_2}x_{j_1}y_{j_2}$ with $j_1 \neq j_2$ must be zero since in other terms x_i always appears together with y_i, so that all terms with $j_1 \neq j_2$ give rise to zero in expectation. Hence, the p^2 terms of the sum only contain p nonzero terms in expectation (those with $j_1 = j_2$). The arbitrary (absolute) moments of x and y being finite, the first αp terms must be divided into $\lceil\alpha\rceil/2$ groups of size two (containing $O(p)$ terms) so that, with the normalization by p^{-1} for each group of size two, the associated expectation does not vanish. We thus discuss the following two cases:

1. α even: the α terms in the sum form $\alpha/2$ groups with different indices each and also different from $j_1 = j_2$. Therefore we have $\mathbb{E}_{x_j}[\phi] = 0$.

2. α odd: the α terms in the sum form $(\alpha-1)/2$ groups with indices different from each other and the remaining one goes with the last term containing $\mathbf{F}$ and one has $\mathbb{E}[\phi] = \frac{C\alpha!!}{p}\operatorname{tr}(\mathbf{F})$ by a combinatorial argument.

The case $\beta = 2$ follows exactly the same line of arguments except that j_1 may not equal j_2 to give rise to non-vanishing terms.

III. Concentration of Hadamard product. It now remains to treat $k(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\circ(k-1)} \circ \mathbf{L}$ and show it also has an asymptotically deterministic behavior (as $\mathbf{\Phi}$). It can be shown that

$$\|\mathbf{N} \circ \mathbf{L}\| \to 0, \quad n, p \to \infty$$

with $\mathbf{N} = (\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\odot(k-1)} - (k-2)!!\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ for k odd and $\mathbf{N} = (\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})^{\odot(k-1)}$ for k even.

To prove this, note that, depending on the parameter $a_0 = \mathbb{E}[f(\xi)]$, the operator norm of $f(\mathbf{Z}^\mathsf{T}\mathbf{Z}/\sqrt{p})/\sqrt{p}$ is either of order $O(\sqrt{p})$ for $a_0 \neq 0$ or $O(1)$ for $a_0 = 0$. In particular, for monomials $f(x) = x^k$ under study here, we have $a_0 = \mathbb{E}[\xi^k] = 0$ for k odd and $a_0 = \mathbb{E}[\xi^k] = (k-1)!! \neq 0$ for k even, $\xi \sim \mathcal{N}(0,1)$. To control the operator norm of the Hadamard product between matrices, we introduce the following lemma.

Lemma A.1. *For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$, we have $\|\mathbf{A} \circ \mathbf{B}\| \leq \sqrt{p}\|\mathbf{A}\|_\infty\|\mathbf{B}\|$.*

Proof of Lemma A.1. Let $\mathbf{e}_1, \ldots, \mathbf{e}_p$ be the canonical basis vectors of $\mathbb{R}^p$, then for all $1 \leq i \leq p$,

$$\|(\mathbf{A} \circ \mathbf{B})\mathbf{e}_i\| \leq \max_{i,j} |\mathbf{A}_{ij}| \|\mathbf{B}\mathbf{e}_i\| = \|\mathbf{A}\|_\infty\|\mathbf{B}\mathbf{e}_i\| \leq \|\mathbf{A}\|_\infty\|\mathbf{B}\|.$$

As a consequence, for any $\mathbf{v} = \sum_{i=1}^p v_i\mathbf{e}_i$, we obtain

$$\|(\mathbf{A} \circ \mathbf{B})\mathbf{v}\| \leq \sum_{i=1}^p |v_i| \|(\mathbf{A} \circ \mathbf{B})\mathbf{e}_i\| \leq \sum_{i=1}^p |v_i| \|\mathbf{A}\|_\infty\|\mathbf{B}\|$$

which, by Cauchy-Schwarz inequality further yields $\sum_{i=1}^p |v_i| \leq \sqrt{p}\|\mathbf{v}\|$. This concludes the proof of Lemma A.1. □

Lemma A.1 tells us that the Hadamard product between a matrix with $o(p^{-1/2})$ entry and a matrix with bounded operator norm is of vanishing operator norm, as $p \to \infty$. As such, since $\|\mathbf{N}\| = O(1)$ and $\mathbf{L}$ has $O(p^{-1})$ entries, we have $\|\mathbf{N} \circ \mathbf{L}\| \to 0$. This concludes the proof of Proposition 3.2.

A.1.5 Proof of Theorem 3.5

We aim to prove Theorem 3.5, which states that the resolvent $\mathbf{Q}(z) \equiv \left(\frac{1}{n}\sigma(\mathbf{W}\mathbf{X})^\mathsf{T}\sigma(\mathbf{W}\mathbf{X}) - z\mathbf{I}_n\right)^{-1}$ admits the following deterministic equivalent (see Definition 7)

$$\bar{\mathbf{Q}}(z) \equiv (\check{\mathbf{K}} - z\mathbf{I}_n)^{-1}, \quad \check{\mathbf{K}} \equiv \frac{N}{n}\frac{\mathbf{K}}{1+\delta}, \quad \mathbf{K} \equiv \mathbb{E}_\mathbf{w}[\sigma(\mathbf{X}^\mathsf{T}\mathbf{w})\sigma(\mathbf{w}^\mathsf{T}\mathbf{X})]$$

with δ the unique solution of $\delta = \frac{1}{n}\operatorname{tr}(\mathbf{K}\bar{\mathbf{Q}})$.

