

Random matrix methods for high-dimensional machine learning models

Présentée le 26 janvier 2024

Faculté informatique et communications
Laboratoire de théorie des communications
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Antoine Philippe Michel BODIN

Acceptée sur proposition du jury

Dr O. Lévêque, président du jury
Dr N. Macris, directeur de thèse
Prof. Y. Lu, rapporteur
Prof. J. Barbier, rapporteur
Prof. L. Chizat, rapporteur

Acknowledgements

I am deeply grateful to my PhD supervisor, Nicolas Macris, for providing me with the opportunity to pursue my doctoral research in the exact laboratory I had envisioned before embarking on this journey. I have gained invaluable knowledge under his guidance, and our numerous engaging scientific discussions have been a source of great joy throughout my PhD. I am certain that these mathematics, physics, quantum and philosophical debates we shared during my PhD will be cherished memories in the future.

Next, I would like to express my sincere thanks to Jean Barbier, whose inspiration and boundless energy propelled both my research and passion for the field. I am also grateful to Yue M. Lu - first, for graciously hosting me in his Harvard laboratory, and second, for his profound insights and astute ability to convey ideas and mathematical methods. I have gained a wealth of knowledge from our interactions, and his guidance has been invaluable to my academic journey. It was an honor to have Lénaïc Chizat as a jury member, and I extend my gratitude for his insightful follow-up discussions on my work. I would also like to express my appreciation to Olivier Lévêque for his insights in random matrix theory and keen interest in my research, as well as for our memorable discussions and enjoyable games of Go that we shared in the laboratory.

I extend my sincere gratitude to my fellow colleagues in the lab. In particular with professors and researchers who were always present throughout my PhD journey: Emre Telatar, Rüdiger Urbanke, Michael Gastpar, and Yanina Shkel who have enriched my journey as a researcher. A special thank you is due to Muriel Bardet for her unwavering help and support throughout my PhD. An acknowledgment is owed to Darius Šidlauskas and Wassila Ouerdane, whose advice played a pivotal role in guiding me through the challenges of pursuing a PhD. I am also thankful for my office mates, who are also my friends, and with whom I share so many good moments: Farzad, Emanuele, Raffaele, Diego, Nathan, Joon, Jean, Rodrigo, Anastasia, Anand, Perrine, Dina, Saleh, and many others.

I would like to thank my friends who have played influential roles in shaping many aspects of my life decisions. I express my gratitude to Sergey, a former colleague and also a compassionate friend with whom I could freely share any concerns and seek valuable advice, as well as the EPFL Team Rocket, comprising Louis, Damien, Laurent, and Christian, with whom I have shared countless precious memories with our master's studies at EPFL, Sat and beyond. I also wish to thank my friends from *classe prépa* and Centrale, particularly Thomas and Franck, for providing valuable feedback on the manuscript of this thesis and for our unforgettable skiing

Acknowledgements

sessions in Switzerland.

I extend my deepest gratitude to my family, in particular my mother, Sabine, and my father, Éric. None of my academic achievements would have been possible without their support at each step of my studies and my life. A special thanks goes to my sister, Laetitia, for always being there for me. I am also grateful to my in-laws, Zdenka and Alain, who welcomed me with open arms and contributed to creating a supportive environment.

Lastly, my heartfelt thanks go to my partner, Carole, who has been a constant pillar of support at every stage of my PhD journey. I am deeply grateful for her unwavering love and encouragement, and I eagerly anticipate the shared future that lies ahead for both of us.

Lausanne, January 8, 2024

Antoine

Abstract

In the rapidly evolving landscape of machine learning research, neural networks stand out with their ever-expanding number of parameters and reliance on increasingly large datasets. The financial cost and computational resources required for the training phase have sparked debates and raised concerns regarding the environmental impact of this process. As a result, it has become paramount to construct a theoretical framework that can provide deeper insights into how model performance scales with the size of the data, number of parameters, and training epochs.

This thesis is concerned with the analysis of such large machine learning models through a theoretical lens. The sheer sizes considered in these models make them suitable for the application of statistical methods in the limit of high dimensions, akin to the thermodynamic limit in the context of statistical physics. Our approach is based on different results from random matrix theory, which involves large matrices with random entries. We will make a deep dive into this field and use a spectrum of tools and techniques that will underpin our investigations of these models across various settings.

Throughout our journey, we begin by constructing a model starting from a linear regression. We then extend and build upon it to allow for a wider range of architectures, culminating in a model that closely resembles the structure of a multi-layer neural network. With the gradient-flow dynamics, we further develop analytical formulas predicting the learning curves of both the training and generalization errors. The equations derived in the process reveal several underlying phenomena emerging from the dynamics such as the double descent, and specific descent structures over time.

We then take a detour to explore the dynamics of the rank-one matrix estimation problem, commonly referred to as the spiked Wigner model. This model is particularly intriguing due to the presence of a phase transition with respect to the signal-to-noise ratio, as well as challenges related to the non-convexity of the loss function and non-linear learning equations. Subsequently, we address the extensive-rank matrix denoising problem which is an extension of the previous model. It holds particular interest in the context of sample covariance matrix estimation, and presents other challenges stemming from the initialization and the tracking of eigenvectors alignment.

Keywords: *Random matrix theory, machine learning, random feature, matrix denoising, gradient flow, high-dimensions, spiked Wigner, double descent, phase transition*

Résumé

Dans le paysage en constante évolution de la recherche en apprentissage automatique, les réseaux de neurones se distinguent par leur nombre de paramètres toujours croissant et leur dépendance à l'égard d'ensembles de données de plus en plus volumineux. Le coût et les ressources informatiques nécessaires à la phase d'entraînement suscitent des débats et soulèvent des préoccupations quant à l'impact environnemental de ce processus. Par conséquent, il devient essentiel de construire un cadre théorique capable de fournir des perspectives plus approfondies sur la manière dont les performances de ces modèles évoluent en fonction de la taille des données, du nombre de paramètres et d'étapes réalisées lors de l'entraînement.

Cette thèse se consacre à l'analyse de ces grands modèles d'apprentissage automatique à travers une perspective théorique. La taille considérable de ces modèles les rend aptes à l'application de méthodes statistiques dans la limite des dimensions élevées, semblable à la limite thermodynamique dans le contexte de la physique statistique. Notre approche repose sur des résultats de la théorie des matrices aléatoires, qui implique des matrices de grande dimension avec des entrées aléatoires. Nous approfondirons ce domaine et utiliserons un éventail d'outils et de techniques qui soutiendront nos investigations de ces modèles dans divers contextes.

Tout au long de notre parcours, nous commencerons par construire un modèle à partir d'une régression linéaire que nous développerons par la suite pour permettre l'analyse d'une plus grande variété d'architectures. Nous aboutirons à un modèle dont la structure se calque étroitement à celle d'un réseau de neurones multicouche. Grâce à la dynamique du *gradient-flow*, nous développerons des formules analytiques prédisant les courbes d'apprentissage des erreurs d'entraînement et de généralisation. Les équations obtenues au cours de cette analyse révèlent plusieurs phénomènes sous-jacents émergeant lors de la dynamique, tels que la "double-descente", ainsi que des structures de descente plus spécifiques au cours de l'apprentissage.

Nous ferons ensuite un détour pour explorer la dynamique du problème d'estimation de matrice de rang un, communément appelé le modèle spiked Wigner. Celui-ci est particulièrement intrigant en raison de la présence d'une transition de phase en fonction du rapport signal/bruit, et présente des difficultés liés à la non-convexité de la fonction d'objectif et la non-linéarité des équations d'apprentissage. Nous aborderons finalement le problème de débruitage de matrices de rang extensif, qui peut se concevoir comme une extension du

Résumé

modèle précédent. Il suscite un intérêt particulier dans le contexte de l'estimation de matrice de covariance dans un échantillon de données, et présente d'autres défis liés à l'initialisation ainsi qu'à l'alignement des vecteurs propres.

Mots clés : *Théorie des matrices aléatoires, apprentissage automatique, random-feature, débruitage de matrices, gradient-flow, hautes dimensions, spiked Wigner, double-descente, transition de phase*

Contents

Acknowledgements	i
Abstract (English/Français)	iii
1 Introduction	1
1.1 The class of linear models	2
1.2 Multivariate gaussian structure of the data	4
1.3 The random feature model	6
1.4 High-dimensional systems	7
1.5 Training and generalization curves	12
1.6 Non-convex optimization with Spiked Wigner model	16
1.7 Matrix denoising and extensive rank models	20
1.8 Organization and main contributions	21
I Methods in random matrix theory	25
2 Preliminaries with random matrices	27
2.1 Random matrices and their spectral distribution	27
2.2 Semicircular law and Marchenko-Pastur law	29
2.3 Holomorphic functional calculus	34
2.4 Algebraic expressions of random matrices	35
3 The Linear-pencil method	37
3.1 Introduction	37
3.2 Main Example: high-dimensional case	41
3.3 Application with finite size	44
3.4 The additivity law of the \mathcal{R} -transform	45
3.5 Derivation of result 3.2 in finite-dimension	49
3.6 Derivation of result 3.1: Three methods	52
Appendices	61
3.A Derivation of the Laguerre polynomials	61

II	High-dimensional estimations in linear models	63
4	Linear regression estimator	65
4.1	High-dimensional test error and double descent	65
4.2	Training and test error in the high-dimensional limit	66
4.3	Time evolution and learning curves	69
5	A framework: the gaussian covariate model	73
5.1	Introduction	73
5.2	Main results	77
5.3	Applications and examples	80
5.4	Conclusion	84
	Appendices	87
5.A	Gradient flow calculations	87
5.B	Test error and training error limits with linear pencils	91
5.C	Other limiting expressions	97
5.D	Applications and calculation details	99
6	The Random feature model	111
6.1	Introduction	111
6.2	Random feature model	114
6.3	Results and insights	117
6.4	Sketch of proofs and analytical derivations	121
6.5	Conclusion	124
	Appendices	125
6.A	Test Error substitutions	125
6.B	Cauchy's integral representation formula	127
6.C	High-dimensional limit	131
6.D	Linear Pencil	132
6.E	Numerical results	140
III	Beyond the linear setting with matrix completion	149
7	The rank-one model: a non-convex setting	151
7.1	Introduction	151
7.2	Analytical solutions and illustrations	154
7.3	Integro-differential equations	159
7.4	Concentration results	160
7.5	Solution of integro-differential equations and overlap	162
7.6	Conclusion and future work	165
	Appendices	167

7.A Analysis of the cost	167
7.B Proof of propositions 7.2 and 7.3	167
7.C Proof of proposition 7.4	170
7.D Laplace Transform applicability	173
7.E Enforcing the spherical constraint in gradient dynamics	175
7.F Strict saddle property	175
7.G Analysis of the stationary equation	176
7.H Intermediate identities	177
7.I Asymptotic analysis of \bar{q}	179
7.J Additional experiments	193
8 Matrix denoising: an extensive rank model	197
8.1 Introduction	197
8.2 Results	200
8.3 Sketch of Proof	204
8.4 Conclusion	207
9 Conclusion and future research directions	209
 Bibliography	 213
 Curriculum Vitae	 227

1 Introduction

As the saying goes, “*All models are wrong*” (Box, 1976), and machine learning models are no exception. At the root of this statement lies the fact that these models are a mathematical construction with a *particular structure* that seeks to replicate the functioning of a “system” whose underlying full description and mechanism is not only fundamentally unknown, but quite often beyond reach. Observations of the true model - or more commonly called samples of *data* - is the raw ingredient to mimic its functioning with the expectation to accurately match it on future unseen observations. This is all achieved with an *optimization algorithm* that seeks to align underlying free parameters of this mathematical structure to better fit with these observations. There are the three pillars around which revolves a machine learning model: the data, the model and the optimization method.

As a general principle stated in Box’s article, since the true model is unknown, the scientist “*should seek an economical description of natural phenomena*” thereby following William Occam’s law of parsimony. Hence the Occam’s razor rule, which states that the scientist should choose simplicity over complexity when faced with two competing models that explain the sample data equally well. And yet, at the time of the writing of this work, current machine learning models seem to defy this principle. Indeed, the current trend in machine learning is to build more and more complex models that are able to fit the data with an ever increasing fidelity. Current models approach a trillion number of parameters (*e.g.* large language models in Brown et al. (2020)) and even at fixed number of data samples, current empirical observations seem to suggest evidence for better generalization performances with both an optimally increased number of parameters and training computations (Hoffmann et al., 2022). This seems to suggest that the notion of simplicity fostered by Occam’s razor is neither solely determined by the number of parameters of the model, nor by the number of training epochs.

While these empirical evidence steer our understanding towards the idea that the most economical representation of the data distribution is enabled by more parameters and its selection can be operated by more learning epochs, an overall growing concern regarding these structures are their interpretability. This is in stark contrast with the classical approach in physical sciences which puts at the forefront the interpretability of the model and simplicity of the

equations. However, physics models are no stranger to modeling complex interactions between a large number of variables. In fact, history in thermodynamic is a successful example where empirical laws were devised first (such as Boyle-Mariotte law in the 17th century) until modern approaches enabled by statistical physics allowed to derive these laws from first principles (for instance with the work of Maxwell-Boltzmann in the 19th century). Sometimes, physical sciences elaborate simplified models of reality that enable us to have an understanding of the interactions at work for the emergence of more complex phenomena. This is the case for instance with the 1-dimensional Ising model in statistical physics that provides hints of the emergence of ferromagnetism in a material, and in particular the existence of a phase transition with respect to the Curie temperature (Ising, 1925).

In this thesis, we propose to follow a similar path and explore different machine learning models through diverse settings for the data, the model design and the optimization algorithm. Although the models that we will scrutinize are simplifications and idealizations of more intricate models commonly used in practice, they offer a precise framework in which the emergence of complex phenomena can be studied. Specifically, our investigation unfolds across several sections: from a detailed exploration of linear models (Sections 1.1 to 1.3), where we assess their performance in high-dimensional scenarios (Sections 1.4 and 1.5), to an examination of non-convex settings with a focus on a rank-one matrix factorization model (Section 1.6). And finally, we address the extensive-rank case in Section 1.7.

1.1 The class of linear models

In supervised learning, a model learns to make predictions based on a set of real-valued labels y generated from a d -dimensional real input vector x . One of the most ubiquitous and simplest models that fit in this class are the linear models. They have been widely described in many text-books (Hastie et al., 2001; James et al., 2013), and although they strike by their simplicity compared to the aforementioned highly-parameterized generative models in the introduction, they serve as a building block of many different models. They can also be considered as the embryo of a neural network as we will see later on. Despite their simplicity, they display many features encountered with more complex models (training error, generalization error, overfitting, etc). They are therefore a convenient starting point and will be a major topic of this thesis. Specifically, the second part of the thesis is dedicated to investigating the emergence of complex phenomena in the high-dimensional limit.

In the most elementary description of these linear models, the labels are assumed to have a linear relationship with the data. Thus, the distribution of y is constructed with a hidden d -dimensional vector β^* such that the output conditional on the vector x follows a normal distribution:

$$\mathbb{E}[y|x] \sim \mathcal{N}(\beta_0^* + x^T \beta^*, \sigma^2) \tag{1.1}$$

Equivalently, the true labels are given by an underlying linear-function $f^*(x) = \beta_0^* + x^T \beta^*$ while the output y is precisely this label to which some additional Gaussian noise with tunable

variance σ^2 is added, say $\epsilon \sim \mathcal{N}(0, \sigma^2)$:

$$y(x) = f^*(x) + \epsilon \quad (1.2)$$

In practice in the supervised learning setting, the model usually only has access to a finite number of samples (x_i, y_i) , for $i = 1, \dots, n$ such that all the x_i are independent with each other and drawn from a true distribution \mathcal{P}_x . We write conveniently $X \in \mathbb{R}^{n \times d}$ the data matrix where each line i corresponds to a sample x_i , and Y such that Y_i corresponds to y_i . For the sake of simplicity, we assume that the labels are centered, leaving $\beta_0^* = 0$. With this condition, the optimization problem consists in finding an estimator $\hat{\beta}$ such that the function $\hat{y}(x) = x^T \hat{\beta}$ effectively captures the data-points and has the potential to generalize to future new points. The measure of fitness is probed with a loss function: in the regression setting, a standard loss is the quadratic loss $\|Y - \hat{Y}\|^2$. Therefore, $\hat{\beta}$ is chosen so as to minimize this loss: $\hat{\beta} = \operatorname{argmin}_{\beta} \|Y - \hat{Y}\|^2$. From a Bayesian perspective, this corresponds to the maximum likelihood estimator (MLE) from the likelihood of Y given the parameters β and the data X :

$$P(Y|X, \beta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i^T \beta)^2}{2\sigma^2}\right) \quad (1.3)$$

In order to mitigate a range of different issues that are discussed in the subsequent sections, a regularization term is added to the loss function. This is equivalent to imposing a prior-distribution on $\beta \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} I\right)$ and maximizing the distribution $P(Y, \beta|X)$:

$$P(Y, \beta|X) = P(Y|X, \beta)P(\beta|X) = \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i^T \beta)^2}{2\sigma^2}\right)\right) \frac{1}{\sqrt{d} \sqrt{2\pi \frac{\sigma^2}{\lambda}}} \exp\left(-\frac{\lambda \|\beta\|^2}{2\sigma^2}\right) \quad (1.4)$$

Hence this corresponds to calculating the ridge regression estimator:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \quad (1.5)$$

Solving $\nabla_{\beta} \mathcal{L}(\hat{\beta}) = 0$ yields an explicit formula known as the ridge regression estimator:

$$\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T Y \quad (1.6)$$

At this stage, we start with the description of our first model 1.1 that will serve as a foundation to introduce the methods used for further complex models. This is referred to as the random ridge regression model, wherein the data is sampled from a Gaussian distribution. This model will be studied in Chapter 4 and follows from a range of different results such as in (Hastie et al., 2019; Belkin et al., 2020a; Advani et al., 2020a).

Model 1.1. (*Random ridge regression*). In this model, we consider the ridge regression estimator with a regularization term λ for the input matrix $X \in \mathbb{R}^{n \times d}$ and the output vector $Y \in \mathbb{R}^n$ related by the linear relation $Y = X\beta^* + \xi$ for some noise vector $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ and a hidden signal $\beta^* \in \mathbb{R}^d$. The data matrix X is a random matrix with independent entries such that

$X_{ij} \sim \mathcal{N}(0, \frac{1}{d})$ for all (i, j) . The true signal β^* is any deterministic vector whose norm satisfies the relation $r^2 = \frac{1}{d} \|\beta^*\|^2$.

When a learning mechanism is specified, instead of the ridge regression estimator, another estimator $\beta(k)$ arises from the learning dynamics at each step k . A common method for this purpose is the gradient descent algorithm which starts with a random initialization $\beta(0)$ and iterates through new values of the estimator using an update rule based on the gradient of the loss function and parameterized by a learning rate η . More precisely, it generates a sequence of estimate vectors $\beta(k)$ with the following formula:

$$\beta(k+1) = \beta(k) - \eta \nabla_{\beta} \mathcal{L}(\beta(k)) \quad (1.7)$$

As shown in Hastie et al. (2019), a sufficiently small learning rate guarantees that when $k \rightarrow +\infty$, the gradient descent algorithm converges to the ridge regression estimator $\hat{\beta}$ that will be denoted as $\beta(+\infty)$ in this case. This algorithm can also be understood in terms of a discretization scheme of the *gradient-flow* method, where β evolves as a continuous function of time t with:

$$\frac{d\beta_t}{dt} = -\nabla_{\beta} \mathcal{L}(\beta_t) \quad (1.8)$$

The gradient-flow method is often regarded as a viable approximation of the gradient-descent method while providing a set of differential ordinary differential equations that are more readily amenable to analysis. Consequently, this approach will be followed in this thesis when examining the dynamics of our machine learning models.

1.2 Multivariate gaussian structure of the data

To introduce the next model that will serve as more general framework to cover many different cases, we will further consider a situation where the data is generated from two different centered-normal distributions sharing a covariance matrix Σ with a specific structure:

$$\begin{pmatrix} x \\ \hat{x} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V^* & \Sigma \\ \Sigma^T & U^* \end{pmatrix} \right) \quad (1.9)$$

In this setting, the true model (the teacher) and the underlying learning model (the student) have access to two distinct data distributions, albeit with specific correlations. Given a data point x for the teacher jointly distributed with a data point \hat{x} for the student, the teacher model outputs $y(x) = x^T \beta^*$ while the student assumes the existence of a linear relation with \hat{x} , so $\hat{y}(\hat{x}) = \hat{x}^T \beta$. This structure is introduced in full generality in Loureiro et al. (2021) and is referred to as the Gaussian covariate model.

Adjusting the variance profile of the data-points enables to capture a broader range of different models. For instance, when $\sigma = 0$, the fundamental case defined in model 1.1 corresponds to $U^* = V^* = \Sigma = I_d$. Otherwise, a substitute of the noise term $\epsilon \sim \mathcal{N}(0, \sigma^2)$ can be constructed

by setting $U^* = I_d$ as before, but adding one extra dimension for x acting as a surrogate of the noise:

$$V^* = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (d+1)} \quad \Sigma = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{(d+1) \times d} \quad (1.10)$$

This way, by setting $\beta_{d+1}^* = 1$, we effectively have defined $y(x) = x_{[1:d]}^T \beta_{[1:d]}^* + x_{d+1}$ with $x_{d+1} \sim \mathcal{N}(0, \sigma^2)$ while $x_{[1:d]}$ is the truncated vector x to its first d elements. This structure can be leveraged more extensively to investigate various models such as a misspecified model where the student only has access to a subset of the teacher vector x as described in Belkin et al. (2020a), or more general kernel methods (Loureiro et al., 2021). This motivates us to introduce our second model which will be thoroughly examined in Chapter 5.

Model 1.2. (*Gaussian Covariate Model*). *In this model, we consider a random matrix $Z \in \mathbb{R}^{n \times d}$ where each element are independent with $Z_{ij} \sim \mathcal{N}(0, \frac{1}{d})$. We let $A \in \mathbb{R}^{d \times p_A}$ and $B \in \mathbb{R}^{d \times p_B}$ be two deterministic matrices that can have different dimension p_A and p_B .*

1. *The teacher is given the data matrix $X = ZB$ and generates the output $Y = X\beta^*$ for a deterministic vector $\beta^* \in \mathbb{R}^{p_B}$*
2. *The student has the data matrix $\hat{X} = ZA$ and generates $Y_t = \hat{X}\beta_t$ by learning a vector $\beta_t \in \mathbb{R}^{p_A}$ using the gradient-flow optimization method with β_0 drawn independently from a normal distribution $[\beta_0]_i \sim \mathcal{N}(0, r_0^2)$.*

Note that this is an alternative but equivalent description of the covariance profile given before. The relation to the former representation results from $U^* = A^T A$, $V^* = B^T B$ and $\Sigma = B^T A$. The opposite relation is displayed in more details in Chapter 5. With this view, the data matrix of the teacher and the student is generated from the same source of randomness Z although the matrix A and B can be set to project the rows of Z to different subspaces.

Furthermore, instead of considering the covariance structure of X , it is also possible to see this model through an alternative angle where the structure is placed on β^* . Since $Y = X\beta^* = ZB\beta^*$, it is possible to investigate this model as a ridge regression $Y = Z\tilde{\beta}^*$ with a signal $\tilde{\beta}^* = B\beta^*$. As an illustrative remark highlighting the capabilities of this model, the Fourier model described in (Belkin et al., 2020a) falls within its scope when we choose $B = F_\omega$, where F_ω is the Fourier matrix of size $d \times d$ with $[F_\omega]_{ij} = \frac{1}{\sqrt{d}} e^{-2\pi i \frac{(i-1)(j-1)}{d}}$.

Despite the apparent simplicity due to the inherent linearity, we will see how model 1.2 can be used to investigate various models where the data is mapped to a feature vector through non-linear functions. These models are commonly referred to as the Kernel ridge regression model (Murphy, 2012). In the subsequent section, we will delve into a specific subclass of these feature-maps, which yields the so-called random feature model.

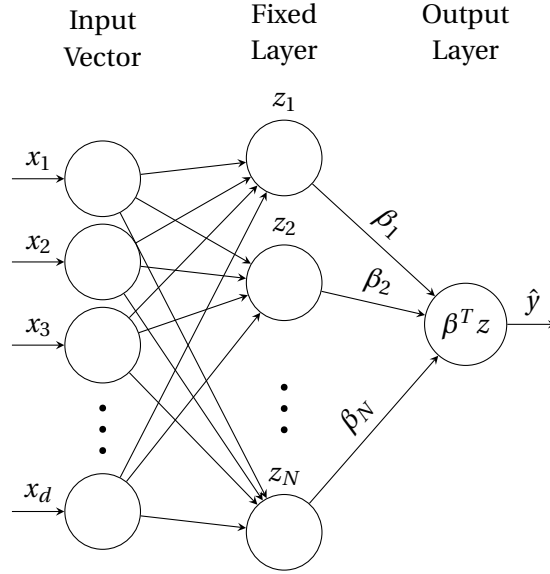


Figure 1.1: Graphical representation of the random feature model

1.3 The random feature model

The final cornerstone of this thesis concerning linear models is the random feature model, initially proposed by Rahimi and Recht (2008). It introduces a weight matrix, denoted as $\Theta \in \mathbb{R}^{N \times d}$, in conjunction with a non-linear activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$. This combination results in the mapping of a data point x to a feature vector z , which can be expressed as $z = \sigma(\Theta x)$ with the point-wise application of σ at each vector element. The student estimator becomes $\hat{y}_t(x) = \sigma(\Theta x)^T \beta_t$ which can either be interpreted as a specific version of a kernel ridge regression, or alternatively as a 2-layer neural network with a first layer fixed. Figure 1.1 illustrates the two dense layers within the model's structure. Therefore, it makes it a model of choice to study and capture neural network behaviors, and it is extensively described in the literature (Hastie et al., 2019; Mei and Montanari, 2019; Jacot et al., 2020a; Dhifallah and Lu, 2020).

The improvements over the previously considered linear-models are twofolds: first the introduction of an additional size parameter, denoted as N , provides greater control over the model complexity, enabling thereby the selection of either an over-parameterized or under-parameterized regime. Second, the introduction of a non-linear activation function, represented as σ , facilitates the modeling of more complex interactions between the data points. We will further introduce a scaling parameter within the activation function to keep the values standardized (of mean 0 and variance 1) in the specifications of model 1.3, as will become clear later in Section 1.4 and further in Chapter 6.

Model 1.3. (*Random feature Model*). In this model, we consider the random weight matrix $\Theta \in \mathbb{R}^{N \times d}$ and the random data matrix $X \in \mathbb{R}^{n \times d}$ with both independent and identically distributed entries $X_{ij} \sim \mathcal{N}(0, 1)$ and $\Theta_{ij} \sim \mathcal{N}(0, 1)$. We define further the matrix $Z = \sigma(d^{-\frac{1}{2}} X \Theta^T) \in \mathbb{R}^{n \times N}$

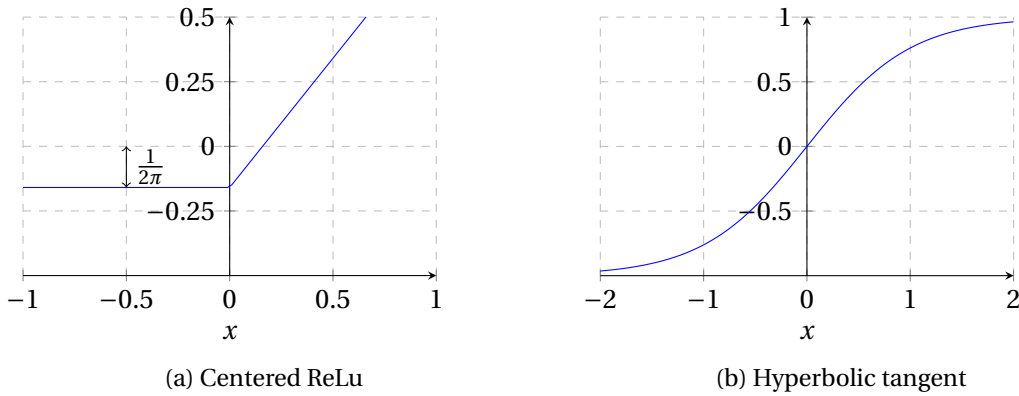


Figure 1.2: Two examples of centered activation functions

with the point-wise application of the activation function $\sigma \in L^2(e^{-\frac{x^2}{2}} dx)$. In particular, for the sake of simplification, we add the constraint that σ is centered, meaning that $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)}[\sigma(\epsilon)] = 0$.

The teacher outputs a vector $Y = X\beta^* + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ and the student learns a vector $\beta_t \in \mathbb{R}^N$ using the gradient-flow optimization method on the regression problem $\hat{Y}_t = Z\beta_t$. The initial value β_0 is drawn independently from a normal distribution $[\beta_0]_i \sim \mathcal{N}(0, \frac{r_0^2}{d})$.

Two classical examples of activation functions are displayed in Figure 1.2 with the centered rectified linear unit (ReLU on the left), which is the function $x \mapsto \max(x, 0)$ shifted downward to account for the null mean, and the hyperbolic tangent (tanh on the right). The introduction of a non-linearity adds complexity to the examination of the evolution of β_t . As for the other models, we will resort to further assumptions in Chapter 6 in order to allow for a tractable analysis. One of these assumptions is the high-dimensional regime, a topic that will be briefly addressed in the subsection Section 1.4.

1.4 High-dimensional systems

1.4.1 Random Matrices for Machine Learning

In this thesis, our focus lies in tracking the average behavior of the systems outlined within each model. And while the models that have been introduced may seem deceptively simple at first sight, only model 1.1 can be comprehensively analyzed for finite values of d and n . Indeed, model 1.2 would require to resort to complex calculations and model 1.3 remains challenging to tractably explore unless operating within a *high-dimensional regime*. This is sometimes referred to as the thermodynamic limit of the system in reference to statistical physics language. In this regime, the dimensions of the problem, represented as d and n (and N for the random feature model) are "infinite", but the ratio $\frac{n}{d}$ (and $\frac{N}{d}$) is fixed and becomes a constant of the model. The consideration of infinite-dimensional approximations becomes especially pertinent in the era of large-scale machine learning models, as evidenced by the

sheer number of parameters employed in the most recent large language models (billions of parameters in Brown et al. (2020)). This is particularly valuable to get analytical insights: analyzing these models in finite size is challenging due to the number of interacting parameters involved. By taking the thermodynamic limit, the analysis is simplified and yields tractable mathematical expressions and insights into the model's behavior. Note that other lines of work examines other possible settings such as the online-learning where each data-point is sampled at each step of a learning process (Wang et al., 2017), or infinitely wide neural network such as the NTK (Jacot et al., 2020b) and mean-field view analysis (Mei et al., 2018). While other analyses investigate various relationships between n and d , such as the polynomial relationship discussed in Hu and Lu (2022), this thesis focuses specifically on cases where $\frac{n}{d}$ (and $\frac{N}{d}$) is of finite order as $n, d, N \rightarrow +\infty$.

This specific setting where both the number of samples n and the number of parameters d go to infinity at the same rate puts us in the realm of random matrix theory. Historically, in the 1950's, Eugene Wigner faced analogous challenges while working on understanding the intricate organizational structure of heavy nuclei in nuclear physics. Instead of solving the Schrödinger Equation for n strongly interacting particles, Wigner proposed an innovative approach where their interactions are approximated by a Hamiltonian matrix with elements independently sampled from a specific distribution. In this way, Wigner postulated that by constructing this simplified model with a random matrix, it could effectively capture the energy spectrum of heavy nuclei, which corresponds to the eigenvalues of the Hamiltonian. This gave rise to the celebrated Wigner surmise as stated in Mehta (2004).

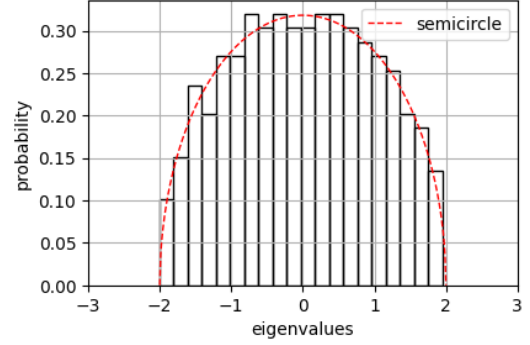
Since then, a multitude of random matrices have been devised and explored, some with specific characteristics that will be further discussed in Section I. A fundamental example, as illustrated in Figure 1.3, is the Gaussian Orthogonal Ensemble (GOE). In the GOE, the elements of an $n \times n$ real symmetric matrix are independently sampled from a normal distribution $\mathcal{N}(0, \frac{1}{n})$. The choice of variance $\frac{1}{n}$ ensures that, in the limit of large n , the eigenvalues take finite values of order 1. As n increases, the eigenvalues of such matrices tend to arrange themselves to follow the well-known Wigner semicircle law, characterized by the probability density function $\rho(x) = \frac{1}{2\pi} \sqrt{4 - x^2}$, thereby following the shape of a semicircle.

It is worth noting that matrix elements need not be normally distributed, just as datasets in the field of machine learning may not conform to such a distribution either. In fact, Wigner was considering larger sets of random matrices in Wigner (1958). Fairly recent research is still being conducted on similar distributional universalities (Bai, 1997; Tao and Vu, 2008), with a focus on a related concept known as "circular law", and demonstrating that the distribution requirements can indeed be relaxed and remain quite general while still leading to the emergence of the same law. In this thesis, our models will primarily resort to Gaussian random matrices, however, the possibility of exploring alternative distributions remains open for future research.

Another ubiquitous random matrix model in machine learning is the class of Wishart matrices,

$$M = \begin{pmatrix} 0.084 & 0.082 & \dots & 0.028 \\ 0.082 & 0.055 & \dots & -0.086 \\ \vdots & \vdots & \ddots & \vdots \\ 0.028 & -0.086 & \dots & -0.032 \end{pmatrix}$$

(a) A real symmetric random matrix with independent entries sampled from a normal distribution $\mathcal{N}(0, \frac{1}{n})$ with $n = 300$.



(b) Eigenvalues histogram for M

Figure 1.3: The "rise" of the semicircle law

which is also a fundamental example described by the product of a random matrix with its transpose (or transconjugate in the complex case). In the context of linear models 1.1 and 1.2 presented before, the Wishart matrix is the Gram matrix $X^T X$ (or sometimes referred to as the kernel when considering XX^T) of the data matrix X . Let's consider $X \in \mathbb{R}^{n \times d}$ with independent entries sampled from $X_{ij} \sim \mathcal{N}(0, \frac{1}{d})$. The variance of the entries is adjusted to ensure that the eigenvalue distribution is well-defined in the limit of large d and n with the fixed ratio $\phi = \frac{n}{d}$. In the high-dimensional limit, akin to the semicircle law for Wigner matrix, the eigenvalues of the Wishart matrix are distributed according to the Marchenko-Pastur distribution as established by these authors in Marčenko and Pastur (1967). This distribution is characterized by the following probability density function ρ with $\lambda_{\pm} = \left(1 \pm \frac{1}{\sqrt{\phi}}\right)^2$:

$$\rho(\lambda) = \frac{\phi}{2\pi} \sqrt{\left(\frac{\lambda_+}{\lambda} - 1\right) \left(1 - \frac{\lambda_-}{\lambda}\right)} \mathbb{1}_{\lambda_- \leq \lambda \leq \lambda_+} + (1 - \phi) \delta_0(\lambda) \mathbb{1}_{\phi < 1} \quad (1.11)$$

As in the Wigner case, in the limit of large n and d , individual matrix elements lose their significance for our purposes, but the calculation reveals valuable analytical insights through the limiting spectral density. For instance, when $\phi < 1$, the formula indicates the presence of a proportion $1 - \phi$ zero eigenvalues - as expected from the matrix $X^T X$ which is not full-rank when $n < d$.

Note that the spectral density is not the primary focus of the calculations in this thesis, but rather can be derived as a byproduct of the resulting equations. In some instances, analytical expressions for the spectral density may not even exist, necessitating the use of numerical methods for its computation. Instead, the central focus lies in computing traces involving large-dimensional random matrices. As an illustrative example, consider equation (1.6) which we will encounter further in Chapter 4. In the noiseless setting when $Y = X\beta^*$, we can quantify

the "dissimilarity" between the estimated parameter $\hat{\beta}$ and the true parameter β^* as follows:

$$\frac{1}{d} \|\hat{\beta} - \beta^*\|^2 = \frac{1}{d} \|(X^T X + \lambda I_d)^{-1} X^T X - I_d\| \beta^* \|^2 = \frac{\lambda^2}{d} \beta^{*T} (X^T X + \lambda I_d)^{-2} \beta^* \quad (1.12)$$

When averaging this dissimilarity over $\beta^* \sim \mathcal{N}(0, I_d)$, the term on the right-hand side can be simplified by using the cyclicity of the trace operator and the independence of β^* and X :

$$\mathbb{E}_{\beta^*} [\beta^{*T} (X^T X + \lambda I_d)^{-2} \beta^*] = \text{Tr} [(X^T X + \lambda I_d)^{-2} \mathbb{E}[\beta^* \beta^{*T}]] = -\frac{\partial}{\partial \lambda} \text{Tr} [(X^T X + \lambda I_d)^{-1}] \quad (1.13)$$

This term is thus related to the derivative of the *Stieltjes transform* of the spectral density $\rho_{X^T X}$ of the Gram matrix $X^T X$. In some contexts, in the limit of large d , such expressions can be averaged and analytical results can be derived, while offering a more elegant and manipulable form than working with the spectrum itself. For instance, consider the distribution ρ_M that emerges with the Wigner matrix M defined as in Figure 1.3, a case which will be further investigated in Chapter 2. We can establish the following relationship between the Stieltjes transform and the trace of $(M - zI_n)^{-1}$ in the limit of large n for any $z \in \mathbb{C} \setminus \mathbb{R}$:

$$\int_{\mathbb{R}} \frac{\rho_M(\lambda) d\lambda}{\lambda - z} = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \text{Tr} [(M - zI_n)^{-1}] \quad (1.14)$$

This Stieltjes transform of ρ_M , denoted as $g(z)$ for convenience, is a solution of an algebraic expression:

$$g(z)^2 + zg(z) + 1 = 0 \quad (1.15)$$

As we consider increasingly complex models and associated random matrices in our different models, the expressions describing the traces tend to become more intricate systems of algebraic equations. Deriving these expressions will be addressed in different chapters of this thesis. We will rely essentially on the so-called *linear pencil method* that was initially described in Rashidi Far et al. (2006) and will be comprehensively discussed in Chapter 3.

1.4.2 Activation functions on matrix elements

As discussed thus far, random matrix theory has a rich literature and set of tools to calculate closed-form expressions involving traces of Gaussian random matrices. We have seen that Marchenko-Pastur law provides an analytic expression that describes the spectral density of the Gram matrix $X^T X$ or the *kernel* $K = XX^T$ in the high-dimensional limit for the Gaussian design matrix X . But the situation is more complicated with an additional non-linear activation function σ , such as in the description of model 1.3 where the kernel becomes $K = ZZ^T$. Here, the elements of Z are not Gaussian and they also exhibit interactions with one another. As the reader may anticipate, a new law leading to further simplifications arise when the sizes of the involved random matrices are increasingly large. An initial study that sheds light on this phenomenon is presented in Pennington and Worah (2017) where the complete kernel spectral distribution is derived analytically. In this context, the class of applicable activation

functions is reduced to the contribution of two specific coefficients μ and ν that are related to σ by the following formula:

$$\mu = \mathbb{E}[\sigma(\epsilon)\epsilon] \quad \nu^2 = \mathbb{E}[\sigma(\epsilon)^2] - \mu^2 \quad (1.16)$$

As an example, we find for the shifted ReLU on the left of Figure 1.2 that $\mu = \frac{1}{2}$ and $\nu = \frac{1}{2}\sqrt{1 - \frac{2}{\pi}}$. Alternatively, in the high-dimensional limit, the impact of the centered activation function in the random feature model can be entirely characterized by the ensemble of activation functions spanned by the linear combination of the two Hermite Polynomials $H_{e_1}(x) = x$ and $H_{e_2}(x) = x^2 - 1$. As a seminal example of this application highlighted in the same paper, the authors describe a phenomenon through an analytical result in which the coefficient ν , controlling the non-linearity of σ , turns out to control the *memorization capacity* of a random feature model using only the Stieltjes transform of the kernel. This capacity is assessed through the training error achieved with random labels.

The exploration of this principle which characterizes the random feature kernel using non-linear activation functions, has spurred the development of an alternative approach involving what are described as *Gaussian equivalents*. This simplification arises whenever there is the application of such activation functions and asserts that the random matrix $Z = \sigma(d^{-\frac{1}{2}}X\Theta)$ can be equivalently represented as $Z = \mu d^{-\frac{1}{2}}X\Theta + \nu\Omega$, where Ω is a newly introduced independent random matrix with independent entries. This alternative representation yields the same results in the random feature model and is extensively applied in Adlam and Pennington (2020a). The rigorousness of the application of such equivalents continues to be an active area of research, as is exemplified by recent work such as (Lu and Yau, 2022; Goldt et al., 2022). In this thesis, we will employ it as a fundamental principle - even in more intricate algebraic expressions of random matrices.

1.4.3 Sample covariance matrix and realistic datasets

A final aspect to address concerns realistic datasets. As for the Wigner's Hamiltonian, it may be tempting to assume that once normalized, datasets can be treated as Gaussian random matrices equally well. This assumption is actually partly motivated by our earlier discussion, which showed that even non-Gaussian independent entries exhibit a universal eigenvalue distribution. However, even for common datasets such as MNIST, the independence of the entries is not guaranteed. In this thesis, we adopt an alternative approach by approximating these datasets under the assumption that any vector x sampled from the data is a Gaussian vector with a hidden sample covariance matrix, as discussed in Potters and Bouchaud (2020). This corresponds to considering that there exists a matrix B such that any vector x is sampled from a standard Gaussian vector z such that $x = Bz$ (or $X = ZB$ in matrix form). This assumption aligns with the specification of our model 1.2. Estimating the covariance matrix can be challenging and is contingent on the number of samples of the data. Therefore, in this setting, we will resort to estimating the covariance matrix using the full dataset when

the sample size permits it ($B^T B \simeq X^T X = B Z^T Z B$). We will then develop some predictive formula for a sub-sample of the data (with a size considerably smaller than n). The eigenvalue distribution of the sub-samples can be shown to follow a formula that will depend on the spectral distribution of the covariance matrix. This will be treated as a theoretical example in Chapter 3 using the previously mentioned linear-pencil method, and will be used in greater length in Chapter 5 to demonstrate the predictive capabilities of the equations derived in this chapter.

1.5 Training and generalization curves

While the model 1.1, 1.2 and 1.3 are optimized - or trained - using the mean square error introduced in equation (1.5), this loss has to be scaled to converge to a finite value in the limit $n, d \rightarrow +\infty$. To achieve this convergence, an additional scaling-factor $\frac{1}{n}$ (or $\frac{1}{d}$) is added to both terms specified in the loss function (1.5). We shall refer to this modified loss as the training error, denoted as $\mathcal{E}_{\text{train}}^\lambda(\beta)$.

Besides, as the model is only trained on a sample of size n drawn from \mathcal{P}_x , perfect generalization to unobserved data points is neither guaranteed nor expected. So in addition to the training error, we assess model performances using the generalization error, denoted as \mathcal{E}_{gen} . This metric represents the expected loss on new data points x sampled from the true distribution \mathcal{P}_x , and still in the limit $n \rightarrow +\infty$:

$$\mathcal{E}_{\text{train}}^\lambda(\beta) = \lim_{n, d \rightarrow \infty} \left\{ \frac{1}{n} \|Y - X\beta\|^2 + \frac{\lambda}{d} \|\beta\|^2 \right\} \quad (1.17)$$

$$\mathcal{E}_{\text{gen}}(\beta) = \lim_{n, d \rightarrow \infty} \mathbb{E}_{x \sim \mathcal{P}_x} [(y(x) - x^T \beta)^2] \quad (1.18)$$

Notably, the training error $\mathcal{E}_{\text{train}}^\lambda(\beta)$ is still a random variable as it depends on the randomness of the matrix X . Furthermore, β_t itself also depends on X and thus, so does $\mathcal{E}_{\text{gen}}(\beta)$ which is consequently a random variable. However, it is often the case that these errors concentrate around their mean value, also referred to as *self-averaging*. Roughly speaking, this means that the probability of $\mathcal{E}_{\text{train}}^\lambda(\beta)$ to deviate from its mean is typically bounded by a decreasing function of n . Note that the same self-averaging phenomenon also holds with respect to β^* and ξ when they are sampled from an appropriate distributions as those described in the previous models. This concentration is typically difficult to prove, and won't be the primary focus of part 4 in which it will usually just be assumed.

It is quite common to further decompose these terms into smaller irreducible quantities which are often referred to as bias and variance. The general representation is given by the identity:

$$\mathbb{E}[(\hat{y}(x) - y(x))^2] = \mathbb{E}[\hat{y}(x) - y(x)]^2 + \text{Var}(\hat{y}(x)) + \text{Var}(y(x)) \quad (1.19)$$

where the first term is the bias, the second term is the variance of the estimator and the third term another irreducible noise from the ground truth. The expectation is made on the

different sources of randomness, here X, ξ, ϵ . But also β_0 when using gradient-flow. However, the same formula also applies conditional on some chosen random variables, so that different decompositions can be made. For instance, we can rewrite the previous identity conditional on the data matrix X and the new input x :

$$\mathbb{E}[(\hat{y}(x) - y(x))^2] = \mathbb{E}[\mathbb{E}[\hat{y}(x) - y(x)|X, x]^2] + \mathbb{E}[\text{Var}(\hat{y}(x)|X, x)] + \mathbb{E}[\text{Var}(y(x)|X, x)] \quad (1.20)$$

To give an example, let's consider an estimator $\hat{\beta}$ (which could be β_t or β_∞) such that $\hat{y}(x) = x^T \hat{\beta}$. Let's also assume that $\mathcal{P}_x \sim \mathcal{N}(0, \frac{1}{d} I_d)$ so that $\mathbb{E}_x[xx^T] = \frac{1}{d} I_d$:

$$\mathbb{E}[\text{Var}(\hat{y}(x)|X, x)] = \mathbb{E}\left[(x^T \hat{\beta} - \mathbb{E}[x^T \hat{\beta}|X, x])^2\right] \quad (1.21)$$

$$= \mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}|X])^T (xx^T) (\hat{\beta} - \mathbb{E}[\hat{\beta}|X])\right] \quad (1.22)$$

$$= \frac{1}{d} \mathbb{E}\left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}|X]\|^2\right] \quad (1.23)$$

For the bias term, assuming further that $y(x) = x^T \beta^* + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and β^* is a deterministic vector, first we find the irreducible noise $\mathbb{E}[\text{Var}(y(x)|X, x)] = \sigma^2$, and the bias term:

$$\mathbb{E}[\mathbb{E}[\hat{y}(x) - y(x)|X, x]^2] = \mathbb{E}\left[(x^T \beta^* - x^T \mathbb{E}[\hat{\beta}|X])^2\right] \quad (1.24)$$

$$= \mathbb{E}\left[(\beta^* - \mathbb{E}[\hat{\beta}|X])^T xx^T (\beta^* - \mathbb{E}[\hat{\beta}|X])\right] \quad (1.25)$$

$$= \frac{1}{d} \mathbb{E}\left[\|\beta^* - \mathbb{E}[\hat{\beta}|X]\|^2\right] \quad (1.26)$$

So in the limit $n, d \rightarrow \infty$, with the self-averaging assumption as before, we find the following generalization error decomposition:

$$\mathcal{E}_{\text{gen}}(\hat{\beta}) = \mathbb{E}[\mathcal{E}_{\text{gen}}(\hat{\beta})] = \sigma^2 + \lim_{n, d \rightarrow \infty} \underbrace{\frac{1}{d} \mathbb{E}\left[\|\beta^* - \mathbb{E}[\hat{\beta}|X]\|^2\right]}_{\mathcal{B}_X(\hat{\beta})} + \lim_{n, d \rightarrow \infty} \underbrace{\frac{1}{d} \mathbb{E}\left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}|X]\|^2\right]}_{\mathcal{V}_X(\hat{\beta})} \quad (1.27)$$

In other words, in the high-dimensional regime, the generalization error can still be decomposed in two irreducible and finite quantities which are the *squared-bias* $\mathcal{B}_X(\hat{\beta})$ and the *variance* $\mathcal{V}_X(\hat{\beta})$. As these terms represent irreducible positive quantities contributing to the total error, they often exhibit a counterbalancing effect. Minimizing one of these terms tends to result in the increase in the other, giving rise to the tradeoff phenomenon. This tradeoff is illustrated as a diagram in Figure 1.4. More specifically, these two errors are known to evolve differently depending on the "*model complexity*". In general, on the one hand, in the case of "simple" models, the bias is typically high because the model lacks the capability to properly fit the data. However, they exhibit relative stability when trained on different datasets, which is referred to as the *under-fitting* regime. On the other hand, in the case of "complex" models, the bias decreases as these models can be fine-tuned enough to fit the training set. However, this comes at the cost of increased sensitivity to the training set, leading to phenomenon known

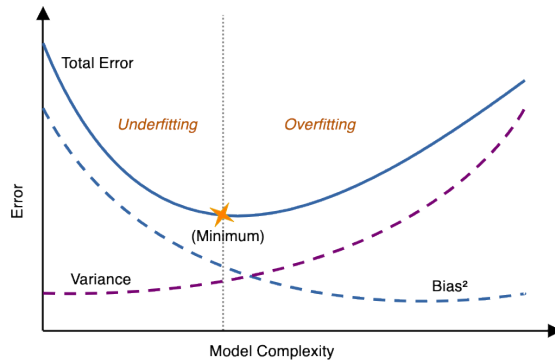


Figure 1.4: Bias-variance tradeoff diagram

as *over-fitting*. We raise the awareness to the reader that different conventions can be taken regarding the bias-variance decomposition as described in Adlam and Pennington (2020b). The convention chosen in this regard is called the *classical* bias-variance decomposition.