The proof of Theorem 3.5 generally follows that of the Marčenko-Pastur law in Section 2.2.2. We provide only the main steps as follows.

Denote the random (column) vector $\boldsymbol{\sigma}_i \equiv \sigma(\mathbf{X}^\mathsf{T}\mathbf{w}_i) \in \mathbb{R}^n$ for $\|\mathbf{X}\| = O(1)$ and $\mathbf{w}$ the i-th row of $\mathbf{W}$ such that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. By rewriting $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}$ as

$$\mathbf{\Sigma}^\mathsf{T} = \left[\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N\right], \quad \frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^N \boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}$$

we obtain, from the resolvent identity (Lemma 2.1) that

$$\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}} = \mathbb{E}\left[\mathbf{Q}\left(\frac{N}{n}\frac{\mathbf{K}}{1+\delta} - \frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}\right)\right]\bar{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}]\frac{N}{n}\frac{\mathbf{K}}{1+\delta}\bar{\mathbf{Q}} - \mathbb{E}\left[\mathbf{Q}\frac{1}{n}\sum_{i=1}^N \boldsymbol{\sigma}_i\boldsymbol{\sigma}_i^\mathsf{T}\right]\bar{\mathbf{Q}}.$$

Working on the expectation inside the second term, we further get, with Sherman-Morrison identity (Lemma 2.8), that

$$\sum_{i=1}^N \mathbb{E}\left[\mathbf{Q}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}\right] = \sum_{i=1}^N \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}}{1+\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i}\right]$$
$$= \sum_{i=1}^N \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}}{1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})}\right] + \sum_{i=1}^N \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}\left(\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})-\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i\right)}{\left(1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})\right)\left(1+\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i\right)}\right]$$

where we denote $\mathbf{Q}_{-i} \equiv \left(\frac{1}{n}\sum_{j\neq i}\sigma_j\sigma_j^\mathsf{T} + \lambda\mathbf{I}_n\right)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - \frac{1}{n}\sigma_i\sigma_i^\mathsf{T}$. Intuitively, the second expectation above should asymptotically vanish, as a result of Lemma 3.3. To show this, let us rewrite again with Sherman-Morrison identity (Lemma 2.8)

$$\sum_{i=1}^N \frac{\mathbf{Q}_{-i}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}\left(\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})-\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i\right)}{\left(1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})\right)\left(1+\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i\right)} = \sum_{i=1}^N \frac{\mathbf{Q}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}\left(\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})-\frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i\right)}{1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})}$$
$$= \sum_{i=1}^N \frac{\mathbf{Q}\frac{1}{n}\mathbf{\Sigma}^\mathsf{T}\mathbf{D}\mathbf{\Sigma}}{1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})}$$

where we denote $\mathbf{D}$ the diagonal matrix with its i-th entry equal to $\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i}) - \frac{1}{n}\sigma_i^\mathsf{T}\mathbf{Q}_{-i}\sigma_i$. Then with the bounds $\|\mathbf{Q}\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma}/n\| = O(1)$, $\|\mathbf{Q}_{-i}\| = O(1)$ and $\|\mathbf{D}\| = O(n^{-1/2})$ from Lemma 3.3, imply that the matrix has asymptotically vanishing operator norm. This gives

$$\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}]\frac{N}{n}\frac{\mathbf{K}}{1+\delta}\bar{\mathbf{Q}} - \sum_{i=1}^N \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\sigma_i\sigma_i^\mathsf{T}}{1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})}\right]\bar{\mathbf{Q}} + o(1)$$
$$= \mathbb{E}[\mathbf{Q}]\frac{N}{n}\frac{\mathbf{K}}{1+\delta}\bar{\mathbf{Q}} - \frac{1}{n}\sum_{i=1}^N \mathbb{E}\left[\frac{\mathbf{Q}_{-i}}{1+\frac{1}{n}\operatorname{tr}(\mathbf{K}\mathbf{Q}_{-i})}\right]\mathbf{K}\bar{\mathbf{Q}} + o(1).$$

Following similar arguments as above, we have, from the rank one update $\mathbf{Q} - \mathbf{Q}_{-i} = -\frac{1}{n}\mathbf{Q}\sigma_i\sigma_i^\mathsf{T}\mathbf{Q}_{-i}$ and $\|\mathbf{K}\bar{\mathbf{Q}}\| = O(1)$ that

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0$$

as $n, p, N \to \infty$.

A.1.6 Computation details for Table 3.3

Here we only provide the case of the popular ReLU nonlinearity $\sigma(t) = \mathrm{ReLU}(t) \equiv \max(t, 0)$. Other functions $\sigma(\cdot)$ in Table 3.3 can be treated similarly.

We first write the expectation as an integral on $\mathbb{R}^p$, which is then reduced to an integral on $\mathbb{R}^2$ with the classical Gram-Schmidt process:

$$\mathbf{K}(\mathbf{a},\mathbf{b}) = \mathbb{E}_\mathbf{w}[\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})] = (2\pi)^{-\frac{p}{2}}\int_{\mathbb{R}^p}\sigma(\mathbf{w}^\mathsf{T}\mathbf{a})\sigma(\mathbf{w}^\mathsf{T}\mathbf{b})e^{-\frac{1}{2}\|\mathbf{w}\|^2}d\,\mathbf{w}$$
$$= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}}\sigma(\tilde{w}_1\tilde{a}_1)\sigma(\tilde{w}_1\tilde{b}_1+\tilde{w}_2\tilde{b}_2)e^{-\frac{1}{2}(\tilde{w}_1^2+\tilde{w}_2^2)}d\,\tilde{w}_1 d\,\tilde{w}_2$$
$$= \frac{1}{2\pi}\int_{\mathbb{R}^2}\sigma(\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{a}})\sigma(\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{b}})e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2}d\,\tilde{\mathbf{w}}$$
$$= \frac{1}{2\pi}\int_{\min(\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{a}},\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{b}})\geq 0}\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{a}}\cdot\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{b}}\cdot e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2}d\,\tilde{\mathbf{w}}$$

where $\tilde{a}_1 = \|\mathbf{a}\|$, $\tilde{b}_1 = \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|}$, $\tilde{b}_2 = \|\mathbf{b}\|\sqrt{1-\frac{(\mathbf{a}^\mathsf{T}\mathbf{b})^2}{\|\mathbf{a}\|^2\|\mathbf{b}\|^2}}$ and we denote $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2]^\mathsf{T}$, $\tilde{\mathbf{a}} = [\tilde{a}_1, 0]^\mathsf{T}$ and $\tilde{\mathbf{b}} = [\tilde{b}_1, \tilde{b}_2]^\mathsf{T}$.

With a simple geometric representation we observe

$$\{\tilde{\mathbf{w}} \mid \min(\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{a}}, \tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{b}}) \geq 0\} = \left\{ r\cos(\theta) + r\sin(\theta) \mid r \geq 0,\ \theta \in \left[\theta_0 - \frac{\pi}{2}, \frac{\pi}{2}\right] \right\}$$

with $\theta_0 \equiv \arccos\left(\frac{\tilde{b}_1}{\|\mathbf{b}\|}\right) = \frac{\pi}{2} - \arcsin\left(\frac{\tilde{b}_1}{\|\mathbf{b}\|}\right)$. Therefore with a polar coordinate change of variable we deduce, for $\sigma(t) = \mathrm{ReLU}(t)$, that

$$\begin{aligned}\mathbf{K}(\mathbf{a},\mathbf{b}) &= \tilde{a}_1 \frac{1}{2\pi} \int_{\theta_0-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\theta)\left(\tilde{b}_1\cos(\theta) + \tilde{b}_2\sin(\theta)\right) d\,\theta \int_{\mathbb{R}^+} r^3 e^{-\frac{1}{2}r^2} d\,r \\ &= \frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\| \left(\sqrt{1-\angle(\mathbf{a},\mathbf{b})^2} + \angle(\mathbf{a},\mathbf{b}) \arccos\left(-\angle(\mathbf{a},\mathbf{b})\right) \right)\end{aligned}$$

with $\angle(\mathbf{a},\mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ as Table 3.3.

A.2 Proofs in Chapter 4

A.2.1 Proofs of Theorem 4.1 and 4.2

We start with the proof of Theorem 4.1 which characterizes the (asymptotic) test performance of the classifier $\mathbf{w}(t)$ given by

$$\mathbf{w}(t) = e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{w}_0 + \left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)\mathbf{w}_{LS}$$

on an unseen test datum $\hat{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$. To this end, we shall evaluate subsequently the random variable $\mathbf{w}(t)^\mathsf{T}\boldsymbol{\mu}$ and $\|\mathbf{w}(t)\|^2$ as below.

Since

$$\begin{aligned}\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t) &= \boldsymbol{\mu}^\mathsf{T} e^{-\frac{\alpha t}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}}\mathbf{w}_0 + \boldsymbol{\mu}^\mathsf{T}\left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}}\right)\mathbf{w}_{LS} \\ &= -\frac{1}{2\pi i}\oint_\gamma f_t(z)\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\, dz - \frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n\, dz\end{aligned}$$

with $\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} + \begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}$, we have

$$\begin{aligned}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} &= \mathbf{Q}(z) \\ &\quad - \mathbf{Q}(z)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\boldsymbol{\mu} & 1+\frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{1}_n\\ * & -1+\frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Q}(z)\frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z).\end{aligned}$$

We thus resort to the computation of the bilinear form $\mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{b}$. By plugging in the deterministic equivalent of $\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p$ we obtain the following estimations

$$\begin{aligned}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\boldsymbol{\mu} &\simeq \|\boldsymbol{\mu}\|^2 m(z) \\ \frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{1}_n &\simeq 0 \\ \frac{1}{n^2}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{1}_n &\simeq \frac{1}{n^2}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{Q}}(z)\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{Q}}(z)\left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n + z\mathbf{I}_n\right)\mathbf{1}_n \\ &= 1 + z\frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{Q}}(z)\mathbf{1}_n = 1 + z\frac{1}{n}\operatorname{tr}\tilde{\mathbf{Q}}(z) \simeq 1 + z\tilde{m}(z)\end{aligned}$$

with the *co-resolvent* $\tilde{\mathbf{Q}}(z) = \left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n\right)^{-1}$ such that $\tilde{\mathbf{Q}}(z) \leftrightarrow \tilde{m}(z)\mathbf{I}_n$ with $\tilde{m}(z)$ given by

$$cm(z) = \tilde{m}(z) + \frac{1}{z}(1-c)$$

for $m(z)$ the *unique* solution of the Marčenko–Pastur equation (4.1). The above relation is a direct consequence of the fact that $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and $\mathbf{Z}\mathbf{Z}^\mathsf{T}$ have the same eigenvalues, except for the additional zeros eigenvalues for the larger matrix.