It may be tempting to assume that the notion of complexity boils down to a simple parameter count, such as in the parameter N of the random feature model 1.3. However, recent works revealed the inadequacy of such a simplistic understanding of the nascent neural networks structures in the random feature model as proved in Belkin et al. (2019a); Hastie et al. (2019); Mei and Montanari (2019). Indeed, the landscape of the generalization curve proves to be far more intricate than initial intuition might suggest, as depicted in the diagram 1.5.

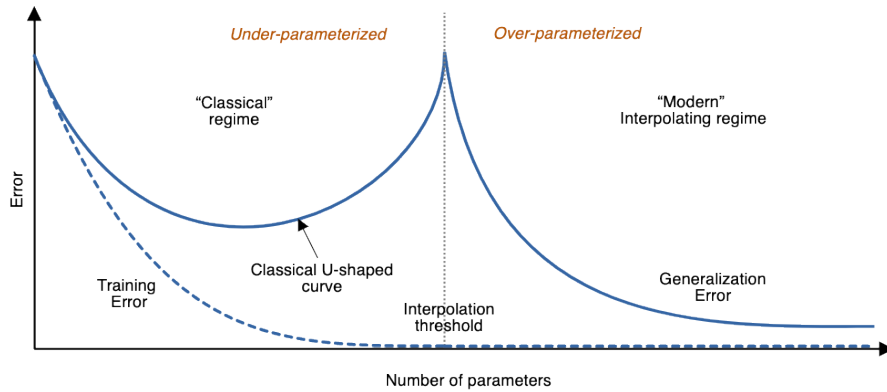


Figure 1.5: Double descent diagram

This diagram continues to feature the classical U-shaped curve previously described, where the over-fitting phase results in an increase of the generalization error. However, it also demonstrates a second descent phase, in which the generalization error decreases again. This phenomenon is known as the *double descent*. It has a long history, with initial observations dating back to 1989, but has attracted a lot of attention in recent years. For a detailed historical overview, we refer the reader to (Loog et al., 2020). Surprisingly, this observation is not confined to the random feature model, but it also manifests in the basic model referred in 1.1.

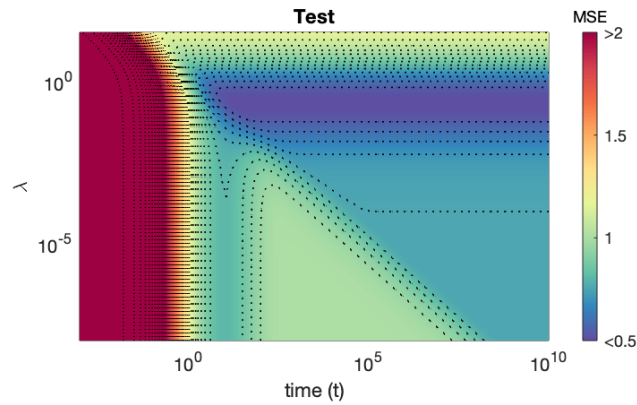


Figure 1.6: Double descent during the training phase in the random feature model - theoretical curves for shifted ReLu and $\phi = \frac{n}{d} = 2$ and $\psi = \frac{N}{d} = 6$

However, what is particularly interesting in model 1.3 is that the model can clearly achieve better generalization performances in this second phase rather known as the "interpolating regime". In other words, adding more parameters in the 2-layer neural network benefits the generalization error instead of exacerbating it. What is also intriguing is that this phenomenon occurs even as the training error becomes virtually zero numerically, a point often referred to as the interpolation threshold. At this location, the model doesn't seem to self-improve purely based on the value given by the training error. However, it remains possible to enhance the generalization performance by introducing more parameters. One interpretation of this phenomenon given in (Belkin et al., 2019a) is that the increase of parameters enriches the function class encompassed by all potential model parameterizations: richer models leads to richer class of learnable functions. Consequently, the model has improved its capacity to discover a more suitable and "simpler" data fit. This perspective not only aligns with the Occam's razor principle but also revises our understanding of complexity, suggesting that it is not merely proportional to the number of parameters.

Quite remarkably, the double-descent phenomena is not an idea confined to theoretical realms, rather, it manifests as a general phenomenon. First and foremost, it has been observed empirically in Nakkiran et al. (2020a) for deep neural network (specifically, ResNet18) and real datasets (such as CIFAR10), and bolstered the notion that an increase in parameters can be advantageous in the domain of deep learning. Furthermore, the double descent phenomenon extends beyond the dimension of the parameter count, it is also observed with the number of training samples (sample-wise double descent) and epochs (epoch-wise double descent). In this thesis, we aim to establish precise generalization error curves for the generalization error as in Figure 1.6. On this picture, we can see that a epoch-wise double-descent structure emerges as the number of parameters N increase compared to d (with the ratio ψ).

Incidentally, even more surprising structures can emerge such as the parameter-wise triple descent observed in (d'Ascoli et al., 2020) and also empirically in (Nakkiran et al., 2020b). We will see that in fact, by tuning the structure of matrix B in model 1.2, any number of descents

can be engineered.

1.6 Non-convex optimization with Spiked Wigner model

The class of linear models trained with the mean-squared loss is highly effective to demonstrate complex phenomena observed empirically such as the double-descent, yet it exhibits at least two main limitations. Firstly, the convex nature of the loss function and the linearity of its gradient renders the model more tractable although this is usually not the case with deep learning models. Secondly, the setting that has been described doesn't allow to consider other scalings with the number of parameters which typically grows linearly with the number of samples or dimensions. For instance, if the first layer of the random feature model is not fixed, the number of parameters would grow quadratically with the number of samples because the weight matrix Θ would need to be learned. Some work has been pursued in this direction (Ba et al., 2022), but the analysis remains challenging.

To address the first issue, we will explore another optimization problem and examine a different setting where the loss function is non-convex and the gradient is non-linear, yet the model remains tractable. This model is known as the Spiked Wigner model and is defined as follows:

Model 1.4 (Spiked Wigner Model). *In this model we consider a hidden signal θ^* sampled uniformly on the hypersphere $\mathbb{S}^{n-1}(\sqrt{n})$, that is such that $\|\theta^*\|^2 = n$ and a real symmetric random matrix $\xi \in \mathbb{R}^{n \times n}$ with independent entries sampled from a normal distribution $\mathcal{N}(0, 1)$. The matrix of observations is generated as follows:*

$$Y = \theta^* \theta^{*T} + \sqrt{\frac{n}{\lambda}} \xi \quad (1.28)$$

for a predefined signal to noise ratio $\lambda > 0$. In this setting, we learn a vector θ by minimizing the Frobenius norm $\mathcal{H}(\theta) = \|Y - \theta\theta^T\|_F^2$ using a gradient-flow method on the hypersphere with a given θ_0 initialized at random with $\theta_0 \sim \mathcal{U}(\mathbb{S}^{n-1}(\sqrt{n}))$.

Contrary to the first three models, the nature of the setting is not to learn a response $y(x)$ given a set of inputs (Y, X) , but instead to recover a signal vector θ^* from a noisy input matrix Y . We can find some connections in machine learning with different methods and problems, such as PCA, low-rank matrix factorization, matrix completion and other lines of research Ge et al. (2017a); Bhojanapalli et al. (2016); Ge et al. (2017b); De Sa et al. (2015).

In the outlined model 1.4, the primary metric of interest is the overlap $q(\theta) = \frac{\theta^T \theta^*}{n}$. As both vectors have norm \sqrt{n} , this quantity characterizes the cosine, ranging from -1 and 1 , of the angle between the true signal and the estimated signal. Note also that $-\theta^*$ yields the exact same observation matrix Y as θ^* , so in the best case it is only possible to recover the signal vector up to a sign. Alternatively, a natural choice is to consider the mean-square-error between the same vectors, but for any estimator $\theta \in \mathbb{S}^{n-1}(\sqrt{n})$, we find the relation $\frac{1}{n} \|\theta - \theta^*\|^2 = 2(1 - q(\theta))$. Consequently, the overlap is self-sufficient to describe both quantities.

1.6 Non-convex optimization with Spiked Wigner model

The statistical limits of the model when $n \rightarrow +\infty$ have been elucidated in great detail in the Bayesian framework in a series of works where the minimum mean-square error is rigorously computed (Korada and Macris, 2009; barbier et al., 2016; Lelarge and Miolane, 2018; Miolane, 2017). It has also prompted many different investigations uncovering intriguing phenomena, especially when the prior distribution on the hidden signal θ^* deviates from the uniform distribution on the hypersphere, which can yield discontinuous overlaps. With the uniform prior on the signal, as is the case in model 1.4, it is shown rigorously that for a fixed $\lambda > 0$, the optimal achievable overlap is $\pm\sqrt{1-\frac{1}{\lambda}}$ when $\lambda \geq 1$, and 0 otherwise. In other words, as this is illustrated in Figure 1.7, there exists a critical limit for $\lambda = 1$ in which the model undergoes a *phase transition*. From an algorithmic perspective, this optimal theoretical overlap can be achieved by selecting the eigenvector associated with the largest eigenvalue of the matrix Y in absolute value (P  ch  , 2004; Baik et al., 2005b).

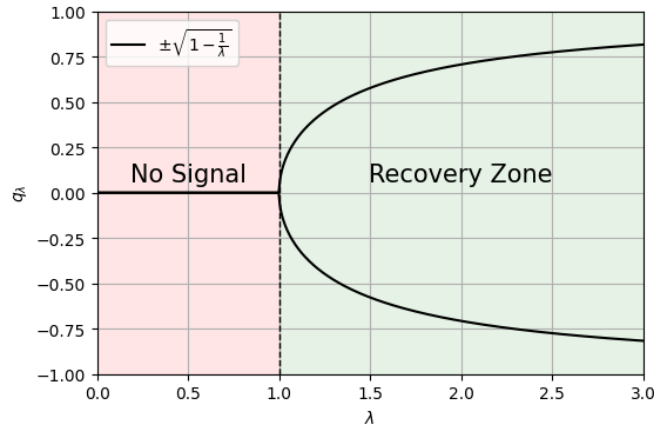


Figure 1.7: Phase transition for the overlap in the Spike-Wigner model with respect to λ

The existence of this transition can be explained by examining the eigenvalue distribution of Y in large dimension. To illustrate this, let us first reformulate the problem as $M_1 = \sqrt{\lambda}\frac{\theta^*\theta^{*T}}{n} + M_0$ with $M_1 = \frac{\sqrt{\lambda}}{n}Y$ and $M_0 = \frac{\xi}{\sqrt{n}}$. In this way, M_0 is the standard Wigner matrix whose spectral distribution becomes the semicircle law in the limit of large n . Now for finite n , when λ is large enough, the eigenvalue distribution of M_1 will essentially match that of M_0 if n is sufficiently large, but with an additional outlier eigenvalue coming from the rank-one perturbation $\frac{\sqrt{\lambda}}{n}\theta^*\theta^{*T}$. Note that as $n \rightarrow +\infty$, the mass of this additional eigenvalue tends to 0 as there are infinitely many more eigenvalues distributed on the support $(-2, 2)$ of the semicircle law. Nevertheless, it remains possible to calculate its exact location in this limit. Intuitively, if λ is not sufficiently large, the additional eigenvalue will be absorbed by the bulk of the semicircle law and won't manifest as an outlier. This is depicted in Figure 1.8 for $\lambda = 3$ where we observe that the outlier eigenvalue is already in close proximity to the bulk.

As an illustrative example using key concepts of random matrix theory, let us show the exact recovery of the aforementioned results. Using the spectral theorem, let $\lambda_1, \dots, \lambda_n$ be the

Chapter 1. Introduction

sorted eigenvalues of M_1 and u_1, \dots, u_n their normalized associated eigenvectors, so that $M_1 = \sum_{i=1}^n \lambda_i u_i u_i^T$. Then, the *resolvent* of M_1 can be expressed as:

$$(M_1 - zI_n)^{-1} = \frac{u_1 u_1^T}{\lambda_1 - z} + \dots + \frac{u_n u_n^T}{\lambda_n - z} \quad (1.29)$$

Depending on the largest value between $|\lambda_1|$ and $|\lambda_n|$, either $\sqrt{n}u_1$ or $\sqrt{n}u_n$ is selected as the estimator θ . Let's assume this is λ_n . In large dimension, we thus expect $\lambda_1, \dots, \lambda_{n-1}$ to follow closely the bulk distribution of a Wigner matrix and λ_n will be an outlier. Now if we examine the *resolvent* using Shermann-Morrison formula:

$$(M_1 - zI_n)^{-1} = (M_0 - zI_n)^{-1} - \frac{\frac{\sqrt{\lambda}}{n} (M_0 - zI_n)^{-1} \theta^* \theta^{*T} (M_0 - zI_n)^{-1}}{1 + \frac{\sqrt{\lambda}}{n} \theta^{*T} (M_0 - zI_n)^{-1} \theta^*} \quad (1.30)$$

we thus expect to find an outlier pole at $z_0 = \lambda_n$ when $1 + \frac{\sqrt{\lambda}}{n} \theta^{*T} (M_0 - zI_n)^{-1} \theta^* = 0$. In the high-dimensional limit, using for instance the results of (Erdős et al., 2008), this becomes $1 + \sqrt{\lambda}g(z_0) = 0$ with $g(z)$ given in equation (1.15). After some algebraic reductions, we can determine that the outlier eigenvalue has the precise location $z_0 = \lambda_n = \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}$ for $\lambda \geq 1$.

In order to recover the overlap, let us set $\theta = \sqrt{n}u_n$ and compute the limiting value of $q(\theta) = \frac{1}{\sqrt{n}} u_n^T \theta^*$. By using again the resolvent of M_1 , notice that from equation (1.29), the quadratic form $\theta^{*T} (M_1 - zI_n)^{-1} \theta^*$ yields an expression with the overlap between θ^* and each eigenvector:

$$\theta^{*T} (M_1 - zI_n)^{-1} \theta^* = \sum_{i=1}^n \frac{1}{\lambda_i - z} u_i^T \theta^* u_i^T \theta^* \quad (1.31)$$

By further multiplying by $\frac{1}{n}(\lambda_n - z)$ and taking the limit when $z \rightarrow \lambda_n$, we find the squared overlap q_λ^2 :

$$\frac{1}{n} \lim_{z \rightarrow \lambda_n} \{(\lambda_n - z) \theta^{*T} (M_1 - zI_n)^{-1} \theta^*\} = \frac{1}{n} (u_n^T \theta^*)^2 = q_\lambda^2 \quad (1.32)$$

Now, by proceeding in a similar way but using the right-hand side of the formula (1.30):

$$\frac{1}{n} \lim_{z \rightarrow \lambda_n} \{(\lambda_n - z) \theta^{*T} (M_1 - zI_n)^{-1} \theta^*\} = \lim_{z \rightarrow \lambda_n} \sqrt{\lambda} (z - \lambda_n) \frac{(\frac{1}{n} \theta^{*T} (M_0 - zI_n)^{-1} \theta^*)^2}{1 + \sqrt{\lambda} \frac{1}{n} \theta^{*T} (M_0 - zI_n)^{-1} \theta^*} \quad (1.33)$$

So in the limit of large n , we expect to find:

$$q_\lambda^2 = \lim_{z \rightarrow z_0} \frac{(z - z_0) \sqrt{\lambda} g(z)^2}{1 + \sqrt{\lambda} g(z)} \quad (1.34)$$

Using l'Hôpital's rule results in $q_\lambda^2 = \frac{g(z_0)^2}{g'(z_0)}$. With the derivative of equation (1.15) with respect to z (which yields the formula $2g'(z)g(z) + zg'(z) + g(z) = 0$) we conclude with the expected result:

$$q_\lambda^2 = \frac{g(\lambda_n)^2}{g'(\lambda_n)} = -g(\lambda_n)(2g(\lambda_n) + \lambda_n) = \frac{1}{\sqrt{\lambda}} \left(-2 \frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda} \right) = 1 - \frac{1}{\lambda} \quad (1.35)$$

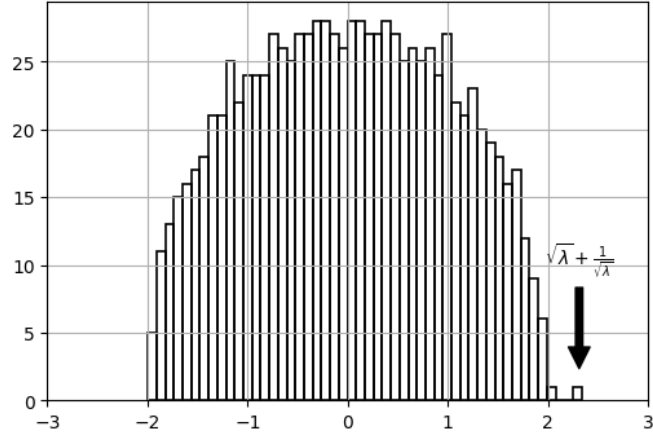


Figure 1.8: Histogram of the eigenvalues of a data matrix sampled according to the Spike Wigner model 1.4 for $\lambda = 3$ and $n = 1000$

As a side remark, it is also possible to achieve a smaller mean-square error between θ and θ^* when the condition on the norm of θ is relaxed. To achieve it, it suffices to rescale the given largest eigenvalue with a coefficient α , so $\tilde{\theta} = \alpha\theta$. In this case, we find the quadratic equation $\frac{1}{n} \|\tilde{\theta} - \theta^*\|^2 = \alpha^2 - 2\alpha q(\theta) + 1$ whose minimum is given at $\alpha_0 = q(\theta)$. Therefore, in this situation, a smaller overlap is obtained with $q(\tilde{\theta}) = q(\theta)^2$ but also a smaller error: $\frac{1}{n} \|\tilde{\theta} - \theta^*\|^2 = 1 - q(\theta)^2$

Another focal point is the best achievable solutions for various algorithms, such as Approximate Message Passing (AMP), as explored in (barbier et al., 2016; Lesieur et al., 2015). Others have investigated the behaviour of AMP under spectral initialization (Montanari and Venkataramanan, 2021). In this thesis, the highlight is again on the gradient-flow algorithm. Specifically, we are interested in the evolution of this optimization algorithm in the precise setting of model 1.4 while constraining the norm of the estimator. The gradient-flow method needs to be adjusted with an additional term ensuring that $\frac{d\|\theta_t\|^2}{dt} = 0$ at all time:

$$\frac{d\theta_t}{dt} = -\nabla_{\theta} \mathcal{H}(\theta_t) + \frac{1}{n} (\theta_t^T \nabla_{\theta} \mathcal{H}(\theta_t)) \theta_t \quad (1.36)$$

In contrast to the linear models, we no longer have a linear differential equation in the parameter θ_t . However, we will demonstrate rigorously that the evolution of the model is still tractable in the high-dimensional limit $n \rightarrow +\infty$, albeit exhibiting different challenges compared to the previous models. In particular, the full time evolution of $q_{\lambda}(t)$ can be expressed analytically although it displays some computationally challenging characteristics which involve multiple integrals and Bessel functions. Nevertheless, the solution is expressive enough to derive the asymptotic behaviour of the model in the limit $t \rightarrow +\infty$, including the first order term. As it can be anticipated, our analysis retrieves the phase transition phenomenon previously discussed in this asymptotic limit. However, quite surprisingly at first sight, we also find that the model exhibits a *critical time* at which the overlap reaches another maximum value. The theory remains consistent as this phenomenon is linked to the initial value of $q_{\lambda}(t = 0)$ which

must not be set to zero in order to initiate the gradient-flow dynamics. The tradeoff is that the algorithm starts with a prior information on the overlap, which results in the existence of this critical time. This will be looked in greater detail in Chapter 7.

1.7 Matrix denoising and extensive rank models

An important characteristic of the former model is that the data matrix grows quadratically with n compared to the former models. However, the hidden signal that is being learned is still of order n . In fact, even the linear models that have been described so far have a number of parameters which is always proportional to n , the number of samples. Therefore, a first step towards quadratic number of parameters is to naturally increase the dimension of the spike and consider the learning of a matrix instead of a vector. This is the setting of model 1.5 which is described as follows:

Model 1.5 (Symmetric positive defined matrix denoising). *In this model, we consider a hidden signal matrix $X^* \in \mathbb{R}^{n \times d}$ with independent entries and $X_{ij}^* \sim \mathcal{N}(0, \frac{1}{n})$, and a symmetric Gaussian noise matrix $\xi \in \mathbb{R}^{n \times n}$ with independent entries and $\xi_{ij} \sim \mathcal{N}(0, 1)$. The matrix of observations is generated as follows:*

$$Y = X^* X^{*T} + \frac{1}{\sqrt{\lambda}} \xi \quad (1.37)$$

The estimator $X(t) \in \mathbb{R}^{n \times m}$ is learned by minimizing the following loss function with a free parameter $\mu > 0$ using the gradient flow algorithm with a random initialization with iid matrix elements from $\mathcal{N}(0, \frac{1}{n})$:

$$\mathcal{H}(X) = \frac{1}{4d} \|Y - XX^T\|_F^2 + \frac{\mu}{2d} \|X\|_F^2 \quad (1.38)$$

Compared to the former model 1.4, this is also referred to as an extensive rank model as the dimension of the signal is of a rank m that grows with n with a fixed ratio $m = \psi n$. And similarly with the signal X where $d = \phi n$.

Certain aspects related to overparameterization have already been elucidated using a similar model in (Tarmoun et al., 2021). However an essential aspect that will be treated in Chapter 8 is the study of the gradient-flow algorithm in the high-dimensional limit for a random Gaussian initialization of $X(0)$. In particular, we do not constrain the initial estimator to be aligned with the eigenvectors of Y , resulting effectively solely in the evolution of the n eigenvalues of XX^T while it is not clear a priori how the system would evolve when this is not the case. To measure the performance of the algorithm, we use the matrix mean-square error:

$$\mathcal{E} = \frac{1}{d} \|X^* X^{*T} - XX^T\|_F^2 \quad (1.39)$$

1.8 Organization and main contributions

The thesis is broadly divided into three main parts. The initial part is dedicated to random matrix theory and serves as a general toolbox for different methods and results employed in the other parts. The next two parts can be read independently. Their primary objective is to establish analytic formulas that describe the evolution of different models in the high-dimensional limit. The second part focuses on the general constructions described in models 1.1, 1.2, and 1.3, all falling under the same scope as the Gaussian covariate model. It will also provide a more comprehensive examination of the dynamics of the random feature model. The last part is centered on matrix denoising problems described in the rank one setting in model 1.4 and the extensive rank in model 1.5.

In Chapter 2, we will briefly review standard definitions and results in random matrix theory that will be used throughout the thesis. We formulate common text-books results and methods along with a few typical examples to illustrate their applications. This chapter is not intended to be comprehensive, rather, it aims to provide a brief reference for the reader.

In Chapter 3, we delve deeper into the realm of random matrix theory. Our focus will be on introducing a stronger result that addresses a broader class of random matrices than those discussed in the preceding chapter. This will be achieved through the use of the *linear pencil method* which expresses a fixed point equations in the form of algebraic equations. These results are known and proved rigorously for specific contexts. A main contribution of this chapter is the presentation of three distinct approaches to obtain the same results with greater simplicity, albeit in a non-rigorous manner. In particular, one approach uses the Replica method, a powerful tool in statistical physics, and another is based on Brownian motions. This, in turn, allows us to derive even stronger results compared to existing literature, to the best of our knowledge. In our approach, the expression being calculated may involve deterministic matrices that can be evaluated at a later stage. This will be particularly valuable in Chapter 5. As an additional byproduct of these methods, we also derive a similar results for the characteristic polynomials of random matrices in finite dimensions. Although this particular result won't be directly applied in the remaining chapters of this thesis, it opens up a potential avenue for further investigation into our various models within the context of finite-size settings.

Chapter 4 introduces the second part of this thesis and examines a simplistic setting described in model 1.1. This chapter does not present new contributions. Instead, it will lay the groundwork for the following chapters by describing the spirit of the methods employed for the machine learning models that will be used later, albeit in a simplified context.

Chapter 5 lays the foundation of a general framework for investigating the teacher-student model, as described in model 1.2, but also 1.1 and 1.3 which are specific sub-cases. We present analytical results to track the evolution of both training and generalization errors over time. We provide examples and heat-maps to display the landscape of the learning curves while varying different parameters. The framework can be customized by specifying two general

Chapter 1. Introduction

data structures for both the teacher and the student through the definition of two covariance matrices. This enables to consider a wide range of different models. In this way, we show how an astute choice of structure can be used to construct a contrived learning problem that can effectively yield as many descent structures as desired. These multiple descents can be derived and analyzed precisely asymptotically. Furthermore, we validate our approach through experiments, showing that the analytical formulas provided by the framework even have a certain predictive capability in tracking learning curves over time with real datasets, such as MNIST.

In Chapter 6, we will have a closer look at the random feature model, which can eventually be regarded as a particular instance of the model described in the previous Chapter 5 in accordance with the Gaussian equivalence principle. Note that this chapter can be read independently and doesn't rely on the former results as we derive the learning curves from scratch. As discussed in the previous sections of this introduction, the random feature model has become a model of choice in theoretical machine learning. It can be described as an "embryonic" deep neural network since it features 2 fully connected dense layers with a non-linear activation function while still remaining analytically tractable when the first layer is fixed. This chapter focuses on the time evolution of the model. In the journey, we establish a set of algebraic equations. The solutions can subsequently be computed through contour integrals to calculate these evolutions in the high-dimensional regime. This approach enables the generation of diverse fine-grained heat-maps and curves without conducting any empirical simulations. However, we do include a series of experiments to compare and validate the theoretical findings. One of the challenge of this derivation consists in reducing the number of equations produced by the linear-pencil under consideration to make it numerically computable. This study reveals the presence of distinct structures reminiscent of the double-descent phenomenon - albeit on an epoch-wise basis instead of parameter-wise. Our observations suggest that these structures are associated with the initial conditions on the model, and in particular with the norm of the initial vector β_0 .

Chapter 7 introduces the last part of this thesis with the rank-one matrix estimation problem described in model 1.4. In contrast to the preceding chapters, the random matrix methods employed here are more conventional. However, this model introduces its unique set of challenges. Due to the structure of the mean-square-error and the constraints imposed on the signal, the gradient flow differential equations are non-linear and the approach to derive the time evolution of the model is more involved. We derive the precise asymptotic behavior of the overlap evolution over time in the high-dimensional limit and retrieve the phase transition phenomenon in the limit $t \rightarrow +\infty$. Furthermore, the analytical expressions can be used to derive the first order correction within this limit. This correction enables us to show the existence of a critical time at which the overlap surpasses that achieved in the limit $t \rightarrow +\infty$. This critical time is intricately linked to the initial condition of the model.

Chapter 8 is the last chapter of this thesis, focusing on addressing the matrix denoising problem as described in model 1.5. This model extends the rank-one model by representing

both the hidden signal and the estimator as matrices, rather than vectors. These matrices have finite aspect ratios in the high-dimensional limit and the objective function exhibits infinitely many minima instead of a finite set. However, the constraints on gradient flow method are more relaxed when compared to the rank-one problem, where, as we recall, the hidden signal was confined to the hypersphere. The technical challenge lies in the initial matrix whose eigenvectors can be unaligned with the observation matrix. We derive the full-time evolution of the gradient-flow method using the principles derived in Chapter 3 with the linear-pencil method and show that the evolution remains tractable despite this initial condition.

In conclusion, this thesis aims to provide a general framework to study the time evolution of different models in the high-dimensional limit.

Methods in random matrix theory **Part I**

2 Preliminaries with random matrices

This chapter primarily focuses on revisiting essential properties and notions from random matrix theory, which will be extensively employed throughout the thesis. It is not intended to provide a comprehensive introduction to the field but rather offers a concise introduction to the main concepts and results essential for the subsequent chapters. Random matrix theory is a rich subject, and readers seeking in-depth insights can delve further into many references (Tao, 2012; Potters and Bouchaud, 2020; Mehta, 2004; Benaych-Georges and Knowles, 2016a).

In the first section, we will introduce fundamental concepts of random matrices and review the relation between the *resolvent* of random matrices and the Stieltjes transform of their spectral distribution that has been briefly discussed in the introduction. We will then inspect two classical examples, namely the semicircle law and the Marchenko-Pastur law, and examine their derivations. The next section will be dedicated to some references on the Cauchy integration formula and its application with random matrices. Finally, we will present various tools for performing operations on multiple random matrices and computing their resulting spectral densities. This will motivate the next chapter dedicated to the linear pencil method resulting in fixed-point equations.

2.1 Random matrices and their spectral distribution

Numerous results in random matrix theory are concerned with the spectral distribution of these matrices, and in particular the limiting behavior of these eigenvalues as the matrix size approaches infinity. As it will be the case in this thesis, it is often convenient to work with the Stieltjes transform of the spectral distribution rather than the distribution itself:

Definition 2.1. *The Stieltjes transform, denoted as g_μ , of a probability density function μ is defined for all $z \in \mathbb{C} \setminus \mathbb{R}$, the set of complex numbers with non-zero imaginary part, as follows:*

$$g_\mu(z) = \int_{\mathbb{R}} \frac{\mu(\lambda)d\lambda}{\lambda - z} \tag{2.1}$$

Chapter 2. Preliminaries with random matrices

Note that the integral is well-defined since z has a non-zero imaginary part and $|\frac{1}{\lambda-z}| \leq \frac{1}{|\operatorname{Im} z|} < \infty$. Furthermore, if the support of μ is not the whole set \mathbb{R} , the domain of g_μ can be extended on some subsets of the real line. For instance, in the case of the distribution of the eigenvalues of positive definite matrices, the support is located on \mathbb{R}^+ and thus g_μ is also defined on \mathbb{R}_-^* , the set of all strictly negative real numbers.

Conversely, the original distribution μ yielding the Stieltjes transform g_μ can be retrieved using the following formula for any $\lambda_0 \in \mathbb{R}$:

$$\mu(\lambda_0) = \lim_{\epsilon \rightarrow 0^+} \frac{1}{\pi} \operatorname{Im} g_\mu(\lambda_0 + i\epsilon) \quad (2.2)$$

This result is a consequence of the Poisson kernel $\eta_\epsilon(x) = \frac{1}{\pi} \operatorname{Im} \left(\frac{1}{x-i\epsilon} \right)$, which in the limit $\epsilon \rightarrow 0^+$, is a representation of the Dirac delta function. In particular, we have:

$$\frac{1}{\pi} \operatorname{Im} g_\mu(\lambda_0 + i\epsilon) = \int_{\mathbb{R}} \frac{1}{\pi} \operatorname{Im} \left(\frac{1}{(\lambda - \lambda_0) - i\epsilon} \right) \mu(\lambda) d\lambda = \int_{\mathbb{R}} \eta_\epsilon(\lambda - \lambda_0) \mu(\lambda) d\lambda \quad (2.3)$$

In this thesis, the distributions of interest are those that arise from the spectral density of a matrix $M \in \mathbb{C}^{n \times n}$. An interesting connection emerges with the trace of $(M - zI_n)^{-1}$, which represents the *resolvent* of M :

Definition 2.2. Let M be a self-adjoint matrix of size $n \times n$ with the eigenvalues $\lambda_1, \dots, \lambda_n$. The empirical spectral distribution of M is defined as the probability measure μ_M such that for all $A \in \mathcal{B}(\mathbb{R})$, the Borel sets of \mathbb{R} :

$$\mu_M(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\lambda_i \in A} \quad (2.4)$$

The Stieltjes transform of the eigenvalue distribution of an Hermitian matrix M (or simply the Stieltjes transform of M) is denoted as g_M such that for all $z \in \mathbb{C} \setminus \operatorname{Sp} M$,

$$g_M(z) = \int_{\mathbb{R}} \frac{d\mu_M(\lambda)}{\lambda - z} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i - z} = \frac{1}{n} \operatorname{Tr}[(M - zI_n)^{-1}] \quad (2.5)$$

As we will later observe, this object plays a pivotal role in many calculations in random matrix theory. In this thesis, it will in fact be directly the quantity of interest instead of the spectral distribution itself when expanding the relevant metrics such as the generalization error of our models.

As a side note, another ubiquitous object in linear algebra is the characteristic polynomial $\chi_M(z) = \det(M - zI_n)$, which has a direct connection with g_M through the following derivative:

$$g_M(z) = -\frac{1}{n} \frac{\partial \ln \chi_M(z)}{\partial z} = -\frac{1}{n} \frac{\chi'_M(z)}{\chi_M(z)} \quad (2.6)$$

The key difference is that in χ_M , the eigenvalues of M are the roots of the polynomial, while in

g_M , they are the given as the poles of a rational function. As it will become apparent later in the next chapter, χ_M can prove to be easier to manipulate than g_M in the finite dimensional case as we are dealing with moments of the matrix elements. However, when investigating the asymptotic behavior of sequences of random matrices when $n \rightarrow +\infty$, the trace of the resolvent often proves to be more convenient and well-defined. Hence we define the limiting Stieltjes transform as follows:

Definition 2.3. *Given a sequence of Hermitian random matrices $M^{(n)}$, we define the limiting Stieltjes transform M when it existst as:*

$$g(z) = \text{Tr}_n [(M^{(n)} - zI_n)^{-1}] := \lim_{n \rightarrow +\infty} \frac{1}{n} \mathbb{E} [\text{Tr} [(M^{(n)} - zI_n)^{-1}]] \quad (2.7)$$

In the cases where there is no ambiguity, we will drop the upperscript $M^{(n)}$ and simply write M .

The major subject in random matrix theory and in this thesis is to derive the limiting expression of $g(z)$. A straightforward approach to determine this function is to use the relation with the moments of M , in particular when $|z| > \|M\|_{\text{op}}$ we find:

$$g(z) = \sum_{k=0}^{+\infty} \frac{-1}{z^{k+1}} \text{Tr}_n [M^k] \quad (2.8)$$

Calculating the limiting traces of the moments of M leads to combinatorics methods that often become involved if not impractical when M is a complex expression with multiple random matrices. In the next section, we will see a different method that can be seen as the groundwork for the linear pencil method which will be the subject of the next chapter.

2.2 Semicircular law and Marchenko-Pastur law

2.2.1 Wigner Matrix

In this section, we will first investigate the case of a Wigner matrix, denoted as $M \in \mathbb{R}^{n \times n}$ and such that $M_{ij} = M_{ji} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ with each element (ij) with $i \leq j$ being independently distributed. We will provide a short outline of a proof to derive the semi-circle law that describes the spectral density in the limit $n \rightarrow +\infty$.

We will use the matrix inversion formula, which frequently finds application in the subsequent chapters. To describe it, let $L \in \mathbb{R}^{n \times n}$ be a matrix that can be partitioned into four blocks $A \in \mathbb{R}^{d \times d}$, $D \in \mathbb{R}^{(n-d) \times (n-d)}$, $B \in \mathbb{R}^{d \times (n-d)}$ and $C \in \mathbb{R}^{(n-d) \times d}$. Let's further assume that D is invertible and the Schur complement of D in L , that is $A - BD^{-1}C$, is also invertible. Then:

$$L = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \implies L^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix} \quad (2.9)$$

Chapter 2. Preliminaries with random matrices

Instead, if A is invertible and the Schur complement of A in L is also invertible, then:

$$L^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix} \quad (2.10)$$

For the Wigner matrix case, we will consider $L = M^{(n)} - zI_n$ and fix the dimension $d = 1$. In such case, A is a real number and C is a column vector and $B = C^T$ is a row vector. The structure of D is similar to a realization of $\sqrt{\frac{n-1}{n}}M^{(n-1)} - zI_{n-1}$. By applying the matrix inversion formula (2.9), we obtain a relation on the element (11) of the inverse of L :

$$(L^{-1})_{11} = \left(M_{11}^{(n)} - z - \sum_{i,j=1}^{n-1} C_i (D^{-1})_{ij} C_j \right)^{-1} \quad (2.11)$$

In this scenario, $M_{11}^{(n)}$ concentrates to 0 since it has a variance $\frac{\sigma^2}{n}$. Thus the fluctuations of this term are of order $O(\frac{1}{\sqrt{n}})$. The same goes for the fluctuations of the last term. To see this, let us use the spectral theorem on D with O an orthogonal basis and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{n-1})$ a diagonal matrix such that $D = O^T \Lambda O$. Then:

$$\sum_{i,j=1}^{n-1} C_i (D^{-1})_{ij} C_j = C^T D^{-1} C = (OC)^T \Lambda^{-1} (OC) = \sum_{i=1}^{n-1} \frac{(OC)_i^2}{\lambda_i} \quad (2.12)$$

Notice first that $\frac{1}{|\lambda_i|} \leq \frac{1}{|\text{Im}(z)|}$ for the reasons stated when arguing that g_μ is well-defined. And secondly, we notice that C is a gaussian vector of covariance matrix $\frac{\sigma^2}{n}I_{n-1}$. because O is orthogonal, the distribution of OC is the same as C . So we find:

$$\text{Var}(C^T D^{-1} C) \leq \frac{1}{|\text{Im} z|^2} \sum_{i=1}^{n-1} \text{Var}((OC)_i^2) \leq \frac{2(n-1)\sigma^4}{n^2 |\text{Im} z|^2} = O\left(\frac{1}{n}\right) \quad (2.13)$$

Note that more general concentration results can be derived with other distributions, we refer to (Vershynin, 2018) for instance for more details. The main focus here is that the right-hand side term of equation (2.10) concentrates towards its mean. As we have already mentioned, D has some similar structure as L so when n is large enough:

$$\lim_{n \rightarrow +\infty} \mathbb{E}[C^T D^{-1} C] = \lim_{n \rightarrow +\infty} \frac{\sigma^2}{n} \mathbb{E}[\text{Tr}[D^{-1}]] = \lim_{n \rightarrow +\infty} \sigma^2 \sqrt{\frac{n-1}{n}} \mathbb{E}\left[g_{n-1}\left(z\sqrt{\frac{n}{n-1}}\right) \right] \quad (2.14)$$

Assuming further the existence of a point-wise limit $\lim_{n \rightarrow +\infty} \mathbb{E}[g_n(z)] = g(z)$, we thus expect that $(L^{-1})_{11}$ concentrates towards $(z + \sigma^2 g(z))^{-1}$.

To pursue further with the proof, a crucial remark is that M is rotationally invariant, in the sense that the distribution of the elements of M are the same as those of $S^T M S$ for any orthogonal matrix S ($SS^T = I_n$). This can be seen from the probability distribution and using

the cyclicity of the trace:

$$P(M) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left\{\frac{-n}{4\sigma^2} \text{Tr}[M^2]\right\} \quad (2.15)$$

The same remark applies for L .¹ In other words, all the diagonal elements of L^{-1} shares the same distribution with a mean $(z + \sigma^2 g(z))^{-1}$ and a bounded variance $O(\frac{1}{n})$. To conclude, we can use the relation with the trace of L^{-1} :

$$\frac{1}{n} \sum_i (L^{-1})_{ii} = \frac{1}{n} \text{Tr}[L^{-1}] = g_n(z) \quad (2.16)$$

So that, in the limit of large n , and by averaging on both sides, we can derive the fixed-point equation:

$$g(z) = -\frac{1}{z + \sigma^2 g(z)} \quad (2.17)$$

This expression can be replaced by its quadratic version $\sigma^2 g(z)^2 + zg(z) + 1 = 0$ which is often the form that we will manipulate. In this situation, we can retrieve the spectral density of M by decomposing: $g = \text{Re } g + i \text{Im } g$ and replacing $z = \lambda + i\epsilon$ with $\epsilon = 0$ to find the limit of $\text{Im } g(z + i\epsilon)$ as $\epsilon \rightarrow 0^+$. This yields:

$$\sigma^2 (\text{Re } g + i \text{Im } g)^2 + \lambda (\text{Re } g + i \text{Im } g) + 1 = 0 \quad (2.18)$$

which results in a system of equations:

$$\begin{cases} \sigma^2 ((\text{Re } g)^2 - (\text{Im } g)^2) + \lambda \text{Re } g + 1 = 0 \\ 2\sigma^2 (\text{Re } g)(\text{Im } g) + \lambda \text{Im } g = 0 \end{cases} \quad (2.19)$$

After reducing it further and replacing $\frac{1}{\pi} \text{Im } g = \rho(\lambda)$ we find:

$$\rho(\lambda) (4\sigma^4 \pi^2 \rho(\lambda)^2 + \lambda^2 - 4\sigma^2) = 0 \quad (2.20)$$

thus, we retrieve the semicircle law:

$$\rho(\lambda) = \frac{1}{2\sigma^2\pi} \sqrt{4\sigma^2 - \lambda^2} \mathbf{1}_{\lambda \in [-2\sigma, 2\sigma]} \quad (2.21)$$

As a side note, it is also possible to express $g(z)$ as one of the two solutions of the quadratic equation:

$$g(z) = \frac{-z \pm z \sqrt{1 - \frac{4\sigma^2}{z^2}}}{2\sigma^2} \quad (2.22)$$

A question arises as to which of the two solution selects the correct limiting value of the trace of the resolvent. This is, in fact, a more general problem that will emerge in the next chapters

¹Letting S_i the orthogonal matrix that leaves the canonical basis vectors with e_1, \dots, e_n unchanged except e_1 and e_i with $S_i e_i = e_1$ and $S_i e_1 = e_i$, we find $(S_i L S_i^{-1})_{11}^{-1} = (S_i L^{-1} S_i)_{11} = (L^{-1})_{ii}$. But $S_i L S_i$ has the same distribution as L , so there is nothing particular about the location (11).

for more complicated results. In some cases, we can resort on certain characteristics that are expected from the trace of the resolvent. In particular, we expect that:

$$\lim_{|z| \rightarrow \infty} -zg(z) = \lim_{|z| \rightarrow \infty} \text{Tr}[-z(M - zI_n)^{-1}] = 1 \quad (2.23)$$

Such a condition is only satisfied by the solution with the "plus" sign, which thus gives us the correct form. With this expression, we can now retrieve the limiting moments of the Wigner matrix and recognize the generating functions of the Catalan numbers $C_k = \frac{1}{k+1} \binom{2k}{k}$ with the series expansion for large enough z :

$$g(z) = \frac{z}{2\sigma^2} \left(\sqrt{1 - \frac{4\sigma^2}{z^2}} - 1 \right) = -\frac{1}{z} \sum_{k=0}^{+\infty} C_k \left(\frac{\sigma}{z} \right)^{2k} \quad (2.24)$$

So by identification with (2.8), we find a relation between the k th moment of M and the k th Catalan number which can also be derived directly from combinatorics methods instead:

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \mathbb{E} \text{Tr} [M^{2k}] = C_k \sigma^{2k} \quad \text{and} \quad \lim_{n \rightarrow +\infty} \frac{1}{n} \mathbb{E} \text{Tr} [M^{2k+1}] = 0 \quad (2.25)$$

2.2.2 Marchenko-Pastur law

Another prominent example of large random matrices is the Gram matrix. It corresponds to the matrix XX^T with $X \in \mathbb{R}^{p \times d}$ where the matrix elements of X are gaussian and independently distributed with $X_{ij} \sim \mathcal{N}(0, \frac{1}{p})$. In this setting, the aspect ratio $\frac{p}{d}$ of the matrix is fixed, denoted as ϕ , and in the limit of large p and d , the spectrum of the Gram matrix follows a distribution commonly referred to as the Marchenko-Pastur distribution (Marchenko and Pastur, 1967). We will adopt a similar approach as the Wigner matrix and specify a variance profile for the previously defined matrix M , now with dimension $n = p + d$. The variance profile is defined as follow: consider a deterministic matrix $\sigma_{ij} \in \mathbb{R}^{n \times n}$ with $\frac{1}{n} \sigma_{ij}^2 = \mathbb{E}[M_{ij}^2]$. For each element ij in the area $1 \leq i, j \leq p$ and $p \leq i, j \leq n$, we let $\sigma_{ij} = 0$. In the remaining domain of (ij) , we set $\sigma_{ij}^2 = \frac{n}{p}$. This corresponds to regarding X as the sub-blocks of M within the domain of non-zero variance profile. We can further analyze and compute the inverse of $L = M - zI_n$ and have:

$$L = M - zI_n = \begin{pmatrix} -zI_p & X \\ X^T & -zI_d \end{pmatrix} \implies L^{-1} = \begin{pmatrix} z(-z^2I + XX^T)^{-1} & \frac{1}{z}(-zI + \frac{1}{z}XX^T)^{-1}X \\ \frac{1}{z}X^T(-zI + \frac{1}{z}XX^T)^{-1} & z(-z^2I + X^TX)^{-1} \end{pmatrix} \quad (2.26)$$

Unlike the previous case, the elements of the diagonal of L^{-1} do not share all the same distribution. Rather, only the first p elements have similar distribution. And using the same procedure as before, now with careful considerations on the variance profile with the block-matrix inversion formula and the different concentration assumptions in the limit of large

p, d , with $\phi = \frac{p}{d}$ we find:

$$\lim_{p,d \rightarrow +\infty} \frac{1}{p} \text{Tr} [z(-z^2 I + XX^T)^{-1}] = \frac{1}{-z - \lim_{p,d \rightarrow +\infty} \frac{1}{p} \text{Tr} [z(-z^2 I + X^T X)^{-1}]} \quad (2.27)$$

With $g_{XX^T}(z) = \text{Tr}_p [(XX^T - zI_p)^{-1}]$ and $g_{X^T X}(z) = \text{Tr}_d [(X^T X - zI_d)^{-1}]$, we thus have:

$$z g_{XX^T}(z^2) = \frac{1}{-z - \frac{z}{\phi} g_{X^T X}(z^2)} \quad (2.28)$$

As is known, g_{XX^T} can be related to $g_{X^T X}$ as XX^T and $X^T X$ share the same non-zero eigenvalues and only differ by the multiplicity of the null eigenvalue. Another way to see this is to calculate first:

$$\text{Tr} [(XX^T - zI_p)^{-1} XX^T] = \text{Tr} [(XX^T - zI_p)^{-1} (XX^T - zI_p + zI_p)] \quad (2.29)$$

$$= p + z \text{Tr} [(XX^T - zI_p)^{-1}] \quad (2.30)$$

and secondly, using the push-through identity and the cyclicity of the trace:

$$\text{Tr} [(XX^T - zI_p)^{-1} XX^T] = \text{Tr} [X(X^T X - zI_d)^{-1} X^T] \quad (2.31)$$

$$= \text{Tr} [(X^T X - zI_d)^{-1} X^T X] \quad (2.32)$$

$$= d + z \text{Tr} [(X^T X - zI_d)^{-1}] \quad (2.33)$$

So we find the expected additional pole at $z = 0$:

$$\text{Tr} [(X^T X - zI_d)^{-1}] = \frac{p-d}{z} + \text{Tr} [(XX^T - zI_p)^{-1}] \quad (2.34)$$

In the high-dimensional limit, this corresponds to $g_{X^T X}(z) = \frac{\phi-1}{z} + \phi g_{XX^T}(z)$. Consequently, we have the closed-form equation:

$$z g_{XX^T}(z^2) \left(z + z \left(\frac{1-\frac{1}{\phi}}{z^2} + g_{XX^T}(z^2) \right) \right) + 1 = 0 \quad (2.35)$$

Hence we retrieve the Stieltjes transform of the celebrated Marchenko-Pastur distribution (Marchenko and Pastur, 1967) albeit evaluated in z^2 :

$$z^2 g_{XX^T}(z^2)^2 + \left(1 + z^2 - \frac{1}{\phi} \right) g_{XX^T}(z^2) + 1 = 0 \quad (2.36)$$

In the next chapter, we will generalize this approach and formulate a general result that can be applied in a straightforward way.

2.3 Holomorphic functional calculus

In the upcoming section, we will examine matrices that may not necessarily be in the form of a resolvent. They can take various forms, such as terms like $e^{-tX^T X}$ where the variable t represents time, as will be discussed in Chapter 4. In such cases, it is convenient to use the Cauchy integration formula such that for any holomorphic function $f : \mathbb{C} \rightarrow \mathbb{C}$, $\lambda \in \mathbb{C}$ and any contour Γ enclosing λ :

$$f(\lambda) = \frac{-1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{\lambda - z} dz \quad (2.37)$$

In particular, when consider a symmetric matrix, denoted as M , which is expressed using the spectral theorem in the form $M = \sum_{i=1}^n \lambda_i u_i u_i^T$ with $u_i \in \mathbb{R}^n$ the normalized eigenvectors and $\lambda_i \in \mathbb{R}$ the eigenvalues of M , we have for any contour Γ containing all the eigenvalues:

$$f(M) = \sum_{i=1}^n f(\lambda_i) u_i u_i^T = \sum_{i=1}^n \frac{-1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{\lambda_i - z} u_i u_i^T dz = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) (M - zI_n)^{-1} dz \quad (2.38)$$

This is developed in greater extent in Dunford and Schwartz (1988). This, in turn, enables the computation of more general traces involving the random matrix M of the previous section with its associated Stieltjes transform $g(z)$:

$$\text{Tr}_n [f(M)] = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) \text{Tr}_n [(M - zI_n)^{-1}] dz = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) g(z) dz \quad (2.39)$$

This time, because we are in the limit of large dimensions, it becomes necessary for Γ to enclose all the branch-cuts of $g(z)$ corresponding to the support of the spectral density ρ defined before, rather than a set of individual eigenvalues. Note that in this scenario, it is always possible to use the other formula $\text{Tr}_n [f(M)] = \int_{\mathbb{R}} f(\lambda) \rho(\lambda) d\lambda$ that uses the spectral density ρ instead of g . The connection can be demonstrated by choosing a contour Γ that closely enlases the branch-cuts produced by the spectral support. For instance, let's assume the support is $(-2, 2)$ (which is the case for the Wigner matrix) and consider Γ representing the perimeter of the rectangle extending from the point at the bottom-left corner $-2 - \delta - i\epsilon$ to the point at the top-right corner $2 + \delta + i\epsilon$, with $\delta > 0$ and $\epsilon > 0$. Thus:

$$\text{Tr}_n [f(M)] = \frac{-1}{2\pi i} \left(\int_{-2-\delta-i\epsilon}^{2+\delta-i\epsilon} + \int_{2+\delta-i\epsilon}^{2+\delta+i\epsilon} + \int_{2+\delta+i\epsilon}^{-2-\delta+i\epsilon} + \int_{-2-\delta+i\epsilon}^{-2-\delta-i\epsilon} \right) f(z) g(z) dz \quad (2.40)$$

$$= \frac{-1}{2\pi i} \left(\int_{-2-\delta-i\epsilon}^{2+\delta-i\epsilon} - \int_{-2-\delta+i\epsilon}^{2+\delta+i\epsilon} + \int_{2+\delta-i\epsilon}^{2+\delta+i\epsilon} + \int_{2+\delta+i\epsilon}^{-2-\delta-i\epsilon} \right) f(z) g(z) dz \quad (2.41)$$

and:

$$\frac{-1}{2\pi i} \left(\int_{-2-\delta-i\epsilon}^{2+\delta-i\epsilon} - \int_{-2-\delta+i\epsilon}^{2+\delta+i\epsilon} \right) f(z) g(z) dz = \int_{-2-\delta}^{2+\delta} \frac{+1}{2\pi i} (f(\lambda + i\epsilon) g(\lambda + i\epsilon) - f(\lambda - i\epsilon) g(\lambda - i\epsilon)) d\lambda \quad (2.42)$$

So in the limit $\epsilon \rightarrow 0^+$, with the continuity of f and the definition of $g(z)$ we find:

$$\lim_{\epsilon \rightarrow 0^+} (f(\lambda + i\epsilon)g(\lambda + i\epsilon) - f(\lambda - i\epsilon)g(\lambda - i\epsilon)) = 2if(\lambda) \lim_{\epsilon \rightarrow 0^+} \text{Im}(g(\lambda + i\epsilon)) \quad (2.43)$$

So in fact:

$$\lim_{\epsilon, \delta \rightarrow 0^+} \frac{-1}{2\pi i} \left(\int_{-2-\delta-i\epsilon}^{2+\delta-i\epsilon} - \int_{-2-\delta+i\epsilon}^{2+\delta+i\epsilon} \right) f(z)g(z)dz = \int_{-2}^2 f(\lambda) \lim_{\epsilon \rightarrow 0^+} \frac{1}{\pi} \text{Im} g(\lambda + i\epsilon) d\lambda = \int_{-2}^2 f(\lambda)\rho(\lambda)d\lambda \quad (2.44)$$

As the other integral terms vanishes when $\delta \rightarrow 0^+$ we find exactly the result that we would obtain by using the distribution ρ instead of g :

$$\text{Tr}_n [f(M)] = \int_{\mathbb{R}} f(\lambda)\rho(\lambda)d\lambda \quad (2.45)$$

Arguably, this last expression may, in some cases, be easier to handle when running numerical calculations as the path of the integral is on the real line. However, some situations that we will encounter are provided with more complex expressions. For instance, we will have to tackle the form $Rf(M)\Omega f(M)R$ with R and Ω two matrices having potential element-wise dependencies with M . For some contour Γ_x and Γ_y , we will extend the former development to decouple the calculations involving random matrices from the application of f with:

$$\text{Tr}_n [Rf(M)\Omega f(M)R] = \frac{-1}{4\pi^2} \oint_{\Gamma_x \times \Gamma_y} f(x)f(y)\text{Tr}_n [R(M-xI)^{-1}\Omega(M-yI)^{-1}R] dx dy \quad (2.46)$$

The precise derivation of this trace in the limit $n \rightarrow +\infty$ inside the double integral will be the main subject of different chapters presented in this thesis.

2.4 Algebraic expressions of random matrices

As stated on the previous formula, while the decoupling of the application of f is applied, there remains a general algebraic expression involving the matrices M, R, Ω . A classical problem arising in random matrix theory is the addition of two random matrices $A, B \in \mathbb{R}^{n \times n}$, so say $g_{A+B}(z) = \text{Tr}_n [(A+B-zI_n)^{-1}]$. A common approach is to use the additivity of the \mathcal{R} -transform, a function that will be further investigated in the next chapter. This transform is applicable with some conditions on A and B , we refer the reader to the free-probability section of the books mentioned in the introduction for more details on the subject. As an example, when A is deterministic and B is a wigner matrix, we can derive the following result:

$$\mathcal{R}_{A+B}(g) = \mathcal{R}_A(g) + \mathcal{R}_B(g) \quad (2.47)$$

where we define $\mathcal{R}_A(g)$ the function such that:

$$\mathcal{R}_A(g_A(z)) = z + \frac{1}{g_A(z)} \quad (2.48)$$

Chapter 2. Preliminaries with random matrices

A similar approach exists with the \mathcal{S} -transform:

$$\mathcal{S}_{AB}(g) = \mathcal{S}_A(g)\mathcal{S}_B(g) \quad (2.49)$$

where we define $\mathcal{S}_A(g)$ the function such that:

$$\mathcal{S}_A(-zg_A(z) - 1) = \frac{g_A(z)}{zg_A(z) + 1} \quad (2.50)$$

Both the \mathcal{R} and \mathcal{S} transforms are powerful tools but the calculations can be involved when dealing with a complex expressions. Another idea that will be explored in the next chapter is to extend the former construction with the variance profile to yield the expressions of interest. For instance, consider the following block-matrix and its inverse:

$$M = \begin{pmatrix} -zI & A \\ B & -I \end{pmatrix} \implies M^{-1} = \begin{pmatrix} (AB - zI)^{-1} & (AB - zI)^{-1}A \\ -B(AB - zI)^{-1} & I + B(AB - zI)^{-1}A \end{pmatrix} \quad (2.51)$$

so clearly, the partial trace of M^{-1} in the first block is precisely the trace of the resolvent of the product of A and B . This is the main concept behind the linearization method that will be presented in the next chapter.

3 The Linear-pencil method

In this chapter, we extend our focus beyond the Wigner case, which deals with a single symmetric random matrix with independent Gaussian entries, and the Wishart matrix, resulting from the multiplication two random matrices. We will introduce the linear pencil method, an approach that enables the computation of traces for more general rational expressions involving multiple independent standard symmetric or non-symmetric random matrices. Albeit not required for simple enough models as the ones described in model 1.1 and 1.4 (and further investigated in chapters 4 and 7), it will be of great use in the others and will be employed extensively in chapters 5, 6 and 8.

This method yields a fixed-point equation, the solutions of which are associated with the partial trace of the inverse of a large random block matrix. Some of these blocks can remain partially deterministic such that these traces can be calculated at a later stage. We will show various sketches of proof to derive this fixed-point equation.

3.1 Introduction

As we have seen in previous Chapter 2, instead of calculating directly the distribution of the spectral density $\rho(\lambda)$ of a large random matrix A , the Stieltjes transform of the spectrum $g(z) = \int_{\lambda} \frac{\rho(\lambda) dz}{\lambda - z}$ is usually the object that yields the most practical formulations and is directly related to the trace of the Resolvent $(A - zI)^{-1}$. The spectral density can be recovered from the Stieltjes transform using the inverse formula $\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \rightarrow 0^+} \text{Im } g(\lambda + i\epsilon)$ and both objects can have analytical expressions in high-dimension depending on the matrix A into consideration. Here we are interested in the situation where we deal with traces of intricate algebraic expressions involving one or many large random matrices. A simple common example is the Marchenko-Pastur distribution derived from $A = XX^T$ where X is a gaussian random matrix. With more intricate algebraic expressions, finding such analytical formulas remain cumbersome even with the help of different methods such as the R-transform and S-transform for adding and multiplying random matrices respectively (see for instance Potters and Bouchaud (2020)). To circumvent this issue, another method using so-called *linear pencil* matrices was

developed first in (Rashidi Far et al., 2006) and since then, has been refined and described as a linearization trick and analyzed rigorously using free probability with operator-valued convolutions in (Mingo and Speicher, 2017) and (Helton et al., 2018). Afterwards, this technique has been further employed successfully in the machine learning community and has been gaining more attraction, for instance in (Adlam and Pennington, 2020a) or (Bodin and Macris, 2021a) for the asymptotic behavior of the training and generalization error in simple machine learning models. New proofs avoiding free-convolutions have been proposed such as perturbative methods in (Cui et al., 2020). In this chapter, we propose to revisit again the linear pencil construction with the help of stochastic calculus and Dyson's brownian motion (Dyson, 1962). We show that based on the intrinsic relation between the determinant from the heat-equation and the trace from the Burger equation (with Hopf-Cole transform already mentioned in Kardar et al. (1986)). We further derive some common results for complex Wigner matrices as well as complex Wishart matrices with the Marchenko pastur distribution (Marchenko and Pastur, 1967). In addition, there is a growing interest in results for finite-dimensional matrices, as demonstrated in (Marcus et al., 2022). We show that our method also applies to some extent to the finite-dimensional case using the same linearization technique.

3.1.1 Preliminaries

For the sake of simplicity, we will develop our results in the Gaussian Unitary Ensemble (GUE). We define $\mathcal{F}_{n,d}$ the set of complex random matrices of size $n \times d$ such that the real part and imaginary part of the entries are all independent standard gaussian distribution. Similarly, in the same spirit, we let \mathcal{F}_n^S be the set of self-adjoint random matrices of size $n \times n$. We will see the definitions of such ensembles later in more detail in Section 3.5.1.

Let us consider an invertible self-adjoint block matrix L with $N \times N$ blocks and denote $L^{(ij)}$ the block (ij) of size $N_i \times N_j$. One can write $L = \sum_{ij} E_{ij} \otimes L^{(ij)}$ where¹ $E_{ij} = e_i e_j^T$ are the basis matrices of $\mathbb{C}^{N \times N}$. We will assume that the size of the blocks of L can be deterministically increased with a linear relation with respect to some parameter n , that is to say for each i there exists some fixed coefficient γ_i such that $N_i(n) = \gamma_i n$. We say that L is a linear-pencil if each block $L^{(ij)}$ is the sum of a deterministic matrix $L_0^{(ij)}$ and a random matrix $W^{(ij)}$ such that $W^{(ij)}$ can be described as a linear combination with real coefficients of some elements of \mathcal{F}_{N_i, N_j} , and $\mathcal{F}_{N_i}^S$ (when $N_i = N_j$) with coefficients proportional to $\frac{1}{\sqrt{n}}$. This means in particular that every element of the block (ij) is normally distributed with a variance proportional to $\frac{1}{n}$. We will define more generally $L_0 = \mathbb{E}L$ and the covariance structure such that for $\mathbb{S} = \{ij | N_i = N_j\}$ is the set of indices indexing the square blocks:

$$\forall (il, jk) \in \mathbb{S}^2, (\sigma_{ij}^{kl})^2 = \frac{1}{\gamma_i \gamma_j} \left\| \mathbb{E} W^{(ij)} \odot (W^{(kl)})^T \right\|_F^2 \quad (3.1)$$

Where $A \odot B$ denotes the Hadamard product of two matrices such that $(A \odot B)_{ij} = A_{ij} B_{ij}$ and $\|\cdot\|_F$ is the Frobenius norm. When $N_i \neq N_l$ or $N_j \neq N_k$, we set $\sigma_{ij}^{kl} = 0$. Another equivalent

¹with e_1, \dots, e_N is the canonical basis of \mathbb{C}^N

definition of σ is as follows: given $(il, jk) \in \mathbb{S}^2$ and (uv) with $1 \leq u \leq N_i$ and $1 \leq v \leq N_j$:

$$\sigma_{ij}^{kl} = n \cdot \text{Cov}(L_{uv}^{(ij)}, L_{vu}^{(kl)}) = n \cdot \mathbb{E} \left[W_{uv}^{(ij)} W_{vu}^{(kl)} \right] \quad (3.2)$$

In the following, we will use the linear operator $\eta_L : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}^{N \times N}$ as the map acting on a matrix g such that

$$[\eta_L(g)]_{il} = \sum_{jk \in \mathbb{S}} \sigma_{ij}^{kl} g_{jk} \quad (3.3)$$

For technical reasons that will become clear later in the following sections, we will assume that this linear operator restricted on its image is invertible. As an example, this is the case for self-adjoint linear pencils when for each element $(jk) \in \mathbb{S}$, there exists a unique $(il) \in \mathbb{S}$ such that $\sigma_{ij}^{kl} \neq 0$. In this case, we find that $\eta_L(E_{jk}) = \sigma_{ij}^{kl} E_{il}$. By construction, this also implies $\eta_L(E_{il}) = \sigma_{ji}^{lk} E_{jk} = \sigma_{ij}^{kl} E_{jk}$, so there is a one-to-one mapping between each elements of the canonical basis of $\mathbb{C}^{N \times N}$ on the image of η_L , so η_L is invertible. In this scenario, for any non-zero random block within the 'block-row' j at a given location (jk) , there is precisely one non-zero random block within the 'block-column' k with element-wise dependencies on the other.

Finally, we define the operator $(\eta_L \otimes I)(\cdot)$ such that:

$$\forall (ij), [(\eta_L \otimes I)(g)]^{(ij)} = \begin{cases} [\eta_L(g)]_{ij} I_{N_i} & \text{if } (ij) \in \mathbb{S} \\ O_{N_i, N_j} & \text{otherwise} \end{cases} \quad (3.4)$$

where I_{N_i} is the identity matrix of size $N_i \times N_i$, and O_{N_i, N_j} the all-zero matrix of size $N_i \times N_j$. Similarly, with J the $N \times N$ the matrix $J_{il} = \delta_{\mathbb{S}}(il)$, we define the partial-trace operator $(J \otimes \text{Tr})[\cdot]$ such that given a $N \times N$ block matrix G with a structure similar to L above ($G = \sum_{ij} E_{ij} \otimes G^{(ij)}$) with:

$$\forall (ij), [(J \otimes \text{Tr})[G]]_{ij} = \begin{cases} \text{Tr}[G^{(ij)}] & \text{if } (ij) \in \mathbb{S} \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

As an example, let's assume $N = 2$ with $N_1 \neq N_2$. Then $\mathbb{S} = \{(00), (11)\}$ and $J = I_2$ the identity matrix, and we have a linear-pencil:

$$L = E_{11} \otimes L^{(11)} + E_{12} \otimes L^{(12)} + E_{21} \otimes L^{(21)} + E_{22} \otimes L^{(22)} \quad (3.6)$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes L^{(11)} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \otimes L^{(12)} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \otimes L^{(21)} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes L^{(22)} \quad (3.7)$$

$$= \begin{pmatrix} L^{(11)} & L^{(12)} \\ L^{(21)} & L^{(22)} \end{pmatrix} \quad (3.8)$$

Similarly for G :

$$G = \begin{pmatrix} G^{(11)} & G^{(12)} \\ G^{(21)} & G^{(22)} \end{pmatrix} \quad (3.9)$$

and we have the partial trace operator:

$$(J \otimes \text{Tr}) [G] = \begin{pmatrix} \text{Tr} [G^{(11)}] & \mathbf{0} \\ \mathbf{0} & \text{Tr} [G^{(22)}] \end{pmatrix} \quad (3.10)$$

and the operator η_L operates on a given matrix g as:

$$\eta_L(g) = \begin{pmatrix} \sigma_{11}^{11} g_{11} + \sigma_{12}^{21} g_{22} & \mathbf{0} \\ \mathbf{0} & \sigma_{21}^{12} g_{11} + \sigma_{22}^{22} g_{22} \end{pmatrix} \quad (3.11)$$

finally, we can apply the operator $\eta_L \otimes I$ on g as:

$$(\eta_L \otimes I)(g) = \begin{pmatrix} (\sigma_{11}^{11} g_{11} + \sigma_{12}^{21} g_{22}) I_{N_1} & O_{N_1, N_2} \\ O_{N_2, N_1} & (\sigma_{21}^{12} g_{11} + \sigma_{22}^{22} g_{22}) I_{N_2} \end{pmatrix} \quad (3.12)$$

3.1.2 Main Statement

The main statement of this chapter holds in the high-dimensional limit $n \rightarrow +\infty$. We define the operator $\text{Tr}_n [\cdot]$ such that for a sequence of random matrices $A(n)$, if the limit exists we define

$$\text{Tr}_n [A(n)] := \lim_{n \rightarrow +\infty} \frac{1}{n} \mathbb{E} \text{Tr} [A(n)] \quad (3.13)$$

In order to simplify the notations, we will discard the reference to n in the sequence $A(n)$ and simply write A . Our first result holds in the asymptotic limit $n \rightarrow +\infty$:

Result 3.1. *Under the assumption of the existence in the limit $n \rightarrow +\infty$ of a deterministic matrix $g \in \mathbb{C}^{N \times N}$ with g_{ij} such that:*

$$g_{ij} = \text{Tr}_n [(L^{-1})^{(ij)}] \quad (3.14)$$

we claim that:

$$g = (J \otimes \text{Tr}_n) [(L_0 - (\eta_L \otimes I)(g))^{-1}] \quad (3.15)$$

In particular, when $L_0 = Z$ where $Z = \mathcal{Z} \otimes I$ is such that its sub-blocks are only scalar matrices (ie, proportional to the identity), that is where $Z^{(ij)} = \mathcal{Z}_{ij} I_{N_i}$ for $(ij) \in \mathbb{S}$ and $Z^{(ij)} = O_{N_i, N_j}$ otherwise, we find:

$$g_{ij} = \gamma_i [(\mathcal{Z} - \eta_L(g))^{-1}]_{ij} \quad (3.16)$$

Upon a rescaling of the elements of g in \tilde{g} with $\tilde{g}_{ij} = \frac{1}{\gamma_i} g_{ij}$ we have the equations:

$$\tilde{g}(\mathcal{Z} - \eta_L(g)) = (\mathcal{Z} - \eta_L(g)) \tilde{g} = I \quad (3.17)$$

These equations are only polynomial in the matrix elements of g , and can thus be further solved or reduced using techniques from algebraic geometry (for instance computing Gröbner basis with Buchberger algorithm Buchberger (1965)).

Next, we present a result in finite dimension n that enables to calculate the average character-

istic polynomial for any linear pencil. It is worth noting that this result will not be employed in the subsequent chapters, instead, it arises as a by-product of our analysis.

Result 3.2. *In finite dimension n , the average determinant of L is given by the formula*

$$\mathbb{E}_W[\det(L)] = \mathbb{E}_{\mathcal{Z}}[\det(L_0 + \mathcal{Z} \otimes I)] \quad (3.18)$$

such that for all $(il) \in \mathbb{S}$, \mathcal{Z}_{il} has a complex normal distribution and for all $(il), (jk) \in \mathbb{S}^2$ we have:

$$\text{Cov}(\mathcal{Z}_{il}, \mathcal{Z}_{kj}) = -\frac{1}{n}[VV^T]_{il,kj} = -\frac{1}{n}\sigma_{ij}^{kl} \quad (3.19)$$

$$\text{Cov}(\mathcal{Z}_{il}, \tilde{\mathcal{Z}}_{kj}) = \frac{1}{n}[V\tilde{V}]_{il,kj} \quad (3.20)$$

for any decomposition of σ such that $\sigma_{ij}^{kl} = \sum_{pq} V_{il,pq} V_{kj,pq}$

Remark: To see that σ admits such a decomposition, one can proceed as follows. First consider the "matrixisation" of σ with $(\Sigma_{il,kj}) = (\sigma_{ij}^{kl} | N_i = N_l, N_k = N_j) \in \mathbb{C}^{|\mathbb{S}| \times |\mathbb{S}|}$. Notice now that $\Sigma_{il,kj} = \sigma_{ij}^{kl} = \sigma_{ji}^{lk} = \Sigma_{kj,il}$, so $\Sigma = \Sigma^T$ and Σ is real, so the spectral theorem provides a decomposition of the form $\sigma = O^T D O$ with O an orthonormal matrix and D a real diagonal matrix. Thus, with $V = O^T D^{\frac{1}{2}} O$, Σ admits a decomposition of the form $\Sigma = VV^T$ (note that V can be a complex matrix because D can have negative values).

3.2 Main Example: high-dimensional case

Let $X \in \mathbb{C}^{n \times d}$ and $Y \in \mathbb{C}^{n \times n}$ be two complex Gaussian random matrices with independent entries, and Y a self-adjoint matrix. The variance is set to be $\frac{1}{n}$ for all entries in X and Y . We take a quenched self-adjoint matrix $C \in \mathbb{C}^{d \times d}$ and we want to calculate the eigenvalue distribution of the matrix $M = X C X^T + \lambda Y$. On the one hand, when $C = 0$ and $\lambda = 1$, it reduces to the complex Wigner matrix Y . On the other hand, when $C = I$ and $\lambda = 0$, it becomes the Wishart matrix XX^T . When C is left undefined, this is the sample covariance matrix problem. The conjunction of $\lambda \neq 0$ and $C \neq 0$ features an example of a linear pencil problem with two kinds of operations (sums and products) acting on random matrices with an additional quenched matrix. We will use this matrix M as a canonical example of the linear pencil method.

In the first subsection 3.2.1, we will construct a linear pencil that will allow us to calculate the Stieltjes transform of the eigenvalue distribution of M which we define as $g_M(z) = \text{Tr}_n[(M - zI)^{-1}]$. In the second subsection 3.2.2 we will use the linear pencil to calculate the effect of the random matrices X and Y .

3.2.1 Constructing a linear pencil

The goal is to determine the eigenvalue distribution of the matrix M using the resolvent $U_0 = (XCX^T + \lambda Y - zI)^{-1}$. We linearize the problem by introducing auxiliary matrices such that the U_0 is given by a sub-block of a larger block-matrix construction.

In this particular case, we notice that $(\lambda Y + XCX^T - zI)U_0 = I$. No inverse are remaining in the former expression. Let's introduce $U_1 = X^T U_0$. We have $XCX^T U_0 + (\lambda Y - zI)U_0 = I$. Then let's introduce $U_2 = CU_1$. We now have $XU_2 + (\lambda Y - zI)U_0 = I$. At this point, we have the linearized system of block-matrices:

$$\underbrace{\begin{pmatrix} \lambda Y - zI & 0 & X \\ 0 & C & I \\ X^T & I & 0 \end{pmatrix}}_L \begin{pmatrix} U_0 \\ -U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix} \quad (3.21)$$

Therefore taking the inverse of L yields the solution

$$U_0 = \begin{pmatrix} I & 0 & 0 \end{pmatrix} \begin{pmatrix} U_0 \\ -U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda Y - zI & 0 & X \\ 0 & C & I \\ X^T & I & 0 \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix} = (L^{-1})^{(11)} \quad (3.22)$$

At this point, we have constructed a linear pencil L . Although the linearization method outlined here is more or less systematic, it is important to realize that these constructions are not unique and other linear pencils can be proposed. However, some may be more or less convenient to use depending on the problem at hand. For instance, the construction that has been displayed here presents some symmetries that can be easier to use for computing the inverses in the next section.

3.2.2 Calculating the interaction of the random matrices

We find $\sigma_{11}^{11} = \lambda^2$ due to Y being self-adjoint and $\sigma_{13}^{31} = \sigma_{31}^{13} = 1$ with X . We find:

$$\eta_L(g) = \begin{pmatrix} \lambda^2 g_{11} + g_{33} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & g_{11} \end{pmatrix} \quad (3.23)$$

Thus we are left to calculate:

$$(L_0 - (\eta_L \otimes I)(g))^{-1} = \begin{pmatrix} -(\lambda^2 g_{11} + g_{33} + z)I_n & 0 & 0 \\ 0 & C & I_d \\ 0 & I_d & -g_{11}I_d \end{pmatrix}^{-1} \quad (3.24)$$

$$= \left(\begin{array}{cc|cc} -(\lambda^2 g_{11} + g_{33} + z)^{-1}I_n & & 0 & 0 \\ & & C & I_d \\ \hline 0 & & I_d & -g_{11}I_d \end{array} \right)^{-1} \quad (3.25)$$

Using the block-matrix inversion formula:

$$\begin{pmatrix} C & I_d \\ I_d & -g_{11}I_d \end{pmatrix}^{-1} = \begin{pmatrix} (C + g_{11}^{-1})^{-1} & (g_{11}C + I_d)^{-1} \\ (g_{11}C + I_d)^{-1} & -g_{11}^{-1}(I - g_{11}^{-1}(C + g_{11}^{-1})^{-1}) \end{pmatrix} \quad (3.26)$$

finally, with $\phi = \frac{n}{d}$ and the fixed point equation of result 3.1, we find the following closed system of equations where $g_C(z) = \text{Tr}_d[(C - zI_d)^{-1}]$ is the Stieltjes transform of C :

$$g_{11} = -(\lambda^2 g_{11} + g_{33} + z)^{-1} \quad (3.27)$$

$$g_{33} = -g_{11}^{-1} \text{Tr}_n [I_d - (g_{11}C + I_d)^{-1}] = \frac{1}{\phi} \frac{1}{g_{11}} \left(\frac{1}{g_{11}} g_C \left(-\frac{1}{g_{11}} \right) - 1 \right) \quad (3.28)$$

With $g_M = g_{11}$, this can be reduced to a single equation for g_M :

$$\lambda^2 g_M^2(z) + \frac{1}{\phi} \left(\frac{1}{g_M(z)} g_C \left(-\frac{1}{g_M(z)} \right) - 1 \right) + z g_M(z) + 1 = 0 \quad (3.29)$$

3.2.3 Special cases

$\lambda = 1$ and $C = 0$: The Stieltjes transform of C gives $g_C(z) = -\frac{1}{z}$ and the equation (3.29) reduces as expected to the well-known quadratic equation for the Stieltjes transform of the Wigner matrix Y :

$$g_M^2(z) + z g_M(z) + 1 = 0 \quad (3.30)$$

$\lambda = 0$ and $C = I$: The Stieltjes transform of C is $g_C(z) = \frac{1}{1-z}$ and the equation (3.29) reduces to:

$$\frac{1}{\phi} \left(\frac{1}{g_M(z)} \frac{1}{1 + \frac{1}{g_M(z)}} - 1 \right) + z g_M(z) + 1 = 0 \quad (3.31)$$

After rearranging the terms, we find the celebrated Marchenko-Pastur equation:

$$z g_M^2(z) + \left(1 + z - \frac{1}{\phi} \right) g_M(z) + 1 = 0 \quad (3.32)$$

$\lambda = 0$ and C is quenched: The Stieltjes transform of C is $g_C(z) = \text{Tr}_d[(C - zI_d)^{-1}]$ and the equation (3.29) reduces to:

$$\frac{-1}{g_M(z)} g_C \left(\frac{-1}{g_M(z)} \right) = z \left(\phi g_M(z) + \frac{\phi - 1}{z} \right) \quad (3.33)$$

Note that if we define $E = C^{\frac{1}{2}} X^T X C^{\frac{1}{2}} \in \mathbb{C}^{d \times d}$ with $g_E(z) = \text{Tr}_d[(E - zI_d)^{-1}]$, we find the relation $g_E(z) = \phi g_M(z) + \frac{\phi - 1}{z}$ and defining $z_1 = \frac{-1}{g_M(z_0)}$, we find $z_1 g_C(z_1) = z_0 g_E(z_0)$ as stated for instance in equation (17.5) in (Potters and Bouchaud, 2020)

3.3 Application with finite size

3.3.1 Wigner Matrix and Hermite polynomials

With a standard 1×1 linear pencil $L = C + W - zI_d$, with C a deterministic matrix and W a Wigner matrix, using the formula we find in result 3.2:

$$\mathbb{E}_W \det(L) = \mathbb{E}_{u \sim \mathcal{N}(0, \frac{1}{d})} [\det(C + (iu - z)I_d)] = \mathbb{E}_{u \sim \mathcal{N}(0, \frac{1}{d})} [\mathcal{X}_C(z - iu)] \quad (3.34)$$

In particular, when C is the all-zero matrix, this is $\mathbb{E}_W \det(L) = \mathbb{E}_u [(iu - z)^d]$ and it is easy to see that it is related to the Hermite polynomials. Indeed, with $h_d(z) = \mathbb{E}_{s \sim \mathcal{N}(0,1)} [(is + z)^d]$, we have using integration by part:

$$h_{d+1}(z) = \mathbb{E}_s [(is + z)(is + z)^d] \quad (3.35)$$

$$= i \mathbb{E}_s [u(is + z)^d] - zh_d(z) \quad (3.36)$$

$$= -dh_{d-1}(z) + zh_d(z) \quad (3.37)$$

which is exactly the recurrence relation of the Hermite polynomials (with $h_0(z) = 1$ and $h_1(z) = z$). Finally we have find the well-kown relation:

$$\mathbb{E}_W \det(W - zI_d) = \mathbb{E}_s \left[\left(i \frac{1}{\sqrt{d}} s - z \right)^d \right] = \frac{1}{\sqrt{d}^d} h_d(-\sqrt{d}z) \quad (3.38)$$

3.3.2 Wishart Matrix and Laguerre polynomials

For this case we consider the example given in (3.21) in finite-dimension but we will only consider the situation where $\lambda = 0$ for the general case is too cumbersome to be written down explicitly here. In this scenario, we define the matrix Σ with the indices in $\{11, 33\}^2$ such that:

$$\Sigma = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (3.39)$$

So one can calculate a (non-unique) potential root V for $\Sigma = VV^T$:

$$V = \frac{1}{\sqrt{2}} \begin{pmatrix} -i & 1 \\ i & 1 \end{pmatrix} \quad (3.40)$$

so with $\mathcal{Z} = iVU$ where $U \sim \mathcal{N}(0, \frac{1}{n}I_2)$ we find:

$$\mathcal{Z}_{11} = \frac{1}{2}(U_1 + iU_2) \quad \text{and} \quad \mathcal{Z}_{33} = -\frac{1}{2}(U_1 - iU_2) \quad (3.41)$$

Next:

$$\det(\bar{L}_0 + \mathcal{Z} \otimes I) = \det((\mathcal{Z}_{11} - z)I_n) \det(\mathcal{Z}_{33}C - I_d) \quad (3.42)$$

so using the block matrix determinant formula:

$$\mathbb{E}_X \det(L) = (-1)^{d+1} \mathbb{E}_X \det(XCX^T - zI_n) \quad (3.43)$$

$$= (-1)^d \mathbb{E}_U \left[\left(\frac{1}{2}(U_1 + iU_2) - z \right)^n \det \left(\frac{1}{2}(U_1 - iU_2)C + 1 \right) \right] \quad (3.44)$$

In particular, when $C = I_d$ we find the formula:

$$\mathbb{E}_X \det(XX^T - zI_n) = -\mathbb{E}_U \left[\left(\frac{1}{2}(U_1 + iU_2) - z \right)^n \left(\frac{1}{2}(U_1 - iU_2) + 1 \right)^d \right] \quad (3.45)$$

Note that this expression can be precisely calculated using a Gauss-Hermite quadrature as this is the expectation of a complex polynomial of degree $d + n$ in $\mathbb{C}[U_1, U_2]$. However, for this particular case, this expression is related to the generalized Laguerre polynomials $\mathcal{L}_k^{(\alpha)}$ as mentioned in (Potters and Bouchaud, 2020). We refer the reader to the appendix 3.A for the details of the calculation. The final result is:

$$\mathbb{E}_X \det(XX^T - zI_n) = \begin{cases} -\frac{n!}{n^n} \mathcal{L}_n^{(d-n)}(nz) & \text{if } n \leq d \\ -\frac{n!}{n^n} (-nz)^{n-d} \mathcal{L}_d^{(0)}(nz) & \text{if } n \geq d \end{cases} \quad (3.46)$$

3.4 The additivity law of the \mathcal{R} -transform

The \mathcal{R} -transform is a convenient tool to compute the limiting eigenvalue distribution of the sum of two matrices say A and B where B is rotated with a Haar-distributed random matrix O . In other words, it allows to compute the eigenvalue distribution of $C = A + O^T B O$ in large dimensions. In particular, we can show that:

$$\mathcal{R}_C(g) = \mathcal{R}_A(g) + \mathcal{R}_B(g) \quad (3.47)$$

where $\mathcal{R}_A(g), \mathcal{R}_B(g), \mathcal{R}_C(g)$ are the \mathcal{R} -transform of A, B and C respectively. The \mathcal{R} -transform can be defined through the equation:

$$\mathcal{R}_A(g_A(z)) = z + \frac{1}{g_A(z)} \quad (3.48)$$

where $g_A(z) = \text{Tr}_n[(A - zI)^{-1}]$ is the Stieltjes transform of A . Sometimes, calculating the \mathcal{R} -transform can be cumbersome. When we already have access to g_B , a more convenient formulation of this law can be expressed that directly gives g_C :

$$g_C(z) = g_B(z - \mathcal{R}_A(g_C(z))) \quad (3.49)$$

To see the connection, let's first remark that that (3.48) can be rewritten with g_A^{-1} is the reciprocal of g_A :

$$\mathcal{R}_A(g) = g_A^{-1}(g) + \frac{1}{g} \quad (3.50)$$

Chapter 3. The Linear-pencil method

Applying the definition in the relation (3.47) gives:

$$g_C^{-1}(g) + \frac{1}{g} = \mathcal{R}_A(g) + g_B^{-1}(g) + \frac{1}{g} \quad (3.51)$$

So that with $g = g_C(z)$ (so that $g_C^{-1}(g_C(z)) = z$) it remains:

$$z = \mathcal{R}_A(g_C(z)) + g_B^{-1}(g_C(z)) \quad (3.52)$$

Finally, by rearranging the terms and applying g_B on both sides of the equation, we can conveniently express the Stieltjes transform of C as displayed in (3.49).

In this section, we will show how the \mathcal{R} -transform emerges from the linear-pencil method for B any polynomial p_d of finite degree d of a random Wigner matrix X : $B = p_d(X)$. Because X is invariant by rotation, and henceforth also $p_d(X)$, equation (3.47) is expected to boil down to:

$$\mathcal{R}_{A+p_d(X)}(g) = \mathcal{R}_A(g) + \mathcal{R}_{p_d(X)}(g) \quad (3.53)$$

Or following the convenient form (3.49) for any $x \in \mathbb{C}^+$:

$$g_{A+p_d(X)}(x) = g_A(x - \mathcal{R}_{p_d(X)}(g_{A+p_d(X)}(x))) \quad (3.54)$$

Let's see first that we can always construct a linear-pencil to compute $g_{p_d(X)}(z)$, the Stieltjes-transform of $p_d(X)$. Let's assume that $p_d(X) = \lambda_d X^d + \dots + \lambda_1 X + \lambda_0$. We can define:

$$(p_d(X) - zI_n)U_1 = I_n \quad (3.55)$$

$$U_2 = XU_1 \quad (3.56)$$

$$U_3 = XU_2 = X^2U_1 \quad (3.57)$$

$$\vdots \quad (3.58)$$

$$U_d = XU_{d-1} = X^dU_1 \quad (3.59)$$

The first matrix U_1 is precisely the resolvent $U_1 = (p_d(X) - zI_n)^{-1}$ and we have:

$$(-zI_n + \lambda_d X^d + \dots + \lambda_1 X + \lambda_0 I_n)U_1 = (-zI_n + \lambda_1 X + \lambda_0 I_n)U_1 + \lambda_2 XU_2 + \dots + \lambda_d XU_d \quad (3.60)$$

This leads to the construction of the following linear-pencil:

$$\underbrace{\begin{pmatrix} \lambda_0 I + \lambda_1 X - zI & \lambda_2 X & \lambda_3 X & \dots & \lambda_d X \\ X & -I & 0 & \dots & 0 \\ 0 & X & -I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -I \end{pmatrix}}_L \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_d \end{pmatrix} = \begin{pmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (3.61)$$

consequently, $(L^{-1})^{(11)} = U_1$. Let's introduce the following matrices:

$$F = \begin{pmatrix} \lambda_0 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix} \in \mathbb{R}^{d \times d} \quad W = \begin{pmatrix} \lambda_1 X & \lambda_2 X & \lambda_3 X & \dots & \lambda_d X \\ X & 0 & 0 & \dots & 0 \\ 0 & X & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad (3.62)$$

Then our linear-pencil can be written as $L = L_0 + W$ where L_0 is made of scalar blocks: $L_0 = -z(e_1 e_1^T) \otimes I + F \otimes I$. We know from result 3.1 that for $h(z) = (J \otimes \text{Tr}_n)[L^{-1}]$ we find:

$$h(z) = (-ze_1 e_1^T + F - \eta_L(h(z)))^{-1} \quad (3.63)$$

Interestingly, with defining $R_L(h) := F - \eta_L(h)$ the fixed-point equation can also be written as:

$$R_L(h(z)) = ze_1 e_1^T + h(z)^{-1} \quad (3.64)$$

This equation for R_L shows some similarity with the \mathcal{R} -transform equation (3.48). However, in this case R_L is a $d \times d$ matrix. To make the connection with $\mathcal{R}_{p_d(X)}$, we can use Sherman-Morrison formula to extract $-ze_1 e_1^T$ the rank-one term form (3.63) and compute $h_{11}(z)$:

$$h_{11}(z) = e_1^T h(z) e_1 \quad (3.65)$$

$$= e_1^T (F - \eta_L(h(z)))^{-1} e_1 + \frac{ze_1^T (F - \eta_L(h(z)))^{-1} e_1 e_1^T (F - \eta_L(h(z)))^{-1} e_1}{1 - ze_1^T (F - \eta_L(h(z)))^{-1} e_1} \quad (3.66)$$

$$= \frac{e_1^T (F - \eta_L(h(z)))^{-1} e_1}{1 - ze_1^T (F - \eta_L(h(z)))^{-1} e_1} \quad (3.67)$$

So this leads to the formula:

$$(R_L(h(z))^{-1})^{(11)} = e_1^T (F - \eta_L(h(z)))^{-1} e_1 = \left(z + \frac{1}{h_{11}(z)} \right)^{-1} \quad (3.68)$$

Let's recall that the because h_{11} is the Stieltjes transform of $p_d(X)$, equation (3.48) states that $\mathcal{R}_{p_d(X)}(h_{11}(z)) = z + \frac{1}{h_{11}(z)}$, so we have:

$$\mathcal{R}_{p_d(X)}(g_{p_d(X)}(z)) = \mathcal{R}_{p_d(X)}(h_{11}(z)) = ((R_L(h(z))^{-1})^{(11)})^{-1} \quad (3.69)$$

Let's go back to the main problem: we want to have an expression for $\text{Tr}_n[G]$ with $G = (A - xI + p_d(X))^{-1}$ for some deterministic matrix A . This time, we consider the linear-pencil $L' = (e_1 e_1^T) \otimes (-xI + A) + F \otimes I + W$ and take $g = (J \otimes \text{Tr}_n)[(L')^{-1}]$. By construction, we have $((L')^{-1})^{(11)} = G$ so we will be interested in g_{11} . Also we can apply again result 3.1 and find:

$$g = (J \otimes \text{Tr}_n)[((e_1 e_1^T) \otimes (-xI + A) + F \otimes I - \eta_{L'}(g) \otimes I)^{-1}] \quad (3.70)$$

Chapter 3. The Linear-pencil method

Let's assume that λ is a random variable sampling the spectrum of A . With a careful look, because all the identity matrices in the right-hand side of the tensor expressions can be written in the basis of the eigenvectors of A , the former expression can be simplified further as:

$$g = \mathbb{E}_\lambda \left[((\lambda - x)e_1 e_1^T + F - \eta_{L'}(g))^{-1} \right] \quad (3.71)$$

The final point is to remark that $\eta_{L'}$ and η_L are in fact exactly the same operators because they operate based on the same structure σ which is set by W only. So we can write:

$$g = \mathbb{E}_\lambda \left[((\lambda - x)e_1 e_1^T + R_L(g))^{-1} \right] \quad (3.72)$$

And again, as we are interested only in g_{11} , we can use Sherman-Morrison formula to find:

$$g_{11} = \mathbb{E}_\lambda \left[(R_L(g)^{-1})^{(11)} - \frac{(\lambda - x)((R_L(g)^{-1})^{(11)})^2}{1 + (\lambda - x)((R_L(g)^{-1})^{(11)})} \right] \quad (3.73)$$

$$= \mathbb{E}_\lambda \left[\frac{((R_L(g)^{-1})^{(11)})}{1 + (\lambda - x)(R_L(g)^{-1})^{(11)}} \right] \quad (3.74)$$

$$= \mathbb{E}_\lambda \left[((\lambda - x) + ((R_L(g)^{-1})^{(11)})^{-1})^{-1} \right] \quad (3.75)$$

So because of equation (3.69) we can expect $((R_L(g)^{-1})^{(11)})^{-1} = \mathcal{R}_{p_d(X)}(g_{11})$. This is non-trivial and will be proved below. So we find:

$$g_{11} = \mathbb{E}_\lambda \left[((\lambda - x) + \mathcal{R}_{p_d(X)}(g_{11}))^{-1} \right] \quad (3.76)$$

Finally, we can rewrite this equation in terms of the Stieltjes transform g_A of A and we find back (3.54):

$$g_{11} = g_A(x - \mathcal{R}_{p_d(X)}(g_{11})) \quad (3.77)$$

It remains to show the relation $((R_L(g)^{-1})^{(11)})^{-1} = \mathcal{R}_{p_d(X)}(g_{11})$. To be true, it requires the existence of a $z_1 \in \mathbb{C}$ such that $g = h(z_1)$. To get the intuition of this, first remark that z_1 can and has to be chosen such that $g_{11} = h_{11}(z_1)$. So there remains to see that for one of these z_1 , this implies that $g = h(z_1)$ (in other words, equality at the elements of index (1, 1) implies equality at each element between the matrices). First define the sub-blocks:

$$R_L(g) = \begin{pmatrix} A_R(g) & B_R(g) \\ C_R(g) & D_R(g) \end{pmatrix} \quad (3.78)$$

With $A_R(g) \in \mathbb{C}^{1 \times 1}$, $B_R(g) \in \mathbb{C}^{1 \times n-1}$, $C_R(g) \in \mathbb{C}^{n-1 \times 1}$, $D_R(g) \in \mathbb{C}^{(n-1) \times (n-1)}$. Therefore (3.72) can be written as:

$$g = \mathbb{E}_\lambda \left[\left(\begin{pmatrix} (\lambda - x) + A_R(g) & B_R(g) \\ C_R(g) & D_R(g) \end{pmatrix}^{-1} \right) \right] \quad (3.79)$$

Using the block-inversion formula, we can derive a simpler expression of the inverse of the

matrix on the right-hand side with:

$$g = \begin{pmatrix} K'(g) & -K'(g)B_R(g)D_R(g)^{-1} \\ -C_R(g)D_R(g)^{-1}K'(g) & D_R(g)^{-1} + D_R(g)^{-1}C_R(g)K'(g)B_R(g)D_R(g)^{-1} \end{pmatrix} \quad (3.80)$$

where $K'(g)$ is given by the expectation:

$$K'(g) = \mathbb{E}_\lambda [((\lambda - x) + A_R(g) - B_R(g)D_R(g)^{-1}C_R(g))^{-1}] \quad (3.81)$$

On the other hand, a similar equation holds for $h(z)$ with:

$$h(z) = \begin{pmatrix} -z + A_R(h(z)) & B_R(h(z)) \\ C_R(h(z)) & D_R(h(z)) \end{pmatrix}^{-1} \quad (3.82)$$

So the block-matrix inversion formula yields a similar simpler expression of the inverse:

$$h = \begin{pmatrix} K(h) & -K(h)B_R(h)D_R(h)^{-1} \\ -C_R(h)D_R(h)^{-1}K(h) & D_R(h)^{-1} + D_R(h)^{-1}C_R(h)K(h)B_R(h)D_R(h)^{-1} \end{pmatrix} \quad (3.83)$$

This time with another expression for $K(h)$:

$$K(h(z)) = (-z + A_R(h) - B_R(h)D_R(h)^{-1}C_R(h))^{-1} \quad (3.84)$$

But now for $z = z_1$, and thus $h_{11}(z_1) = g_{11}$, this imply that $K(h(z_1)) = K'(g)$ by construction. In this situation, (3.80) and (3.83) are exactly the same, so h and g satisfies the same fixed-point equation and hence we expect $g = h$.