We thus get

$$\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} \simeq \mathbf{Q}(z) - \mathbf{Q}(z)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z) & 1 \\ 1 & z\tilde{m}(z)\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z)$$
$$\simeq \mathbf{Q}(z) - \frac{\mathbf{Q}(z)}{z\|\boldsymbol{\mu}\|^2 m(z)\tilde{m}(z) - 1}\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}z\tilde{m}(z) & -1 \\ -1 & \|\boldsymbol{\mu}\|^2 m(z)\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z)$$

and the term $\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n$ is consequently given by

$$\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n \simeq \|\boldsymbol{\mu}\|^2 m(z) - \frac{\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z) & 0\end{bmatrix}}{z\|\boldsymbol{\mu}\|^2 m(z)\tilde{m}(z) - 1}\begin{bmatrix}z\tilde{m}(z) & -1 \\ -1 & \|\boldsymbol{\mu}\|^2 m(z)\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z) \\ 1 + z\tilde{m}(z)\end{bmatrix}$$
$$\simeq \frac{\|\boldsymbol{\mu}\|^2 m(z) z\tilde{m}(z)}{\|\boldsymbol{\mu}\|^2 m(z) z\tilde{m}(z) - 1} \simeq \frac{\|\boldsymbol{\mu}\|^2(zm(z)+1)}{1 + \|\boldsymbol{\mu}\|^2(zm(z)+1)} \simeq \frac{\|\boldsymbol{\mu}\|^2 m(z)}{(\|\boldsymbol{\mu}\|^2 + c)m(z) + 1}$$

where we use the relations $\tilde{m}(z) = cm(z) - \frac{1}{z}(1-c)$ and $(zm(z)+1)(cm(z)+1) = m(z)$ from (4.1), while the term $\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0 = O(n^{-1/2})$ due to the independence of $\mathbf{w}_0$ with respect to $\mathbf{Z}$.

Following the same arguments, we have

$$\begin{aligned}\|\mathbf{w}(t)\|^2 = &-\frac{1}{2\pi i}\oint_\gamma f_t^2(z)\mathbf{w}_0\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\, dz \\ &- \frac{1}{\pi i}\oint_\gamma \frac{f_t(z)(1-f_t(z))}{z}\mathbf{w}_0\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n\, dz \\ &- \frac{1}{2\pi i}\oint_\gamma \frac{(1-f_t(z))^2}{z^2}\frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{X}}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n\, dz\end{aligned}$$

together with

$$\begin{aligned}\mathbf{w}_0\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0 &\simeq \sigma^2 m(z) \\ \mathbf{w}_0\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{y}^\mathsf{T} &\simeq 0 \\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\tilde{\mathbf{X}}^\mathsf{T}\left(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\tilde{\mathbf{X}}\mathbf{1}_n &\simeq 1 - \frac{1}{(\|\boldsymbol{\mu}\|^2 + c)m(z) + 1}.\end{aligned}$$

We now wish to replace the terms in $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$ and $\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t)$ by their asymptotic approximations to reach an almost sure convergence of the right-hand side contour integration. Note that, both $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$ and $\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t)$ are functionals of the resolvent $\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}$ that is only well defined for z not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. More generally, the aforementioned approximations can be summarized by the fact that, for a generic analytic $h(z)$, we have, as $n \to \infty$,

$$h(z) - \bar{h}(z) \to 0$$

almost surely for all z not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. To reach a convergence result for $\oint_\gamma h(z)dz - \oint_\gamma \bar{h}(z)dz \to 0$, we first show that there exists a probability one set Ω_z on which $h(z)$ is uniformly bounded for all large n, with a bound independent of z. Then by the "no eigenvalues outside the support" theorem (see for example [PS09]) we know that, with probability one, for all n, p large, no eigenvalue of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ appears outside the interval $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$. As such, the intersection set $\Omega = \cap_{z_i}\Omega_{z_i}$ for a finitely many z_i, is still a probability one set. Finally by Lebesgue's dominant convergence theorem [Bil12], together with the analyticity of the function under consideration, we conclude the proof of Theorem 4.1. The proof of Theorem 4.2 follows exactly the same line of arguments and is thus omitted here.

A.2.2 Detailed Derivation of (4.3)-(4.6)

Our objective here is to further simplify the contour integrations obtained above, by carefully choosing the path γ. First note that, $\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T}$, as a low rank perturbation of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ such that

$$\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\mathsf{T} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} + \begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}$$

may have (asymptotically) one isolated eigenvalue that jumps out the support of the (limiting) Marčenko–Pastur distribution.

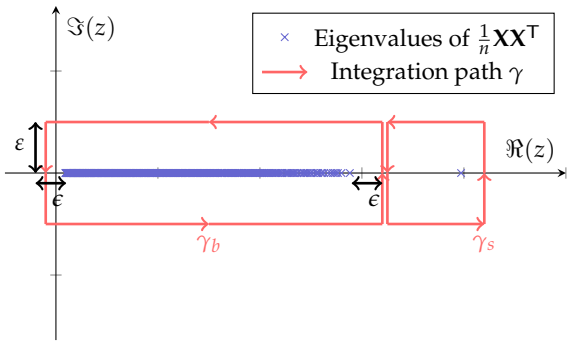


Figure A.1: Eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\boldsymbol{\mu} = [1.5; \mathbf{0}_{p-1}]$, $p = 512$, $n = 1\,024$ and $c_1 = c_2 = 1/2$.

We first determine the location of the isolated eigenvalue λ. More concretely, we wish to find λ an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ that lies outside the support of the Marčenko–Pastur distribution (more precisely, not an eigenvalue of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$). Solving the following equation

for $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \lambda \mathbf{I}_p\right) = 0 \\
&\Leftrightarrow \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda \mathbf{I}_p + \begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda \mathbf{I}_p\right)\det\left(\mathbf{I}_p + \mathbf{Q}(\lambda)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\left(\mathbf{I}_2 + \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{1}_n^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(\lambda)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{1}_n\end{bmatrix}\right) \simeq 0 \\
&\Leftrightarrow \det\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(\lambda) + 1 & 1 + z\tilde{m}(\lambda)\\ \|\boldsymbol{\mu}\|^2 m(\lambda) & 1\end{bmatrix} \simeq 0 \\
&\Leftrightarrow 1 + (\|\boldsymbol{\mu}\|^2 + c)m(\lambda) \simeq 0
\end{aligned}$$

where we recall that $\mathbf{Q}(\lambda) \equiv \left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda \mathbf{I}_p\right)^{-1}$ and we used the fact that $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ as well as the Sylvester's determinant identity $\det(\mathbf{I}_p + \mathbf{AB}) = \det(\mathbf{I}_n + \mathbf{BA})$ for $\mathbf{A}, \mathbf{B}$ of appropriate dimensions (Lemma 2.3). Together with (4.1) we deduce the (empirical) isolated eigenvalue $\lambda = \lambda_s + o(1)$ with

$$\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2}$$

which gives the asymptotic location of the isolated eigenvalue as $n \to \infty$. In the following, we may thus use λ_s instead of λ throughout the computation. By splitting the path γ into $\gamma_b + \gamma_s$ that circles respectively around the main bulk between $[\lambda_- \equiv (1-\sqrt{c})^2, \lambda_+ \equiv (1+\sqrt{c})^2]$ and the isolated eigenvalue λ_s, we deduce, with the residual theorem that $E = E_{\gamma_b} + E_{\gamma_s}$ where

$$\begin{aligned}
E_{\gamma_s} &= -\frac{1}{2\pi i}\oint_{\gamma_s} \frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1+(\|\boldsymbol{\mu}\|^2+c)m(z)}\,dz = -\mathrm{Res}\frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1+(\|\boldsymbol{\mu}\|^2+c)m(z)} \\
&= -\lim_{z\to\lambda_s}(z-\lambda_s)\frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1+(\|\boldsymbol{\mu}\|^2+c)m(z)} = -\frac{1-f_t(\lambda_s)}{\lambda_s}\frac{\|\boldsymbol{\mu}\|^2 m(\lambda_s)}{(\|\boldsymbol{\mu}\|^2+c)m'(\lambda_s)} \\
&= -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+c}\frac{1-f_t(\lambda_s)}{\lambda_s}\frac{1-c-\lambda_s-2c\lambda_s m(\lambda_s)}{cm(\lambda_s)+1} = \left(\|\boldsymbol{\mu}\|^2 - \frac{c}{\|\boldsymbol{\mu}\|^2}\right)\frac{1-f_t(\lambda_s)}{\lambda_s}
\end{aligned} \tag{A.25}$$

with $m'(z)$ the derivative of $m(z)$ with respect to z and is obtained by taking the derivative of (4.1).

We now move on to the contour integration γ_b in the computation of E_{γ_b}. We follow the idea in [BS08] and choose γ_b to be a rectangle with sides parallel to the axes, intersecting the real axis at 0 and λ_+ (in fact at $-\epsilon$ and $\lambda_+ + \epsilon$ so that the functions under consideration remain analytic) and the horizontal sides being a distance $\varepsilon \to 0$ away from the real axis. Since for nonzero $x \in \mathbb{R}$, the limit $\lim_{z\in\mathbb{Z}\to x} m(z) \equiv \check{m}(x)$ exists [SC95] and is given by

$$\check{m}(x) = \frac{1-c-x}{2cx} \pm \frac{i}{2cx}\sqrt{4cx-(1-c-x)^2} = \frac{1-c-x}{2cx} \pm \frac{i}{2cx}\sqrt{(x-\lambda_-)(\lambda_+-x)}$$

with the branch of $\pm$ is determined by the imaginary part of z such that $\Im(z)\cdot\Im m(z) > 0$. For simplicity we denote

$$\Re\check{m} = \frac{1-c-x}{2cx},\ \Im\check{m} = \frac{1}{2cx}\sqrt{(x-\lambda_-)(\lambda_+-x)}$$