3.5 Derivation of result 3.2 in finite-dimension

3.5.1 Derivation of a heat-equation

As stated in the preliminaries, we will consider the case of the Gaussian Unitary Ensemble (GUE). We consider the linear-pencil $L_t = L_0 + W_t$ with L_1 the random-matrix into consideration, and L_0 the deterministic matrix at $t = 0$. The elements of W_t are brownian motions, for which we impose that W_t is self-adjoint and that $\text{Re } W_t$ and $\text{Im } W_t$ are two independent random matrices.

The blocks of L_t can be seen as Dyson brownian motions with the covariance σ embedded into the stochastic covariation for $u \neq v$ and $N_i = N_l$ and $N_j = N_k$:

$$d[\text{Re } L_{uv}^{(ij)}, \text{Re } L_{uv}^{(lk)}] = d[\text{Im } L_{uv}^{(ij)}, \text{Im } L_{uv}^{(lk)}] = \frac{1}{2n} \sigma_{ij}^{kl} dt \quad (3.85)$$

Specifically, this implies what one would expect when dealing simply with the real random

Chapter 3. The Linear-pencil method

matrices scenario for $N_i = N_l$ and $N_j = N_k$:

$$d[\operatorname{Re} L_{uv}^{(ij)}, \operatorname{Re} L_{vu}^{(kl)}] = d[\operatorname{Re} L_{uv}^{(ij)}, \operatorname{Re} \bar{L}_{uv}^{(lk)}] = d[\operatorname{Re} L_{uv}^{(ij)}, \operatorname{Re} L_{uv}^{(lk)}] = \frac{1}{2n} \sigma_{ij}^{kl} dt \quad (3.86)$$

With the additional imaginary part, we also get for $N_i = N_l$ and $N_j = N_k$:

$$d[\operatorname{Im} L_{uv}^{(ij)}, \operatorname{Im} L_{vu}^{(kl)}] = d[\operatorname{Im} L_{uv}^{(ij)}, \operatorname{Im} \bar{L}_{uv}^{(lk)}] = -d[\operatorname{Im} L_{uv}^{(ij)}, \operatorname{Im} L_{uv}^{(lk)}] = -\frac{1}{2n} \sigma_{ij}^{kl} dt \quad (3.87)$$

This way, using $L_{uv}^{(ij)} = \operatorname{Re} L_{uv}^{(ij)} + i \operatorname{Im} L_{uv}^{(ij)}$ and the independence between the real and imaginary part, we find for $u \neq v$ and for $N_i = N_l$ and $N_j = N_k$:

$$d[L_{uv}^{(ij)}, L_{uv}^{(kl)}] = \frac{1}{n} \left(d[\operatorname{Re} L_{uv}^{(ij)}, \operatorname{Re} L_{uv}^{(kl)}] - d[\operatorname{Im} L_{uv}^{(ij)}, \operatorname{Im} L_{uv}^{(kl)}] \right) = 0 \quad (3.88)$$

$$d[L_{uv}^{(ij)}, L_{vu}^{(kl)}] = \frac{1}{n} \left(d[\operatorname{Re} L_{uv}^{(ij)}, \operatorname{Re} L_{vu}^{(kl)}] - d[\operatorname{Im} L_{uv}^{(ij)}, \operatorname{Im} L_{vu}^{(kl)}] \right) = \frac{1}{n} \sigma_{ij}^{kl} dt \quad (3.89)$$

Note that the first equation is true for the complex case but not always for real random matrices. Considering the complex case thus simplifies the derivations that follow.

Note that these results also apply when $u = v$, except when $dL^{(ij)}$ is self-adjoint. In this case, the diagonal terms of $dL^{(ij)}$ are necessarily real, so we impose for $N_i = N_j = N_k = N_l$:

$$d[\operatorname{Re} L_{uu}^{(ij)}, \operatorname{Re} L_{uu}^{(kl)}] = \frac{1}{n} \sigma_{ij}^{kl} dt \quad d[\operatorname{Im} L_{uu}^{(ij)}, \operatorname{Im} L_{uu}^{(kl)}] = 0 \quad (3.90)$$

In conclusion, we have for any block $(ij), (kl)$ such that $N_i = N_l$ and $N_j = N_k$:

$$d[L_{uv}^{(ij)}, L_{pq}^{(kl)}] = \frac{1}{n} \delta_{uq} \delta_{vp} \sigma_{ij}^{kl} dt \quad (3.91)$$

We further define the block matrix Z such that for all $(ij) \in \mathbb{S}$, $Z_{ij} = z_{ij} I_{N_i}$ with $z_{ij} \in \mathbb{C}$ and 0 otherwise. Then let's define $f_n(L_t, Z) = \det(L_t + Z)$, and $\langle f_n \rangle(t, Z) = \mathbb{E} f_n(L_t, Z)$. Using Itô formula, we have:

$$df_n(L_t, Z) = \sum_{ij} \sum_{uv} \frac{\partial f_n}{\partial L_{uv}^{(ij)}} dL_{uv}^{(ij)} + \frac{1}{2} \sum_{ij, kl} \sum_{uv, pq} \frac{\partial^{(2)} f_n}{\partial L_{uv}^{(ij)} \partial L_{pq}^{(kl)}} d[L_{uv}^{(ij)}, L_{pq}^{(kl)}] \quad (3.92)$$

Using the covariation, the expression can be simplified to:

$$df_n(L_t, Z) = \sum_{ij} \sum_{uv} \frac{\partial f_n}{\partial L_{uv}^{(ij)}} dL_{uv}^{(ij)} + \frac{1}{2n} \sum_{ij, kl} \sum_{uv} \frac{\partial^{(2)} f_n}{\partial L_{uv}^{(ij)} \partial L_{vu}^{(kl)}} \sigma_{ij}^{kl} dt \quad (3.93)$$

Provided invertibility as per the assumptions, let $G = (L + Z)^{-1}$. Straightforward calculations

leads to the two following formulas:

$$\frac{\partial f_n}{\partial L_{uv}^{(ij)}} = f_n \cdot G_{vu}^{(ji)} \quad \frac{\partial G_{vu}^{(ji)}}{\partial L_{pq}^{(kl)}} = -G_{vp}^{(jk)} G_{qu}^{(li)} \quad (3.94)$$

Combining this with the previous formula, we have:

$$\sum_{uv} \frac{\partial^{(2)} f_n}{\partial L_{uv}^{(ij)} \partial L_{vu}^{(kl)}} = \sum_{uv} f_n \cdot \left(G_{vu}^{(ji)} G_{uv}^{(lk)} - G_{vv}^{(jk)} G_{uu}^{(li)} \right) \quad (3.95)$$

$$= f_n \cdot \left(\text{Tr} \left[G^{(ji)} G^{(lk)} \right] - \text{Tr} \left[G^{(jk)} \right] \text{Tr} \left[G^{(li)} \right] \right) \quad (3.96)$$

Similarly, with the Wirtinger derivatives, we find:

$$\frac{\partial f_n}{\partial z_{il}} = f_n \cdot \text{Tr} \left[G^{(li)} \right] \quad \frac{\partial^{(2)} f_n}{\partial z_{il} \partial z_{kj}} = f_n \cdot \left(\text{Tr} \left[G^{(li)} \right] \text{Tr} \left[G^{(jk)} \right] - \text{Tr} \left[G^{(lk)} G^{(ji)} \right] \right) \quad (3.97)$$

As a side remark, the traces of interest are thus given by the formula

$$\frac{1}{n} \text{Tr} \left[G^{(ij)} \right] = \frac{1}{n} \frac{\partial \log f_n}{\partial z_{ji}} \quad (3.98)$$

In conclusion, the Ito formula can be rewritten as:

$$df_n(L_t, Z) = \sum_{ij} \sum_{uv} \frac{\partial f_n}{\partial L_{uv}^{(ij)}} dL_{uv}^{(ij)} - \frac{1}{2n} \sum_{ij,kl} \frac{\partial^{(2)} f_n}{\partial z_{il} \partial z_{kj}} \sigma_{ij}^{kl} dt \quad (3.99)$$

and thus on the average, it yields the reversed-time heat-equation which is exact for all $n \in \mathbb{N}^*$:

$$\frac{\partial \langle f_n \rangle}{\partial t} = -\frac{1}{2n} \sum_{ij,kl} \frac{\partial^{(2)} \langle f_n \rangle}{\partial z_{il} \partial z_{kj}} \sigma_{ij}^{kl} \quad (3.100)$$

Note that this is an extension of the heat-equation from a Dyson brownian motion as described for instance in (Tao, 2012).

3.5.2 Solution of the heat-equation

The heat-equation can be solved exactly at finite n . Let's first consider the "matrixisation" of σ with $(\Sigma_{il,kj}) = (\sigma_{ij}^{kl} | N_i = N_l, N_k = N_l) \in \mathbb{C}^{|\mathbb{S}| \times |\mathbb{S}|}$ as described in the remark in 3.1.2 with a decomposition of the form $\Sigma = VV^T$. Now let's consider a real vector $u \in \mathbb{R}^{|\mathbb{S}|}$ and make the change of variable for $(pq) \in \mathbb{S}$, of $\langle \tilde{f}_n \rangle(t, u) = \langle f_n \rangle(t, iVu)$:

$$\frac{\partial \langle \tilde{f}_n \rangle}{\partial u_{pq}}(t, u) = i \sum_{kj} V_{kj,pq} \frac{\partial \langle f_n \rangle}{\partial z_{kj}}(t, iVu) \quad (3.101)$$

Then:

$$\frac{\partial^{(2)}\langle \tilde{f}_n \rangle}{\partial u_{pq}^2}(t, u) = - \sum_{il, kj} V_{il, pq} V_{kj, pq} \frac{\partial \langle f_n \rangle}{\partial z_{kj} \partial z_{il}}(t, iVu) \quad (3.102)$$

In particular, we have:

$$\sum_{pq \in \mathbb{S}} \frac{\partial^{(2)}\langle \tilde{f}_n \rangle}{\partial u_{pq}^2}(t, u) = - \sum_{il, kj} [VV^T]_{il, kj} \frac{\partial \langle f_n \rangle}{\partial z_{kj} \partial z_{il}}(t, iVu) = - \sum_{il, kj} \Sigma_{il, kj} \frac{\partial \langle f_n \rangle}{\partial z_{kj} \partial z_{il}}(t, iVu) \quad (3.103)$$

Therefore, $\langle \tilde{f}_n \rangle$ satisfies the "diagonal" heat-equation: $\frac{\partial \langle \tilde{f}_n \rangle}{\partial t} = \frac{1}{2n} \Delta_u \langle \tilde{f}_n \rangle$ which has the solution at $s = iVr$ and $t = 1$:

$$\langle f_n \rangle(1, s) = \left(\frac{n}{2\pi} \right)^{|\mathbb{S}|/2} \int_{\mathbb{R}^{|\mathbb{S}|}} \langle f_n \rangle(0, iVu + s) \exp\left(-\frac{n}{2} \|u\|^2\right) du \quad (3.104)$$

So we arrive at the solution of the characteristic polynomial at finite n :

$$\langle f_n \rangle(1, s) = \mathbb{E}_{u \sim \mathcal{N}(0, \frac{1}{n} I_{|\mathbb{S}|})} [\langle f_n \rangle(0, iVu + s)] \quad (3.105)$$

One can also use the complex normal distribution as:

$$\langle f_n \rangle(1, s) = \mathbb{E}_{v \sim \mathcal{CN}(s, \frac{1}{n} V \bar{V}^T, -\frac{1}{n} V V^T)} [\langle f_n \rangle(0, v)] \quad (3.106)$$

3.6 Derivation of result 3.1: Three methods

3.6.1 The method of characteristics applied to the heat-equation

Throughout this Section, we will make the following assumption in the limit $n \rightarrow +\infty$:

$$\frac{1}{n} \mathbb{E}[\log f_n] = \frac{1}{n} \log \langle f_n \rangle + o_n(1) \quad (3.107)$$

Rather than deriving the exact stochastic process for the Stieltjes transform as done in (Bodin and Macris, 2021a), our approach focuses on establishing an equation for the logarithm of f_n , specifically the KPZ equation, followed by the Burger equation. Using equation (3.104), we obtain the following:

$$\frac{\partial \ln \langle \tilde{f}_n \rangle}{\partial t} = \frac{1}{\langle \tilde{f}_n \rangle} \frac{\partial \langle \tilde{f}_n \rangle}{\partial t} = \frac{1}{2n \langle \tilde{f}_n \rangle} \sum_{il} \frac{\partial^{(2)} \langle \tilde{f}_n \rangle}{\partial u_{il}^2} \quad (3.108)$$

and we retrieve the noiseless KPZ equation:

$$\frac{1}{2n} \Delta_u \ln \langle \tilde{f}_n \rangle = \frac{1}{2n} \sum_{il} \frac{\partial^{(2)} \ln \langle \tilde{f}_n \rangle}{\partial u_{il}^2} = \frac{1}{2n} \sum_{il} \left(\frac{1}{\langle \tilde{f}_n \rangle} \frac{\partial \ln \langle \tilde{f}_n \rangle}{\partial u_{il}^2} - \left(\frac{\partial \ln \langle \tilde{f}_n \rangle}{\partial u_{il}} \right)^2 \right) \quad (3.109)$$

$$= \frac{\partial \ln \langle \tilde{f}_n \rangle}{\partial t} - \frac{1}{2n} (\nabla_u \ln \langle \tilde{f}_n \rangle)^2 \quad (3.110)$$

3.6 Derivation of result 3.1: Three methods

So without the change of variable in $\langle f_n \rangle \rightarrow \langle \tilde{f}_n \rangle$, the KPZ equation becomes:

$$\frac{\partial \langle f_n \rangle}{\partial t} = -\frac{1}{2n} \sum_{il,kj} \left(\frac{\partial^{(2)} \ln \langle f_n \rangle}{\partial z_{il} \partial z_{kj}} + \frac{\partial \ln \langle f_n \rangle}{\partial z_{il}} \frac{\partial \ln \langle f_n \rangle}{\partial z_{kj}} \right) \sigma_{ij}^{kl} \quad (3.111)$$

At this point, we define the functions $\langle g_{pq}^n \rangle$ for each $(pq) \in \mathbb{S}$ such that $\langle g_{pq}^n \rangle(t, Z) = \frac{1}{n} \frac{\partial \ln \langle f_n \rangle(t, Z)}{\partial z_{qp}}$. Note that $\langle g_{pq}^n \rangle$ may not coincide with the expectation of the partial trace $\frac{1}{n} \frac{\partial \mathbb{E} \ln f_n}{\partial z_{qp}}$. However, we will assume the equivalence in the limit of large n . Then we have:

$$\frac{\partial \langle g_{pq}^n \rangle}{\partial t} = -\frac{1}{2n} \sum_{il,kj} \frac{\partial^{(2)} \langle g_{pq}^n \rangle}{\partial z_{il} \partial z_{kj}} - \frac{1}{2} \frac{\partial}{\partial z_{qp}} \left(\sum_{il,kj} \langle g_{li}^n \rangle \sigma_{ij}^{kl} \langle g_{jk}^n \rangle \right) \quad (3.112)$$

The second term on the right-hand side of the equation can be simplified as:

$$\frac{\partial}{\partial z_{qp}} \left(\sum_{il,kj} \langle g_{li}^n \rangle \sigma_{ij}^{kl} \langle g_{jk}^n \rangle \right) = \sum_{il,kj} \frac{\partial \langle g_{li}^n \rangle}{\partial z_{qp}} \sigma_{ij}^{kl} \langle g_{jk}^n \rangle + \sum_{il,kj} \langle g_{li}^n \rangle \sigma_{ij}^{kl} \frac{\partial \langle g_{jk}^n \rangle}{\partial z_{qp}} \quad (3.113)$$

$$= \sum_{il,kj} \frac{\partial \langle g_{li}^n \rangle}{\partial z_{qp}} \sigma_{ij}^{kl} \langle g_{jk}^n \rangle + \sum_{il,kj} \langle g_{jk}^n \rangle \sigma_{kl}^{ij} \frac{\partial \langle g_{li}^n \rangle}{\partial z_{qp}} \quad (3.114)$$

$$= 2 \sum_{il} \frac{1}{n} \frac{\partial \ln \langle f_n \rangle}{\partial z_{il} \partial z_{qp}} [\eta(\langle g^n \rangle)]_{il} \quad (3.115)$$

$$= 2 \sum_{il} [\eta(\langle g^n \rangle)]_{il} \frac{\partial \langle g_{pq}^n \rangle}{\partial z_{il}} \quad (3.116)$$

With the assumption that in the large limit n , the quantity $\langle g^n \rangle(t, Z)$ converges to a finite value $g(t, Z)$, we expect the dynamics of g to follow the Burger equation:

$$\frac{\partial g_{pq}}{\partial t} + \sum_{il} [\eta(g)]_{il} \frac{\partial g_{pq}}{\partial z_{il}} = 0 \quad (3.117)$$

Finally, let's consider a trajectory $s \rightarrow (\hat{t}(s), \hat{Z}(s))$ such that $\hat{g}(s) = g(\hat{t}(s), \hat{Z}(s))$ is constant for all s . Then it implies for all $(pq) \in \mathbb{S}$:

$$\frac{d\hat{g}_{pq}}{ds} = \frac{\partial g_{pq}}{\partial t} \frac{\partial \hat{t}}{\partial s} + \sum_{il} \frac{\partial g_{pq}}{\partial z_{il}} \frac{\partial \hat{Z}_{il}}{\partial s} = 0 \quad (3.118)$$

One solution is to match $\frac{\partial \hat{t}}{\partial s} = 1$ and $\frac{\partial \hat{Z}_{il}}{\partial s} = [\eta(\hat{g}(s))]_{il}$ (Note that the right hand side of the second equation is constant as $\hat{g}(s)$ is assumed constant over s). This is achieved by setting $\hat{t}(s) = s$ and $\hat{Z}_{il}(s) = [\eta(\hat{g}(s))]_{il}(s-1)$. Using again that \hat{g} is independent of s , we find $\hat{g}(1) = \hat{g}(0)$, and therefore:

$$g(1, 0) = g(0, -\eta(g(1, 0))) \quad (3.119)$$

At time $t = 0$, the term $g(0, -\eta(g(1, 0)))$ corresponds to the right-hand side of Equation (3.15) while $g(1, 0)$ corresponds to the definition of g on left-hand side. This above relationship thus leads to the desired result.

3.6.2 The method of steepest descent

The equation (3.104) can be formulated as:

$$\langle f_n \rangle(1, s) = \left(\frac{n}{2\pi} \right)^{|\mathbb{S}|/2} \int_{\mathbb{R}^{|\mathbb{S}|}} e^{n\mathcal{C}_n(u, s)} \mathbf{d}u \quad (3.120)$$

Where we define the term $\mathcal{C}_n(u, s)$:

$$\mathcal{C}_n(u, s) = \frac{1}{n} \ln \langle f_n \rangle(0, iVu + s) - \frac{1}{2} \|u\|^2 \quad (3.121)$$

Although not rigorously proved in this work, we use an approach in the same spirit as the method of the steepest descent - that is to transform the integral as a contour integral. This contour can be "twisted" to cross one or multiple saddle points which acts as accumulation points for the integral, which in turn help derive a fixed-point equation. In our scenario, let's first assume in the limit of large n the concentration of $h_n = \frac{1}{n} \ln \langle f_n \rangle$ towards a function h and:

$$\lim_{n \rightarrow \infty} C_n(u, s) \simeq \mathcal{C}(u, s) = h(0, iVu + s) - \frac{1}{2} \|u\|^2 \quad (3.122)$$

Then we can expect the existence of a specific saddle points $\tilde{u}(s)$ solution of $\nabla_u \mathcal{C}_n(\tilde{u}(s), s) = 0$ for which:

$$h(1, s) = \lim_{n \rightarrow +\infty} \frac{1}{n} \ln \langle f_n \rangle(1, s) \simeq \mathcal{C}(\tilde{u}(s), s) \quad (3.123)$$

and in particular, with the same definition as in the previous section with $g_{ij}(1, Z) = \frac{\partial h(1, Z)}{\partial z_{ji}}$, we expect:

$$g_{ij}(1, s) = \frac{\partial \mathcal{C}(\tilde{u}(s), s)}{\partial u_{ji}} \frac{\partial \tilde{u}(s)}{\partial s_{ji}} + \frac{\partial \mathcal{C}(\tilde{u}(s), s)}{\partial s_{ji}} = 0 + \frac{\partial h(0, iV\tilde{u}(s) + s)}{\partial s_{ji}} = g_{ij}(0, iV\tilde{u}(s) + s) \quad (3.124)$$

Therefore at $s = 0$:

$$g(1, 0) = g(0, iV\tilde{u}(0)) \quad (3.125)$$

Now on the other hand at this specific point u we find for all $(il) \in \mathbb{S}$:

$$\nabla_u \mathcal{C}(u(s), s) = 0 \implies \sum_{kj} iV_{kj, il} \frac{\partial h}{\partial z_{kj}}(0, iV\tilde{u}(s) - s) - \tilde{u}_{il}(s) = 0 \quad (3.126)$$

So:

$$V^T g(0, iV\tilde{u}(s) - s) = -i\tilde{u}(s) \quad (3.127)$$

So at $s = 0$:

$$g(0, iVV^T(V^T)^{-1}\tilde{u}(0)) = -i(V^T)^{-1}\tilde{u}(0) \quad (3.128)$$

Combining both equations we have:

$$-i(V^T)^{-1}\tilde{u}(0) = g(1, 0) \quad (3.129)$$

and then:

$$g(0, iVV^T(L^T)^{-1}\tilde{u}(0)) = g(0, -VV^T g(1, 0)) = g(1, 0) \quad (3.130)$$

Finally, $VV^T = \Sigma$ and given a vector w and its corresponding matrix W , it is easy to see that Σw is in fact the vectorization of $\eta(W)$. Hence we find a similar result as equation (3.119):

$$g(1, 0) = g(0, -\eta(g(1, 0))) \quad (3.131)$$

3.6.3 The Replica method

The replica method is another approach leading to similar results and doesn't rely on a stochastic process over time. This approach has been described in Appendix D in (Bodin and Macris, 2021a). In this section, here we propose a different variation which accounts for the possibility of keeping deterministic matrices in the linear-pencil. We will use the notation $f_n(Z)$ instead of $f_N(L_1, Z)$ and use the relation:

$$h_n(Z) = \frac{1}{n} \ln f_n(Z) = \lim_{r \rightarrow 0^+} \frac{1}{n} \frac{1 - f_n(Z)^{-r}}{r} \quad (3.132)$$

Let us define $\tilde{N} = \sum_{i=1}^N N_i$. Using the density function of the complex normal distribution $\mathcal{CN}(0, L^{-1}, 0)$ we have the following relation for the determinant:

$$f_n(Z)^{-1} = |\det(L)|^{-1} = \int_{\mathbb{C}^{\tilde{N}}} e^{-\frac{1}{2}(\tilde{x}^T L \tilde{x} + \tilde{x}^T \tilde{L} \tilde{x})} \frac{d\tilde{x}}{\pi^{\tilde{N}}} \quad (3.133)$$

In the following, we will make the assumption that L is self-adjoint (in particular that L_0 , W and Z in $L = W + L_0 + Z$ are all self-adjoint) - although we will see that this condition can be weakened. When $r \in \mathbb{N}^*$, we have the product of r integrals:

$$f_n(Z)^{-r} = \int_{\mathbb{C}^{r\tilde{N}}} e^{-\sum_{a=1}^r \tilde{x}_a^T L x_a} \prod_{a=1}^r \frac{d\tilde{x}_a}{\pi^{\tilde{N}}} \quad (3.134)$$

After expanding L with $L = W + L_0 + Z$, we find:

$$\sum_a \tilde{x}_a^T L x_a = \sum_a \sum_{ij} \sum_{uv} \tilde{x}_{au}^{(i)} (L)_{uv}^{(ij)} x_{av}^{(j)} \quad (3.135)$$

$$= \sum_{ij} \sum_{uv} (W_{uv}^{(ij)} + (L_0 + Z)_{uv}^{(ij)}) \sum_a \tilde{x}_{au}^{(i)} x_{av}^{(j)} \quad (3.136)$$

$$= \sum_{ij} \sum_{uv} (\operatorname{Re} W_{uv}^{(ij)} + i \operatorname{Im} W_{uv}^{(ij)} + (L_0 + Z)_{uv}^{(ij)}) \sum_a \tilde{x}_{au}^{(i)} x_{av}^{(j)} \quad (3.137)$$

As a reminder, we impose the following covariance structure:

$$\mathbb{E}[\operatorname{Re} W_{uv}^{(ij)} \operatorname{Re} W_{uv}^{(kl)}] = \mathbb{E}[\operatorname{Im} W_{uv}^{(ij)} \operatorname{Im} W_{uv}^{(kl)}] = \frac{1}{2n} \sigma_{ij}^{lk} \quad (3.138)$$

Chapter 3. The Linear-pencil method

So using this structure and the self-adjoint condition, we have:

$$\mathbb{E}[\operatorname{Re} W_{uv}^{(ij)} \operatorname{Re} W_{vu}^{(kl)}] = \mathbb{E}[\operatorname{Re} W_{uv}^{(ij)} \operatorname{Re} \bar{W}_{uv}^{(lk)}] = \frac{1}{2n} \sigma_{ij}^{kl} \quad (3.139)$$

$$\mathbb{E}[\operatorname{Im} W_{uv}^{(ij)} \operatorname{Im} W_{vu}^{(kl)}] = \mathbb{E}[\operatorname{Im} W_{uv}^{(ij)} \operatorname{Im} \bar{W}_{uv}^{(kl)}] = -\frac{1}{2n} \sigma_{ij}^{kl} \quad (3.140)$$

Which leads to the general average:

$$\mathbb{E}[W_{uv}^{(ij)} W_{vu}^{(kl)}] = \frac{1}{n} \sigma_{ij}^{kl} \quad \text{and} \quad \mathbb{E}[W_{uv}^{(ij)} W_{uv}^{(kl)}] = 0 \quad (3.141)$$

So by taking the expectation over W with the moment generating function (which boils down to similar calculations as in Section 3.5.1):

$$\mathbb{E}_W[f_n(Z)^{-r}] = \int_{\mathbb{C}^{rN}} \exp \left\{ \frac{1}{2n} \sum_{ij,kl} \sigma_{ij,kl}^{kl} \sum_{uv} \left(\sum_a \bar{x}_{au}^{(i)} x_{av}^{(j)} \right) \left(\sum_b \bar{x}_{bv}^{(k)} x_{bu}^{(l)} \right) + R(x) \right\} \prod_{a=1}^r \frac{dx_a}{\pi^{\bar{N}}} \quad (3.142)$$

Where

$$R(x) = -\frac{1}{n} \sum_{ij} \sum_{uv} (L_0 + Z)_{uv}^{(ij)} \sum_a \bar{x}_{au}^{(i)} x_{av}^{(j)} \quad (3.143)$$

Notice further that:

$$\sum_{uv} \left(\sum_a \bar{x}_{au}^{(i)} x_{av}^{(j)} \right) \left(\sum_b \bar{x}_{bv}^{(k)} x_{bu}^{(l)} \right) = \sum_{ab} \left(\sum_u \bar{x}_{au}^{(i)} x_{bu}^{(l)} \right) \left(\sum_v x_{av}^{(j)} \bar{x}_{bv}^{(k)} \right) \quad (3.144)$$

So let $\Sigma_{il,kj} = \sigma_{ij}^{kl}$ be the "matrixization" of σ , and for the sake of clarity, let's define $X_{ab}^{(il)}(x) = \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{bu}^{(l)}$. Then we can find an extended version of the Hubbard-Stratonovich transform:

$$\left| \det \left(\frac{n}{2} \Sigma \right) \right| \int \exp \left\{ -\frac{n}{2} \sum_{ij,kl} \left(q_{ab}^{(il)} - X_{ab}^{(il)}(x) \right) \sigma_{ij}^{kl} \left(\bar{q}_{ab}^{(jk)} - \bar{X}_{ab}^{(jk)}(x) \right) \right\} \prod_{il} \frac{dq_{ab}^{(il)}}{\pi} = 1 \quad (3.145)$$

So in fact we have (with extra caution about the Re)

$$e^{\frac{n}{2} \sum_{ij,kl} \sigma_{ij}^{kl} X_{ab}^{(il)}(x) \bar{X}_{ab}^{(jk)}(x)} = \left| \det \left(\frac{n}{2} \Sigma \right) \right| \int e^{-\frac{n}{2} \sum_{ij,kl} \sigma_{ij}^{kl} \left(q_{ab}^{(il)} \bar{q}_{ab}^{(jk)} - 2 \operatorname{Re}(\bar{X}_{ab}^{(il)}(x) q_{ab}^{(jk)}) \right)} \prod_{il} \frac{dq_{ab}^{(il)}}{\pi} \quad (3.146)$$

By grouping the terms in x, \hat{q}, q , the expression can be rewritten as:

$$\mathbb{E}[f_n(Z)^{-r}] = \left| \det \left(\frac{n}{2} \Sigma \right) \right| \int_q e^{n(\hat{Q}_r(q) + \bar{R}_{n,r}(q))} \left(\prod_{\substack{a \leq b \\ (ij)}} \frac{dq_{ab}^{(ij)}}{\pi} \right) \quad (3.147)$$

where we introduced the functions \tilde{Q}_r and $\tilde{R}_{n,r}$ given by:

$$\tilde{Q}_r(q) = -\frac{1}{2} \sum_{a \leq b} \sum_{i,j,k,l} \sigma_{ij}^{kl} q_{ab}^{(il)} \bar{q}_{ab}^{(jk)} \quad (3.148)$$

$$\tilde{R}_{n,r}(q) = \frac{1}{n} \log \int_{\mathbb{C}^{r\tilde{N}}} \exp \left\{ n \left(R(x) + \sum_{i,j,k,l,a \leq b} \sigma_{ij}^{kl} \operatorname{Re} \left(\bar{q}_{ab}^{(jk)} X_{ab}^{(il)}(x) \right) \right) \right\} \left(\prod_{a=1}^r \frac{dx_a}{\pi^{\tilde{N}}} \right) \quad (3.149)$$

\tilde{Q}_r doesn't depend on n but $\tilde{R}_{n,r}$ does. However, we make the assumption that for any given r , it concentrates towards a value $\tilde{R}_{\infty,r}$ in the limit $n \rightarrow +\infty$, and assume that we can perform the saddle-point approximation. This way we find for some constant C :

$$\mathbb{E}[f_n(Z)^{-r}] \simeq C e^{\frac{n \operatorname{Extr}_q(\tilde{Q}_r(q) + \tilde{R}_{\infty,r}(q))}{q}} \quad (3.150)$$

To compute the extremum, we constrain the subset of the possible solutions with the following Ansatz: we assume that the extrema is located at a certain $q_{ab}^{(ij)} = p^{(ij)} \delta_{ab}$. In this way, we can simplify the dependency in r in \tilde{Q}_r and $\tilde{R}_{\infty,r}$, and have the extremum given by:

$$\operatorname{Extr}_q(\tilde{Q}_r(q) + \tilde{R}_{\infty,r}(q)) = r \operatorname{Extr}_{p \in \mathbb{C}^{N \times N}}(\mathcal{Q}(p) + \mathcal{R}_{\infty}(p)) \quad (3.151)$$

where define \mathcal{R} and \mathcal{Q} are related to \tilde{R} and \tilde{Q} by:

$$\tilde{R}_{r,\infty}(q) = r \mathcal{R}_{\infty}(p) \quad \tilde{Q}_r(q) = r \mathcal{Q}(p) \quad (3.152)$$

This lead to the following expressions for $\mathcal{R}_{\infty}(p)$:

$$\mathcal{R}_{\infty}(p) = \lim_n \frac{1}{nr} \log \int_{\mathbb{C}^{r\tilde{N}}} \exp \left\{ n \left(R(x) + \sum_{aijkl} \sigma_{ij}^{kl} \operatorname{Re} \left(\bar{p}^{(jk)} \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{au}^{(l)} \right) \right) \right\} \left(\prod_{a=1}^r \frac{dx_a}{\pi^{\tilde{N}}} \right) \quad (3.153)$$

One can use the following:

$$\sum_{aijkl} \sigma_{ij}^{kl} \operatorname{Re} \left(\bar{p}^{(jk)} \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{au}^{(l)} \right) = \frac{1}{2} \sum_{aijkl} \sigma_{ij}^{kl} \left(\bar{p}^{(jk)} \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{au}^{(l)} + p^{(jk)} \frac{1}{n} \sum_u \bar{x}_{au}^{(l)} x_{au}^{(i)} \right) \quad (3.154)$$

$$= \frac{1}{2} \sum_{ail} \left([\eta(\bar{p})]_{il} \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{au}^{(l)} + [\eta(p)]_{il} \frac{1}{n} \sum_u \bar{x}_{au}^{(l)} x_{au}^{(i)} \right) \quad (3.155)$$

$$= \sum_{ail} \frac{[\eta(\bar{p})]_{il} + [\eta(p)]_{li}}{2} \frac{1}{n} \sum_u \bar{x}_{au}^{(i)} x_{au}^{(l)} \quad (3.156)$$

So:

$$\mathcal{R}_{\infty}(p) = \lim_n \frac{1}{nr} \log \int_{\mathbb{C}^{r\tilde{N}}} \prod_{a=1}^r \left(\exp \left\{ - \sum_{ij} \sum_{uv} ((L_0 + Z)_{uv}^{(il)} - \left[\frac{\eta(\bar{p}) + \eta(p)^T}{2} \right]_{il} \delta_{uv}) \bar{x}_{au}^{(i)} x_{av}^{(l)} \right\} \frac{dx_a}{\pi^{\tilde{N}}} \right) \quad (3.157)$$

$$= \lim_n \frac{1}{n} \log \left| \det \left(L_0 + Z - \frac{1}{2} (\eta(\bar{p}) + \eta(p)^T) \otimes I \right) \right| \quad (3.158)$$

and similarly for \mathcal{Q} :

$$\mathcal{Q}(p) = \frac{1}{2} \sum_{ij,kl} \sigma_{ij}^{kl} p^{(il)} \bar{p}^{(jk)} = \frac{1}{2} \sum_{il} [\eta(\bar{p})]_{il} p^{(il)} \quad (3.159)$$

So:

$$\mathcal{Q}(p) + \mathcal{R}_\infty(p) = \frac{1}{2} \sum_{il} [\eta(\bar{p})]_{il} p^{(il)} + \lim_n \frac{1}{n} \log \left| \det \left(L_0 + Z - \frac{1}{2} (\eta(\bar{p}) + \eta(p)^T) \otimes I \right) \right| \quad (3.160)$$

We can show that we can limit the search of the extremum on the subspace $\text{im}(\eta)$ instead of $\mathbb{C}^{N \times N}$. We will show that for any $p \in \mathbb{C}^{N \times N}$, there always exists $p_i \in \text{im}(\eta)$ such that:

$$\mathcal{Q}(p) + \mathcal{R}_\infty(p) = \mathcal{Q}(p_i) + \mathcal{R}_\infty(p_i) \quad (3.161)$$

First of all, as it is assumed in the introduction, we have that $\eta|_{\text{im}(\eta)}$, the restriction of η on its image $\text{im}(\eta)$ is an invertible operator which we will define as η^{-1} (and implicitly assume that, when invoked, it is defined on $\text{im}(\eta) \rightarrow \text{im}(\eta)$). This assumption implies that $\ker(\eta) + \text{im}(\eta) = \mathbb{C}^{N \times N}$. Therefore, given any $p \in \mathbb{C}^{N \times N}$, there exists a $p_k \in \ker(\eta)$ and a $p_i \in \text{im}(\eta)$ such that $p = p_k + p_i$.

Next, using the linearity of η , we find $\eta(\bar{p}) = \overline{\eta(p_k + p_i)} = \overline{\eta(p_i)} = \eta(\bar{p}_i)$ and $\eta(p)^T = \eta(p_i)^T$ so that settles $\mathcal{R}_\infty(p) = \mathcal{R}_\infty(p_i)$. For \mathcal{Q} , we use the fact that η is a Hermitian operator with the scalar product $\langle A, B \rangle = \text{Tr}[\bar{A}^T B]$. Indeed for the matrix basis $E_{ij} = e_i e_j^T$ we have precisely that $\langle \eta(E_{jk}), E_{il} \rangle = \sigma_{ij}^{kl}$ and $\langle E_{jk}, \eta(E_{il}) \rangle = \sigma_{ji}^{lk}$, and because L is self-adjoint we find the expected result $\sigma_{ij}^{kl} = \sigma_{ji}^{lk}$. Consequently, for any $g, h \in \mathbb{C}^{N \times N}$, we have $\langle \eta(g), h \rangle = \langle g, \eta(h) \rangle$. So we have:

$$\sum_{il} [\eta(\bar{p})]_{il} p^{(il)} = \text{Tr}[\eta(\bar{p})^T p] = \langle \eta(p), p \rangle = \langle \eta(p_k + p_i), p_k + p_i \rangle \quad (3.162)$$

$$= \langle \eta(p_i), p_k + p_i \rangle \quad (3.163)$$

$$= \langle p_i, \eta(p_k + p_i) \rangle \quad (3.164)$$

$$= \langle p_i, \eta(p_i) \rangle \quad (3.165)$$

Therefore, we also have $\mathcal{Q}(p) = \mathcal{Q}(p_i)$, which concludes with Equation (3.161). Therefore, the extremum calculation can be simplified:

$$\text{Extr}_{p \in \mathbb{C}^{N \times N}} (\mathcal{Q}(p) + \mathcal{R}_\infty(p)) = \text{Extr}_{p \in \text{Im}(\eta)} (\mathcal{Q}(p) + \mathcal{R}_\infty(p)) \quad (3.166)$$

$$= \text{Extr}_{u \in \text{Im}(\eta)} (\mathcal{Q}(\eta^{-1}(u)) + \mathcal{R}_\infty(\eta^{-1}(u))) \quad (3.167)$$

where we use the inverse expression:

$$\mathcal{Q}(\eta^{-1}(u)) + \mathcal{R}_\infty(\eta^{-1}(u)) = \frac{1}{2} \sum_{il} \bar{u}_{il} [\eta^{-1}(u)]_{(il)} + \lim_n \frac{1}{n} \log \left| \det \left(L_0 + Z - \frac{\bar{u} + u^T}{2} \otimes I \right) \right| \quad (3.168)$$

3.6 Derivation of result 3.1: Three methods

Let u_* be an extremal point of $\mathcal{Q}(\eta^{-1}(u)) + \mathcal{R}_\infty(\eta^{-1}(u))$ on $\text{Im}(\eta)$. So at $u = u_*$, we expect $\frac{\partial \mathcal{Q}(\eta^{-1}(u)) + \mathcal{R}_\infty(\eta^{-1}(u))}{\partial u_{il}} = \frac{\partial \mathcal{Q}(\eta^{-1}(u)) + \mathcal{R}_\infty(\eta^{-1}(u))}{\partial \bar{u}_{il}} = 0$ for all $il \in \mathbb{S}^2$. Now, because η is a Hermitian operator, so is η^{-1} . So we can write (using carefully the properties of the Wirtinger derivatives):

$$\frac{\partial \langle u, \eta^{-1}(u) \rangle}{\partial u_{il}} = \left\langle \frac{\partial u}{\partial \bar{u}_{il}}, \eta^{-1}(u) \right\rangle + \left\langle u, \eta^{-1} \left(\frac{\partial u}{\partial u_{il}} \right) \right\rangle \quad (3.169)$$

$$= \langle 0, \eta^{-1}(u) \rangle + \langle u, \eta^{-1}(E_{il}) \rangle \quad (3.170)$$

$$= \langle \eta^{-1}(u), E_{il} \rangle = \overline{\langle E_{il}, \eta^{-1}(u) \rangle} = \overline{[\eta^{-1}(u)]_{il}} \quad (3.171)$$

And:

$$\frac{\partial \langle u, \eta^{-1}(u) \rangle}{\partial \bar{u}_{il}} = \left\langle \frac{\partial u}{\partial u_{il}}, \eta^{-1}(u) \right\rangle + \left\langle u, \eta^{-1} \left(\frac{\partial u}{\partial \bar{u}_{il}} \right) \right\rangle \quad (3.172)$$

$$= \langle E_{il}, \eta^{-1}(u) \rangle + 0 = [\eta^{-1}(u)]_{il} \quad (3.173)$$

So we find for \mathcal{Q} and for p_* given such that $u_* = \eta(p_*)$:

$$\frac{\partial \mathcal{Q} \circ \eta^{-1}}{\partial u_{il}}(u_*) = \frac{1}{2} \overline{[\eta^{-1}(u_*)]_{(il)}} = \frac{1}{2} \bar{p}_*^{(il)} \quad (3.174)$$

$$\frac{\partial \mathcal{Q} \circ \eta^{-1}}{\partial \bar{u}_{il}}(u_*) = \frac{1}{2} p_*^{(il)} = \frac{1}{2} (p_*^T)^{(li)} \quad (3.175)$$

Now for \mathcal{R}_∞ , we find:

$$\frac{\partial \mathcal{R}_\infty \circ \eta^{-1}}{\partial u_{il}}(u_*) = -\frac{1}{2} \text{Tr}_n \left[\left((L_0 + Z - \frac{\bar{u}_* + u_*^T}{2} \otimes I)^{-1} \right)^{(il)} \right] \quad (3.176)$$

$$\frac{\partial \mathcal{R}_\infty \circ \eta^{-1}}{\partial \bar{u}_{il}}(u_*) = -\frac{1}{2} \text{Tr}_n \left[\left((L_0 + Z - \frac{\bar{u}_* + u_*^T}{2} \otimes I)^{-1} \right)^{(li)} \right] \quad (3.177)$$

So in fact, we have:

$$\frac{1}{2} \bar{p}_* = \frac{1}{2} (J \otimes \text{Tr}_n) \left[\left((L_0 + Z - \eta \left(\frac{\bar{p}_* + p_*^T}{2} \right) \otimes I)^{-1} \right) \right] \quad (3.178)$$

$$\frac{1}{2} p_*^T = \frac{1}{2} (J \otimes \text{Tr}_n) \left[\left((L_0 + Z - \eta \left(\frac{\bar{p}_* + p_*^T}{2} \right) \otimes I)^{-1} \right) \right] \quad (3.179)$$

So by defining $h_* = \frac{\bar{p}_* + p_*^T}{2}$, and summing both equations we have:

$$h_* = (J \otimes \text{Tr}_n) \left[\left((L_0 + Z - \eta(h_*) \otimes I)^{-1} \right) \right] \quad (3.180)$$

Now let's assume that we can interchange the limit in n and r and have calculate the average:

$$h(Z) = \lim_{n \rightarrow +\infty} \mathbb{E}[h_n(Z)] = \lim_{r \rightarrow 0^+} \lim_{n \rightarrow +\infty} \frac{1}{r} \frac{1 - \mathbb{E}[f_n(Z)^{-r}]}{r} \quad (3.181)$$

²Because $u \in \text{Im}(\eta)$ and by definition of η in (3.3), u can only be defined on the set of indices spanned by \mathbb{S}

Chapter 3. The Linear-pencil method

Then we have $h(Z) = \mathcal{Q}(p_*(Z), Z) + \mathcal{R}_\infty(p_*(Z), Z)$. We can pursue the calculation:

$$g_{ij}(Z) = \frac{\partial h(Z)}{\partial z_{ji}} = \sum_{kl} \frac{\partial p_*^{(kl)}(Z)}{\partial z_{ji}} \frac{\partial \mathcal{Q} + \mathcal{R}_\infty}{\partial p_{kl}}(p_*(Z), Z) + \frac{\partial \mathcal{Q} + \mathcal{R}_\infty}{\partial z_{ji}}(p_*(Z), Z) \quad (3.182)$$

$$= 0 + (J \otimes \text{Tr}_n) [(L_0 + (Z - \eta(h_*(Z))) \otimes I)^{-1}]_{ij} \quad (3.183)$$

$$= h_*^{(ij)}(Z) \quad (3.184)$$

Consequently, we notice that g is self-adjoint (as is h_*). So at $Z = 0$ we retrieve result 3.1:

$$g = (J \otimes \text{Tr}_n) [(L_0 - \eta(g) \otimes I)^{-1}] \quad (3.185)$$

Remark on the non-self-adjoint case:As stated above, given a non-self-adjoint matrix L_0, W, Z such that $L = L_0 + W + Z$ is invertible, one can always construct a bigger self-adjoint matrix $\mathbf{L} = \mathbf{L}_0 + \mathbf{W} + \mathbf{Z}$ with:

$$\mathbf{L} = \begin{pmatrix} 0 & L \\ \bar{L}^T & 0 \end{pmatrix} \quad (3.186)$$

The same results as before will apply, in particular at $\mathbf{Z} = 0$ we find

$$\mathbf{g} = (J \otimes \text{Tr}_n) [(\mathbf{L}_0 - \boldsymbol{\eta}(\mathbf{g}) \otimes I)^{-1}] \quad (3.187)$$

where $\boldsymbol{\eta}$ is the operator applied for the covariance structure $\boldsymbol{\sigma}$ of \mathbf{L} . But the inverse of \mathbf{L} has the following form:

$$\mathbf{L}^{-1} = \begin{pmatrix} 0 & (\bar{L}^T)^{-1} \\ L^{-1} & 0 \end{pmatrix} \quad (3.188)$$

And consequently, \mathbf{g} is expected to be the solution of the following form:

$$\mathbf{g} = \begin{pmatrix} 0 & \bar{g}^T \\ g & 0 \end{pmatrix} \quad (3.189)$$

With careful consideration, using the operator η applied on the covariance structure σ of L , it can be shown that we have the following relation (where g and \bar{g}^T are permuted):

$$\boldsymbol{\eta}(\mathbf{g}) = \begin{pmatrix} 0 & \eta(g) \\ \eta(\bar{g}^T) & 0 \end{pmatrix} \quad (3.190)$$

Hence the formula for g when $Z = 0$ when L is not self-adjoint is still valid:

$$g = (J \otimes \text{Tr}_n) [(L_0 - \eta(g) \otimes I)^{-1}] \quad (3.191)$$

Appendix

3.A Derivation of the Laguerre polynomials

We let

$$P_{n,d}(z) = \mathbb{E} \left[\left(\frac{1}{2} U_1 + i \frac{1}{2} U_2 - z \right)^n \left(\frac{1}{2} U_1 - i \frac{1}{2} U_2 + 1 \right)^d \right] \quad (3.192)$$

We find:

$$P_{n,d}(z) = \mathbb{E} \left[\frac{\partial}{\partial t_1^n \partial t_2^d} \exp \left\{ t_1 \left(\frac{1}{2} U_1 + i \frac{1}{2} U_2 - z \right) + t_2 \left(\frac{1}{2} U_1 - i \frac{1}{2} U_2 + 1 \right) \right\} \right] \Bigg|_{t_1=t_2=0} \quad (3.193)$$

$$= \mathbb{E} \left[\frac{\partial}{\partial t_1^n \partial t_2^d} \exp \left\{ (t_2 - z t_1) + \frac{1}{2} (t_1 + t_2) U_1 + \frac{i}{2} (t_1 - t_2) U_2 \right\} \right] \Bigg|_{t_1=t_2=0} \quad (3.194)$$

$$= \frac{\partial}{\partial t_1^n \partial t_2^d} \left\{ e^{(t_2 - z t_1)} \mathbb{E} \left[e^{\frac{1}{2} (t_1 + t_2) U_1} \right] \mathbb{E} \left[e^{\frac{i}{2} (t_1 - t_2) U_2} \right] \right\} \Bigg|_{t_1=t_2=0} \quad (3.195)$$

$$= \frac{\partial}{\partial t_1^n \partial t_2^d} \left\{ e^{(t_2 - z t_1) + \frac{1}{4n} (t_1 + t_2)^2 - \frac{1}{4n} (t_1 - t_2)^2} \right\} \Bigg|_{t_1=t_2=0} \quad (3.196)$$

$$= \frac{\partial}{\partial t_1^n \partial t_2^d} \left\{ e^{(t_2 - z t_1) + \frac{1}{n} t_1 t_2} \right\} \Bigg|_{t_1=t_2=0} \quad (3.197)$$

First let's consider the case $d \geq n$:

$$P_{n,d}(z) = \frac{\partial}{\partial t_1^n} \left\{ e^{-z t_1} \frac{\partial}{\partial t_2^d} e^{\left(\frac{1}{n} t_1 + 1\right) t_2} \right\} \Bigg|_{t_1=t_2=0} = \frac{\partial}{\partial t_1^n} \left\{ \left(\frac{1}{n} t_1 + 1 \right)^d e^{-z t_1} \right\} \Bigg|_{t_1=0} \quad (3.198)$$

Chapter 3. The Linear-pencil method

Then Using Leibniz formula we find:

$$P_{n,d}(z) = \sum_{k=0}^n \binom{n}{k} (-z)^k \left[\frac{\partial}{\partial t_1^{n-k}} \left(\frac{1}{n} t_1 + 1 \right)^d \right]_{t_1=0} \quad (3.199)$$

$$= \sum_{k=0}^n \frac{(-z)^k}{k!} \frac{n!}{(n-k)!} \frac{d!}{(d-n+k)!} \frac{1}{n^{n-k}} \quad (3.200)$$

$$= \frac{n!}{n^n} \sum_{k=0}^n \binom{n+(d-n)}{n-k} \frac{(-nz)^k}{k!} \quad (3.201)$$

$$= \frac{n!}{n^n} \mathcal{L}_n^{(d-n)}(nz) \quad (3.202)$$

where we used the expression of the generalized Laguerre polynomial $\mathcal{L}_n^{(\alpha)}$ with $\alpha = d - n$.