and therefore

$$\begin{aligned} E_{\gamma_b} &= -\frac{1}{2\pi i}\oint_{\gamma_b}\frac{1-f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1+(\|\boldsymbol{\mu}\|^2+c)m(z)}\,dz \\ &= -\frac{\|\boldsymbol{\mu}\|^2}{\pi i}\int_{\lambda_-}^{\lambda_+}\frac{1-f_t(x)}{x}\Im\left[\frac{\Re\check{m}-i\Im\check{m}}{1+(\|\boldsymbol{\mu}\|^2+c)(\Re\check{m}-i\Im\check{m})}\right]dx \\ &= -\frac{\|\boldsymbol{\mu}\|^2}{\pi i}\int_{\lambda_-}^{\lambda_+}\frac{1-f_t(x)}{x}\Im\left[\frac{\Re\check{m}+\frac{\|\boldsymbol{\mu}\|^2+c}{cx}-i\Im\check{m}}{1+2(\|\boldsymbol{\mu}\|^2+c)\Re\check{m}+\frac{(\|\boldsymbol{\mu}\|^2+c)^2}{cx}}\right]dx \end{aligned}$$

with $z = x \pm i\varepsilon$ and in the limit $\varepsilon \to 0$ (on different sides of the real axis) and with $(\Re\check{m})^2 + (\Im\check{m})^2 = \frac{1}{cx}$. We then take the imaginary part in the right-hand side, which results in

$$\begin{aligned} E_{\gamma_b} &= \frac{\|\boldsymbol{\mu}\|^2}{\pi}\int_{\lambda_-}^{\lambda_+}\frac{1-f_t(x)}{x}\frac{\Im\check{m}}{1+2(\|\boldsymbol{\mu}\|^2+c)\Re\check{m}+\frac{(\|\boldsymbol{\mu}\|^2+c)^2}{cx}}\,dx \\ &= \frac{1}{2\pi}\int_{\lambda_-}^{\lambda_+}\frac{1-f_t(x)}{x}\frac{\sqrt{4cx-(1-c-x)^2}}{\lambda_s-x}\,dx \end{aligned} \tag{A.26}$$

where we recall the definition $\lambda_s \equiv c+1+\|\boldsymbol{\mu}\|^2+\frac{c}{\|\boldsymbol{\mu}\|^2}$. Ultimately we assemble (A.25) and (A.26) to get the expression in (4.3). The derivations of (4.4)-(4.6) follow the same arguments and are thus omitted.

Bibliography

[A+11] Radoslaw Adamczak et al. On the marchenko-pastur and circular laws for some classes of random matrices with dependent entries. *Electronic Journal of Probability*, 16:1065–1095, 2011.

[AAR00] George E Andrews, Richard Askey, and Ranjan Roy. *Special functions*, volume 71. Cambridge university press, 2000.

[AG13] Naum Ilich Akhiezer and Izrail Markovich Glazman. *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

[AGD94] Ludwig Arnold, Volker Matthias Gundlach, and Lloyd Demetrius. Evolutionary formalism for products of positive random matrices. *The Annals of Applied Probability*, 4(3):859–901, 1994.

[ALM+01] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[AMGS12] Konstantin Avrachenkov, Alexey Mishenin, Paulo Gonçalves, and Marina Sokol. Generalized optimization framework for graph-based semi-supervised learning. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 966–974. SIAM, 2012.

[AS65] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

[AS17] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

[ASD96] DJ Albers, JC Sprott, and WD Dechert. Dynamical behavior of artificial neural networks with random weights. *Intelligent Engineering Systems Through Artificial Neural Networks*, 6:17–22, 1996.

[AZLS18] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[BA00] Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BC11] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[BDEL03] Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[BGC16] Florent Benaych-Georges and Romain Couillet. Spectral analysis of the gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237, 2016.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[BGN12] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[BH89] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[Bil12] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, third edition, 2012.

[Bis07] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[BP19] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

[BR89] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[BS06] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[BS08] Zhidong D Bai and Jack W Silverstein. Clt for linear spectral statistics of large-dimensional sample covariance matrices. In *Advances In Statistics*, pages 281–333. World Scientific, 2008.

[BS10] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CBG16] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

[CC15] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[CDV07] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[Cha03] Chein-I Chang. *Hyperspectral imaging: techniques for spectral detection and classification*, volume 1. Springer Science & Business Media, 2003.

[CK16] Romain Couillet and Abla Kammoun. Random matrix improved subspace clustering. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 90–94. IEEE, 2016.

[Cla90] F. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

[CLC18] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[CLM18] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. Classification asymptotics in the random matrix regime. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1875–1879. IEEE, 2018.

[CP11a] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[CP11b] David Cox and Nicolas Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *Face and Gesture 2011*, pages 8–15. IEEE, 2011.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[CS18] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

[CW08] 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*, pages 160–167. ACM, 2008.

[CWSA16] Romain Couillet, Gilles Wainrib, Harry Sevi, and Hafiz Tiomoko Ali. The asymptotic performance of linear echo state neural networks. *The Journal of Machine Learning Research*, 17(1):6171–6205, 2016.

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.

[DK05] Didier D'Acunto and Krzysztof Kurdyka. Explicit bounds for the łojasiewicz exponent in the gradient inequality for polynomials. In *Annales Polonici Mathematici*, volume 1, pages 51–61, 2005.

[DLT+18] Simon S Du, Jason D Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1338–1347, 2018.

[DM16] David Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[DV13] Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.

[DVFRCA14] Fabrizio De Vico Fallani, Jonas Richiardi, Mario Chavez, and Sophie Achard. Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130521, 2014.

[DVSH18] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1308–1317, 2018.

[EK09] Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, 2009.