Now when $d < n$, we use instead:

$$P_{n,d}(z) = \frac{\partial}{\partial t_2^d} \left\{ e^{t_2} \frac{\partial}{\partial t_1^n} e^{\left(\frac{1}{n} t_2 - z\right) t_1} \right\} \Bigg|_{t_1=t_2=0} = \frac{\partial}{\partial t_2^d} \left\{ \left(\frac{1}{n} t_2 - z \right)^n e^{t_2} \right\} \Bigg|_{t_2=0} \quad (3.203)$$

So with Leibniz formula we find:

$$P_{n,d}(z) = \sum_{k=0}^d \binom{d}{k} \frac{1}{n^k} (-z)^{n-k} \frac{n!}{(n-k)!} \quad (3.204)$$

$$= \frac{n!}{n^n} (-nz)^{n-d} \sum_{k=0}^d \binom{d}{k} \frac{(-nz)^{d-k}}{(d-k)!} \quad (3.205)$$

$$= \frac{n!}{n^n} (-nz)^{n-d} \mathcal{L}_d^{(0)}(nz) \quad (3.206)$$

High-dimensional estimations in **Part II** linear models

4 Linear regression estimator

This chapter offers a brief overview of the methodologies used in the subsequent Gaussian covariate model and random feature model. We begin by introducing a standard linear regression model as specified in model 1.1 in the introduction and demonstrate the computation of asymptotic results for its generalization error using the ridge regression estimator. Following this, we will explore further into the process of deriving the training curves by applying contour integration formulas.

4.1 High-dimensional test error and double descent

We consider the data matrix $X \in \mathbb{R}^{n \times d}$ with independent entries, where $X_{ij} \sim \mathcal{N}(0, \frac{1}{d})$. Additionally, we assume that the labels Y are generated by the equation $Y = X\beta^* + \xi$, where $\beta^* \in \mathbb{R}^d$ is a deterministic vector of fixed norm $\frac{1}{d} \|\beta^*\|^2 = r^2$. In this equation, the noise vector ξ is a random variable, independent of X , which follows a uniform distribution on the sphere $\xi \sim \mathcal{U}(\mathbb{S}^{d-1}(\sigma\sqrt{d}))$. It is worth noting that a more common assumption is that $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$. In fact, in the high-dimensional limit, both distributions can be considered interchangeably. For the sake of simplicity, we choose the former distribution. In this context, the ridge-regression estimator with a regularization parameter λ is given by:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T (X\beta^* + \xi) \quad (4.1)$$

Let's compute the generalization error using the classical bias-variance decomposition:

$$\mathcal{E}_{\text{gen}} = \lim_{d, n \rightarrow +\infty} \mathbb{E} \|y(x) - x^T \hat{\beta}\|^2 = \sigma^2 + \lim_{d, n \rightarrow +\infty} \mathcal{B}_X(\hat{\beta}) + \lim_{d, n \rightarrow +\infty} \mathcal{V}_X(\hat{\beta}) \quad (4.2)$$

where, as stated in the introduction in equation (1.21) and (1.24):

$$\mathcal{B}_X(\hat{\beta}) = \frac{1}{d} \mathbb{E} \left[\|\beta^* - \mathbb{E}[\hat{\beta}|X]\|^2 \right] \quad \text{and} \quad \mathcal{V}_X(\hat{\beta}) = \frac{1}{d} \mathbb{E} \left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}|X]\|^2 \right] \quad (4.3)$$

We will show that we can express the generalization error using specifically the trace of the resolvent of $X^T X$, defined as $g_{X^T X}(z) = \frac{1}{d} \text{Tr}[(X^T X - zI_d)^{-1}]$. Starting from the equation $\mathbb{E}[\hat{\beta}|X] = (X^T X + \lambda I)^{-1} X^T X \beta^*$ we can derive an expression for the average of the bias as follows:

$$\mathcal{B}_X(\hat{\beta}) = \frac{1}{d} \mathbb{E}_X \left\| (I - (X^T X + \lambda I)^{-1} X^T X) \beta^* \right\|^2 \quad (4.4)$$

$$= \lambda^2 \frac{1}{d} \mathbb{E}_X \beta^{*T} (X^T X + \lambda I)^{-2} \beta^* \quad (4.5)$$

$$= \lambda^2 r^2 \mathbb{E}_X \left[\frac{1}{d} \text{Tr}[(X^T X + \lambda I)^{-2}] \right] \quad (4.6)$$

Note that, as the random matrix $X^T X$ is invariant under rotation, the direction of the vector β^* has no effect in the preceding equation. Therefore, it can be considered deterministic with a fixed norm. As for the variance, we have:

$$\mathcal{V}_X(\hat{\beta}) = \frac{1}{d} \mathbb{E}_X \left\| ((X^T X + \lambda I)^{-1} X^T \xi) \right\|^2 \quad (4.7)$$

$$= \frac{1}{d} \mathbb{E}_X \xi^T X (X^T X + \lambda I)^{-2} X^T \xi \quad (4.8)$$

$$= \mathbb{E}_X \sigma^2 \frac{1}{d} \text{Tr}[(X^T X + \lambda I)^{-1} - \lambda (X^T X + \lambda I)^{-2}] \quad (4.9)$$

In conclusion we have found:

$$\mathcal{B}_X(\hat{\beta}) = \lambda^2 r^2 \mathbb{E}_X g'_{X^T X}(-\lambda) \quad (4.10)$$

$$\mathcal{V}_X(\hat{\beta}) = \sigma^2 (\mathbb{E}_X g_{X^T X}(-\lambda) - \lambda \mathbb{E}_X g'_{X^T X}(-\lambda)) \quad (4.11)$$

At this point, there remains to calculate the expected value of the Stieltjes-transform of $X^T X$. In the limit of large d , the Stieltjes-transform of $X^T X$ approaches that of the Marchenko-Pastur distribution. In the next section, we will utilize this fact and the properties of this Stieltjes transform to pursue our analysis.

4.2 Training and test error in the high-dimensional limit

In the high-dimensional limit, we can use the Marchenko-Pastur distribution to compute the limiting value of the Stieltjes-transform of $X^T X$ with the aspect ratio $\phi = \frac{n}{d}$. The Marchenko-Pastur distribution is defined as (see Chapter 3):

$$z g^2(z) + (1 + z - \phi) g(z) + 1 = 0 \quad (4.12)$$

Given the bias and variance formula (4.10) and (4.11), we need to evaluate the expression at $z = -\lambda$ (note this value is outside of the spectrum of $X^T X$ and thus $g(z)$ is well defined at this point), and also calculate the derivative g' . So let $f = g(-\lambda)$ and $h = g'(-\lambda)$, by calculating the

4.2 Training and test error in the high-dimensional limit

derivative of the Marchenko-Pastur law, we get the system of two algebraic equations:

$$-\lambda f^2 + (1 - \lambda - \phi) f + 1 = 0 \quad (4.13)$$

$$f^2 - 2\lambda f h + f + (1 - \lambda - \phi) h = 0 \quad (4.14)$$

With these two equations, it is possible to express the value of f , and then that of h and derive the asymptotic limit in $d \rightarrow \infty$ of the bias and variance. However, for the sake of illustration, we will instead consider another approach where we will derive directly two closed-form algebraic equations for the bias and the variance. To this end, let's define b and v the limiting values of the bias and variance, using (4.10) and (4.11) we have:

$$b - \lambda^2 r^2 h = 0 \quad (4.15)$$

$$v - \sigma^2 (f - \lambda h) = 0 \quad (4.16)$$

The method now consists in computing a reduced Gröbner basis (Buchberger, 1965) in the polynomial ring $\mathbb{R}[f, h, v, b]$ with these four equations. Using an appropriate order with the application of the Buchberger algorithm which is illustrated at the end of this paragraph, we find the closed-form equation for b :

$$(\lambda^2 + 2\lambda\phi + 2\lambda + (\phi - 1)^2)(b + r^2(\phi - 1))b - \lambda^2\phi r^4 = 0 \quad (4.17)$$

Hence when $\lambda = 0$, the equation reduces to:

$$(\phi - 1)^2 (b + r^2(\phi - 1))b = 0 \quad (4.18)$$

So either $b = 0$ (when $\phi > 1$) or $b = r^2(1 - \phi)$ (when $\phi < 1$). In the same way, we can find a quadratic equation for v :

$$(\lambda^2 + 2\lambda\phi + 2\lambda + (\phi - 1)^2)(v + \sigma^2)v - \phi\sigma^4 = 0 \quad (4.19)$$

This equation is further simplified when $\lambda = 0$:

$$(\phi - 1)^2 (v + \sigma^2)v - \phi\sigma^4 = 0 \quad (4.20)$$

So when $\lambda = 0$ we find two solutions:

$$v = \frac{-\sigma^2(\phi - 1)^2 \pm \sqrt{\sigma^4(\phi - 1)^4 + 4\phi\sigma^4(\phi - 1)^2}}{2(\phi - 1)^2} = \sigma^2 \frac{\pm(\phi + 1) - (\phi - 1)}{2(\phi - 1)} \quad (4.21)$$

Hence we find that $v = \frac{\sigma^2}{\phi - 1}$ when $\phi > 1$ or $v = \frac{\sigma^2\phi}{1 - \phi}$ when $\phi < 1$. All in all, we have:

$$\mathcal{E}_{\text{gen}}(r, \sigma, \phi, \lambda = 0) = \begin{cases} \sigma^2 + \sigma^2 \frac{1}{\phi - 1} & \text{if } \phi > 1 \\ \sigma^2 + r^2(1 - \phi) + \sigma^2 \frac{\phi}{1 - \phi} & \text{if } \phi < 1 \end{cases} \quad (4.22)$$

Chapter 4. Linear regression estimator

Note that we find back the results of Theorem 1 from (Hastie et al., 2019) when we take $\gamma = \frac{1}{\phi}$. Importantly, this result displays the double-descent phenomena - albeit in a simple situation - where the test error diverges in the neighborhood of $\phi = 1$.

In the following chapters (in particular Chapter 6), we will use these methods more extensively in order to reduce the number of equations. This can be accomplished automatically with the use of a Computer Algebra System. As an example, we can use SymPy (Meurer et al., 2017) in python:

```

from sympy import *
print(__version__)

1.12

f,h,lam,b,v,r,sig,phi = symbols("f,h,lambda,b,v,r,sigma,phi")

eq1 = (-lam)*f**2 + (1-lam-phi)*f + 1
eq2 = f**2 - 2*lam*f*h + f + (1 - lam - phi)*h

eqb = b - (lam**2)*(r**2)*h
eqv = v - (sig**2)*(f-lam*h)

EQ_List = Matrix([
    eq1, eq2, eqb, eqv
])
EQ_List


$$\begin{bmatrix} -f^2\lambda + f(-\lambda - \phi + 1) + 1 \\ f^2 - 2fh\lambda + f + h(-\lambda - \phi + 1) \\ b - h\lambda^2r^2 \\ -\sigma^2(f - h\lambda) + v \end{bmatrix}$$


# change the order of the generators
res_b = groebner(EQ_List, [f,h,v,b])
res_v = groebner(EQ_List, [f,h,b,v])

# Equation for the bias
res_b[-1]


$$b^2(\lambda^2 + 2\lambda\phi + 2\lambda + \phi^2 - 2\phi + 1) + b(\lambda^2\phi r^2 - \lambda^2r^2 + 2\lambda\phi^2r^2 - 2\lambda r^2 + \phi^3r^2 - 3\phi^2r^2 + 3\phi r^2 - r^2) - \lambda^2\phi r^4$$


# Equation for the variance
res_v[-1]

```


$$-\phi\sigma^4 + v^2(\lambda^2 + 2\lambda\phi + 2\lambda + \phi^2 - 2\phi + 1) + v(\lambda^2\sigma^2 + 2\lambda\phi\sigma^2 + 2\lambda\sigma^2 + \phi^2\sigma^2 - 2\phi\sigma^2 + \sigma^2)$$

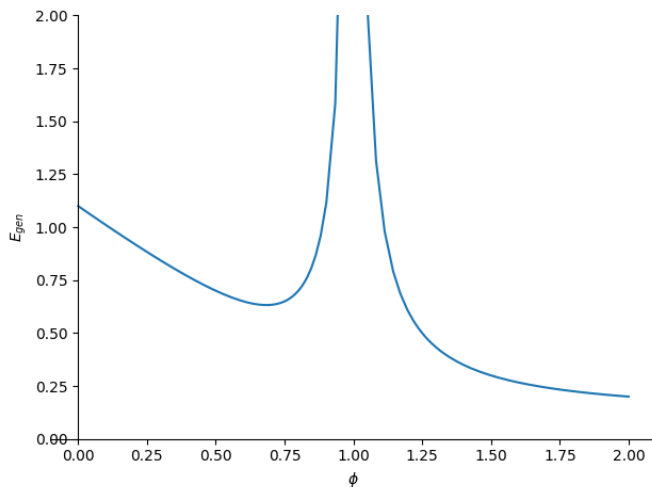
```
# solution for the bias when $\lambda = 0$
Matrix(solve(res_b[-1].subs(lam,0), b))
```

$$\begin{bmatrix} 0 \\ r^2 \cdot (1 - \phi) \end{bmatrix}$$

```
# solution for the variance when $\lambda = 0$
Matrix(solve(res_v[-1].subs(lam,0), v))
```

$$\begin{bmatrix} \frac{\sigma^2}{\phi-1} \\ -\frac{\phi\sigma^2}{\phi-1} \end{bmatrix}$$

```
# a plot for $r=1, \sigma^2 = 0.1$
plot(0.1 + Max(0,1-phi) + 0.1*Max(1/(phi-1),-phi/(phi-1)), (phi, 0, 2), ylim=(0.,2.),
      ylabel="$E_{gen}$")
```



4.3 Time evolution and learning curves

In this section, we explain how we can derive the evolution of the bias and variance in time, thereby following the ideas of (Advani et al., 2020a) but using complex contours of integration as a tool to express the evolution of the bias and variance. We will use the same model as before, but with the gradient-flow algorithm with the training error defined as:

$$\mathcal{E}_{\text{train}}^\lambda(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \tag{4.23}$$

Chapter 4. Linear regression estimator

Using the gradient-flow formula, we find a differential equation for β_t :

$$\frac{\partial \beta_t}{\partial t} = -\nabla_{\beta} \mathcal{E}_{\text{train}}^{\lambda}(\beta_t) = X^T(Y - X\beta_t) - \lambda\beta_t \quad (4.24)$$

In the following, we will assume that we choose an initial vector β_0 sampled on the hypersphere independently of β^* and X , with a norm $\|\beta_0\|^2 = r_0^2 d$. We can express β_t as follows:

$$\beta_t = \left(I - e^{-(X^T X + \lambda I)t} \right) (X^T X + \lambda I)^{-1} X^T Y + e^{-(X^T X + \lambda I)t} \beta_0 \quad (4.25)$$

When β_0 is a random variable, we can use another formulation in equation (4.2) with a conditional bias and conditional variance, as explained in Section 1.5 in the introduction:

$$\mathcal{B}_{X, \beta_0}(\beta_t) = \frac{1}{d} \mathbb{E} \|\beta^* - \mathbb{E}[\beta_t | X, \beta_0]\|^2 \quad (4.26)$$

$$\mathcal{V}_{X, \beta_0}(\beta_t) = \frac{1}{d} \mathbb{E} \|\beta_t - \mathbb{E}[\beta_t | X, \beta_0]\|^2 \quad (4.27)$$

The bias term can be decomposed into smaller terms:

$$\mathcal{B}_{X, \beta_0}(\beta_t) = \frac{1}{d} \mathbb{E} \|\beta^* - \mathbb{E}[\beta_t | X]\|^2 + \frac{1}{d} \mathbb{E} \|\mathbb{E}[\beta_t | X] - \mathbb{E}[\beta_t | X, \beta_0]\|^2 \quad (4.28)$$

In order to use the contour integration formula, let's define a contour $\Gamma \subset \mathbb{C}$ which encloses the spectrum of $X^T X$ without containing the point $-\lambda$. Each term can be expanded:

$$\frac{1}{d} \mathbb{E} \|\beta^* - \mathbb{E}[\beta_t | X]\|^2 = \frac{1}{d} \mathbb{E} \left\| \left\{ I - \left(I - e^{-t(X^T X + \lambda I)} \right) (X^T X + \lambda I)^{-1} X^T X \right\} \beta^* \right\|^2 \quad (4.29)$$

$$= \frac{-1}{2i\pi} \frac{1}{d} \|\beta^*\|^2 \oint_{\Gamma} \left(1 - z \frac{1 - e^{-t(z+\lambda)}}{z + \lambda} \right)^2 \mathbb{E}_X \text{Tr}[(X^T X - zI)^{-1}] dz \quad (4.30)$$

$$= \frac{-r^2}{2i\pi} \oint_{\Gamma} \left(\frac{\lambda + z e^{-t(z+\lambda)}}{z + \lambda} \right)^2 \mathbb{E}_X g_{X^T X}(z) dz \quad (4.31)$$

Which, in the limit $t \rightarrow \infty$, corresponds to equation (4.4) as expected. Note that the minus sign comes from the counter-clockwise integration on the complex contour Γ .

The second term yields an additional error term coming from the initialization:

$$\frac{1}{d} \mathbb{E} \|\mathbb{E}[\beta_t | X] - \mathbb{E}[\beta_t | X, \beta_0]\|^2 = -\frac{r_0^2}{2i\pi} \oint_{\Gamma} e^{-2t(z+\lambda)} \mathbb{E}_X g_{X^T X}(z) dz \quad (4.32)$$

Similarly for the variance, because we fixed $\frac{1}{d} \|\xi\|^2 = \sigma^2$, we get:

$$\mathcal{V}_{X, \beta_0}(\beta_t) = \frac{-\sigma^2}{2i\pi} \oint_{\Gamma} z \left(\frac{1 - e^{-t(z+\lambda)}}{z + \lambda} \right)^2 \mathbb{E}_X g_{X^T X}(z) dz \quad (4.33)$$

We suppose that in the limit $d \rightarrow \infty$, we can take $\lim_{d \rightarrow \infty} \mathbb{E}_X g_{X^T X}(z) = g(z)$ and:

$$\mathcal{E}_{\text{gen}}(t) = \sigma^2 + \frac{-1}{2i\pi} \oint_{\Gamma} \left\{ r^2 \left(\frac{\lambda + ze^{-t(z+\lambda)}}{z + \lambda} \right)^2 + r_0^2 e^{-2t(z+\lambda)} + \sigma^2 z \left(\frac{1 - e^{-t(z+\lambda)}}{z + \lambda} \right)^2 \right\} g(z) dz \quad (4.34)$$

Thus, we have an analytic formula that can be computed numerically to determine the generalization error at any time t for any set of parameters $(r, r_0, \sigma, \phi, \lambda)$.

The analytical approach developed here will serve as a foundation for the methods explored in the remaining chapters for more advanced models.

5 A framework: the gaussian covariate model

A recent line of work has shown remarkable behaviors of the generalization error curves in simple learning models. Even the least-squares regression has shown atypical features such as the model-wise double descent, and further works have observed triple or multiple descents. Another important characteristic are the epoch-wise descent structures which emerge during training. The observations of model-wise and epoch-wise descents have been analytically derived in limited theoretical settings (such as the random feature model) and are otherwise experimental. In this work which is based on the work (Bodin and Macris, 2022), we leverage the model 1.2 from the introduction to provide a full and unified analysis of the whole time-evolution of the generalization curve, in the asymptotic large-dimensional regime and under gradient-flow, within a wider theoretical setting stemming from a gaussian covariate model. In particular, we cover most cases already disparately observed in the literature, and also provide examples of the existence of multiple descent structures as a function of a model parameter or time. Furthermore, we show that our theoretical predictions adequately match the learning curves obtained by gradient descent over realistic datasets. Technically we compute averages of rational expressions involving random matrices using recent developments in random matrix theory based on "linear pencils" as described in Chapter 3.

5.1 Introduction

5.1.1 Preliminaries

With growing computational resources, it has become customary for machine learning models to use a *huge* number of parameters (billions of parameters in Brown et al. (2020)), and the need for scaling laws has become of utmost importance Hoffmann et al. (2022). Therefore it is of great relevance to study the asymptotic (or "thermodynamic") limit of simple models in which the number of parameters and data samples are sent to infinity. A landmark progress made by considering these theoretical limits, is the analytical (oftentimes rigorous) calculation of precise double-descent curves for the generalization error starting with Belkin et al. (2020a); Hastie et al. (2019); Mei and Montanari (2019), Advani et al. (2020a), d'Ascoli et al. (2020),

Gerace et al. (2020a), Deng et al. (2021a), Kini and Thrampoulidis (2020) confirming in a precise (albeit limited) theoretical setting the experimental phenomenon initially observed in Belkin et al. (2019b), Geiger et al. (2019); Spigler et al. (2019b), Nakkiran et al. (2020a). Further derivations of triple or even multiple descents for the generalization error have also been performed d’Ascoli et al. (2020); Nakkiran et al. (2020b); Chen et al. (2021); Richards et al. (2021); Wu and Xu (2020). Other aspects of multiples descents have been explored in Lin and Dobriban (2021); Adlam and Pennington (2020b) also for the Neural tangent kernel in Adlam and Pennington (2020a). The tools in use come from modern random matrix theory Pennington and Worah (2017); Rashidi Far et al. (2006); Mingo and Speicher (2017), and statistical physics methods such as the replica method Engel and Van den Broeck (2001a).

In this chapter we are concerned with a line of research dedicated to the precise *time-evolution* of the generalization error under gradient flow corroborating, among other things, the presence of epoch-wise descents structures Crisanti and Sompolinsky (2018); Bodin and Macris (2021a) observed in Nakkiran et al. (2020a). We consider the gradient flow dynamics for the training and generalisation errors in the setting of a Gaussian Covariate model, and develop analytical methods to track the whole time evolution. In particular, for infinite times we get back the predictions of the *least square estimator* which have been thoroughly described in a similar model by Loureiro et al. (2021).

In the next paragraphs we set-up the model together with a list of special realizations, and describe our main contributions.

5.1.2 Model description

Generative Data Model: In this chapter, we use the so-called Gaussian Covariate model in a teacher-student setting. An observation in our data model is defined through the realization of a gaussian vector $z \sim \mathcal{N}(0, \frac{1}{d}I_d)$. The teacher and the student obtain their observations (or two different views of the world) with the vectors $x \in \mathbb{R}^{p_B}$ and $\hat{x} \in \mathbb{R}^{p_A}$ respectively, which are given by the application of two linear operations on z . In other words there exists two matrices $B \in \mathbb{R}^{d \times p_B}$ and $A \in \mathbb{R}^{d \times p_A}$ such that $x = B^T z$ and $\hat{x} = A^T z$. Note that the generated data can also be seen as the output of a generative 1-layer linear network. In the following, the structure of A and B is pretty general as long as it remains independent of the realization z : the matrices may be random matrices or block-matrices of different natures and structures to capture more sophisticated models. While the models we treat are defined through appropriate A and B , we will often only need the structure of $U = AA^T$ and $V = BB^T$.

A direct connection can be made with the Gaussian Covariate model described in Loureiro et al. (2021) which suggests considering directly observations $\bar{x} = (x^T, \hat{x}^T)^T \sim \mathcal{N}(0, \Sigma)$ for a given covariance structure Σ . The spectral theorem provides the existence of orthonormal matrix O and diagonal D such that $\Sigma = O^T D O$ and D contains d non-zero eigenvalues in a squared block D_1 and $p_A + p_B - d$ zero eigenvalues. We can write $D = J^T D_1 J$ with $J = (I_d | 0_{p_A + p_B - d})$. Therefore if we let $z = \frac{1}{\sqrt{d}} D_1^{-\frac{1}{2}} J O \bar{x}$ which has variance $\frac{1}{d} I_d$, then upon noticing $J J^T = I_d$ and

defining $(A|B)^T = \sqrt{d}O^T J^T D_1^{\frac{1}{2}}$ we find $(A|B)^T z \sim \mathcal{N}(0, \Sigma)$.

The Gaussian Covariate model unifies many different models as shown in Table 5.1.1. These special cases are all discussed in Section 5.3 and Appendix 5.D

Table 5.1.1: Different matrices and corresponding models

Target Matrix B	Estimator Matrix A	Corresponding Model
$\begin{pmatrix} r\sqrt{\frac{d}{p}}I_p & 0 \\ 0 & \sigma\sqrt{\frac{d}{q}}I_q \end{pmatrix}$	$\begin{pmatrix} \sqrt{\frac{d}{p}}I_p \\ O_{q \times p} \end{pmatrix}$	Ridgeless regression with signal r and noise σ
$\begin{pmatrix} \sqrt{\frac{r^2 d}{p}}I_{\gamma p} & 0 & 0 \\ 0 & \sqrt{\frac{r^2 d}{p}}I_{\gamma' p} & 0 \\ 0 & 0 & \sqrt{\frac{\sigma^2 d}{q}}I_q \end{pmatrix}$	$\begin{pmatrix} \sqrt{\frac{d}{\gamma p}}I_{\gamma p} \\ O_{(1-\gamma)p \times \gamma d} \\ O_{q \times \gamma d} \end{pmatrix}$	Mismatched ridgeless regression with signal r and noise σ and mismatch parameter γ with $\gamma + \gamma' = 1$
$\begin{pmatrix} I_{\gamma d} & 0 & 0 & 0 \\ 0 & I_{\gamma d} & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & I_{\gamma d} \end{pmatrix}$	$\begin{pmatrix} \frac{1}{\alpha^0}I_{\gamma d} & 0 & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & \frac{1}{\alpha^{\frac{p-1}{2}}}I_{\gamma d} \end{pmatrix}$	non-isotropic ridgeless regression noiseless with a α polynomial distortion of the inputs scalings
$\begin{pmatrix} r\sqrt{\frac{d}{p}}I_p & O_{p \times q} \\ O_{N \times p} & O_{N \times q} \\ O_{q \times p} & \sigma\sqrt{\frac{d}{q}}I_q \end{pmatrix}$	$\begin{pmatrix} \mu\sqrt{\frac{d}{p}}W \\ \nu\sqrt{\frac{d}{p}}I_N \\ O_{q \times N} \end{pmatrix}$	Random feature regression of a noisy linear function with W the random weights and (μ, ν) describing a non-linear activation function
$\begin{pmatrix} \sqrt{\omega_1} & 0 & \dots & 0 \\ 0 & \sqrt{\omega_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\omega_d} \end{pmatrix}$	$\begin{pmatrix} \sqrt{\omega_1} & 0 & \dots & 0 \\ 0 & \sqrt{\omega_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\omega_d} \end{pmatrix}$	Further Kernel methods

Learning task: We consider the problem of learning a linear teacher function $f_d(x) = \beta^{*T} x$ with x and \hat{x} sampled as defined above, and with $\beta^* \in \mathbb{R}^p$ a column vectors. This hidden vector β^* (to be learned) can potentially be a deterministic vector. We suppose that we have n data-points $(z_i, y_i)_{1 \leq i \leq n}$ with $x_i = Bz_i, \hat{x}_i = Az_i$. This data can be represented as the $n \times d$ matrix $Z \in \mathbb{R}^{n \times d}$ where z_i^T is the i -th row of Z , and the column vector $Y \in \mathbb{R}^n$ with i -th entry y_i . Therefore, we have the matrix notation $Y = ZB\beta^*$. We can also set $X = ZB$ so that $Y = X\beta^*$.

In the same spirit, we define the estimator of the student $\hat{y}_\beta(z) = \beta^T x = z^T A\beta$. We note that in general the dimensions of β and β^* (i.e., p_A and p_B) are not necessarily equal as this depends on the matrices B and A . We have $\hat{Y} = Z A \beta = \hat{X} \beta$ for $\hat{X} = Z A$.

Training and test error: We will consider the training error $\mathcal{E}_{\text{train}}^\lambda$ and test errors \mathcal{E}_{gen} with a regularization coefficient $\lambda \in \mathbb{R}_+^*$ defined as

$$\mathcal{E}_{\text{train}}^\lambda(\beta) = \frac{1}{n} \|\hat{Y} - Y\|_2^2 + \frac{\lambda}{n} \|\beta\|_2^2, \quad \mathcal{E}_{\text{gen}}(\beta) = \mathbb{E}_{z \sim \mathcal{N}(0, \frac{I_d}{d})} [(z^T A \beta - z^T B \beta^*)^2] \quad (5.1)$$

It is well known that the least-squares estimator $\hat{\beta} = \text{argmin} \mathcal{E}_{\text{train}}^\lambda(\beta)$ is given by the Thikonov regression formula $\hat{\beta}^\lambda = (\hat{X}^T \hat{X} + \lambda I)^{-1} \hat{X}^T Y$ and that in the limit $\lambda \rightarrow 0$, this estimator converges towards the $\hat{\beta}^0$ given by the Moore-Penrose inverse $\hat{\beta}^0 = (\hat{X}^T \hat{X})^+ \hat{X}^T Y$.

Gradient-flow: We use the gradient-flow algorithm to explore the evolution of the test error through time with $\frac{\partial \beta_t}{\partial t} = -\frac{n}{2} \nabla_{\beta} \mathcal{E}_{\text{train}}^\lambda(\beta_t)$. In practice, for numerical calculations we use the discrete-time version, gradient-descent, which is known to converge towards the aforementioned least-squares estimator provided a sufficiently small time-step (in the order of $\frac{1}{\lambda_{\text{max}}}$ where λ_{max} is the maximum eigenvalue of $\hat{X}^T \hat{X}$). The upfront coefficient n on the gradient is used so that the test error scales with the dimension of the model and allows for considering the evolution in the limit $n, d, p_A, p_B \rightarrow +\infty$ with a fixed ratios $\frac{n}{d}, \frac{p_A}{d}, \frac{p_B}{d}$. We will note $\phi = \frac{n}{d}$.

5.1.3 Contributions

1. We provide a *general unified framework* covering multiple models in which we derive, in the asymptotic large size regime, the *full time-evolution* under gradient flow dynamics of the training and generalization errors for teacher-student settings. In particular, in the infinite time-limit we check that our equations reduce to those of Loureiro et al. (2021) (as should be expected). But with our results we now have the possibility to explore quantitatively potential advantages of different stopping times: indeed our formalism allows to compute the time derivative of the generalization curve at any point in time.
2. Various special cases are illustrated in Section 5.3, and among these a simpler re-derivation of the whole dynamics of the random feature model Bodin and Macris (2021a), the full dynamics for kernel methods, and situations exhibiting multiple descent curves both as a function of model parameters and time (See Section 5.3.2 and Appendix 5.D.2). In particular, our analysis allows to design multiple descents with respect to the training epochs.
3. We show that our equations can also capture the learning curves over realistic datasets such as MNIST with gradient descent (See Section 5.3.4 and Appendix 5.D.5), extending further the results of Loureiro et al. (2021) to the time dependence of the curves. This could be an interesting guideline for deriving scaling laws for large learning models.
4. We use modern random matrix techniques, namely an improved version of the linear-pencil method - recently introduced in the machine learning community by Adlam et al. (2019) - to derive asymptotic limits of traces of rational expressions involving random matrices. We refer to Chapter 3 for more details about the methods and the fixed point equation that will be employed throughout this work.

Notations: We will use $\text{Tr}_d[\cdot] \equiv \lim_{d \rightarrow +\infty} \frac{1}{d} \text{Tr}[\cdot]$ and similarly for $\text{Tr}_n[\cdot]$. We also occasionally use $N_d(v) = \lim_{d \rightarrow +\infty} \frac{1}{d} \|v\|_2$ for a vector v (when the limit exists).

5.2 Main results

We resort to the high-dimensional assumptions (see Bodin and Macris (2021a) for similar assumptions).

Assumptions 5.1 (High-Dimensional assumptions). *In the high-dimensional limit, i.e, when $d \rightarrow +\infty$ with all ratios $\frac{n}{d}, \frac{p_A}{d}, \frac{p_B}{d}$ fixed, we assume the following*

1. All the traces $\text{Tr}_d[\cdot], \text{Tr}_n[\cdot]$ concentrate on a deterministic value.
2. There exists a sequence of complex contours $\Gamma_d \subset \mathbb{C}$ enclosing the eigenvalues of the random matrix $\hat{X}^T \hat{X} \in \mathbb{R}^{d \times d}$ but not enclosing $-\lambda$, and there exist also a fixed contour Γ enclosing the support of the limiting (when $d \rightarrow +\infty$) eigenvalue distribution of $\hat{X}^T \hat{X}$ but not enclosing $-\lambda$.

With these assumptions in mind, we derive the precise time evolution of the test error in the high-dimensional limit (see result 5.1) and similarly for the training error (see result 5.4). We will also assume that the results are still valid in the case $\lambda = 0$ as suggested in Mei and Montanari (2019).

5.2.1 Time evolution formula for the test error

Result 5.1. *The limiting test error time evolution for a random initialization β_0 such that $N_d(\beta_0) = r_0$ and $\mathbb{E}[\beta_0] = 0$ is given by the following expression:*

$$\bar{\mathcal{E}}_{\text{gen}}(t) = c_0 + r_0^2 \mathcal{B}_0(t) + \mathcal{B}_1(t) \quad (5.2)$$

with $V^* = B\beta^* \beta^{*T} B^T$ and $c_0 = \text{Tr}_d[V^*]$ and:

$$\mathcal{B}_1(t) = \frac{-1}{4\pi^2} \oint_{\Gamma} \oint_{\Gamma} \frac{(1 - e^{-t(x+\lambda)})(1 - e^{-t(y+\lambda)})}{(x+\lambda)(y+\lambda)} f_1(x, y) dx dy + \frac{1}{i\pi} \oint_{\Gamma} \frac{1 - e^{-t(z+\lambda)}}{z+\lambda} f_2(z) dz \quad (5.3)$$

$$\mathcal{B}_0(t) = \frac{-1}{2i\pi} \oint_{\Gamma} e^{-2t(z+\lambda)} f_0(z) dz \quad (5.4)$$

where $f_1(x, y) = f_2(x) + f_2(y) + \tilde{f}_1(x, y) - c_0$ and:

$$\tilde{f}_1(x, y) = \text{Tr}_d[(\phi U + \zeta_x I)^{-1} (\zeta_x \zeta_y V^* + \tilde{f}_1(x, y) \phi U^2) (\phi U + \zeta_y I)^{-1}] \quad (5.5)$$

$$f_2(z) = c_0 - \text{Tr}_d[\zeta_z V^* (\phi U + \zeta_z I)^{-1}] \quad (5.6)$$

$$f_0(z) = -\left(1 + \frac{\zeta_z}{z}\right) \quad (5.7)$$

and ζ_z given by the self-consistent equation:

$$\zeta_z = -z + \text{Tr}_d [\zeta_z U (\phi U + \zeta_z I)^{-1}] \quad (5.8)$$

The former result can be expressed in terms of expectations w.r.t the joint limiting eigenvalue distributions of U and V^* when they commute with each other.

Result 5.2. Besides, when U and V^* commute, let u, v^* be jointly-distributed according to U and V^* eigenvalues respectively. Then:

$$\tilde{f}_1(x, y) = \mathbb{E}_{u, v^*} \left[\frac{\zeta_x \zeta_y v^* + \tilde{f}_1(x, y) \phi u^2}{(\phi u + \zeta_x)(\phi u + \zeta_y)} \right], \quad f_2(z) = c_0 - \mathbb{E}_{u, v^*} \left[\frac{\zeta_z v^*}{\phi u + \zeta_z} \right] \quad (5.9)$$

$$\zeta_z = -z + \mathbb{E}_u \left[\frac{\zeta_z u}{\phi u + \zeta_z} \right] \quad (5.10)$$

Notice also that in the limit $t \rightarrow \infty$:

$$\mathcal{B}_1(+\infty) = f_1(-\lambda, -\lambda) - 2f_2(-\lambda) = \tilde{f}_1(-\lambda, -\lambda) - c_0, \quad \mathcal{B}_0(+\infty) = 0 \quad (5.11)$$

which leads to the next result.

Result 5.3. In the limit $t \rightarrow \infty$, the limiting test error is given by $\bar{\mathcal{E}}_{\text{gen}}(+\infty) = \tilde{f}_1(-\lambda, -\lambda)$.

Remark 1 Notice that the matrix V^* is of rank one depending on the hidden vector β^* . However, it is also possible to calculate the average generalization (and training) error over a prior distribution $\beta^* \sim \mathcal{P}^*$. Averaging $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [\bar{\mathcal{E}}_{\text{gen}}]$ propagates the expectation within $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [\mathcal{B}_0(t)]$ and $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [\mathcal{B}_1(t)]$, which propagates it further into the traces of $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [\tilde{f}_1]$ and $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [f_2]$. In fact we find:

$$\mathbb{E}_{\mathcal{P}^*} [\tilde{f}_1(x, y)] = \text{Tr}_d [(\phi U + \zeta_x I)^{-1} (\zeta_x \zeta_y \mathbb{E}_{\mathcal{P}^*} [V^*] + \mathbb{E}_{\mathcal{P}^*} [\tilde{f}_1(x, y)] \phi U^2) (\phi U + \zeta_y I)^{-1}] \quad (5.12)$$

$$\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [f_2(z)] = c_0 - \text{Tr}_d [\zeta_z \mathbb{E}_{\mathcal{P}^*} [V^*] (\phi U + \zeta_z I)^{-1}] \quad (5.13)$$

In conclusion, we find that $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [\bar{\mathcal{E}}_{\text{gen}}]$ follows the same equations as $\bar{\mathcal{E}}_{\text{gen}}$ in result 5.1 with $\mathbb{E}_{\beta^* \sim \mathcal{P}^*} [V^*]$ instead of V^* . In the following, we will consider V^* without any distinction whether it comes from a specific vector β^* or averaged through a sample distribution \mathcal{P}^* .

Remark 2 In the particular case where U is diagonal, the matrix V^* can be replaced by the following diagonal matrix \tilde{V}^* which, in fact, commutes with U :

$$\tilde{V}^* = \begin{pmatrix} [V^*]_{11}[\beta^*]_1^2 & 0 & \dots & 0 \\ 0 & [V^*]_{22}[\beta^*]_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & [V^*]_{dd}[\beta^*]_d^2 \end{pmatrix} \quad (5.14)$$

This comes essentially from the fact that given a diagonal matrix D and a non-diagonal matrix A , then $[DA]_{ii} = [D]_{ii}[A]_{ii}$. This is particularly helpful, and shows that in many cases the calculations of \tilde{f}_1 or f_2 remain tractable even for a deterministic β^* (see the example in Appendix 5.D.3).

Remark 3 Sometimes $U = AA^T$ and $V = BB^T$ are more difficult to handle than their dual counterparts $U_\star = \phi A^T A$ and $V_\star = \phi B^T B$ together with the additional matrix $\Xi = \phi A^T B$. The following expressions are thus very useful (See Appendix 5.C):

$$f_1(x, y) = \text{Tr}_n \left[(U_\star + \zeta_x I)^{-1} (\Xi \beta^* \beta^{*T} \Xi^T) + \tilde{f}_1(x, y) U_\star (U_\star + \zeta_y I)^{-1} \right] \quad (5.15)$$

$$f_2(z) = \text{Tr}_n \left[(\Xi \beta^* \beta^{*T} \Xi^T) (U_\star + \zeta_z I)^{-1} \right] \quad (5.16)$$

$$\zeta_z = -z + \text{Tr}_n \left[\zeta_z U_\star (U_\star + \zeta_z I)^{-1} \right] \quad (5.17)$$

In fact, when $x = y = -\lambda$ (which corresponds to the limit when $t \rightarrow \infty$), these are the same expressions as (59) in Loureiro et al. (2021) with the appropriate change of variable $\lambda(1+V) \rightarrow \zeta$ and $\tilde{f}_1 \rightarrow \rho + q - 2m$.

5.2.2 Time evolution formula for the training error

Result 5.4. *The limiting training error time evolution is given by the following expression:*

$$\bar{\mathcal{E}}_{train}^0(t) = c_0 + r_0^2 \mathcal{H}_0(t) + \mathcal{H}_1(t) \quad (5.18)$$

with:

$$\mathcal{H}_1(t) = \frac{-1}{4\pi^2} \oint_{\Gamma} \oint_{\Gamma} \frac{(1 - e^{-t(x+\lambda)})(1 - e^{-t(y+\lambda)})}{(x+\lambda)(y+\lambda)} h_1(x, y) dx dy + \frac{1}{i\pi} \oint_{\Gamma} \frac{1 - e^{-t(z+\lambda)}}{z+\lambda} h_2(z) dz \quad (5.19)$$

$$\mathcal{H}_0(t) = \frac{-1}{2i\pi} \oint_{\Gamma} e^{-2t(z+\lambda)} h_0(z) dz \quad (5.20)$$

where $h_1(x, y) = h_2(x) + h_2(y) + \tilde{h}_1(x, y) - c_0$ and with $\eta_z = \frac{-z}{\zeta_z}$:

$$\tilde{h}_1(x, y) = \eta_x \eta_y \tilde{f}_1(x, y), \quad h_2(z) = \eta_z (c_0 f_0(z) + f_2(z)), \quad h_0(z) = \eta_z f_0(z) \quad (5.21)$$

Eventually, in the limit $t \rightarrow \infty$ we find:

$$\mathcal{H}_1(+\infty) = h_1(-\lambda, -\lambda) - 2h_2(-\lambda) = \tilde{h}_1(-\lambda, -\lambda) - c_0, \quad \mathcal{H}_0(+\infty) = 0 \quad (5.22)$$

Result 5.5. *In the limit $t \rightarrow \infty$, we have the relation $\bar{\mathcal{E}}_{train}^0(+\infty) = \eta_{-\lambda}^2 \bar{\mathcal{E}}_{gen}^0(+\infty)$*

We notice the same proportionality factor $\eta_{-\lambda}^2 = \left(\frac{\lambda}{\zeta(-\lambda)}\right)^2$ as already stated in Loureiro et al. (2021), however interestingly, in the time evolution of the training error, such a factor is not valid as we have $h_2(z) \neq \eta_z f_2(z)$.

5.3 Applications and examples

We discuss some of the models provided in table 5.1.1 and some others in Appendix 5.D.

5.3.1 Ridgeless regression of a noisy linear function

Target function Consider the following noisy linear function $y(x) = rx^T \beta_0^* + \sigma \epsilon$ for some constant $\sigma \in \mathbb{R}^+$ and $\epsilon \sim \mathcal{N}(0, 1)$, and a hidden vector $\beta_0^* \sim \mathcal{N}(0, I_p)$. Assume we have a data matrix $X \in \mathbb{R}^{n \times p}$. In order to incorporate the noise in our structural matrix B , we consider an additional parameter $q(d)$ that grows linearly with d and such that $d = p + q$. Let $\phi_0 = \frac{n}{p}$. Therefore $\phi = \frac{n}{d} = \frac{n}{p} \frac{p}{d} = \phi_0 \psi$. Also, we let $\beta^{*T} = (\beta_0^{*T} | \beta_1^T) \sim \mathcal{N}(0, I_{p+q})$ and we consider an average V^* over β^* . We construct the following block-matrix B and compute the averaged V^* as follow:

$$B = \begin{pmatrix} r\sqrt{\frac{d}{p}}I_p & 0 \\ 0 & \sigma\sqrt{\frac{d}{q}}I_q \end{pmatrix} \implies V^* = \begin{pmatrix} r^2 \frac{1}{\psi} I_p & 0 \\ 0 & \sigma^2 \frac{1}{1-\psi} I_q \end{pmatrix} \quad (5.23)$$

Now let's consider the random matrix $Z \in \mathbb{R}^{n \times d}$ and split it into two sub-blocks $Z = \left(\sqrt{\frac{p}{d}}X | \sqrt{\frac{q}{d}}\Sigma\right)$. The framework of the chapter yields the following output vector:

$$Y = ZB\beta^* = rX\beta_0^* + \sigma\xi \quad (5.24)$$

where $\xi = \Sigma\beta_1^*$ is used as a proxy for the noise ϵ .

Estimator Now let's consider the linear estimator $\hat{y}_t = x^T \beta_t$. To capture the structure of this model, we use the following block-matrix A and compute the resulting matrix U :

$$A = \begin{pmatrix} \sqrt{\frac{d}{p}}I_p \\ 0_{q \times p} \end{pmatrix} \implies U = \begin{pmatrix} \frac{1}{\psi}I_p & 0 \\ 0 & 0_{q \times q} \end{pmatrix} \quad (5.25)$$

Therefore, it is straightforward to check that we have indeed: $\hat{Y}_t = ZA\beta_t = X\beta_t$.