[EK10a] Noureddine El Karoui. On Information Plus Noise Kernel Random Matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

[EK10b] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[EKBB+13] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[EPPP00] Theodoros Evgeniou, Massimiliano Pontil, Constantine Papageorgiou, and Tomaso Poggio. Image representations for object detection using kernel classifiers. In *Asian Conference on Computer Vision*, pages 687–692. Citeseer, 2000.

[FB16] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

[FSA99] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[Gel93] Erol Gelenbe. Learning in the recurrent random neural network. *Neural computation*, 5(1):154–164, 1993.

[GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GIP+18] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798, 2018.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org, 2017.

[GR14] Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

[GSB16] Raja Giryes, Guillermo Sapiro, and Alexander M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions Signal Processing*, 64(13):3444–3457, 2016.

[GSL+02] T Van Gestel, Johan AK Suykens, Gert Lanckriet, Annemie Lambrechts, B De Moor, and Joos Vandewalle. Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel fisher discriminant analysis. *Neural computation*, 14(5):1115–1147, 2002.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HLN07] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

[HPS06] Morris W Hirsch, Charles Chapman Pugh, and Michael Shub. *Invariant manifolds*, volume 583. Springer, 2006.

[HWH16] Kun He, Yan Wang, and John Hopcroft. A powerful generative model using random weights for the deep image representation. In *Advances in Neural Information Processing Systems*, pages 631–639, 2016.

[HZDZ12] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, 2012.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[Jae01] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

[JKL09] Kevin Jarrett, Koray Kavukcuoglu, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE, 2009.

[Kar09] Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, pages 2362–2405, 2009.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances In Neural Information Processing Systems*, pages 586–594, 2016.

[KC17] Abla Kammoun and Romain Couillet. Subspace kernel spectral clustering of large dimensional data. 2017.

[KK12] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial Intelligence and Statistics*, pages 583–591, 2012.

[KMM+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LC18a] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3063–3071. PMLR, 2018.

[LC18b] Cosme Louart and Romain Couillet. A random matrix and concentration inequalities framework for neural networks analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4214–4218. IEEE, 2018.

[LC19a] Zhenyu Liao and Romain Couillet. Inner-product kernels are asymptotically equivalent to binary discrete kernels. 2019.

[LC19b] 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.

[LC19c] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. 2019.

[LCPB00] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397, 2000.

[Led05] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[LML+14] Zhiyun Lu, Avner May, Kuan Liu, Alireza Bagheri Garakani, Dong Guo, Aurélien Bellet, Linxi Fan, Michael Collins, Brian Kingsbury, and Michael Picheny. How to scale up kernel methods to be as good as deep neural nets. *arXiv preprint arXiv:1411.4000*, 2014.

[Loj82] S Lojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1982.

[LP09] Anna Lytova and Leonid Pastur. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability*, 37(5):1778–1840, 2009.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[MDH11] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

[MK87] Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[MP67] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[MRW+99] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.*, pages 41–48. Ieee, 1999.

[MSV09] Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pages 1049–1056, 2009.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[NP06] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[O'R12] Sean O'Rourke. A note on the marchenko-pastur law for a class of random matrices with dependent entries. *Electronic Communications in Probability*, 17, 2012.

[Pas05] Leonid A Pastur. A simple approach to the global regime of gaussian ensembles of random matrices. *Ukrainian Mathematical Journal*, 57(6):936–966, 2005.

[Pau07] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[PDDC09] Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11):e1000579, 2009.

[PS09] Debashis Paul and Jack W Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

[PS11] Leonid Andreevich Pastur and Mariya Shcherbina. *Eigenvalue distribution of large random matrices*. Number 171. American Mathematical Soc., 2011.

[PSG17] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in neural information processing systems*, pages 4785–4795, 2017.

[PW17] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2634–2643, 2017.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[RR08] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[RRR67] IUrii Anatol'evich Rozanov, IUriui Anatol'evich Rozanov, and Yu A Rozanov. *Stationary random processes*. Holden-Day, 1967.

[Rud62] Walter Rudin. *Fourier analysis on groups*, volume 121967. Wiley Online Library, 1962.

[Rud64] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[RVC+04] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[SB95] Jack W Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

[SC95] Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

[SC18] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[SKC+11] Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *ICML*, volume 2, page 6, 2011.

[SKD92] Wouter F Schmidt, Martin A Kraaijveld, and Robert PW Duin. Feed forward neural networks with random weights. In *International Conference on Pattern Recognition*, pages 1–1. IEEE COMPUTER SOCIETY PRESS, 1992.

[SKZ14] Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems*, pages 406–414, 2014.

[SMG13] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[SS04] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[Ste81] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[SV99] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[SW17] Simone Scardapane and Dianhui Wang. Randomness in neural networks: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), 2017.

[UVL18] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[Vap92] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[VOPT09] Gabriel Vasile, Jean-Philippe Ovarlez, Frederic Pascal, and Céline Tison. Coherency matrix estimation of heterogeneous clutter in high-resolution polarimetric sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4):1809–1826, 2009.

[VVLGS12] Steven Van Vaerenbergh, Miguel Lázaro-Gredilla, and Ignacio Santamaría. Kernel recursive least-squares tracker for time-varying regression. *IEEE transactions on neural networks and learning systems*, 23(8):1313–1326, 2012.

[VZ12] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[Wil97] Christopher KI Williams. Computing with infinite networks. *Advances in neural information processing systems*, pages 295–301, 1997.