Analytic result In this specific example, U and V^* are both diagonal-matrices, so Result 5.2

applies. Let's define (u, v^*) a random-variable sampling uniformly the eigenvalues of (U, V^*) . The structures of U and V^* give the following joint-distribution:

$$\mathcal{P}\left(u = \frac{1}{\psi}, v^* = \frac{r^2}{\psi}\right) = \psi \quad \mathcal{P}\left(u = 0, v^* = \frac{\sigma^2}{1-\psi}\right) = 1 - \psi \quad (5.26)$$

In this specific example, we focus only on rederiving the high-dimensional generalization error without any regularization term ($\lambda = 0$) for the minimum least-squares estimator. So we calculate $\zeta = \zeta(0)$ as follows: $\zeta = \frac{\frac{\zeta}{\psi}}{\frac{\phi}{\psi} + \zeta} + 0$ implies $\zeta^2 + \phi_0\zeta = \zeta$ so $\zeta \in \{0, 1 - \phi_0\}$. For \tilde{f}_1 we get:

$$\tilde{f}_1 = \psi \frac{\tilde{f}_1 \frac{\phi}{\psi^2}}{\left(\frac{\phi}{\psi} + \zeta\right)^2} + \psi \frac{r^2}{\psi} \frac{\zeta^2}{\left(\frac{\phi}{\psi} + \zeta\right)^2} + (1 - \psi) \frac{\sigma^2}{1 - \psi} \frac{\zeta^2}{(\zeta)^2} \quad (5.27)$$

In fact, the expression can be simplified as follow (without the constants ϕ, ψ):

$$\left(1 - \frac{\phi_0}{(\phi_0 + \zeta)^2}\right) \tilde{f}_1 = r^2 \frac{\zeta^2}{(\phi_0 + \zeta)^2} + \sigma^2 \quad (5.28)$$

Using both solutions $\zeta = 0$ or $\zeta = 1 - \phi_0$ yields the same results as in Hastie et al. (2019); Belkin et al. (2020a) using 5.3:

$$\mathcal{E}_{\text{gen}}(+\infty) = \begin{cases} \sigma^2 \frac{\phi_0}{\phi_0 - 1} & (\zeta = 0) \\ r^2(1 - \phi_0) + \sigma^2 \frac{1}{1 - \phi_0} & (\zeta = 1 - \phi_0) \end{cases} \quad (5.29)$$

5.3.2 Non-isotropic ridgeless regression of a noiseless linear model

Non-isotropic models have been studied in Dobriban and Wager (2018) and then also Wu and Xu (2020); Richards et al. (2021); Nakkiran et al. (2020b); Chen et al. (2021) where multiple-descents curve have been observed or engineered. In this section, we extend this idea to show that any number of descents can be generated and derive the precise curve of the generalization error as in Figure 5.3.1.

Target function We use the standard linear model $y(z) = z^T \beta^*$ for a random $\beta^* \sim \mathcal{N}(0, I_d)$. Therefore, we consider the matrix $B = I_d$ and thus $V^* = I_d$ such that $Y = ZB\beta^* = Z\beta^*$.

Estimator: Following the structure provided in table 5.1.1, the design a matrix A is a scalar matrix with $p \in \mathbb{N}^*$ sub-spaces of different scales spaced by a polynomial progression $\alpha^{-\frac{1}{2}i}$. In other words, the student is trained on a dataset with different scalings. We thus have $U = A^2$ and $\hat{Y}_t = \hat{X}\beta_t = Z A \beta_t$, such that for a given data-sample \hat{x} and any $0 \leq i \leq p - 1$ and $1 \leq k \leq \frac{d}{p}$, we have $\text{Var}\left(\hat{x}_{i\frac{d}{p}+k}\right) = \alpha^{-i}$.

Analytic results We refer the reader to the Appendix 5.D.2 for the calculation. Depending if ϕ is above or below 1, ζ is the solution of the following equations: $\zeta = 0$ or $1 = \frac{1}{p} \sum_{i=0}^{p-1} \frac{1}{\phi + \alpha^i \zeta}$. In

the over-parameterized regime ($\phi < 1$), the generalisation error is fully characterized by the equation:

$$\bar{\mathcal{E}}_{\text{gen}}(+\infty) = \phi(1 - \phi) \left(\frac{1}{p} \sum_{i=0}^{p-1} \frac{\alpha^i \zeta}{(\phi + \alpha^i \zeta)^2} \right)^{-1} - \phi \quad (5.30)$$

In the asymptotic limit $\alpha \rightarrow \infty$, ζ can be approximated and thus we can derive an asymptotic expansion of $\bar{\mathcal{E}}_{\text{gen}}(+\infty)$ for $\phi \in [0, 1] \setminus \frac{k}{p}\mathbb{Z}$ where clearly, the multiple descents appear as roots of the denominator of the sum:

$$\bar{\mathcal{E}}_{\text{gen}}(+\infty) = \frac{1}{p} \sum_{k=0}^{p-1} \frac{\phi(1 - \phi)}{\left(\phi - \frac{k}{p}\right)\left(\frac{k+1}{p} - \phi\right)} \mathbb{1}_{\left] \frac{k}{p}, \frac{k+1}{p} \right[} (\phi) - \phi + o_{\alpha}(1) \quad (5.31)$$

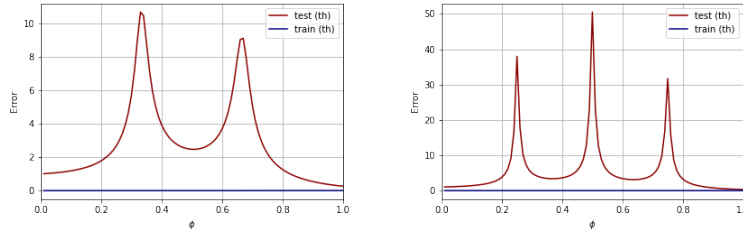


Figure 5.3.1: Example of theoretical multiple descents in the least-squares solution for the non-isotropic ridgeless regression model with $p = 3, \lambda = 10^{-7}$ (left) and $p = 4, \lambda = 10^{-13}$ (right), and $\alpha = 10^4$ in both of them.

Interestingly, we can see how these peaks are being formed with the time-evolution of the gradient flow as in Figure 5.3.2 with one peak close to $\phi = \frac{1}{3}$ and the second one at $\phi = \frac{2}{3}$. (Note that small λ requires more computational resources to have finer resolution at long times, hence here the second peak develops fully after $t = 10^4$). It is worth noticing also the existence of multiple time-descent, in particular at $\phi = 1$ with some "ripples" that can be observed even in the training error.

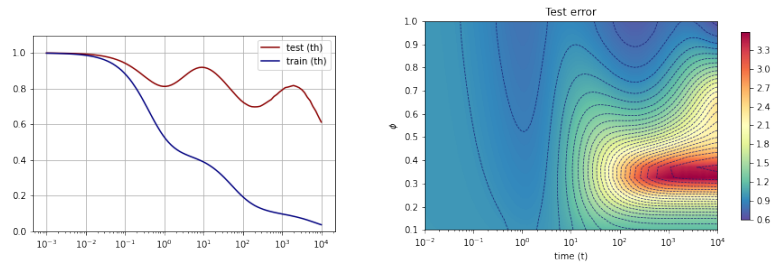


Figure 5.3.2: Example of theoretical multiple descents evolution in the non-isotropic ridgeless regression model with $p = 3, \lambda = 10^{-5}, \alpha = 100$ with $\phi = 1$ on the left and a range $\phi \in (0, 1)$ on the right heatmap.

The eigenvalue distribution (See Appendix 5.D.2) provides some insights on the existence of these phenomena. The emergence of a new spike when ϕ increases in Figure 5.3.1 when $p = 3$ coincides with the rise of a new "bulk" in the eigenvalue distribution. This can be seen in Figure 5.3.3 around $\phi = \frac{1}{3}$ and $\phi = \frac{2}{3}$. Note the analogy with the generic double-descent phenomena discussed in Hastie et al. (2019), where, instead of two distinct bulks, there is only

one bulk but with a mass concentrated at 0. Furthermore, the existence of multiple bulks allow for multiple evolution at different scales (with the $e^{-(z+\lambda)t}$ terms) and thus enable the emergence of multiple epoch-wise peaks.

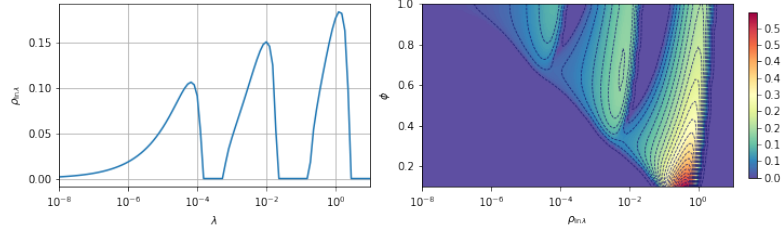


Figure 5.3.3: Theoretical (log-)eigenvalue distribution in the non-isotropic ridgeless regression model with $p = 3, \lambda = 10^{-5}, \alpha = 100$ with $\phi = 1$ on the left and a range $\phi \in (0, 1)$ on the right heatmap.

5.3.3 Random feature regression

In this section, we show that we can derive the learning curves for the random feature model introduced in Rahimi and Recht (2008), and we consider the setting described in Bodin and Macris (2021a). In this setting, we define the random weight-matrix $W \in \mathbb{R}^{p \times N}$ where $\psi_0 = \frac{N}{p}$ such that $W_{ij} \sim \mathcal{N}(0, \frac{1}{p})$ and $d = p + N + q$ and $\phi = \frac{n}{d}, \psi = \frac{p}{d}$, and $\phi_0 = \frac{n}{p} = \frac{n}{d} \frac{d}{p} = \frac{\phi}{\psi}$ (thus $\frac{q}{d} = 1 - (1 + \psi_0)\psi$). So with $Z = \left(\sqrt{\frac{p}{d}}X | \sqrt{\frac{p}{d}}\Omega | \sqrt{\frac{q}{d}}\xi \right)$, using the structures A and B from table 5.1.1 we have: $ZA = \mu XW + \nu\Omega$ and $ZB = X + \sigma\xi$, hence the model:

$$\hat{Y} = ZA\beta = (\mu XW + \nu\Omega)\beta \quad (5.32)$$

$$Y = ZB\beta^* = X\beta_0^* + \sigma\xi\beta_1^* \quad (5.33)$$

With further calculation that can be found in Appendix 5.D.4, a similar complete time derivation of the random feature regression can be performed with a much smaller linear-pencil than the one suggested in Bodin and Macris (2021a). As stated in this former work, the curves derived from this formula track the same training and test error in the high-dimensional limit as the model with the point-wise application of a centered non-linear activation function $f \in L^2(e^{-\frac{x^2}{2}}dx)$ with $\hat{Y} = \frac{1}{\sqrt{p}}f(\sqrt{p}XW)\beta$. More precisely, with the inner-product defined such that for any function $g \in L^2(e^{-\frac{x^2}{2}}dx)$, $\langle f, g \rangle = \mathbb{E}_{x \sim \mathcal{N}(0,1)}[f(x)g(x)]$, we derive the equivalent model parameters (μ, ν) with $\mu = \langle f, H_{e_1} \rangle$, $\nu^2 = \langle f, f \rangle - \mu^2$ while having the centering condition $\langle f, H_{e_0} \rangle = 0$ where (H_{e_n}) is the Hermite polynomial basis.

This transformation is dubbed the Gaussian equivalence principle and has been observed and rigorously proved under weaker conditions in Pennington and Worah (2017); Péché (2019); Hu and Lu (2022), and since then has been applied more broadly for instance in Adlam and Pennington (2020a).

5.3.4 Towards realistic datasets

As stated in Loureiro et al. (2021), the training and test error of realistic datasets can also be captured. In this example we track the MNIST dataset and focus on learning the parity of the images ($y = +1$ for even numbers and $y = -1$ for odd-numbers). We refer to Appendix 5.D.5 for thorough discussions of Figures 5.3.4 and 5.3.5 as well as technical details to obtain them, and other examples. Besides the learning curve profile at $t = +\infty$, the full theoretical time evolution is predicted and matches the experimental runs. In particular, the rise of the double-descent phenomenon is observed through time.

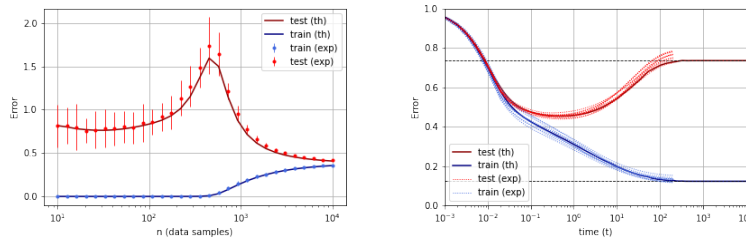


Figure 5.3.4: Comparison between the analytical and experimental learning profiles for the minimum least-squares estimator at $\lambda = 10^{-3}$ on the left (20 runs) and the time evolution at $\lambda = 10^{-2}$, $n = 700$ on the right (10 runs).

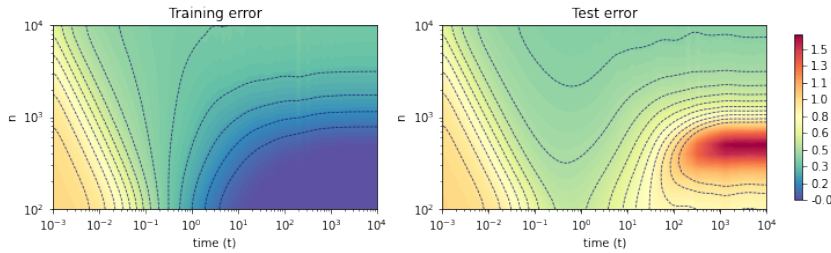


Figure 5.3.5: Analytical training error and test error heat-maps for the theoretical gradient flow for $\lambda = 10^{-3}$.

5.4 Conclusion

The time-evolution can also be investigated using the dynamical mean field theory (DMFT) from statistical mechanics. We refer the reader to the book Parisi et al. (2020) and a series of recent works Sompolinsky et al. (1988); Crisanti and Sompolinsky (2018); Agoritsas et al. (2018); Mignacco et al. (2020, 2021) for an overview of this tool. This method is a priori unrelated to ours and yields a set of non-linear integro-differential equations for time correlation functions which are in general not solvable analytically and one has to resort to a numerical solution. It would be interesting to understand if for the present model the DMFT equations can be reduced to our set of algebraic equations. We believe it can be a fruitful endeavor to compare in detail the two approaches: the one based on DMFT and the one based on random matrix theory tools and Cauchy integration formulas.

Another interesting direction which came to our knowledge recently is the one taken in Lu and Yau (2022); Hu and Lu (2022) and in Misiakiewicz (2022); Xiao and Pennington (2022), who study the high-dimensional polynomial regime where $n \propto d^\kappa$ for a fixed κ . In particular, it is becoming notorious that changing the scaling can yield additional descents. This regime is out of the scope of the present work but it would be desirable to explore if the linear-pencils and the random matrix tools that we extensively use in this work can extend to these cases.

Appendix

5.A Gradient flow calculations

In this section, we derive the main equations for the gradient flow algorithm, and derive and set of Cauchy integration formula involving the limiting traces of large matrices. The calculation factoring out Z in the limit $d \rightarrow \infty$ is pursued in the next section. First, we recall and expand the training error function in 5.1:

$$\mathcal{E}_{\text{train}}^\lambda(\beta_t) = \frac{1}{n} \|Y - \hat{X}\beta_t\|_2^2 + \frac{\lambda}{n} \|\beta_t\|_2^2 \quad (5.34)$$

$$= \frac{1}{n} \|Y\|_2^2 - \frac{2}{n} Y^T \hat{X}\beta_t + \frac{1}{n} \beta_t^T \hat{X}^T \hat{X}\beta_t + \frac{\lambda}{n} \|\beta_t\|_2^2 \quad (5.35)$$

$$= \frac{1}{n} \|ZB\beta^*\|_2^2 - \frac{2}{n} \beta^{*T} B^T Z^T ZA\beta_t + \frac{1}{n} \beta_t^T A^T Z^T ZA\beta_t + \frac{\lambda}{n} \|\beta_t\|_2^2 \quad (5.36)$$

Let $K = (\hat{X}^T \hat{X} + \lambda I)^{-1} = (A^T Z^T ZA + \lambda I)^{-1}$ which is invertible for $\lambda > 0$. Therefore, we can write the gradient of the training error for any β as:

$$\frac{n}{2} \nabla_{\beta} \mathcal{E}_{\text{train}}(\beta) = \hat{X}^T (\hat{X}\beta - Y) + \lambda\beta = (\hat{X}^T \hat{X} + \lambda I)\beta - \hat{X}^T Y = K^{-1}\beta - \hat{X}^T Y \quad (5.37)$$

The gradient flow equations reduces to a first order ODE

$$\frac{\partial \beta_t}{\partial t} = -\frac{n}{2} \nabla_{\beta} \mathcal{E}_{\text{train}}^\lambda(\beta_t) = \hat{X}^T Y - K^{-1}\beta_t \quad (5.38)$$

The solution can be completely expressed using $L_t = (I - \exp(-tK^{-1}))$ as

$$\beta_t = \exp(-tK^{-1})\beta_0 + (I - \exp(-tK^{-1}))K\hat{X}^T Y \quad (5.39)$$

$$= (I - L_t)\beta_0 + L_t K \hat{X}^T X\beta^* \quad (5.40)$$

In the following two subsections, we will focus on deriving an expression of the time evolution of the test error and training error using these equations averaged over the a centered random vector β_0 such that $r_0^2 = N_d(\beta_0)^2$.

5.A.1 Test error

As above, the test error can be expanded using the fact that on $\mathcal{N}_0 = \mathcal{N}(0, \frac{1}{d})$, we have the identity $\mathbb{E}_{z \sim \mathcal{N}_0}[zz^T] = \frac{1}{d}I_d$:

$$\mathcal{E}_{\text{gen}}(\beta_t) = \mathbb{E}_{z \sim \mathcal{N}_0} [(z^T A \beta_t - z^T B \beta^*)^2] \quad (5.41)$$

$$= (A \beta_t - B \beta^*)^T \mathbb{E}_{z \sim \mathcal{N}_0}[zz^T] (A \beta_t - B \beta^*) \quad (5.42)$$

$$= \frac{1}{d} \beta_t^T (A^T A) \beta_t - \frac{2}{d} \beta^{*T} B^T A \beta_t + \frac{1}{d} \beta^{*T} B^T B \beta^* \quad (5.43)$$

So expanding the first term yields

$$\beta_t^T (A^T A) \beta_t = (\beta_0^T (I - L_t) + \beta^{*T} X^T \hat{X} K L_t) (A^T A) ((I - L_t) \beta_0 + L_t K \hat{X}^T X \beta^*) \quad (5.44)$$

$$= \beta_0^T (I - L_t) (A^T A) (I - L_t) \beta_0 \quad (5.45)$$

$$+ \beta^{*T} (B^T Z^T Z A) K L_t U (A^T A) L_t K (A^T Z^T Z B) \beta^* \quad (5.46)$$

$$+ 2 \beta_0^T (I - L_t) (A^T A) L_t K (A^T Z^T Z B) \beta^* \quad (5.47)$$

while the second term yields

$$\beta^{*T} B^T A \beta_t = \beta^{*T} B^T A ((I - L_t) \beta_0 + L_t K \hat{X}^T X \beta^*) \quad (5.48)$$

$$= \beta^{*T} B^T A (I - L_t) \beta_0 + \beta^{*T} L_t K (A^T Z^T Z B) \beta^* \quad (5.49)$$

Let's consider now the high-dimensional limit $\bar{\mathcal{E}}_{\text{gen}}(t) = \lim_{d \rightarrow +\infty} \mathcal{E}_{\text{gen}}(\beta_t)$. We further make the underlying assumption that the generalisation error concentrates on its mean with β_0 , that is to say: $\bar{\mathcal{E}}_{\text{gen}}(t) = \lim_{d \rightarrow +\infty} \mathbb{E}_{\beta_0}[\mathcal{E}_{\text{gen}}(\beta_t)]$. Let $V^* = B \beta^* \beta^{*T} B^T$ and $c_0 = \text{Tr}_d[V^*]$, then using the former expanded terms in 5.41 we find the expression

$$\bar{\mathcal{E}}_{\text{gen}}(t) = c_0 + r_0^2 \text{Tr}_d [A(I - L_t)^2 A^T] \quad (5.50)$$

$$+ \text{Tr}_d [Z^T Z A K L_t A^T A L_t K A^T Z^T Z V^*] - 2 \text{Tr}_d [A L_t K A^T Z^T Z V^*] \quad (5.51)$$

So $\bar{\mathcal{E}}_{\text{gen}}(t) = c_0 + r_0^2 \mathcal{B}_0(t) + \mathcal{B}_1(t)$ with:

$$\mathcal{B}_0(t) = \text{Tr}_d [A^T (I - L_t)^2 A] \quad (5.52)$$

$$\mathcal{B}_1(t) = \text{Tr}_d [Z^T Z A K L_t A^T A L_t K A^T Z^T Z V^*] - 2 \text{Tr}_d [A L_t K A^T Z^T Z V^*] \quad (5.53)$$

Let $K(z) = (\hat{X}^T \hat{X} - zI)^{-1}$ the resolvent of $\hat{X}^T \hat{X}$, and let's have the convention $K = K(-\lambda)$ to remain consistent with the previous formula. Then for any holomorphic functional $f: \mathbb{U} \rightarrow \mathbb{C}$ defined on an open set \mathbb{U} which contains the spectrum of $\hat{X}^T \hat{X}$, with Γ a contour in \mathbb{C} enclosing the spectrum of $\hat{X}^T \hat{X}$ but not the poles of f , we have with the extension of f onto $\mathbb{C}^{n \times n}$:

$f(\hat{X}^T \hat{X}) = \frac{-1}{2i\pi} \oint_{\Gamma} f(z)K(z)dz$. For instance, we can apply it for the following expression:

$$KL_t = L_tK = (I - \exp(-t\hat{X}^T \hat{X} - t\lambda I))(\hat{X}^T \hat{X} - \lambda I)^{-1} \quad (5.54)$$

$$= \frac{-1}{2i\pi} \oint_{\Gamma} \frac{1 - e^{-t(z+\lambda)}}{z + \lambda} (\hat{X}^T \hat{X} - zI)^{-1} dz \quad (5.55)$$

$$= \frac{-1}{2i\pi} \oint_{\Gamma} \frac{1 - e^{-t(z+\lambda)}}{z + \lambda} K(z) dz \quad (5.56)$$

So we can generalize this idea to each trace and rewrite $\mathcal{B}_1(t)$ and $\mathcal{B}_0(t)$ with

$$\mathcal{B}_1(t) = \frac{-1}{4\pi^2} \oint_{\Gamma} \oint_{\Gamma} \frac{(1 - e^{-t(x+\lambda)})(1 - e^{-t(y+\lambda)})}{(x + \lambda)(y + \lambda)} f_1(x, y) dx dy + \frac{1}{i\pi} \oint_{\Gamma} \frac{1 - e^{-t(z+\lambda)}}{z + \lambda} f_2(z) dz \quad (5.57)$$

$$\mathcal{B}_0(t) = \frac{-1}{2i\pi} \oint_{\Gamma} e^{-2t(z+\lambda)} f_0(z) dz \quad (5.58)$$

where we introduce the set of functions $f_1(x, y)$, $f_2(z)$ and $f_0(z)$

$$f_1(x, y) = \text{Tr}_d [Z^T Z A K(x) A^T A K(y) A^T Z^T Z V^*] \quad (5.59)$$

$$f_2(z) = \text{Tr}_d [A K(z) A^T Z^T Z V^*] \quad (5.60)$$

$$f_0(z) = \text{Tr}_d [A K(z) A^T] \quad (5.61)$$

Let $G(x) = (UZ^T Z - xI)^{-1}$, using the push-through identity, it is straightforward that $AK(z)A = G(z)U = UG(z)^T$. This help us reduce further the expression of f_1 into smaller terms which will be easier to handle with linear-pencils later on

$$f_1(x, y) = \text{Tr}_d [Z^T Z U G(x)^T G(y) U Z^T Z V^*] \quad (5.62)$$

$$= \text{Tr}_d [(G(x)^{-1} + xI)^T G(x)^T G(y) (G(y)^{-1} + yI) V^*] \quad (5.63)$$

$$= \text{Tr}_d [(I + yG(y)) V^* (I + xG(x))^T] \quad (5.64)$$

$$= c_0 + y \text{Tr}_d [G(y) V^*] + x \text{Tr}_d [G(x) V^*] + xy \text{Tr}_d [G(x) V^* G(y)^T] \quad (5.65)$$

Similarly with f_2 and f_0 , they can be rewritten as

$$f_2(z) = \text{Tr}_d [G(z) U Z^T Z V^*] \quad (5.66)$$

$$= \text{Tr}_d [G(z) (G(z)^{-1} + zI) V^*] \quad (5.67)$$

$$= c_0 + z \text{Tr}_d [G(z) V^*] \quad (5.68)$$

$$f_0(z) = \text{Tr}_d [G(z) U] \quad (5.69)$$

Hence in fact the definition $\tilde{f}_1(x, y) = xy \text{Tr}_d [G(x) V^* G(y)^T]$ such that

$$f_1(x, y) = f_2(x) + f_2(y) + \tilde{f}_1(x, y) - c_0 \quad (5.70)$$

At this point, the equations provided by 5.57 are valid for any realization Z in the limit $d \rightarrow \infty$. We will see in the next section how to simplify these terms by factoring out Z .

5.A.2 Training error

Similar formulas can be derived for the training error. For the sake of simplicity, we provide a formula to track the training error without the regularization term, that is to say $\mathcal{E}_{\text{train}}^0(\beta_t)$ (as in Loureiro et al. (2021)) while still minimizing the loss $\mathcal{E}_{\text{train}}^\lambda(\beta_t)$. So using the expanded expression 5.34, and considering the high-dimensional assumption with concentration $\bar{\mathcal{E}}_{\text{train}}^0(t) := \lim_{d \rightarrow +\infty} \mathcal{E}_{\text{train}}(\beta_t) = \lim_{d \rightarrow +\infty} \mathbb{E}_{\beta_0}[\mathcal{E}_{\text{train}}(\beta_t)]$ we have

$$\bar{\mathcal{E}}_{\text{train}}^0(t) = \text{Tr}_n [Z^T Z V^*] + r_0^2 \text{Tr}_n [A^T Z^T Z A (I - L_t)^2] \quad (5.71)$$

$$+ \text{Tr}_n [Z^T Z A K L_t A^T Z^T Z A L_t K A^T Z^T Z V^*] \quad (5.72)$$

$$- 2 \text{Tr}_n [Z^T Z A L_t K A^T Z^T Z V^*] \quad (5.73)$$

First of all, standard random matrix results (for instance see Rubio and Mestre (2011)) assert the result $\text{Tr}_d [Z^T Z V^*] = \text{Tr}_d [Z^T Z] \text{Tr}_d [V^*] = \phi c_0$. This result can also be derived under our random matrix theory framework, for completeness we provide this calculation in 5.C.2. Therefore, we can define $\mathcal{H}_0(t)$ and $\mathcal{H}_1(t)$ such that

$$\bar{\mathcal{E}}_{\text{train}}^0(t) = c_0 + r_0^2 \mathcal{H}_0(t) + \mathcal{H}_1(t) \quad (5.74)$$

where we have the traces

$$\mathcal{H}_0(t) = \text{Tr}_n [A^T Z^T Z A (I - L_t)^2] \quad (5.75)$$

$$\mathcal{H}_1(t) = \text{Tr}_n [Z^T Z A K L_t (A^T Z^T Z A) L_t K A^T Z^T Z V^*] - 2 \text{Tr}_n [Z^T Z A L_t K A^T Z^T Z V^*] \quad (5.76)$$

And using the functional calculus argument with Cauchy integration formula over the same contour Γ we find

$$\mathcal{H}_1(t) = \frac{-1}{4\pi^2} \oint_{\Gamma} \oint_{\Gamma} \frac{(1 - e^{-t(x+\lambda)})(1 - e^{-t(x+\lambda)})}{(x+\lambda)(y+\lambda)} h_1(x, y) dx dy + \frac{1}{i\pi} \oint_{\Gamma} \frac{(1 - e^{-t(z+\lambda)})}{(z+\lambda)} h_2(z) dz \quad (5.77)$$

$$\mathcal{H}_0(t) = \frac{-1}{2i\pi} \oint_{\Gamma} e^{-2t(z+\lambda)} h_0(z) dz \quad (5.78)$$

Where we use the traces (which only contain algebraic expression of matrices):

$$h_1(x, y) = \text{Tr}_n [Z^T Z A K(x) A Z^T Z A^T K(y) A^T Z^T Z V^*] \quad (5.79)$$

$$h_2(z) = \text{Tr}_n [Z^T Z A K(z) A^T Z^T Z V^*] \quad (5.80)$$

$$h_0(z) = \text{Tr}_n [Z^T Z A^T K(z) A^T] \quad (5.81)$$

The expression of h_1 can be reduced to smaller terms as before with f_1

$$\phi h_1(x, y) = \text{Tr}_d [Z^T ZUG(x)^T Z^T ZG(y)UZ^T ZV^*] \quad (5.82)$$

$$= \text{Tr}_d [(G(x)^{-1} + xI)^T G(x)^T Z^T ZG(y)(G(y)^{-1} + yI)V^*] \quad (5.83)$$

$$= \text{Tr}_d [Z^T ZV^*] + x\text{Tr}_d [G(x)^T Z^T ZV^*] + y\text{Tr}_d [Z^T ZG(y)V^*] \quad (5.84)$$

$$+ xy\text{Tr}_d [Z^T ZG(y)V^* G(x)^T] \quad (5.85)$$

$$= c_0\phi + x\text{Tr}_d [Z^T ZG(x)V^*] + y\text{Tr}_d [Z^T ZG(y)V^*] \quad (5.86)$$

$$+ xy\text{Tr}_d [Z^T ZG(y)V^* G(x)^T] \quad (5.87)$$

and similarly with h_2

$$\phi h_2(z) = \text{Tr}_d [Z^T ZG(z)UZ^T ZV^*] \quad (5.88)$$

$$= \text{Tr}_d [Z^T ZG(z)(G(z)^{-1} + zI)V^*] \quad (5.89)$$

$$= \text{Tr}_d [Z^T ZV^*] + z\text{Tr}_d [Z^T ZG(z)V^*] \quad (5.90)$$

$$= c_0\phi + z\text{Tr}_d [Z^T ZG(z)V^*] \quad (5.91)$$

and similarly with h_0

$$\phi h_0(z) = \text{Tr}_d [Z^T ZG(z)U] \quad (5.92)$$

$$= \text{Tr}_d [G(z)(G(z)^{-1} + zI)] \quad (5.93)$$

$$= 1 + z\text{Tr}_d [G(z)] \quad (5.94)$$

We can also define the term $\tilde{h}_1(x, y) = xy\text{Tr}_d [ZG(y)V^* G(x)^T Z^T]$ so that:

$$h_1(x, y) = h_2(x) + h_2(y) + \tilde{h}_1(x, y) - c_0 \quad (5.95)$$

5.B Test error and training error limits with linear pencils

In this section we compute a set of self-consistent equation to derive the high-dimensional evolution of the training and test error. We refer to Chapter 3 for the definition and result statements concerning the linear pencils.

We will derive essentially two linear-pencils of size 6×6 and 4×4 which will enable us to calculate the limiting values for $\tilde{f}_1, \tilde{f}_2, f_0$ for the test error, and \tilde{h}_1, h_2, h_0 for the training error. Note that these block-matrices are derived essentially by observing the recursive application of the block-matrix inversion formula and manipulating it so as to obtain the desired result.

Compared to other works such as Bodin and Macris (2021a); Adlam and Pennington (2020a), our approach yields smaller sizes of linear-pencils to handle, which in turn yields a smaller set of algebraic equations. One of the ingredient of our method consists in considering a multiple-stage approach where the trace of some random blocks can be calculated in different parts (See the random feature model for example in Appendix 5.D.4). However, the question

of finding the simplest linear-pencil remains open and interesting to investigate.

5.B.1 Limiting traces of the test error

Limiting trace for \tilde{f}_1 and f_0

We construct a linear-pencil M_1 as follow (with Z the random matrix into consideration)

$$M_1 = \begin{pmatrix} 0 & 0 & 0 & -yI & 0 & Z^T \\ 0 & 0 & 0 & 0 & Z & I \\ 0 & 0 & 0 & U & I & 0 \\ -xI & 0 & U & -xyV^* & 0 & 0 \\ 0 & Z^T & I & 0 & 0 & 0 \\ Z & I & 0 & 0 & 0 & 0 \end{pmatrix} \quad (5.96)$$

The inverse of this block-matrix contains the terms in the traces of \tilde{f}_1 and f_0 . To see this, let's calculate the inverse of M_1 by splitting it first into other "flattened" blocks:

$$M_1 = \begin{pmatrix} 0 & B_y^T \\ B_x & D \end{pmatrix} \implies M_1^{-1} = \begin{pmatrix} -B_x^{-1}DB_y^{T-1} & B_x^{-1} \\ B_y^{T-1} & 0 \end{pmatrix} \quad (5.97)$$

Where B_x and D are given by

$$B_x = \begin{pmatrix} -xI & 0 & U \\ 0 & Z^T & I \\ Z & I & 0 \end{pmatrix} \quad D = \begin{pmatrix} -xyV^* & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (5.98)$$

then to calculate the inverse of B_x , notice first its lower right-hand sub-block has inverse

$$\begin{pmatrix} Z^T & I \\ I & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & I \\ I & -Z^T \end{pmatrix} \quad (5.99)$$

Which lead us to the following inverse using the block-matrix inversion formula (the dotted terms aren't required):

$$B_x^{-1} = \begin{pmatrix} G(x) & -G(x)U & G(x)UZ^T \\ -ZG(x) & \dots & I_n - ZG(x)UZ^T \\ Z^T ZG(x) & \dots & \dots \end{pmatrix} \quad (5.100)$$

5.B Test error and training error limits with linear pencils

With $g_d^{(ij)}$ the trace of the squared sub-block $(M_1^{-1})^{(ij)}$ divided by the size of the block (ij) , we find the desired functions

$$\tilde{f}_1(x, y) = \lim_{d \rightarrow +\infty} -g_d^{(11)} \quad (5.101)$$

$$f_0(x) = \lim_{d \rightarrow +\infty} -g_d^{(15)} \quad \text{OR} \quad f_0(y) = \lim_{d \rightarrow +\infty} -g_d^{(51)} \quad (5.102)$$

Let's now consider g the limiting value of g_d , and calculate the mapping $\eta(g)$:

$$\eta(g) = \begin{pmatrix} 0 & 0 & 0 & 0 & \phi g^{(26)} & 0 \\ 0 & 0 & 0 & 0 & 0 & g^{(15)} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \phi g^{(62)} & 0 & 0 & 0 & \phi g^{(22)} & 0 \\ 0 & g^{(51)} & 0 & 0 & 0 & g^{(11)} \end{pmatrix} \quad (5.103)$$

So we can calculate the matrix $\Pi(M_1)$ such that the elements of g are the limiting trace of the squared sub-blocks of $(\Pi(M_1))^{-1}$ (divided by the block-size) following the steps of the result in Chapter 3:

$$\Pi(M_1) = \begin{pmatrix} 0 & 0 & 0 & -yI & -\phi g^{(26)}I & 0 \\ 0 & 0 & 0 & 0 & 0 & (1 - g^{(15)})I \\ 0 & 0 & 0 & U & I & 0 \\ -xI & 0 & U & -xyV^* & 0 & 0 \\ -\phi g^{(62)}I & 0 & I & 0 & -\phi g^{(22)}I & 0 \\ 0 & (1 - g^{(51)})I & 0 & 0 & 0 & -g^{(11)}I \end{pmatrix} \quad (5.104)$$

Therefore, there remains to compute the inverse of $\Pi(M_1)$. We split again $\Pi(M_1)$ as flattened sub-blocks to make the calculation easier

$$\Pi(M_1) = \begin{pmatrix} 0 & \tilde{B}_y^T \\ \tilde{B}_x & \tilde{D} \end{pmatrix} \Rightarrow \Pi(M_1)^{-1} = \begin{pmatrix} -\tilde{B}_x^{-1}\tilde{D}(\tilde{B}_y^{-1})^T & \tilde{B}_x^{-1} \\ (\tilde{B}_y^{-1})^T & 0 \end{pmatrix} \quad (5.105)$$

With the three block-matrices

$$\tilde{B}_x = \begin{pmatrix} -xI & 0 & U \\ -g^{(62)}\phi I & 0 & I \\ 0 & (1 - g^{(51)})I & 0 \end{pmatrix} \quad \tilde{B}_y = \begin{pmatrix} -xI & 0 & U \\ -g^{(26)}\phi I & 0 & I \\ 0 & (1 - g^{(15)})I & 0 \end{pmatrix} \quad (5.106)$$

$$\tilde{D} = \begin{pmatrix} -xyV^* & 0 & 0 \\ 0 & -g^{(22)}\phi I & 0 \\ 0 & 0 & -g^{(11)}I \end{pmatrix} \quad (5.107)$$

A straightforward application of the block-matrix inversion formula yields inverse of \tilde{B}_x

$$\tilde{B}_x^{-1} = \begin{pmatrix} (\phi g^{(62)} U - xI)^{-1} & -U(\phi g^{(62)} U - xI)^{-1} & 0 \\ 0 & 0 & (1 - g^{(51)})^{-1} I \\ \phi g^{(62)} (\phi g^{(62)} U - xI)^{-1} & -x(\phi g^{(62)} U - xI)^{-1} & 0 \end{pmatrix} \quad (5.108)$$

Therefore, we retrieve the following close set of equations:

$$g^{(11)} = \text{Tr}_d [(g^{(62)} \phi U - xI)^{-1} (xyV^* + g^{(22)} \phi U^2) (g^{(26)} \phi U - yI)^{-1}] \quad (5.109)$$

$$g^{(22)} = g^{(11)} (1 - g^{(15)})^{-1} (1 - g^{(51)})^{-1} \quad (5.110)$$

$$g^{(26)} = (1 - g^{(51)})^{-1} \quad (5.111)$$

$$g^{(15)} = -\text{Tr}_d [U(g^{(62)} \phi U - xI)^{-1}] \quad (5.112)$$

These equations can be simplified slightly by removing $g^{(22)}, g^{(26)}$ and introducing $q^{(15)}$:

$$g^{(11)} = \text{Tr}_d [(\phi U - xq^{(15)} I)^{-1} (xyq^{(15)} q^{(51)} V^* + g^{(11)} \phi U^2) (\phi U - yq^{(51)} I)^{-1}] \quad (5.113)$$

$$q^{(15)} = \text{Tr}_d [(\phi U - xq^{(15)} I + q^{(15)} U) (\phi U - xq^{(15)} I)^{-1}] \quad (5.114)$$

$$g^{(15)} = 1 - q^{(15)} \quad (5.115)$$

Let $\zeta_x = -xq^{(15)}$, or by symmetry $\zeta_y = -yq^{(51)}$, then using the fact that $\tilde{f}_1(x, y) = -g^{(11)}$ and $f_0(x) = -g^{(15)}$ we find the system of equations

$$\tilde{f}_1(x, y) = \text{Tr}_d [(\phi U + \zeta_x I)^{-1} (\zeta_x \zeta_y V^* + \tilde{f}_1(x, y) \phi U^2) (\phi U + \zeta_y I)^{-1}] \quad (5.116)$$

$$f_0(x) = - \left(1 + \frac{\zeta_x}{x} \right) \quad (5.117)$$

$$\zeta_z = -z + \text{Tr}_d [\zeta_z U (\phi U + \zeta_z I)^{-1}] \quad (5.118)$$

Remark: As a byproduct of this analysis, notice the term $g^{(62)} = (q^{(15)})^{-1} = \frac{-x}{\zeta_x}$. In fact we have:

$$g^{(62)} = \text{Tr}_n [I_n - ZG(x)UZ^T] \quad (5.119)$$

$$= 1 - \text{Tr}_n [Z(AA^T Z^T Z - xI)^{-1} AA^T Z^T] \quad (5.120)$$

$$= 1 - \text{Tr}_n [(ZAA^T Z^T - xI)^{-1} ZAA^T Z^T] \quad (5.121)$$

$$= 1 - \text{Tr}_n [(\hat{X}\hat{X}^T - xI)^{-1} (\hat{X}\hat{X}^T - xI_n + xI_n)] \quad (5.122)$$

$$= -x \text{Tr}_n [(\hat{X}\hat{X}^T - xI)^{-1}] \quad (5.123)$$

So if we let $m(x) = \text{Tr}_n [(\hat{X}\hat{X}^T - xI)^{-1}]$ the trace of the resolvent of the student data matrix, we find that $m(x) = \zeta_x^{-1}$. This can be useful for analyzing the eigenvalues as in Appendix 5.D.2.

Limiting trace for f_2

As before, we construct a second linear-pencil M_2 with Z the random matrix component into consideration

$$M_2 = \begin{pmatrix} I & 0 & 0 & 0 \\ -zV^* & -zI & 0 & U \\ 0 & 0 & Z^T & I \\ 0 & Z & I & 0 \end{pmatrix} \quad (5.124)$$

The former flattened block B_z can be recognized in the lower right-hand side of M_2 , thus we can use the block matrix-inversion formula and get:

$$M_2^{-1} = \left(\begin{array}{ccc|ccc} I & & & 0 & 0 & 0 \\ \hline zG(z)V^* & & & & & \\ -zZG(z)V^* & & & & & \\ zZ^T ZG(z)V^* & & & & & \\ \hline & & & & B_z^{-1} & \end{array} \right) \quad (5.125)$$

Now it is clear that we can express $f_2(z) = c_0 + \lim_{d \rightarrow +\infty} g_d^{(21)}$. Following the steps of Chapter 3 we calculate the mapping

$$\eta(g) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \phi g^{(34)} & 0 & 0 \\ 0 & 0 & g^{(23)} & 0 \end{pmatrix} \quad (5.126)$$

Which in returns enable us to calculate $\Pi(M_2)$

$$\Pi(M_2) = \begin{pmatrix} I & 0 & 0 & 0 \\ -zV^* & -zI & 0 & U \\ 0 & -g^{(34)}\phi I & 0 & I \\ 0 & 0 & (1 - g^{(23)})I & 0 \end{pmatrix} \quad (5.127)$$

To compute the inverse of $\Pi(M_2)$, the block-matrix is first split with the sub-block \tilde{B}_z defined as follow

$$\tilde{B}_z = \begin{pmatrix} -zI & 0 & U \\ -g^{(34)}\phi I & 0 & I \\ 0 & (1 - g^{(23)})I & 0 \end{pmatrix} \quad \Pi(M_2) = \left(\begin{array}{ccc|ccc} I & & & 0 & 0 & 0 \\ \hline -zV^* & & & & & \\ 0 & & & & & \\ 0 & & & & & \\ \hline & & & & \tilde{B}_z & \end{array} \right) \quad (5.128)$$

A straightforward application of the block-matrix inversion formula yields the inverse of \tilde{B}_z :

$$\tilde{B}_z^{-1} = \begin{pmatrix} (g^{(34)}\phi U - zI)^{-1} & -U(g^{(34)}\phi U - zI)^{-1} & 0 \\ 0 & 0 & (1 - g^{(23)})^{-1}I \\ g^{(34)}\phi(g^{(34)}\phi U - zI)^{-1} & -z(g^{(34)}\phi U - zI)^{-1} & 0 \end{pmatrix} \quad (5.129)$$

Hence we can derive the inverse

$$\Pi(M_2)^{-1} = \left(\begin{array}{c|ccc} I & 0 & 0 & 0 \\ \hline z(g^{(34)}\phi U - zI)^{-1}V^* & & & \\ 0 & & \tilde{B}_z^{-1} & \\ zg^{(34)}\phi(g^{(34)}\phi U - zI)^{-1}V^* & & & \end{array} \right) \quad (5.130)$$

Eventually, using the fixed-point result on linear-pencils, we derive the set of equations

$$g^{(21)} = \text{Tr}_d [zV^*(g^{(34)}\phi U - zI)^{-1}] \quad (5.131)$$

$$g^{(34)} = (1 - g^{(23)})^{-1} \quad (5.132)$$

$$g^{(23)} = -\text{Tr}_d [U(g^{(34)}\phi U - zI)^{-1}] \quad (5.133)$$

$$g^{(41)} = \text{Tr}_d [zg^{(34)}\phi(g^{(34)}\phi U - zI)^{-1}V^*] \quad (5.134)$$

$$g^{(22)} = \text{Tr}_d [(g^{(34)}\phi U - zI)^{-1}] \quad (5.135)$$

$$(5.136)$$

In fact, it is a straightforward to see that $g^{(23)}, g^{(34)}$ follows the same equations as the former $g^{(15)}, g^{(26)}$ in the previous subsection, therefore $g^{(23)} = g^{(15)} = 1 - q^{(15)} = 1 + \frac{\zeta_z}{z}$, and thus $g^{(34)} = -\frac{z}{\zeta_z}$. Eventually we get $g^{(21)} = -\text{Tr}_d [\zeta_z V^*(\phi U + \zeta_z I)^{-1}]$ so in the limit $d \rightarrow \infty$:

$$f_2(z) = c_0 - \text{Tr}_d [\zeta_z V^*(\phi U + \zeta_z I)^{-1}] \quad (5.137)$$

5.B.2 Limiting traces for the training error

Limiting trace for h_1

A careful attention to the linear-pencil M_1 shows that the terms in the trace of \tilde{h}_1 are actually given by the location $g^{(22)}$. We have to be careful also of the fact that $(M_1^{-1})^{(22)}$ is a block matrix of size $n \times n$, so it is already divided by the size n (and not d). Hence we simply have with $\eta_z = \frac{-z}{\zeta_z}$:

$$\tilde{h}_1(x, y) = -g^{(22)} = \frac{-x}{\zeta_x} \frac{-y}{\zeta_y} f_1(x, y) = \eta_x \eta_y f_1(x, y) \quad (5.138)$$

Limiting trace for h_2

In the case of h_2 , we need the specific term provided by the linear-pencil M_2 by the location $g^{(41)}$ with $\phi h_2(z) = c_0 \phi + g^{(41)}$

For h_2 we use the linear pencil for f_2 , but instead of using $g^{(21)}$ we use $h_2 = c_0 \phi + g^{(41)}$. We

find:

$$g^{(41)} = z\phi \text{Tr}_d [V^*(\phi U + \zeta_z I)^{-1}] \quad (5.139)$$

$$= \phi \frac{z}{\zeta_z} \text{Tr}_d [\zeta_z V^*(\phi U + \zeta_z I)^{-1}] \quad (5.140)$$

$$= \phi \frac{z}{\zeta_z} (c_0 - f_2(z)) \quad (5.141)$$

Hence:

$$h_2(z) = c_0 \left(1 - \frac{-z}{\zeta_z}\right) + \frac{-z}{\zeta_z} f_2(z) = \eta_z (c_0 f_0(z) + f_2(z)) \quad (5.142)$$

Limiting trace for h_0

Finally for h_0 we use again the linear pencil M_2 with:

$$\text{Tr}_d [zG(z)] = zg^{(22)} = -\text{Tr}_d [\zeta_z(\phi U + \zeta_z I)^{-1}] \quad (5.143)$$

$$= -\text{Tr}_d [(\zeta_z + \phi U - \phi U)(\phi U + \zeta_z I)^{-1}] \quad (5.144)$$

$$= -1 + \phi \text{Tr}_d [U(\phi U + \zeta_z I)^{-1}] \quad (5.145)$$

$$= -1 + \frac{\phi}{\zeta_z} \text{Tr}_d [\zeta_z U(\phi U + \zeta_z I)^{-1}] \quad (5.146)$$

$$= -1 + \frac{\phi}{\zeta_z} (\zeta_z + z) \quad (5.147)$$

Therefore:

$$h_0(z) = \left(1 - \frac{-z}{\zeta_z}\right) = -\left(1 + \frac{\zeta_z}{z}\right) \frac{-z}{\zeta_z} = \eta_z f_0(z) \quad (5.148)$$

5.C Other limiting expressions

In this section we bring the sketch of proofs of additional expressions seen in the main results.