[WZWD13] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2013.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[Zan69] Willard I Zangwill. Convergence conditions for nonlinear programming algorithms. *Management Science*, 16(1):1–13, 1969.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[Zhu05] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Théorie des matrices aléatoires pour l'apprentissage automatique en grande dimension et les réseaux de neurones

Mots clés : Apprentissage automatique, théorie des matrices aléatoires, réseaux de neurones

Résumé : Le "Big Data" et les grands systèmes d'apprentissage sont omniprésents dans les problèmes d'apprentissage automatique aujourd'hui. Contrairement à l'apprentissage de petite dimension, les algorithmes d'apprentissage en grande dimension sont sujets à divers phénomènes contre-intuitifs et se comportent de manière très différente des intuitions de petite dimension sur lesquelles ils sont construits. Cependant, en supposant que la dimension et le nombre des données sont à la fois grands et comparables, la théorie des matrices aléatoires (RMT) fournit une approche systématique pour évaluer le comportement statistique de ces grands systèmes d'apprentissage, lorsqu'ils sont appliqués à des données de grande dimension. L'objectif principal de cette thèse est de proposer un schéma d'analyse basé sur la RMT, pour une grande famille de systèmes d'apprentissage automatique: d'évaluer leurs performances, de mieux les comprendre et finalement les améliorer, afin de mieux gérer les problèmes de grandes dimensions aujourd'hui.

Précisément, nous commençons par exploiter la connexion entre les grandes matrices à noyau, les projection aléatoires non-linéaires et les réseaux de neurones aléatoires simples. En considérant que les données sont tirées indépendamment d'un modèle de mélange gaussien, nous fournissons une caractérisation précise des performances de ces systèmes d'apprentissage en grande dimension, exprimée en fonction des statistiques de données, de la dimensionnalité et, surtout, des hyper-paramètres du problème. Lorsque des algorithmes d'apprentissage plus complexes sont considérés, ce schéma d'analyse peut être étendu pour accéder à de systèmes d'apprentissage qui sont définis (implicitement) par des problèmes d'optimisation convexes, lorsque des points optimaux sont atteints. Pour trouver ces points, des méthodes d'optimisation telles que la descente de gradient sont régulièrement utilisées. À cet égard, dans le but d'avoir une meilleur compréhension théorique des mécanismes internes de ces méthodes d'optimisation et, en particulier, leur impact sur le modèle d'apprentissage, nous évaluons aussi la dynamique de descente de gradient dans les problèmes d'optimisation convexes et non convexes.

Ces études préliminaires fournissent une première compréhension quantitative des algorithmes d'apprentissage pour le traitement de données en grandes dimensions, ce qui permet de proposer de meilleurs critères de conception pour les grands systèmes d'apprentissage et, par conséquent, d'avoir un gain de performance remarquable lorsqu'il est appliqué à des jeux de données réels. Profondément ancré dans l'idée d'exploiter des données de grandes dimensions avec des informations répétées à un niveau "global" plutôt qu'à un niveau "local", ce schéma d'analyse RMT permet une compréhension renouvelée et la possibilité de contrôler et d'améliorer une famille beaucoup plus large de méthodes d'apprentissage automatique, ouvrant ainsi la porte à un nouveau schéma d'apprentissage automatique pour l'intelligence artificielle.

Title : A random matrix framework for large dimensional machine learning and neural networks

Keywords : Machine learning, random matrix theory, neural networks

Abstract : Large dimensional data and learning systems are ubiquitous in modern machine learning. As opposed to small dimensional learning, large dimensional machine learning algorithms are prone to various counterintuitive phenomena and behave strikingly differently from the low dimensional intuitions upon which they are built. Nonetheless, by assuming the data dimension and their number to be both large and comparable, random matrix theory (RMT) provides a systematic approach to assess the (statistical) behavior of these large learning systems, when applied on large dimensional data. The major objective of this thesis is to propose a full-fledged RMT-based framework for various machine learning systems: to assess their performance, to properly understand and to carefully refine them, so as to better handle large dimensional problems that are increasingly needed in artificial intelligence applications.

Precisely, we exploit the close connection between kernel matrices, random feature maps, and single-hidden-layer random neural networks. Under a simple Gaussian mixture modeling for the input data, we provide a precise characterization of the performance of these large dimensional learning systems as a function of the data statistics, the dimensionality, and most importantly the hyperparameters (e.g., the choice of the kernel function or activation function) of the problem. Further addressing more involved learning algorithms, we extend the present RMT analysis framework to access large learning systems that are implicitly defined by convex optimization problems (e.g., logistic regression), when optimal points are assumed reachable. To find these optimal points, optimization methods such as gradient descent are regularly used. Aiming to have a better theoretical grasp of the inner mechanism of optimization methods and their impact on the resulting learning model, we further evaluate the gradient descent dynamics in training convex and non-convex objects.

These preliminary studies provide a first quantitative understanding of the aforementioned learning algorithms when large dimensional data are processed, which further helps propose better design criteria for large learning systems that result in remarkable gains in performance when applied on real-world datasets. Deeply rooted in the idea of mining large dimensional data with repeated patterns at a global rather than a local level, the proposed RMT analysis framework allows for a renewed understanding and the possibility to control and improve a much larger range of machine learning approaches, and thereby opening the door to a renewed machine learning framework for artificial intelligence.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France