5.C.1 Expression with dual counterpart matrices U_\star and V_\star

The former functionals f_2 and \tilde{f}_1 can be rewritten as:

$$f_2(z) = c_0 - \text{Tr}_d [\zeta_z V^*(\phi U + \zeta_z I)^{-1}] \quad (5.149)$$

$$= c_0 - \text{Tr}_d [(\zeta_z I + \phi U - \phi U)V^*(\phi U + \zeta_z I)^{-1}] \quad (5.150)$$

$$= c_0 - \text{Tr}_d [V^*] + \text{Tr}_d [\phi A^T V^*(\phi U + \zeta_z I)^{-1} A^T] \quad (5.151)$$

$$= c_0 - c_0 + \text{Tr}_d [\phi A^T B \beta^* \beta^{*T} B^T A (U_\star + \zeta_z I)^{-1}] \quad (5.152)$$

$$= \text{Tr}_n [(\Xi \beta^* \beta^{*T} \Xi^T)(U_\star + \zeta_z I)^{-1}] \quad (5.153)$$

With similar steps using:

$$\zeta_x V^* \zeta_y = -(\zeta_x I + \phi U) V^* (\zeta_y I + \phi U) + \zeta_x V^* (\zeta_y I + \phi U) + (\zeta_x I + \phi U) V^* \zeta_y + \phi^2 U V^* U \quad (5.154)$$

We find:

$$\tilde{f}_1(x, y) = -c_0 + \text{Tr}_d [\zeta_x V^* (\zeta_y I + \phi U)^{-1}] + \text{Tr}_d [(\zeta_x I + \phi U)^{-1} V^* \zeta_y] \quad (5.155)$$

$$+ \text{Tr}_d [(\phi U + \zeta_x I)^{-1} (\phi^2 U V^* U + \tilde{f}_1(x, y) \phi U^2) (\phi U + \zeta_y I)^{-1}] \quad (5.156)$$

$$= c_0 - f_2(x) - f_2(y) \quad (5.157)$$

$$+ \text{Tr}_n [(U_\star + \zeta_x I)^{-1} ((\Xi \beta^* \beta^{*T} \Xi^T) + \tilde{f}_1(x, y) U_\star) U_\star (U_\star + \zeta_y I)^{-1}] \quad (5.158)$$

Hence in fact:

$$f_1(x, y) = \text{Tr}_n [(U_\star + \zeta_x I)^{-1} ((\Xi \beta^* \beta^{*T} \Xi^T) + \tilde{f}_1(x, y) U_\star) U_\star (U_\star + \zeta_y I)^{-1}] \quad (5.159)$$

Finally, we have using the push-through identity and the cyclicity of the trace:

$$\zeta_z = -z + \text{Tr}_d [\zeta_z A A^T (\phi A A^T + \zeta_z I)^{-1}] \quad (5.160)$$

$$= -z + \text{Tr}_d [\zeta_z A (\phi A^T A + \zeta_z I)^{-1} A^T] \quad (5.161)$$

$$= -z + \text{Tr}_n [\zeta_z U_\star (U_\star + \zeta_z I)^{-1}] \quad (5.162)$$

5.C.2 Limiting trace of $Z^T Z V^*$

Here we show another way in which our random matrix result can be used to infer the result on the limiting trace $\text{Tr}_d [Z^T Z V^*]$. To this end, we can design the linear-pencil:

$$M_3 = \begin{pmatrix} I & -V^* & 0 & 0 \\ 0 & I & Z^T & 0 \\ 0 & 0 & I & Z \\ 0 & 0 & 0 & I \end{pmatrix} \quad (5.163)$$

It is straightforward to calculate the inverse of the sub-matrix:

$$\begin{pmatrix} I & Z^T & 0 \\ 0 & I & Z \\ 0 & 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -Z^T & Z^T Z \\ 0 & I & -Z \\ 0 & 0 & I \end{pmatrix} \quad (5.164)$$

So that:

$$M_3^{-1} = \begin{pmatrix} I & V^* & -Z^T V^* & V^* Z^T Z \\ 0 & I & -Z^T & Z^T Z \\ 0 & 0 & I & -Z \\ 0 & 0 & 0 & I \end{pmatrix} \quad (5.165)$$

At this point, it is clear that the quantity of interest is provided by the term $g^{(14)}$ of the linear-pencil M_3 . We find calculate further:

$$\eta(g) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi g^{(33)} \\ 0 & 0 & g^{(42)} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (5.166)$$

Based on the inverse of M_3 , we can already predict that $g^{(33)} = 1$ and $g^{(42)} = 0$. Hence:

$$\Pi(M_3) = \begin{pmatrix} I & -V^* & 0 & 0 \\ 0 & I & 0 & -\phi I \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \Rightarrow \Pi(M_3)^{-1} = \begin{pmatrix} I & V^* & 0 & \phi V^* \\ 0 & I & 0 & \phi I \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \quad (5.167)$$

Finally we obtain $g^{(14)} = \text{Tr}_d[\phi V^*]$, and hence $\text{Tr}_d[Z^T Z V^*] = \phi \text{Tr}_d[V^*]$.

5.D Applications and calculation details

5.D.1 Mismatched Ridgeless regression of a noisy linear function

Target function Here we consider a slightly more complicated version of the former example where we let $y(x_0, x_1) = r(x_0^T \beta_0^* + x_1^T \beta_1^*) + \sigma \epsilon$ and still averaged over $\beta_0 \sim \mathcal{N}(0, I_{\gamma p})$ and $\beta_1 \sim \mathcal{N}(0, I_{(1-\gamma)p})$ with $x_0 \in \mathbb{R}^{\gamma p}$, $x_1 \in \mathbb{R}^{(1-\gamma)p}$. We let again $d = p + q$ and $\psi = \frac{p}{d}$ and $\phi_0 = \frac{p}{q}$. Therefore the former relation still holds $\phi = \frac{n}{d} = \frac{n}{p} \frac{p}{d} = \phi_0 \psi$. Similarly, we derive a block-matrix B and compute V^* :

$$B = \begin{pmatrix} r\sqrt{\frac{d}{p}}I_{\gamma p} & 0 & 0 \\ 0 & r\sqrt{\frac{d}{p}}I_{(1-\gamma)p} & 0 \\ 0 & 0 & \sigma\sqrt{\frac{d}{q}}I_q \end{pmatrix} \Rightarrow V^* = \begin{pmatrix} \frac{r^2}{\psi}I_{\gamma p} & 0 & 0 \\ 0 & \frac{r^2}{\psi}I_{(1-\gamma)p} & 0 \\ 0 & 0 & \frac{\sigma^2}{1-\psi}I_q \end{pmatrix} \quad (5.168)$$

So that with the splitting $Z = \left(\sqrt{\frac{p}{d}}X_0 | \sqrt{\frac{p}{d}}X_1 | \sqrt{\frac{q}{d}}\Sigma\right)$, and $\beta^{*T} = (\beta_0^{*T} | \beta_1^{*T} | \beta_2^{*T})$, and with $\xi = \Sigma\beta_2^*$:

$$Y = ZB\beta^* = r(X_0\beta_0^* + X_1\beta_1^*) + \sigma\xi \quad (5.169)$$

Estimator Following the same steps, we construct A and U with

$$A = \begin{pmatrix} \sqrt{\frac{d}{\gamma p}}I_{\gamma p} \\ 0_{(1-\gamma)p \times \gamma d} \\ 0_{q \times \gamma d} \end{pmatrix} \Rightarrow U = \begin{pmatrix} \frac{1}{\gamma\psi}I_{\gamma p} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (5.170)$$

So that we get the linear estimator \hat{Y}_t

$$\hat{Y}_t = Z A \beta_t = \frac{1}{\sqrt{Y}} X_0 \beta_t \quad (5.171)$$

Analytic result as U and V^* commute again, the joint probability distribution can be derived:

$$\mathcal{P}\left(u = \frac{1}{\gamma\psi}, v = \frac{r^2}{\psi}\right) = \gamma\psi \quad (5.172)$$

$$\mathcal{P}\left(u = 0, v = \frac{r^2}{\psi}\right) = (1-\gamma)\psi \quad (5.173)$$

$$\mathcal{P}\left(u = 0, v = \frac{\sigma^2}{(1-\psi)}\right) = 1-\psi \quad (5.174)$$

Therefore, in the regime $\lambda = 0$, with $\kappa = \frac{\phi_0}{\gamma}$, a calculation leads to the following result (dubbed the "mismatched model" in Hastie et al. (2019))

$$\mathcal{E}_{\text{gen}(+\infty)} = \tilde{f}_1 = \begin{cases} \frac{\kappa}{\kappa-1}(\sigma^2 + (1-\gamma)r^2) & (\kappa > 1) \\ \frac{1}{1-\kappa}\sigma^2 + r^2\gamma(1-\kappa) & (\kappa < 1) \end{cases} \quad (5.175)$$

5.D.2 Non isotropic model

We have the joint probabilities $P(u = \alpha^{-i}, v = 1) = \frac{1}{p} = \gamma$ for $i \in \{0, \dots, p-1\}$ and $\lambda = 0$. Then:

$$\tilde{f}_1 = \frac{1}{p} \sum_{i=0}^{p-1} \frac{\tilde{f}_1 \phi + (\alpha^i \zeta)^2}{(\phi + \alpha^i \zeta)^2} \quad (5.176)$$

$$\zeta = \frac{1}{p} \sum_{i=0}^{p-1} \frac{\zeta}{\phi + \alpha^i \zeta} \quad (5.177)$$

$$f_2 = c_0 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\zeta \alpha^i}{\phi + \alpha^i \zeta} \quad (5.178)$$

So either $\zeta = 0$ and thus $\tilde{f}_1 = 0$, or $\zeta \neq 0$ and:

$$\tilde{f}_1 = \left(1 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\phi}{(\phi + \alpha^i \zeta)^2}\right)^{-1} \frac{1}{p} \sum_{i=0}^{p-1} \frac{(\alpha^i \zeta)^2}{(\phi + \alpha^i \zeta)^2} \quad (5.179)$$

$$1 = \frac{1}{p} \sum_{i=0}^{p-1} \frac{1}{\phi + \alpha^i \zeta} \quad (5.180)$$

Writing further down $(\alpha^i \zeta)^2 = (\alpha^i \zeta + \phi - \phi)^2 = (\alpha^i \zeta + \phi)^2 - 2\phi(\alpha^i \zeta + \phi) + \phi^2$ we get:

$$\frac{1}{p} \sum_{i=0}^{p-1} \frac{(\alpha^i \zeta)^2}{(\phi + \alpha^i \zeta)^2} = 1 - 2\phi \frac{1}{p} \sum_{i=0}^{p-1} \frac{1}{\phi + \alpha^i \zeta} + \phi^2 \frac{1}{p} \sum_{i=0}^{p-1} \frac{1}{(\phi + \alpha^i \zeta)^2} \quad (5.181)$$

$$= 1 - 2\phi + \phi^2 \frac{1}{p} \sum_{i=0}^{p-1} \frac{1}{(\phi + \alpha^i \zeta)^2} \quad (5.182)$$

$$= (1 - \phi) - \phi \left(1 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\phi}{(\phi + \alpha^i \zeta)^2} \right) \quad (5.183)$$

So:

$$\tilde{f}_1 = (1 - \phi) \left(1 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\phi}{(\phi + \alpha^i \zeta)^2} \right)^{-1} - \phi \quad (5.184)$$

Now injecting the expression for ζ :

$$1 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\phi}{(\phi + \alpha^i \zeta)^2} = \frac{1}{p} \sum_{i=0}^{p-1} \left(\frac{1}{\phi + \alpha^i \zeta} - \frac{\phi}{(\phi + \alpha^i \zeta)^2} \right) \quad (5.185)$$

$$= \frac{1}{p} \sum_{i=0}^{p-1} \frac{\alpha^i \zeta}{(\phi + \alpha^i \zeta)^2} \quad (5.186)$$

Hence the formula

$$\mathcal{E}_{\text{gen}}(\infty) = (1 - \phi) \left(\frac{1}{p} \sum_{i=0}^{p-1} \frac{\alpha^i \zeta}{(\phi + \alpha^i \zeta)^2} \right)^{-1} - \phi \quad (5.187)$$

Asymptotic limit: Let's consider the behavior of the generalisation error when $\alpha \rightarrow \infty$. Let's consider the potential solution for some $k \in \{0, \dots, p-1\}$:

$$\zeta^k = \frac{c_k}{\alpha^k} (1 + o_\alpha(1)) \quad (5.188)$$

for some constant c_k . Then:

$$p = \sum_{i=0}^{p-1} \frac{1}{\phi + c_k \alpha^{i-k} (1 + o_\alpha(1))} = \frac{1}{\phi + c_k} + \frac{k}{\phi} + o_\alpha(1) \quad (5.189)$$

Hence we choose:

$$c_k = \phi \left(\frac{1}{p\phi - k} - 1 \right) \quad (5.190)$$

Because $\mathcal{E}_{\text{gen}}(\infty) \geq 0$, we need to enforce $\zeta^k > 0$ which leads to the condition $\frac{1}{p\phi - k} - 1 \geq 0$, that is $1 \geq p\phi - k > 0$. So in fact it implies $\phi \in \left] \frac{k}{p}, \frac{k+1}{p} \right]$, so ζ^k can only be a solution for ϕ in this range. Therefore we can consider the solution $\zeta(\phi) = \sum_{i=0}^{p-1} \mathbb{1}_{\left] \frac{k}{p}, \frac{k+1}{p} \right]}(\phi) \zeta^k(\phi)$. Then notice:

$$\sum_{i=0}^{p-1} \frac{\alpha^i \zeta^k}{(\phi + \alpha^i \zeta^k)^2} = \frac{c_k}{(c_k + \phi)^2} + o_\alpha(1) = -p^2 \left(\phi - \frac{k}{p} \right) \left(\phi - \frac{k+1}{p} \right) + o_\alpha(1) \quad (5.191)$$

and thus for $\phi \in [0, 1] \setminus \frac{k}{p}\mathbb{Z}$:

$$\mathcal{E}_{\text{gen}}(\infty) = \sum_{k=0}^{p-1} \frac{\phi(1-\phi)}{p \left(\phi - \frac{k}{p}\right) \left(\frac{k+1}{p} - \phi\right)} \mathbb{1}_{\left] \frac{k}{p}, \frac{k+1}{p} \right[} (\phi) - \phi + o_{\alpha}(1) \quad (5.192)$$

So we clearly see that in the limit of α large, the test error approaches a function with two roots at the denominator.

Evolution:

$$\tilde{f}_1(x, y) = \frac{1}{p} \sum_{i=0}^{p-1} \frac{\tilde{f}_1(x, y) \phi + \alpha^{2i} \zeta_x \zeta_y}{(\phi + \alpha^i \zeta_x)(\phi + \alpha^i \zeta_y)} \quad (5.193)$$

$$\zeta_z = -z + \frac{1}{p} \sum_{i=0}^{p-1} \frac{\zeta_z}{\phi + \alpha^i \zeta_z} \quad (5.194)$$

$$f_2(z) = c_0 - \frac{1}{p} \sum_{i=0}^{p-1} \frac{\alpha^i \zeta_z}{\phi + \alpha^i \zeta_z} \quad (5.195)$$

In particular f_2 is given by:

$$f_2(z) = c_0 - 1 + \frac{\phi}{p} \sum_{i=0}^{p-1} \frac{1}{\phi + \alpha^i \zeta_z} = c_0 - 1 + \phi \zeta_z \left(1 + \frac{z}{\zeta_z}\right) \quad (5.196)$$

and \tilde{f}_1 is given by:

$$\tilde{f}_1(x, y) = \frac{\frac{1}{p} \sum_{i=0}^{p-1} \frac{\alpha^{2i} \zeta_x \zeta_y}{(\phi + \alpha^i \zeta_x)(\phi + \alpha^i \zeta_y)}}{1 - \frac{\phi}{p} \sum_{i=0}^{p-1} \frac{1}{(\phi + \alpha^i \zeta_x)(\phi + \alpha^i \zeta_y)}} \quad (5.197)$$

Eigenvalue distribution

In our figures, we look at the log-eigenvalue distribution of the student data $\rho_{\log \lambda}$ as it provides the most natural distributions on a log-scale basis. So in fact, if we plot the curve $y(x) = \rho_{\log \lambda}(x)$ we have:

$$y(x) = \rho_{\log \lambda}(x) = \frac{\partial}{\partial x} \mathcal{P}(\log \lambda \leq x) \quad (5.198)$$

$$= \frac{\partial}{\partial x} \mathcal{P}(\lambda \leq e^x) \quad (5.199)$$

$$= e^x \rho_{\lambda}(e^x) \quad (5.200)$$

So in a log-scale basis we have $\rho_{\log \lambda}(\log x) = x \rho_{\lambda}(x)$. It is interesting to notice the connection with η_x for running computer simulations:

$$\rho_{\log \lambda}(\log x) = \frac{x}{\pi} \lim_{\epsilon \rightarrow 0^+} m(x + i\epsilon) = \frac{1}{\pi} \lim_{\epsilon \rightarrow 0^+} \frac{x + i\epsilon}{\zeta(x + i\epsilon)} = -\frac{1}{\pi} \lim_{\epsilon \rightarrow 0^+} \eta_{x+i\epsilon} \quad (5.201)$$

It is work mentioning that the bulks are further "detached" as α grows as it can be seen in figure 5.D.1. Furthermore, bigger α makes the spike more distinguisable.

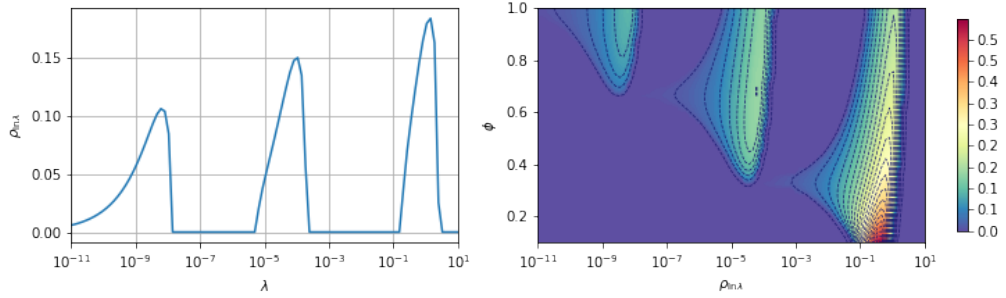


Figure 5.D.1: Theoretical (log-)eigenvalue distribution in the non-isotropic ridgeless regression model with $p = 3$, $\lambda = 10^{-5}$, $\alpha = 10^4$ with $\phi = 1$ on the left and a range $\phi \in (0, 1)$ on the right heatmap.

5.D.3 Kernel Methods

Kernel methods are equivalent to solving the following linear regression problem:

$$\beta = \arg \min_{\beta} \sum_{i=1}^n (\theta_0^T \phi(x_i) - \beta^T \phi(x_i))^2 + \lambda \|\beta\|^2 \quad (5.202)$$

Where $\phi(x) = (\phi_i(x))_{i \in \mathbb{N}} = (\sqrt{\omega_i} e_i(x))$ for some orthogonal basis $(e_i)_{i \in \mathbb{N}}$. In fact we can consider:

$$A = B = \begin{pmatrix} \sqrt{\omega_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\omega_d} \end{pmatrix} \quad (5.203)$$

and $z_i = (e_1(x_i), \dots, e_d(x_i))$. Then let's consider the following linear regression problem:

$$\hat{\beta} = \arg \min_{\beta} \|Z(B\beta^* - A\beta)\|^2 + \lambda \|\beta\|^2 \quad (5.204)$$

$$\mathcal{E}_{\text{gen}}(\hat{\beta}) = \mathbb{E}_z \left[(z^T (B\beta^* - A\hat{\beta}))^2 \right] \quad (5.205)$$

This problem is identical to the kernel methods in the situation with a specific $\beta^{*T} = (\theta_{01}, \dots, \theta_{0d})$. Although V^* and U don't commute with each other, Notice that with $x = y = -\lambda$, due to the

diagonal structure of U :

$$\tilde{f}_1 = \text{Tr}_d [(\phi U + \zeta I)^{-1} (\zeta^2 V^* + \tilde{f}_1 \phi U^2) (\phi U + \zeta I)^{-1}] \quad (5.206)$$

$$= \frac{1}{d} \sum_{i=1}^d [(\zeta^2 V^* + \tilde{f}_1 \phi U^2) (\phi U + \zeta I)^{-2}]_{ii} \quad (5.207)$$

$$= \frac{1}{d} \sum_{i=1}^d (\zeta^2 [V^*]_{ii} + \tilde{f}_1 \phi [U^2]_{ii}) (\phi [U]_{ii} + \zeta)^{-2} \quad (5.208)$$

So in fact we find the self-consistent set of equation with $\mathcal{E}_{\text{gen}}(+\infty) = \tilde{f}_1$:

$$\zeta = \lambda + \frac{1}{d} \sum_{i=1}^d \frac{\zeta \omega_i}{\phi \omega_i + \zeta} \quad (5.209)$$

$$\tilde{f}_1 = \frac{1}{d} \sum_{i=1}^d \frac{\tilde{f}_1 \phi \omega_i^2 + \zeta^2 \theta_{0i}^2 \omega_i}{(\phi \omega_i + \zeta)^2} \quad (5.210)$$

This is precisely the results from equation (78) in Loureiro et al. (2021) (see also Bordelon et al. (2020)) with the change of variables $\lambda(1 + V) \rightarrow \zeta$ and $\rho + q - 2m \rightarrow \tilde{f}_1$.

5.D.4 Random feature example

We get the following matrices U, V with $\tilde{\mu}^2 = \frac{\mu^2}{\psi}, \tilde{\nu}^2 = \frac{\nu^2}{\psi}, \tilde{r}^2 = \frac{r^2}{\psi}, \tilde{\sigma}^2 = \frac{\sigma^2}{1 - (1 + \psi_0)\psi}$:

$$U = \begin{pmatrix} \tilde{\mu}^2 W W^T & \tilde{\mu} \tilde{\nu} W & 0 \\ \tilde{\mu} \tilde{\nu} W^T & \tilde{\nu}^2 I_N & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad V = \begin{pmatrix} \tilde{r}^2 I_p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \tilde{\sigma}^2 I_q \end{pmatrix} \quad (5.211)$$

In fact, the matrices U and V do not commute with each other, so we have more involved calculations. First we consider the subspace $F = \text{Ker}(V - \tilde{\sigma}^2 I_q)^\perp$. Let's define the matrices:

$$U_F = \begin{pmatrix} \tilde{\mu}^2 W W^T & \tilde{\mu} \tilde{\nu} W \\ \tilde{\mu} \tilde{\nu} W^T & \tilde{\nu}^2 I_N \end{pmatrix} \quad V_F = \begin{pmatrix} \tilde{r}^2 I_p & 0 \\ 0 & 0 \end{pmatrix} \quad (5.212)$$

$$U_{F^\perp} = \begin{pmatrix} 0 \end{pmatrix} \quad V_{F^\perp} = \begin{pmatrix} \tilde{\sigma}^2 I_q \end{pmatrix} \quad (5.213)$$

Then, although U and V can't be diagonalized in the same basis, they are still both block-diagonal matrices in the same direct-sum space $\mathbb{R}^d = F \oplus F^\perp$, so in fact the following split between the two subspaces F and F^\perp holds:

$$\tilde{f}_1 = \text{Tr}_d [(\phi U_F + \zeta_x I)^{-1} \zeta_x \zeta_y V_F (\phi U_F + \zeta_y I)^{-1}] \quad (5.214)$$

$$+ \text{Tr}_d [(\phi U_{F^\perp} + \zeta_x I)^{-1} \zeta_x \zeta_y V_{F^\perp} (\phi U_{F^\perp} + \zeta_y I)^{-1}] \quad (5.215)$$

$$+ \text{Tr}_d [(\phi U + \zeta_x I)^{-1} \tilde{f}_1 \phi U^2 (\phi U + \zeta_y I)^{-1}] \quad (5.216)$$

Now let's define $\kappa_1, \kappa_2, \kappa_3$ such that:

$$\tilde{f}_1 = r^2 \kappa_1 + \tilde{f}_1 (1 - \kappa_2^{-1}) + \sigma^2 \kappa_3 \quad (5.217)$$

That is to say, we get directly $\tilde{f}_1 = (r^2 \kappa_1 + \sigma^2 \kappa_3) \kappa_2$ and by definition:

$$r^2 \kappa_1 = \text{Tr}_d [(\phi U_F + \zeta_x I)^{-1} \zeta_x \zeta_y V_F (\phi U_F + \zeta_y I)^{-1}] \quad (5.218)$$

$$1 - \frac{1}{\kappa_2} = \text{Tr}_d [(\phi U + \zeta_x I)^{-1} \phi U^2 (\phi U + \zeta_y I)^{-1}] \quad (5.219)$$

$$\sigma^2 \kappa_3 = \text{Tr}_d [(\phi U_{F^\perp} + \zeta_x I_q)^{-1} \zeta_x \zeta_y V_{F^\perp} (\phi U_{F^\perp} + \zeta_y I_q)^{-1}] = \sigma^2 \quad (5.220)$$

So we already know that $\kappa_3 = 1$. Let's focus on κ_1 , we can deal with a linear pencil M such that we would get the desired term. First we define similarly A_F^T , the restriction of A^T on the subspace F :

$$A_F = \begin{pmatrix} \tilde{\mu} W \\ \tilde{\nu} I_N \end{pmatrix} \implies U_F = A_F A_F^T \quad (5.221)$$

Then, following the structure of M_1 we can construct the following linear-pencil M :

$$M = \begin{pmatrix} 0 & 0 & \zeta_y I & A_F \\ 0 & 0 & A_F^T & -\frac{1}{\phi} I \\ \zeta_x I & A_F & -\zeta_x \zeta_y V_F & 0 \\ A_F^T & -\frac{1}{\phi} I & 0 & 0 \end{pmatrix} = \left(\begin{array}{c|c} 0 & B_y \\ B_x & \begin{pmatrix} -\zeta_x \zeta_y V_F & 0 \\ 0 & 0 \end{pmatrix} \end{array} \right) \quad (5.222)$$

So that:

$$M^{-1} = \left(\begin{array}{c|c} B_x^{-1} \begin{pmatrix} -\zeta_x \zeta_y V_F & 0 \\ 0 & 0 \end{pmatrix} B_y^{-1} & B_x^{-1} \\ \hline B_y^{-1} & 0 \end{array} \right) \quad (5.223)$$

where:

$$B_x^{-1} = \begin{pmatrix} (\phi U_F + \zeta_x I)^{-1} & \phi (\phi U_F + \zeta_x I)^{-1} A_F \\ A_F^T \phi (\phi U_F + \zeta_x I)^{-1} & (-\frac{1}{\phi} I - \frac{1}{\zeta_y} A_F^T A_F)^{-1} \end{pmatrix} \quad (5.224)$$

In the above matrices, the sub-blocks A_F and V_F are implicitly flattened, so in fact M is given completely by:

$$M = \begin{pmatrix} 0 & 0 & 0 & \zeta_y I & 0 & \tilde{\mu} W \\ 0 & 0 & 0 & 0 & \zeta_y I & \tilde{\nu} I \\ 0 & 0 & 0 & \tilde{\mu} W^T & \tilde{\nu} I & -\frac{1}{\phi} I \\ \zeta_x I & 0 & \tilde{\mu} W & -\tilde{r}^2 \zeta_x \zeta_y I_p & 0 & 0 \\ 0 & \zeta_x I & \tilde{\nu} I & 0 & 0 & 0 \\ \tilde{\mu} W^T & \tilde{\nu} I & -\frac{1}{\phi} I & 0 & 0 & 0 \end{pmatrix} \quad (5.225)$$

and therefore, one has to pay attention on the quantity of interest which is given by a sum of

Chapter 5. A framework: the gaussian covariate model

two terms:

$$r^2 \kappa_1 = \lim_{d \rightarrow +\infty} \left(\frac{p}{d} g^{(11)} + \frac{N}{d} g^{(22)} \right) = \psi(g^{(11)} + \psi_0 g^{(22)}) \quad (5.226)$$

Using a Computer-Algebra-System, we get the equations with $\gamma_x, \gamma_y, \delta_x, \delta_y$ defined such that $g^{(36)} = -\psi \gamma_x \zeta_x$, $g^{(63)} = -\psi \gamma_y \zeta_y$, $\delta_x = \zeta_x g^{(14)}$, $\delta_y = \zeta_y g^{(41)}$:

$$\psi g^{(11)} = (\zeta_x \zeta_y)^{-1} (\delta_x \delta_y) (r^2 \zeta_x \zeta_y + \mu^2 \psi_0 g^{(33)}) \quad (5.227)$$

$$\psi g^{(22)} = \phi^{-2} (\gamma_x \gamma_y) (\psi g^{(11)} \mu^2 v^2 \phi^2) \quad (5.228)$$

$$g^{(33)} = (\zeta_x \zeta_y) (\gamma_x \gamma_y) (\psi g^{(11)} \mu^2) \quad (5.229)$$

$$\delta_y = (1 + \gamma_y \mu^2 \psi_0)^{-1} \quad (5.230)$$

$$\gamma_y = (\mu^2 \delta_y + \phi_0^{-1} \zeta_y + v^2)^{-1} \quad (5.231)$$

So:

$$(1 - \mu^4 \psi_0 (\delta_x \delta_y) (\gamma_x \gamma_y)) \psi g^{(11)} = (\delta_x \delta_y) (r^2) \quad (5.232)$$

and:

$$\psi g^{(11)} + \psi_0 \psi g^{(22)} = (1 + \psi_0 \mu^2 v^2 (\gamma_x \gamma_y)) (\psi g^{(11)}) \quad (5.233)$$

Hence the result:

$$\kappa_1 = \frac{1 + v^2 \mu^2 \psi_0 (\gamma_x \gamma_y)}{1 - \mu^4 \psi_0 (\delta_x \delta_y) (\gamma_x \gamma_y)} (\delta_x \delta_y) \quad (5.234)$$

Also there remain to use the last equation regarding ζ_x using the fact that:

$$\zeta_y + y = \text{Tr}_d [\zeta_x U (\phi U + \zeta_x I)^{-1}] \quad (5.235)$$

Notice that we have

$$g^{(63)} = -\gamma_y \psi \zeta_y = \text{Tr}_N \left[\left(-\frac{1}{\phi} I - \frac{1}{\zeta_y} A_F^T A_F \right)^{-1} \right] \quad (5.236)$$

So because $A_F^T A_F = A^T A$:

$$\zeta_y \gamma_y = \phi_0 \zeta_y \text{Tr}_N [(\phi A^T A + \zeta_y I)^{-1}] \quad (5.237)$$

$$= \phi_0 \text{Tr}_N [(\phi A^T A + \zeta_y I - \phi A^T A) (\phi A^T A + \zeta_y I)^{-1}] \quad (5.238)$$

$$= \phi_0 \text{Tr}_N [I - \phi A^T A (\phi A^T A + \zeta_y I)^{-1}] \quad (5.239)$$

$$= \phi_0 (1 - \text{Tr}_N [\phi (\phi U + \zeta_y I)^{-1} U]) \quad (5.240)$$

$$= \phi_0 \left(1 - \frac{\phi_0}{\psi_0 \zeta_y} \text{Tr}_d [\zeta_y U (\phi U + \zeta_y I)^{-1}] \right) \quad (5.241)$$

$$= \phi_0 \left(1 - \frac{\phi_0}{\psi_0 \zeta_y} (\zeta_y + y) \right) \quad (5.242)$$

Therefore:

$$\frac{\gamma_y}{\phi_0} \zeta_y = 1 - \frac{\phi_0}{\psi_0} \left(1 + \frac{y}{\zeta_y} \right) \quad (5.243)$$

For κ_2 we can calculate the following expression - which in fact is general and doesn't depend

on the specific design of U :

$$1 - \frac{1}{\kappa_2} = \text{Tr}_d [(\phi U + \zeta_x I)^{-1} \phi U^2 (\phi U + \zeta_y I)^{-1}] \quad (5.244)$$

$$= \text{Tr}_d [(\phi U + \zeta_x I)^{-1} (\phi U + \zeta_x I - \zeta_x I) U (\phi U + \zeta_y I)^{-1}] \quad (5.245)$$

$$= \text{Tr}_d [(I - \zeta_x (\phi U + \zeta_x I)^{-1}) U (\phi U + \zeta_y I)^{-1}] \quad (5.246)$$

$$= \text{Tr}_d [U (\phi U + \zeta_y I)^{-1} - \zeta_x (\phi U + \zeta_x I)^{-1} U (\phi U + \zeta_y I)^{-1}] \quad (5.247)$$

$$= \text{Tr}_d \left[U (\phi U + \zeta_y I)^{-1} - \frac{\zeta_x}{\zeta_y - \zeta_x} (U (\phi U + \zeta_x I)^{-1} - U (\phi U + \zeta_y I)^{-1}) \right] \quad (5.248)$$

$$= \frac{1}{\zeta_y - \zeta_x} \text{Tr}_d [\zeta_y U (\phi U + \zeta_y I)^{-1} - \zeta_x U (\phi U + \zeta_x I)^{-1}] \quad (5.249)$$

$$= \frac{1}{\zeta_y - \zeta_x} (\zeta_y + y - \zeta_x - x) \quad (5.250)$$

$$= 1 + \frac{y - x}{\zeta_y - \zeta_x} \quad (5.251)$$

Hence the general formula:

$$\kappa_2 = -\frac{\zeta_y - \zeta_x}{y - x} \quad (5.252)$$

One can check that the same formula applies for instance for the mismatched ridgeless regression. Also, we assume that it can be replaced by its continuous limit in $y \rightarrow x$ in the situation $x = y$.

Finally for f_2 , we find

$$f_2 = c_0 - \text{Tr}_d [\zeta_z V (\phi U + \zeta_z I)^{-1}] \quad (5.253)$$

$$= c_0 - \text{Tr}_d [\zeta_z V_{F^\perp} (\phi U_{F^\perp} + \zeta_z I)^{-1}] - \text{Tr}_d [\zeta_z V_F (\phi U_F + \zeta_z I)^{-1}] \quad (5.254)$$

$$= c_0 - \sigma^2 - \lim_{d \rightarrow +\infty} \left(\frac{p}{d} \tilde{g}^{(11)} + \frac{N}{d} \tilde{g}^{(22)} \right) \quad (5.255)$$

$$= c_0 - \sigma^2 - \psi(\tilde{g}^{(11)}) + \phi_0 \tilde{g}^{(22)} \quad (5.256)$$

where we use \tilde{g} associated to a slightly different linear-pencil \tilde{M} :

$$\tilde{M} = \begin{pmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ \zeta_z I & A_F & -\zeta_z V_F & 0 \\ A_F^T & -\frac{1}{\phi} I & 0 & 0 \end{pmatrix} \quad (5.257)$$

from which we get using a Compute-Algebra-System

$$\psi \tilde{g}^{(11)} + \psi \phi_0 \tilde{g}^{(22)} = r^2 \delta_z \quad (5.258)$$

Another more straightforward way for obtaining the same result without the need for an additional linear-pencil is to notice that if we let $E_1 = (I_p | 0_{p \times N})$ such that $V_F = \tilde{r}^2 E_1 E_1^T$, then

we have:

$$\text{Tr}_d [\zeta_x V_F (\phi U_F + \zeta_x I)^{-1}] = \text{Tr}_d [\zeta_x \tilde{r}^2 E_1^T (\phi U_F + \zeta_x I)^{-1} E_1] \quad (5.259)$$

$$= \tilde{r}^2 \zeta_x \text{Tr}_p [E_1^T (\phi U_F + \zeta_x I)^{-1} E_1] \quad (5.260)$$

Therefore reusing the definition of δ_x and the former linear-pencil M :

$$\text{Tr}_d [\zeta_x V_F (\phi U_F + \zeta_x I)^{-1}] = \tilde{r}^2 \psi \zeta_x g^{(14)} = r^2 \delta_x \quad (5.261)$$

Conclusion we have the following equations

$$\tilde{f}_1(x, y) = \left(-\frac{\zeta_y - \zeta_x}{y - x} \right) \left(r^2 \frac{1 + v^2 \mu^2 \psi_0 (\gamma_x \gamma_y)}{1 - \mu^4 \psi_0 (\delta_x \delta_y) (\gamma_x \gamma_y)} (\delta_x \delta_y) + \sigma^2 \right) \quad (5.262)$$

$$f_2(z) = c_0 - (r^2 \delta_z + \sigma^2) \quad (5.263)$$

$$\delta_z = (1 + \gamma_z \mu^2 \psi_0)^{-1} \quad (5.264)$$

$$\gamma_z = (\mu^2 \delta_z + \phi_0^{-1} \zeta_z + v^2)^{-1} \quad (5.265)$$

$$\frac{\gamma_y}{\phi_0} \zeta_y = 1 - \frac{\phi_0}{\psi_0} \left(1 + \frac{y}{\zeta_y} \right) \quad (5.266)$$

5.D.5 Realistic datasets

For the realistic datasets, we capture the time evolution for two different datasets: MNIST and Fashion-MNIST. To capture the dynamics over a realistic dataset $X \in \mathbb{R}^{n_{\text{tot}} \times d}$, it is more convenient to use the dual matrices U_\star, V_\star, Ξ . We only need to estimate U_\star and $\Xi \beta^\star$ with $U_\star \simeq \frac{1}{n_{\text{tot}}} X^T X$ and $\Xi \beta^\star \simeq \frac{1}{n_{\text{tot}}} X^T Y$. In both cases, we sill sample a subset of $n < n_{\text{tot}}$ data-samples for the training set. The scope of the theoretical equations is still subject to the high-dimensional limit assumption, in other words we need n and d "large enough", that is to say $1 \ll n$. At the same time, the approximation of U_\star and $\Xi \beta^\star$ hints at n_{tot} sufficiently large compared to the number of considered samples n . Hence we need also $n \ll n_{\text{tot}}$.

Numerically, for the two following datasets and as per assumptions 5.1, the theoretical prediction rely on a contour enclosing the spectrum $\text{Sp}(\hat{X}^T \hat{X})$ of $\hat{X}^T \hat{X}$, but not enclosing $-\lambda$. Therefore, in order to proceed with our computations, we take a symmetric rectangle around the x-axis crossing the axis at the particular values $-\frac{\lambda}{2}$ and $1.2 \max \text{Sp}(\hat{X}^T \hat{X})$ after a preliminary computation of the spectrum. For the need of our experiments, we commonly discretized the contour and ran a numerical integration over the discretized set of points.

MNIST Dataset: We consider the MNIST dataset with $n_{\text{tot}} = 70'000$ images of size 28×28 of numbers between 0 and 9. In our setting, we consider the problem of estimating the parity of the number, that is the vector Y with $Y_i = 1$ if image i represents an even number and $Y_i = -1$ for an odd-number. The dataset $X \in \mathbb{R}^{n_{\text{tot}} \times d}$ is further processed by centering each column to

its mean, and normalized by the global standard-deviation of X (in other words the standard deviation of X seen as a flattened $n_{\text{tot}} \times d$ vector) and further by \sqrt{d} (for consistency with the theoretical random matrix Z).

The results that we obtain are shown in Figure 5.3.4. On the figure on the left side we show the theoretical prediction of the training and test error with the minimum least-squares estimator (or alternatively the limiting errors at $t = +\infty$). We make the following observations which in fact relates to the same ones as in Figure 4 in Loureiro et al. (2021):

- There is an apparent larger deviation in the test error for smaller n which tends to heal with increasing number of data samples
- A bias between the mean observation of the test error and the theoretical prediction emerges around the double-descent peak between $n = 100$ and $n = 1000$, in particular, the experiments are slightly above the given prediction. We notice that this bias is even more pronounced for smaller values of λ .
- Although it is not visible on the figure, increasing n further tends to create another divergence between the theoretical prediction and the experimental runs - as it is expected with n getting closer to n_{tot} .

Besides the limiting error, we chose to draw the time-evolution of the training and test error around at $n = 700$ around the double descent on the right side of Figure 5.3.4. This time, a gradient descent algorithm is executed for each 10 experimental runs with a constant learning-rate $dt = 0.01$. Due to the log-scale of the axis, it is interesting to notice that with such a basic non-adaptive learning-rate, each tick on the graph entails 10 times more computational time to update the weights. By contrast, the theoretical curves can be calculated at any point in time much farther away. Overall we see a good agreement between the evolution of the experimental runs with the theoretical predictions. However, as it is expected around the double-descent spike, learning-curves of the experimental runs appear slightly biased and above the theoretical curves.

Our analytical framework thus offers the possibility to capture the full theoretical evolution of the training and test errors at any given point in time across a variable range of the parameter n . This capability is illustrated in the heat-maps presented in Figure 5.3.5, showcasing the complete theoretical evolution of both errors for the same model and dataset with $\lambda = 10^{-3}$.

Fashion-MNIST Dataset: We provide another example with MNIST-Fashion dataset with $d = 784$ and $n_{\text{tot}} = 70'000$. The dataset X is processed as for the MNIST dataset. We take the output vector Y such that $Y_i = 1$ for items i above the waist, and $Y_i = -1$ otherwise. We provide the results in Figure 5.D.2 where the training set is sampled randomly with n elements in n_{tot} and the test set is sampled in the remaining examples. As it can be seen, the test error is slightly above the prediction for $n < 10^3$ but fits well with the predicted values for larger

Chapter 5. A framework: the gaussian covariate model

n . Furthermore, the learning curves through time in Figure 5.D.3 are different compared to the MNIST dataset in Figure 5.3.4 and we still observe a good match with the theoretical predictions. However the mismatch in the learning curves seems to increase in the specific case when λ is lower, increasing thereby the effect of the double descent.

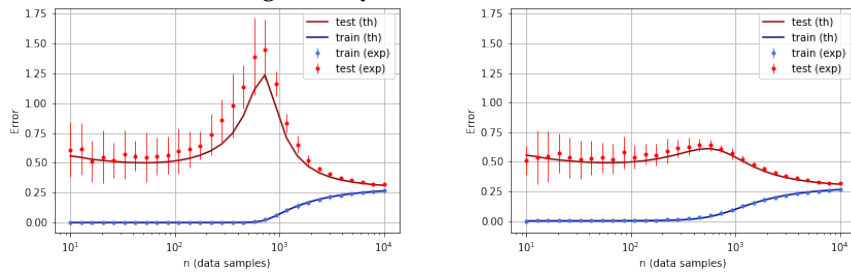


Figure 5.D.2: Comparison between the analytical and experimental learning profiles for the minimum least-squares estimator at $\lambda = 10^{-3}$ on the left (average and ± 2 -standard-deviations over 20 runs) and $\lambda = 10^{-2}$, $n = 700$ on the right.

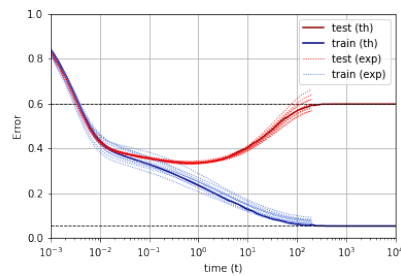


Figure 5.D.3: Comparison between the analytical and experimental learning evolution at $\lambda = 10^{-2}$, $n = 700$ (10 runs).

6 The Random feature model

Recent evidence has shown the existence of a so-called double-descent and even triple-descent behavior for the generalization error of deep-learning models. This important phenomenon commonly appears in implemented neural network architectures, and also seems to emerge in epoch-wise curves during the training process. A recent line of research has highlighted that random matrix tools can be used to obtain precise analytical asymptotics of the generalization (and training) errors of the random feature model. In this chapter which is based on the work (Bodin and Macris, 2021a), we analyze the *whole temporal behavior* of the generalization and training errors under gradient flow for the random feature model which has been described in model 1.3 in the introduction and briefly outlined in Chapter 5, Section 5.3.3. This chapter stands alone as a self-contained unit, reintroducing the random-feature model independently of the Gaussian covariate framework. Furthermore, we conduct a more comprehensive analysis of this model. We show that in the asymptotic limit of large system size the *full time-evolution* path of both errors can be calculated analytically. This allows us to observe how the double and triple descents develop over time, if and when early stopping is an option, and also observe time-wise descent structures. Our techniques are based on Cauchy complex integral representations of the errors together with recent random matrix methods based on linear pencils.

6.1 Introduction

Deep learning models have vastly increased in terms of number of parameters in the architecture and data sample sizes with recent applications using unprecedented numbers with as much as 175 billions parameters trained over billions of tokens Brown et al. (2020). Such massive amounts of data and growing training budgets have spurred research seeking empirical power laws to scale model sizes appropriately with available resources Kaplan et al. (2020), and nowadays it is common wisdom among practitioners that "larger models are better". This ongoing trend has been challenging classical statistical modeling where it is thought that increasing the number of parameters past an interpolation threshold (at which the training error vanishes while the test error usually increases) is doomed to over-fit the data Hastie et al.

(2001). We refer to Zhang et al. (2016) for a recent extensive discussion on this contradictory state of affairs. Progress towards rationalizing this situation came from a series of papers Belkin et al. (2019b, 2018, 2019c, 2020b); Spigler et al. (2019a); Geiger et al. (2020); Advani et al. (2020b) evidencing the existence of phases where increasing the number of parameters beyond the interpolation threshold can actually achieve good generalization, and the characteristic U curve of the bias-variance tradeoff is followed by a "descent" of the generalization error. This phenomenon has been called the *double descent* and was analytically corroborated in linear models Hastie et al. (2019); Dereziński et al. (2020); Muthukumar et al. (2020); Bartlett et al. (2020); Deng et al. (2021b) as well as random feature (RF) (or random feature regression) shallow network models Mei and Montanari (2019); Liao et al. (2020); Gerace et al. (2020b); D'Ascoli et al. (2020). Many of these works provide rigorous proofs with precise asymptotic expressions of double descent curves. Further developments have brought forward rich phenomenology, for example, a triple-descent phenomenon d'Ascoli et al. (2020) linked to the degree of non-linearity of the activation function. Further empirical evidence Nakkiran et al. (2020a) has also shown that a similar effect occurs *while* training (ResNet18s on CIFAR10 trained using Adam) and has been called *epoch-wise double descent*. Moreover the authors of Nakkiran et al. (2020a) extensively test various CIFAR data sets, architectures (CNNs, ResNets, Transformers) and optimizers (SGD, Adam) and classify their observations into three types of double descents: (i) model-wise double descent when the number of network parameters is varied; (ii) sample-wise double descent when the data set size is varied; and (iii) epoch-wise double descent which occurs while training. We wish to note that sample-wise double descent was derived long ago in precursor work on single layer perceptron networks Opper (1998); Engel and Van den Broeck (2001b). An important theoretical challenge is to unravel all these structures in a unified analytical way and understand how generalization error evolves in time.

In this contribution we achieve a detailed analytical analysis of the gradient flow dynamics of the RF model (or regression) in the high-dimensional asymptotic limit. The model was initially introduced in Rahimi and Recht (2008) as an approximation of kernel machines; more recently it has been recognized as an important playground for theoretical analysis of the model-wise double descent phenomenon, using tools from random matrix theory Mei and Montanari (2019); Liao et al. (2020); Jacot et al. (2020b). Following Mei and Montanari (2019) we view the RF model as a 2-layer neural network with fixed-random-first-layer-weights and dynamical second layer learned weights. The data is given by n training pairs constituted of d -dimensional input vectors and output given by a linear function with additive gaussian noise. The data is fed through N neurons with a non-linear activation function and followed by one linear neuron whose weights we learn by gradient descent over a quadratic loss function. The high-dimensional asymptotic limit is defined as the regime $n, d, N \rightarrow +\infty$ while the ratios tend to finite values $\frac{N}{d} \rightarrow \psi$ and $\frac{n}{d} \rightarrow \phi$. As the training loss is convex one expects that the least-squares predictor (with Moore-Penrose inversion) gives the long time behavior of gradient descent. This has led to the calculation of highly non-trivial analytical algebraic expressions for training and generalization errors which describe (model-wise and sample-wise) double and triple descent curves Mei and Montanari (2019); d'Ascoli et al. (2020). However, to the

best of our knowledge, there is no complete analytical derivation of the whole time evolution of the two errors.

We analyze the gradient flow equations in the high-dimensional regime and deduce the whole time evolution of the training and generalization errors. Numerical simulations show that the gradient flow is an excellent approximation of gradient descent in the high-dimensional regime as long as the step size is small enough (see Fig. 6.3.1). Main contributions presented in detail in Sect. 6.3 comprise:

- a.** Results 6.1 and 6.2 give expressions of the time evolution of the errors in terms of *one and two-dimensional integrals over spectral densities whose Stieltjes transforms are given by a closed set of purely algebraic equations*. The expressions lend themselves to numerical computation as illustrated in Fig. 6.1.1 and more extensively in Sect. 6.3 and the supplementary material.
- b.** Model and sample-wise double descents develop after some definite time at the interpolation threshold and are preceded by a *dip or minimum* before the spike develops. This indicates that early stopping is beneficial for some parameter regimes. A similar behavior also occurs for the triple descent. (See Figs. 6.3.2, 6.3.3 and the 3D version Fig. 6.1.1).
- c.** We observe two kinds of epoch-wise "descent" structures. The first is a *double plateau* monotonously descending structure at widely different time scales in the largely overparameterized regime (see Fig. 6.3.2). The second is an *epoch-wise double descent* similar to the one found in Nakkiran et al. (2020a). In fact, as in Nakkiran et al. (2020a), rather than a spike, this double descent appears to be an *elongated bump* over a wide time scale (see Fig. 6.3.4 and the 3D version Fig. 6.1.1).

Let us say a few words about the techniques used in this work. We first translate the gradient flow equations for the learned weights of the second layer into a set of integro-differential equations for generating functions, as in Bodin and Macris (2021b), involving the resolvent of a random matrix (constructed out of the fixed first layer weights, the data, and the non-linear activation). The solution of the integro-differential equations and the time evolution of the errors can then be expressed in terms of Cauchy complex integral representation which has the advantage to decouple the time dependence and static contributions involving traces of algebraic combinations of standard random matrices (see Liao and Couillet (2018) for related methods). This is the content of propositions 6.1 and 6.2. With a natural concentration hypothesis in the high-dimensional regime, it remains to carry out averages over the static traces involving random matrices. This is resolved using traces of sub-blocks from the inverse of a larger nontrivial block-matrix, a so-called *linear pencil*. To the best of our knowledge linear pencils have been introduced in the machine learning community only recently in Adlam and Pennington (2020a). This theory is developed in the context of random matrix theory in Rashidi Far et al. (2006); Helton et al. (2007) and Helton et al. (2018) using operator valued free-probability. The non-linearity of the activation function is addressed using the gaussian equivalence principle Pennington and Worah (2017); P  ch   (2019); Adlam and Pennington (2020a). Finally, our analysis is not entirely mathematically controlled mainly due to the

concentration hypothesis in Sect. 6.2.3 but comparison with simulations (see Fig. 6.3.1 and SM) confirms that the analytical results are exact.

In the conclusion we briefly discuss possible extensions of the present analysis and open problems among which is the comparison with a dynamical mean-field theory approach.

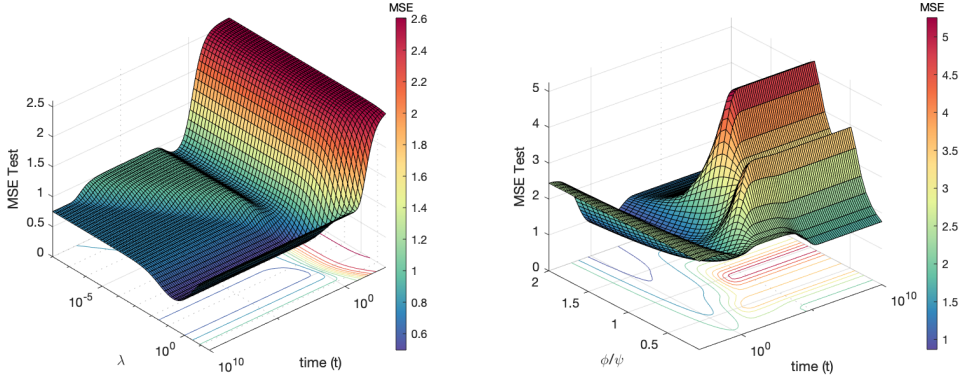


Figure 6.1.1: 3D plot of analytical test error evolution. See Figs. 6.3.4 and 6.3.3 (on the right) for parameter values.

6.2 Random feature model

6.2.1 Model description

Generative model and neural network: We consider the problem of learning a linear function $f_d(x) = d^{-\frac{1}{2}} \beta^*{}^T x$ with $x, \beta^* \in \mathbb{R}^d$ column vectors. The vector x is interpreted as a random input and β^* as a random *hidden* vector; both with distribution $\mathcal{N}(0, I_d)$, I_d the $d \times d$ identity matrix. We assume having access to the hidden function through the noisy outputs $y = f_d(x) + \epsilon$ with additive gaussian noise $\epsilon \sim \mathcal{N}(0, s^2)$, $s \in \mathbb{R}_+$. We suppose that we have n data-points $(x_i, y_i)_{1 \leq i \leq n}$. This data can be represented as the $n \times d$ matrix $X \in \mathbb{R}^{n \times d}$ where x_i^T is the i -th row of X , and the column vector $Y \in \mathbb{R}^n$ with i -th entry y_i . Therefore, we have the matrix notation $Y = d^{-\frac{1}{2}} X \beta^* + \xi$ where $\xi \sim \mathcal{N}(0, s^2 I_n)$ and I_n the $n \times n$ identity matrix.

We learn the data with a shallow 2-layer neural network. There are N hidden neurons with weight (column) vectors $\theta_i \in \mathbb{R}^d$, $i = 1, \dots, N$ each connected to the d input neurons. Out of these we form the matrix (of the first layer connecting input and hidden neurons) $\Theta \in \mathbb{R}^{N \times d}$ where θ_i^T is the i -th row of Θ . Its entries are assumed independent and sampled through a standard gaussian distribution $\mathcal{N}(0, 1)$; they are not learned but fixed once for all. The data-points in X are applied linearly to the parameters Θ , and the output $Z \in \mathbb{R}^{n \times N}$ of the first layer is the *pointwise application* of an activation function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, $Z = \sigma(d^{-\frac{1}{2}} X \Theta^T)$. We use the notation z_i^T to express the i -th row of Z . The second layer consists in a weight (column) vector $\beta_t \in \mathbb{R}^N$ to be learned, indexed by time $t \geq 0$, with components initially sampled at $t = 0$ i.i.d $\mathcal{N}(0, r^2)$, $r \in \mathbb{R}_+$. The prediction vector is expressed as $\hat{Y}_t = N^{-\frac{1}{2}} Z \beta_t$.

We assume that the activation function belongs to $L^2(e^{-\frac{x^2}{2}} dx)$ with inner product denoted $\langle \cdot, \cdot \rangle$. It can be expanded on the basis of Hermite polynomials, so $\sigma \in \text{Span}((H_{e_k})_{k \geq 0})$, where $H_{e_k}(x) = (-1)^k e^{\frac{x^2}{2}} \frac{d^k}{dx^k} e^{-\frac{x^2}{2}}$ (so $H_{e_0}(x) = 1$, $H_{e_1}(x) = x$, $H_{e_2}(x) = x^2 - 1$, $H_{e_3}(x) = x^3 - 3x$, ...). Furthermore we take σ centered with $\langle \sigma, H_{e_0} \rangle = 0$, and set $\mu = \langle \sigma, H_{e_1} \rangle$, $\nu^2 = \langle \sigma, \sigma \rangle - \mu^2$. For instance, $\sigma = \text{id}$ has $(\mu, \nu) = (1, 0)$ while $\sigma = \text{Relu} - \frac{1}{\sqrt{2\pi}}$ has $(\mu, \nu) = (\frac{1}{2}, \frac{1}{2}(1 - \frac{2}{\pi})^{1/2}) \simeq (0.5, 0.3)$. Finally, we recall that we are interested in the high dimensional regime where the parameters N, n, d tend to infinity with the ratios $\frac{N}{d} \rightarrow \psi$ and $\frac{n}{d} \rightarrow \phi$.

Training and test errors: For a new input $x_0 \in \mathbb{R}^d$, we define the predictor $\hat{y}_t(x_0) = \frac{1}{\sqrt{N}} z(x_0)^T \beta_t$ where $z(x_0) = \sigma(\frac{1}{\sqrt{d}} \Theta x_0)$. We will further define the standard training and test errors with a penalization term $\lambda > 0$ and the quadratic loss:

$$\mathcal{E}_{\text{train}}(\beta) = \frac{1}{n} \|Y - \hat{Y}\|^2 + \frac{\lambda}{N} \|\beta\|^2, \quad \mathcal{E}_{\text{gen}}(\beta) = \mathbb{E}_{x_0 \sim \mathcal{N}(0,1)} [(y(x_0) - \hat{y}(x_0))^2] \quad (6.1)$$

Note that because of the λ -penalization term, in this context, the training error can be above the test error in some configurations of parameters. Also, we will slightly abuse this notation throughout the chapter by using $\mathcal{E}_{\text{train}}(t), \mathcal{E}_{\text{gen}}(t)$ to designate $\mathcal{E}_{\text{train}}(\beta_t), \mathcal{E}_{\text{gen}}(\beta_t)$.

Gradient flow: Minimizing the training error of this shallow-network is equivalent to a standard Tikhonov regularization problem with a design matrix Z for which the optimal weights are given by $\beta_\infty = (\frac{Z^T Z}{N} + \frac{n}{N} \lambda I_N)^{-1} \frac{Z^T}{\sqrt{N}} Y$. The errors generated by the predictors with weights β_∞ have been analytically calculated in the high-dimensional regime in Mei and Montanari (2019) and further analyzed in d'Ascoli et al. (2020). Here we study the *whole time evolution* of the gradient flow and thus introduce an additional time dimension in our model. Of course as $t \rightarrow +\infty$ one recovers the errors generated by the predictors with weights β_∞ . The output vector β_t is updated through the ordinary differential equation $\frac{d\beta_t}{dt} = -\eta \nabla_{\beta} \mathcal{E}_{\text{train}}(\beta_t)$ with a fixed learning rate parameter $\eta > 0$. As η can be absorbed in the time parameter, from now on we consider without loss of generality that $\eta = \frac{n}{2}$. Setting $\delta = \lambda \frac{n}{N}$, we find that the gradient flow for β_t is a first order linear matrix differential equation,

$$\frac{d\beta_t}{dt} = - \left(\frac{Z^T Z}{N} + \delta I_N \right) \beta_t + \frac{Z^T Y}{\sqrt{N}}. \quad (6.2)$$

Recall the initial condition β_0 is a vector with i.i.d $\mathcal{N}(0, r^2)$ components.

6.2.2 Cauchy integral representations of the training and test errors

An important step of our analysis is the representation of $\mathcal{E}_{\text{train}}$ and \mathcal{E}_{gen} in terms of Cauchy contour integrals in the complex plane. To this end we decompose both errors in elementary contributions and derive contour integrals for each of them. Details are found in section 6.4 and the SM.

We begin with the test error which is more complicated. We have

$$\mathcal{E}_{\text{gen}}(t) = 1 + s^2 - 2\mu g(t) + \mu^2 h(t) + \nu^2 l(t) + o_d(1) \quad (6.3)$$

where $\lim_{d \rightarrow +\infty} o_d(1) = 0$ with high probability, and $g(t) = \frac{\beta^{*T} \Theta^T \beta_t}{\sqrt{d} \sqrt{d} \sqrt{N}}$, $h(t) = \left\| \frac{\Theta^T \beta_t}{\sqrt{d} \sqrt{N}} \right\|^2$, and $l(t) = \left\| \frac{\beta_t}{\sqrt{N}} \right\|^2$. To describe Cauchy's integral representation of the elementary functions g , h , l we introduce the resolvent $R(z) = \left(\frac{Z^T Z}{N} - z I_N \right)^{-1}$ for all $z \in \mathbb{C} \setminus \text{Sp} \left(\frac{Z^T Z}{N} \right)$.

Proposition 6.1 (Test error). *Let \mathcal{R}_z be the functional acting on holomorphic functions $f : \mathbb{C} \setminus \text{Sp} \left(\frac{Z^T Z}{N} \right) \rightarrow \mathbb{C}$ as $\mathcal{R}_z \{f(z)\} = -\oint_{\Gamma} \frac{dz}{2\pi i} f(z)$ over a contour Γ encircling the spectrum $\text{Sp} \left(\frac{Z^T Z}{N} \right)$ in the counterclockwise direction. Similarly, let $\mathcal{R}_{x,y}$ be the functional acting on two-variable holomorphic functions $f : (\mathbb{C} \setminus \text{Sp} \left(\frac{Z^T Z}{N} \right))^2 \rightarrow \mathbb{C}$ as $\mathcal{R}_{x,y} \{f(x,y)\} = \oint_{\Gamma} \oint_{\Gamma} \frac{dx}{2\pi i} \frac{dy}{2\pi i} f(x,y)$. Let $G_t(z) = \frac{\beta^{*T} \Theta^T}{\sqrt{d} \sqrt{d}} R(z) \frac{\beta_t}{\sqrt{N}}$ and $K(z) = \frac{\beta^{*T} \Theta^T}{\sqrt{d} \sqrt{d}} R(z) \frac{Z^T Y}{N}$. We have for all $t \geq 0$*

$$g(t) = \mathcal{R}_z \left\{ e^{-t(z+\delta)} G_0(z) + \frac{1 - e^{-t(z+\delta)}}{z + \delta} K(z) \right\}. \quad (6.4)$$

Let $L_t(z) = \frac{\beta_t^T}{\sqrt{N}} R(z) \frac{\beta_t}{\sqrt{N}}$ and $U_t(z) = \frac{Y^T Z}{N} R(z) \frac{\beta_t}{\sqrt{N}}$ and $V(z) = \frac{Y^T Z}{N} R(z) \frac{Z^T Y}{N}$. For all $t \geq 0$

$$l(t) = \mathcal{R}_z \left\{ e^{-2t(z+\delta)} L_0(z) + 2e^{-t(z+\delta)} \left(\frac{1 - e^{-t(z+\delta)}}{\delta + z} \right) U_0(z) + \left(\frac{1 - e^{-t(\delta+z)}}{\delta + z} \right)^2 V(z) \right\}. \quad (6.5)$$

Let $H_t(x,y) = \frac{\beta_t^T}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{\beta_t}{\sqrt{N}}$, $Q_t(x,y) = \frac{\beta_t^T}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T Y}{N}$ and $W(x,y) = \frac{Y^T Z}{N} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T Y}{N}$. For all $t \leq 0$

$$h(t) = \mathcal{R}_{x,y} \left\{ e^{-t(2\delta+x+y)} \left(\frac{e^{t(\delta+y)} - 1}{\delta + y} Q_0(x,y) + \frac{e^{t(\delta+x)} - 1}{\delta + x} Q_0(y,x) \right) \right\} \\ + \mathcal{R}_{x,y} \left\{ \frac{1 - e^{-t(x+\delta)}}{x + \delta} \frac{1 - e^{-t(y+\delta)}}{y + \delta} W(x,y) \right\} + \mathcal{R}_{x,y} \left\{ e^{-t(x+y+2\delta)} H_0(x,y) \right\}. \quad (6.6)$$

A similar but much simpler representation holds for the training error.

Proposition 6.2 (Training error). *With the same definitions than in proposition 6.1 we have*

$$\mathcal{E}_{\text{train}}(t) = \frac{\|Y\|^2}{n} + \frac{1}{c} \mathcal{R}_z \left\{ (z + \delta) e^{-2t(z+\delta)} L_0(z) - 2e^{-2t(z+\delta)} U_0(z) - \frac{1 - e^{-2t(\delta+z)}}{\delta + z} V(z) \right\}. \quad (6.7)$$

6.2.3 High-dimensional framework

The Cauchy integral representation involves a set of one-variable functions $\mathcal{S}_1 = \{G_0, K, L_0, U_0, V\} : \mathbb{C} \setminus \text{Sp} \left(\frac{Z^T Z}{N} \right) \rightarrow \mathbb{C}$ and a set of two-variable functions $\mathcal{S}_2 = \{H_0, W, Q_0\} : (\mathbb{C} \setminus \text{Sp} \left(\frac{Z^T Z}{N} \right))^2 \rightarrow \mathbb{C}$ so that g , h , l and thus also \mathcal{E}_{gen} and $\mathcal{E}_{\text{train}}$ are actually functions of $(t; \mathcal{S}_1, \mathcal{S}_2)$. Thus we can

write for instance: $\mathcal{E}_{\text{gen}}(\beta_t) = \mathcal{E}_{\text{gen}}(t; \mathcal{S}_1, \mathcal{S}_2)$. We simplify the problem by considering the high-dimensional regime where $N, n, d \rightarrow \infty$ with ratios $\frac{N}{d} \rightarrow \psi, \frac{n}{d} \rightarrow \phi$ tending to fixed values of order one. In this regime we expect that the functions in \mathcal{S}_1 and \mathcal{S}_2 concentrate and can therefore be replaced by their averages over randomness. These averages can be carried out using recent progress in random matrix theory Rashidi Far et al. (2006), Helton et al. (2007), and we are able to compute pointwise asymptotic values of the functions in $\mathcal{S}_1, \mathcal{S}_2$, and eventually substitute them in the Cauchy integral representations for the training and test error. In general, rigorously showing concentration of the various functions involved is not easy and we will make the following assumptions:

Assumptions 6.1. *In the high dimensional limit with $d, N, n \rightarrow +\infty$ and $\frac{N}{d} \rightarrow \psi, \frac{n}{d} \rightarrow \phi$:*

1. *The random functions in $\mathcal{S}_1, \mathcal{S}_2$ are assumed to concentrate. We let $\bar{\mathcal{S}}_1 = \{\bar{G}_0, \bar{K}, \bar{L}_0, \bar{U}_0, \bar{V}\}$ and $\bar{\mathcal{S}}_2 = \{\bar{H}_0, \bar{W}, \bar{Q}_0\}$ be the pointwise limit of the functions.*
2. *There exists a bounded subset $\mathcal{C} \subset \mathbb{R}^+$ such that the functions in $\bar{\mathcal{S}}_1$ and $\bar{\mathcal{S}}_2$ are holomorphic on $\mathbb{C} \setminus \mathcal{C}$ and $(\mathbb{C} \setminus \mathcal{C})^2$ respectively*
3. *The gaussian equivalence principle (see sect. 6.4.2) can be applied to the limiting quantities.*

It is common that the closure of the spectrum of suitably normalized random matrices concentrates on a deterministic set. Thus the bounded set \mathcal{C} can be understood as the limit of the finite interval $[0, \lim_d \max \text{Sp}(\frac{Z^T Z}{N})]$. In the sequel we will distinguish the theoretical high-dimensional regime from the finite dimensional regime using the upper-bar notation.

Definition 6.1 (High-dimensional framework). *Under the assumptions 6.1, we define the theoretical test error $\bar{\mathcal{E}}_{\text{gen}}(t) = \mathcal{E}_{\text{gen}}(t; \bar{\mathcal{S}}_1, \bar{\mathcal{S}}_2)$ and the theoretical training error $\bar{\mathcal{E}}_{\text{train}}(t) = \mathcal{E}_{\text{train}}(t; \bar{\mathcal{S}}_1, \bar{\mathcal{S}}_2)$*

We conjecture that $\lim_d \mathcal{E}_{\text{train}}(t) = \bar{\mathcal{E}}_{\text{train}}(t)$ and $\lim_d \mathcal{E}_{\text{gen}}(t) = \bar{\mathcal{E}}_{\text{gen}}(t)$ at all times $t \in \mathbb{R}$. We verify that this conjecture stands experimentally for sufficiently large d on different configurations (see additional figures in the SM). This also lends experimental support on the assumption 6.1. Furthermore we conjecture that the $d \rightarrow +\infty$ and $t \rightarrow \infty$ limits commute, namely $\lim_d \lim_t \mathcal{E}_{\text{train}}(t) = \lim_t \bar{\mathcal{E}}_{\text{train}}(t)$ and $\lim_d \lim_t \mathcal{E}_{\text{gen}}(t) = \lim_t \bar{\mathcal{E}}_{\text{gen}}(t)$.

6.3 Results and insights

6.3.1 Main results

In this section we provide the main results of this work: analytical formulas tracking the test and training errors during gradient flow of the random feature model for all times in the high-dimensional theoretical framework.

Chapter 6. The Random feature model

Result 6.1. Under the assumption 6.1, the theoretical test and training errors of definition 6.1 are given for all times $t \geq 0$ by the formulas

$$\bar{\mathcal{E}}_{gen}(t) = 1 + s^2 - 2\mu\bar{g}(t) + \mu^2\bar{h}(t) + v^2\bar{l}(t), \quad (6.8)$$

$$\bar{\mathcal{E}}_{train}(t) = 1 + s^2 + \frac{1}{c} \int_{\mathbb{R}} \left[(\delta + \omega) e^{-2t(\omega+\delta)} \rho_{\bar{L}_0}(\omega) - \frac{1 - e^{-2t(\delta+\omega)}}{\delta + \omega} \rho_{\bar{V}}(\omega) \right] d\omega, \quad (6.9)$$

with $c = \frac{\phi}{\psi}$, $\delta = c\lambda$, and the functions \bar{g} , \bar{h} , \bar{l} given by

$$\bar{g}(t) = \int_{\mathbb{R}} \frac{1 - e^{-t(\omega+\delta)}}{\omega + \delta} \rho_{\bar{K}}(\omega) d\omega, \quad (6.10)$$

$$\bar{l}(t) = \int_{\mathbb{R}} \left[e^{-2t(\omega+\delta)} \rho_{\bar{L}_0}(\omega) + \left(\frac{1 - e^{-t(\omega+\delta)}}{\omega + \delta} \right)^2 \rho_{\bar{V}}(\omega) \right] d\omega, \quad (6.11)$$

$$\bar{h}(t) = \iint_{\mathbb{R}^2} \left[e^{-t(u+v+2\delta)} \rho_{\bar{H}_0}(u, v) + \frac{1 - e^{-t(u+\delta)}}{u + \delta} \frac{1 - e^{-t(v+\delta)}}{v + \delta} \rho_{\bar{W}}(u, v) \right] dudv, \quad (6.12)$$

where the measures $\rho_{\bar{K}}, \rho_{\bar{L}_0}, \rho_{\bar{V}}, \rho_{\bar{H}_0}, \rho_{\bar{W}}$ (are possibly signed) are characterized by their Stieltjes transforms given by $\bar{K}, \bar{L}_0, \bar{V}, \bar{H}_0, \bar{W}$

$$\begin{cases} \bar{K}(x) = t_1^x, & \bar{L}_0(x) = r^2 g_1^x, & \bar{V}(x) = s^2 (1 + x g_1^x) + (c - h_4^x), \\ \bar{H}_0(x, y) = r^2 q_1, & \bar{W}(x, y) = s^2 c q_4 + q_2 \end{cases} \quad (6.13)$$

where for each $x, y \in \mathbb{C}^+$ (the upper half complex plane) $g_1^x, h_4^x, t_1^x, g_1^y, h_4^y, t_1^y$ and q_1, q_2, q_4, q_5 (which depend symmetrically on (x, y) , e.g., $q_1 = q_1^{x,y} = q_1^{y,x}$) are solutions of a purely algebraic system of equations (see SM for the criterion to select the relevant solution)

$$\begin{cases} 0 = \mu\psi g_1^x h_4^x - t_1^x \\ 0 = \mu\psi g_1^y h_4^y - t_1^y \\ 0 = (c - 1 - x g_1^x) (c - \mu^2 \phi g_1^x h_4^x) - c h_4^x \\ 0 = (c - 1 - y g_1^y) (c - \mu^2 \phi g_1^y h_4^y) - c h_4^y \\ 0 = 1 - g_1^x (\mu^2 h_4^x + (c - 1 - x g_1^x) v^2 - x) \\ 0 = 1 - g_1^y (\mu^2 h_4^y + (c - 1 - y g_1^y) v^2 - y) \\ 0 = -\mu^2 g_1^y q_2 + \mu^2 h_4^x q_1 + \mu g_1^y t_1^x + \mu g_1^x t_1^y - c g_1^y q_4 v^2 - g_1^y - q_1 x + q_1 v^2 (c - g_1^x x - 1) \\ 0 = \mu (\phi - \psi g_1^x x - \psi) (-\mu g_1^x q_2 + \mu h_4^y q_1 + g_1^x t_1^y) + c q_4 (1 - \mu t_1^y) - q_2 \\ 0 = -\mu^2 \phi g_1^x (1 - \mu t_1^x) q_4 + \mu^2 q_5 (c - g_1^y y - 1) - v^2 \phi g_1^x q_4 - \phi q_4 + q_1 v^2 (\phi - \psi g_1^y y - \psi) \\ 0 = \psi (\mu^2 \phi g_1^x g_1^y q_4 + \psi g_1^x g_1^y + q_1) (1 - \mu t_1^y) - \mu^2 \psi g_1^x q_5 (c - g_1^x x - 1) - q_5 \end{cases}$$

We can also deduce the limiting training error and test errors in the infinite time limit:

Result 6.2. *In the limit $t \rightarrow \infty$ we find:*

$$\lim_{t \rightarrow \infty} \bar{\mathcal{E}}_{gen}(t) = 1 + s^2 - 2\mu\bar{K}(-\delta) + \mu^2\bar{W}(-\delta, -\delta) + v^2 \frac{d\bar{V}}{dx}(-\delta), \quad \lim_{t \rightarrow \infty} \bar{\mathcal{E}}_{train}(t) = 1 + s^2 - \frac{1}{c}\bar{V}(-\delta)$$

Interestingly, in the limit $t \rightarrow \infty$, the expressions become simpler and completely algebraic in the sense that we do not need to compute integrals (or double-integrals) over the supports of the eigenvalue distributions. It is not obvious to see on the analytical expressions that the result is the same as the algebraic expressions obtained in Mei and Montanari (2019) but Fig. 6.3.1 shows an excellent match with simulation experiments. We note here that checking that two sets of complicated algebraic equations are equivalent is in general a non-trivial problem of computational algebraic geometry Cox et al. (2007).

6.3.2 Insights and illustrations of results

The set of analytical formulas allows to compute numerically the measures $\rho_K, \rho_{L_0}, \rho_V, \rho_{H_0}, \rho_W$ and in turn the full time evolution of the test and training errors. The result matches the simulation of a large random feature model where d is taken large as can be seen on Figs. 6.3.1 for the infinite time limit (experimental check of result 6.2) and additional figures in the SM (experimental check of result 6.1). Below we illustrate numerical computations obtained with analytical formulas of result 6.1 for various sets of parameters $(t, \mu, v, \psi, \phi, r, s, \lambda)$. For instance, we can freely choose two of these parameters and plot the generalization error in 3D as in Fig. 6.1.1, or as a heat-map in the following. We describe three important phenomena which are observed with our analysis.

Double descent and early-stopping benefits: while Mei and Montanari (2019) mostly analyze the minimum least-squares estimator of the random feature model which displays the double-descent at $\psi = \phi$, we are predicting the whole time evolution of the gradient flow as in Fig. 6.3.2. We clearly observe the double-descent curve at $t = 10^{10}$ for $\psi = \phi$; but we now notice that if we stop the training earlier, say at times $1 < t < 10$, the generalization error performs better than the minimum least squares estimator. Actually, in the time interval $t \in (1, 10)$ for $\psi \approx \phi$ the test error even has a *dip or minimum* just before the spike develops. We also notice a two-steps descent structure with the test error which is non-existent in the training error and materializes long after the training error has stabilized in the overparameterized regime $\psi \gg \phi$. This is also reminiscent but not entirely similar to the abrupt *grokking* phenomenon described in Power et al. (2021).

Triple descent: We can observe a triple descent phenomenon materialized by two spikes as seen in Fig. 6.3.1 at $t = \infty$ (we also check that the theoretical result matches very well the empirical prediction of the minimum least squares estimator both for training and test errors). This triple descent phenomenon is already contained in the formulas of Mei and Montanari (2019) (although not discussed in this reference) and has been analyzed in detail in d’Ascoli et al. (2020). The test error contains a so-called *linear spike* for $\phi = 1$ ($n = d$) and a *non-linear*

Chapter 6. The Random feature model

spike for $\psi = \phi$ ($N = n$). The two spikes are often not seen together as this requires certain conditions to be met, and they tend to materialize together for specific values μ, ν of the activation function where $\mu \gg \nu$. Here we further observe the evolution through time of the triple descent and the two spikes and how they develop in Fig. 6.3.3. There, we notice that the linear-spike seems to appear *earlier* than the non-linear one.

Epoch-wise descent structures: Important phenomena that we uncover here are two time-wise "descent structures". (i) As can be seen in Fig. 6.3.2, the test error develops a *double plateau* structure at widely different time scales in the over-parameterized regime ($\psi \gg \phi$) while there seems to be only one time scale for the training error. This kind of double plateau descent is different from the "usual" double-descent. (ii) Moreover, on Fig. 6.3.4 for well chosen parameters (in particular for noises with s and r "larger" and $\psi = 2\phi$), we can also observe an *elongated bump* (rather than a thin spike) for small λ 's. Notice the logarithmic time-scale which clearly shows that here we need to wait exponentially longer to attain the "second descent" after the bump. This is very reminiscent of the epoch-wise double descent described in Nakkiran et al. (2020a) for deep networks (which happens on similar time scales).

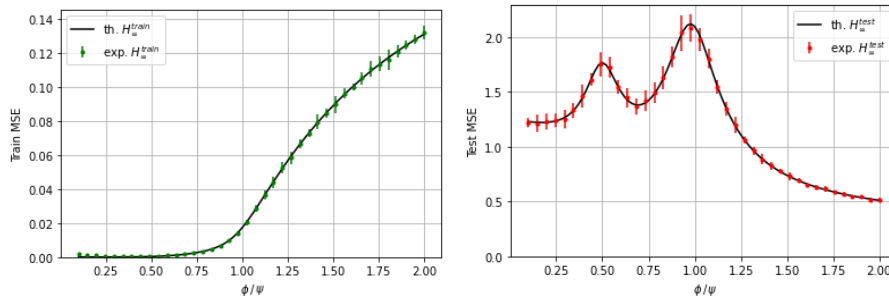


Figure 6.3.1: *Large time limit.* Analytical training error and test error profile with parameters $(\mu, \nu, \psi, r, s, \lambda) = (10, 1, 2, 1, 0.5, 0.01)$ compared with experimental least squares MSE with 40 data-points with $d = 5000$ (average of 10 instances with confidence bar at 2σ)

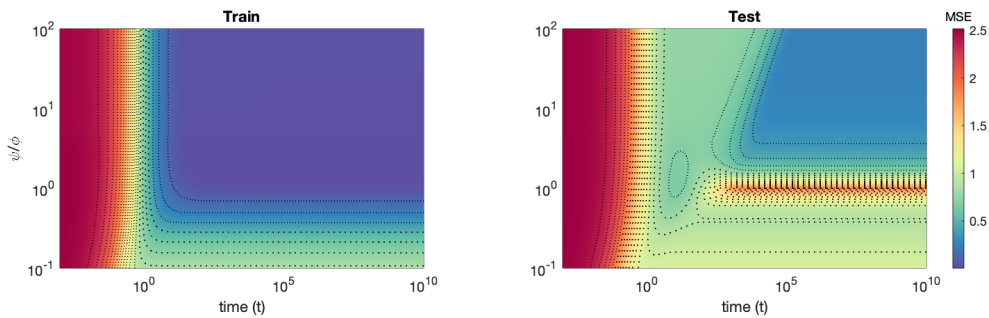


Figure 6.3.2: *Model-wise double descent.* Analytical training error and test error evolution with parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 3, 2., 0.4, 0.001)$. Note that we vary the number of model parameters (ψ).

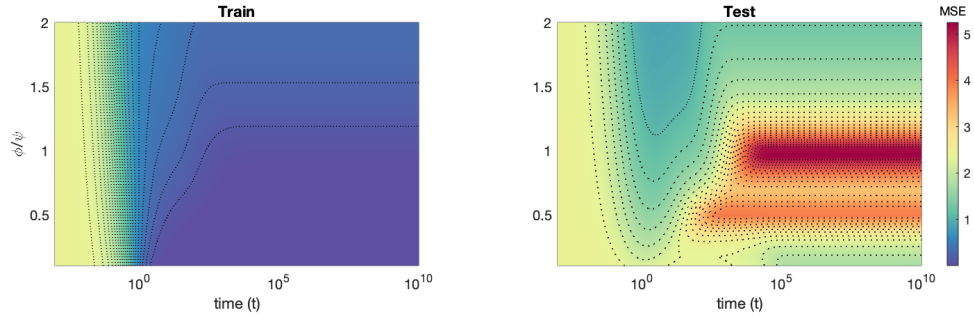


Figure 6.3.3: *Sample-wise descents*. Analytical training error and test error evolution with parameters $(\mu, \nu, \psi, r, s, \lambda) = (0.9, 0.1, 2, 1, 0.8, 0.0001)$. Note that we vary the number of samples (ϕ).

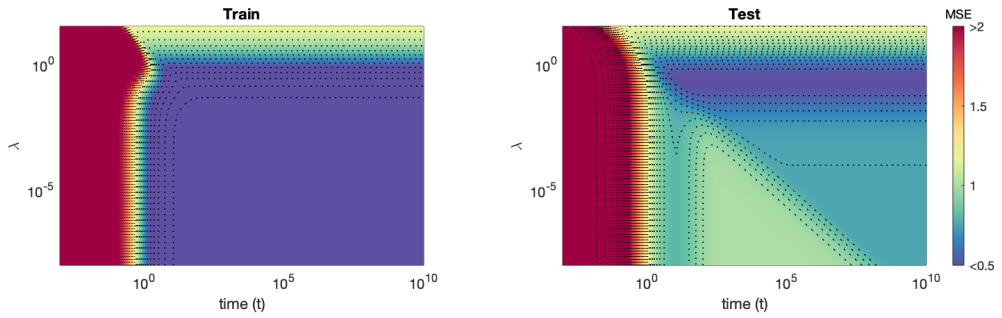


Figure 6.3.4: *Epoch-wise descent structures*. Analytical test error evolution with respect to different values of λ ($\mu, \nu, \psi, \phi, r, s) = (0.5, 0.3, 6, 3, 2.0, 0.5)$. Here the ratio of number of parameters and samples is fixed.

6.4 Sketch of proofs and analytical derivations

The analysis is threefold. Firstly, we decompose the training and test errors in elementary terms and establish Cauchy's integral representation for each of them, as provided in proposition 6.1. A crucial advantage of this form is that it dissociates a scalar time-wise component and static matrix terms. Secondly, we switch to the high-dimensional framework where the matrix terms are substituted by their limit using the gaussian equivalence principle. Thirdly, we can compute the expectations of matrix terms thanks to a random matrix technique based on linear pencils. In this section we only sketch the main ideas for each step and provide details in the supplementary material.

6.4.1 Cauchy's integral representation

We sketch the derivation for the test error and leave details to appendices. The derivation for the training error is entirely found in the SM. Expanding the square in Equ. (6.1) and carrying out averages we find Equ. (6.3) for $\mathcal{E}_{\text{gen}}(t)$ with $g(t) = \frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \frac{\beta_t}{\sqrt{N}}$, $h(t) = \left\| \frac{\Theta^T}{\sqrt{d}} \frac{\beta_t}{\sqrt{N}} \right\|^2$, and $l(t) = \left\| \frac{\beta_t}{\sqrt{N}} \right\|^2$ (see SM for this derivation).

We show how to derive the Cauchy integral representation for $g(t)$. For $h(t)$, $l(t)$ the steps are similar and found in SM. Let us consider the function $(t, z) \mapsto G_t(z)$ as in 6.1. Then we have the relation $g(t) = \frac{-1}{2i\pi} \oint_{\Gamma} dz G_t(z)$ where Γ is a loop in \mathbb{C} enclosing the spectrum of $\frac{Z^T Z}{N}$. This can easily be seen by decomposing the symmetric $\frac{Z^T Z}{N}$ in an orthonormal basis v_1, \dots, v_N with the eigenvalues $\lambda_1, \dots, \lambda_N$: then we have $G_t(z) = \sum_{i=1}^N \frac{1}{\lambda_i - z} \left(\frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} v_i v_i^T \frac{\beta_t}{\sqrt{N}} \right)$ and because λ_i are all encircled by Γ , we find $-\oint_{\Gamma} \frac{dz}{2\pi i} G_t(z) = \sum_{i=1}^N \frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} v_i v_i^T \frac{\beta_t}{\sqrt{N}} = g(t)$. Now, the ODE derived for β_t in (6.2), can be written slightly differently using the fact that $\frac{Z^T Z}{N} = R(z)^{-1} + zI$ for any z outside $\text{Sp}(\frac{Z^T Z}{N})$. Namely, $\frac{d\beta_t}{dt} = \frac{Z^T Y}{\sqrt{N}} - R(z)^{-1} \beta_t - (z + \delta) \beta_t$. Then, we can derive an integro-differential equation for $G_t(z)$ involving $g(t)$ and $K(z)$:

$$\frac{\partial_t G_t(z)}{\partial t} = K(z) - g(t) - (z + \delta) G_t(z) \quad (6.14)$$

In the following, we let \mathcal{L} be the Laplace transform operator $(\mathcal{L}f)(p) = \int_0^{+\infty} dt e^{-pt} f(t)$, $\text{Re } p$ large enough. Note that the contour integral is performed over a compact set Γ so for $\text{Re } p$ large enough, by Fubini's theorem, the operations \mathcal{L} and \mathcal{R}_z commute. Applying \mathcal{L} to (6.14) and rearranging terms we find for $\text{Re}(p + z + \delta) \neq 0$:

$$\mathcal{L}G_p(z) = \frac{G_0(z)}{p + z + \delta} + \frac{K(z)}{p(p + z + \delta)} - \frac{\mathcal{L}g(p)}{p + z + \delta} \quad (6.15)$$

Now, we can always choose Γ such that $-(p + \delta)$ is outside of the contour if we assume $\text{Re}(p + \delta) > 0$ (since $\min_i \lambda_i \geq 0$). Thus, applying \mathcal{R}_z to (6.15) nullifies the last term because the pole is outside Γ , and using commutativity $\mathcal{R}_z \mathcal{L}G_p = \mathcal{L} \mathcal{R}_z G_p$,

$$\mathcal{R}_z \mathcal{L}G_p = \mathcal{L} \mathcal{R}_z \left\{ e^{-t(z+\delta)} G_0(z) + \frac{1 - e^{-t(z+\delta)}}{z + \delta} K(z) \right\} = \mathcal{L}g(p). \quad (6.16)$$

Finally, using the inverse Laplace transform leads to (6.4).

6.4.2 Gaussian equivalence principle

The matrix terms must be estimated in the limit $d \rightarrow \infty$ with $\{\beta^*, \beta_0, \xi, \Theta, X\}$ all independently distributed. As per assumptions 6.1 all the matrix terms in $\mathcal{S}_1, \mathcal{S}_2$ are assumed to concentrate. So for instance we assume that the following limit exists $\bar{K}(z) \equiv \lim_{d \rightarrow \infty} K(z) = \lim_{d \rightarrow \infty} \mathbb{E}_{\beta^*, \xi, \Theta, X} [K(z)]$. Using cyclicity of the trace we easily perform averages over β^*, ξ to find

$$\bar{K}(z) = \lim_d \mathbb{E}_{\beta^*, \Theta, X} \text{Tr} \left[\frac{\Theta^T}{\sqrt{d}} R(z) \frac{Z^T X \beta^* \beta^{*T}}{N} \frac{1}{d} \right] = \lim_d \frac{1}{d} \mathbb{E}_{\beta^*, \Theta, X} \text{Tr} \left[\frac{\Theta^T}{\sqrt{d}} R(z) \frac{Z^T X}{N} \right]. \quad (6.17)$$

After these reductions, the expressions of all functions in $\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2$ essentially involve products of random matrices Θ, X and pointwise applications of the non-linear activation σ . This can be further reduced to simpler algebraic expressions using the *gaussian equivalence prin-*

ciple. This principle states that: *there exists a standard gaussian random matrix $\Omega \in \mathbb{R}^{n \times N}$ independent of $\{X, \Theta\}$ such that in the infinite dimensional limit we can make the substitution $Z = \sigma(d^{-1/2} X \Theta^T) \rightarrow \mu d^{-1/2} X \Theta^T + v \Omega$ in the expressions of all functions in $\tilde{\mathcal{F}}_1, \tilde{\mathcal{F}}_2$.* This approach is quite general and is well described in Adlam and Pennington (2020a); Adlam et al. (2019) (and formerly in Pennington and Worah (2017) and P ech e (2019)). Thus it remains to compute expectations of traces containing only products, and inverses of products and sums, of gaussian matrices.

6.4.3 Expectations over random matrices using linear pencils

We explain how to compute the limit of (6.17) once the gaussian equivalence principle has been applied. A powerful approach is to design a so-called *linear pencil*. In the present context this is a suitable block-matrix containing gaussian random matrices and multiples of the identity matrix, for which full block-inversion gives back the products of terms in the traces that are being sought. This approach has been described in Rashidi Far et al. (2006); Helton et al. (2007); Mingo and Speicher (2017). We have found a suitable linear pencil which contains fortuitously *all* the terms required in $\tilde{\mathcal{F}}_1, \tilde{\mathcal{F}}_2$. It is described by the 13×13 *block-matrix* M , and pursuing with our example, we get for instance with the block (7, 12) that $\lim_d K(y) = \lim_d \frac{1}{d} \text{Tr}[(M^{-1})^{(7,12)}]$

$$M = \begin{pmatrix} -xI & -\mu \frac{\Theta}{\sqrt{d}} & -I & 0 & 0 & 0 & \frac{\Theta}{\sqrt{d}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & \frac{X^T}{\sqrt{N}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & v \frac{\Omega^T}{\sqrt{N}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & \frac{X}{\sqrt{N}} & v \frac{\Omega}{\sqrt{N}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu \frac{\Theta^T}{\sqrt{d}} & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & \frac{\Theta^T}{\sqrt{d}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & v \frac{\Omega^T}{\sqrt{N}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & \frac{X^T}{\sqrt{N}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & v \frac{\Omega}{\sqrt{N}} & \frac{X}{\sqrt{N}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & -I \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & -\mu \frac{\Theta^T}{\sqrt{d}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & \mu \frac{\Theta}{\sqrt{d}} & 0 & 0 & 0 & -yI \end{pmatrix}$$

Next, the great advantage of the linear pencil is that (as described in Rashidi Far et al. (2006); Helton et al. (2007); Mingo and Speicher (2017)) it allows to write a fixed point equation $F(G) = G$ for a "small" 13×13 matrix G with *scalar* matrix elements. We also provide in the SM an independent derivation of the fixed point equations using the replica method (a technique from statistical physics Edwards and Jones (1976)). The components of G are linked to the limiting traces of the blocks of M^{-1} as in $[G]_{2,7} = \bar{K}(z)$. The action of F can be completely described as an algebraic function leading to (a priori) $13 \times 13 = 169$ equations over the matrix elements of G . The number of equations can be immediately reduced to 39 because many elements vanish, and with the help of a computer algebra system the number of equations

can be further brought down to 10. We refer to the SM for all the details about the method.

6.5 Conclusion

We believe that our analysis could be extended to study the learning of non-linear functions, the effect of multilayered structures, and potentially different layers such as convolutions, as long as they are not learned. A challenging task is to extend the present methods to learned multilayers. A further question is the application of our analysis in teacher-student scenarios with realistic datasets (See Loureiro et al. (2021); Adlam et al. (2019)).

Finally we wish to point out that a comparison of the approach of the present chapter (and the similar but simpler one of Bodin and Macris (2021b)) with the dynamical mean-field theory (DMFT) approach of statistical physics remains to be investigated. DMFT has a long history originating in studies of complex systems (turbulent fluids, spin glasses) where one eventually derives a set of complicated integro-differential equations for suitable correlation and response functions capturing the whole dynamics of the system (we refer to the recent book Parisi et al. (2020) and references therein). This is a powerful formalism but the integral equations must usually be solved entirely numerically which itself is not a trivial task. For problems close to the present context (neural networks, generalized linear models, phase retrieval) DMFT has been developed in the recent works Sompolinsky et al. (1988); Crisanti and Sompolinsky (2018); Agoritsas et al. (2018); Mignacco et al. (2020, 2021). We think that comprehensively comparing this formalism with the present approach is an interesting open problem. It would be desirable to connect the DMFT equations to our closed form solutions for the training and generalization errors expressed in terms of a set of algebraic equations of suitable Stieltjes transforms.

Appendix

6.A Test Error substitutions

The test error $\mathcal{E}_{\text{gen}}(t)$ in (6.1) can be expanded into smaller terms

$$\begin{aligned}\mathcal{E}_{\text{gen}}(t) &= \mathbb{E}_{x_0}[y(x_0)^2] - 2\mathbb{E}_{x_0}[y(x_0)\hat{y}_t(x_0)] + \mathbb{E}_{x_0}[\hat{y}_t(x_0)^2] \\ &= \mathbb{E}_{x_0}[y(x_0)^2] - 2\frac{\beta^{*T}}{\sqrt{d}}\mathbb{E}_{x_0}[x_0 z(x_0)^T] \frac{\beta_t}{\sqrt{N}} + \frac{\beta_t^T}{\sqrt{N}}\mathbb{E}_{x_0}[z(x_0)z(x_0)^T] \frac{\beta_t}{\sqrt{N}}.\end{aligned}\quad (6.18)$$

The random noise ϵ from $y(x_0)$ only impacts the first term on the right hand side with $\mathbb{E}_{x_0}[y(x_0)^2] = 1 + s^2$. Using further $q(t) = \frac{\beta^{*T}}{\sqrt{d}}\mathbb{E}_{x_0}[x_0 z(x_0)^T] \frac{\beta_t}{\sqrt{N}}$ and $p(t) = \frac{\beta_t^T}{\sqrt{N}}\mathbb{E}_{x_0}[z(x_0)z(x_0)^T] \frac{\beta_t}{\sqrt{N}}$, we write $\mathcal{E}_{\text{gen}}(t) = 1 + s^2 - 2q(t) + p(t)$.

We provide analytical arguments to justify the formula (6.3) showing that:

$$q(t) = \mu g(t) + o_d(1) \quad (6.19)$$

$$p(t) = \mu^2 h(t) + v^2 l(t) + o_d(1) \quad (6.20)$$

with

$$g(t) = \frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \frac{\beta_t}{\sqrt{N}}, \quad l(t) = \left\| \frac{\beta_t}{\sqrt{N}} \right\|^2, \quad h(t) = \left\| \frac{\Theta^T}{\sqrt{d}} \frac{\beta_t}{\sqrt{N}} \right\|^2 \quad (6.21)$$

and where $\lim_{d \rightarrow +\infty} o_d(1) = 0$ with probability tending to one when $d \rightarrow +\infty$. The arguments below are based further on the prior assumption that the (θ_i/\sqrt{d}) are sampled uniformly on the hyper-sphere of radius 1. We will assume further that these results can be extended in our setting with θ_i sampled from a gaussian distribution. Notice that this is a reasonable assumption because $\|\theta_i\|^2/d$ is a χ^2 distribution of mean 1 and variance $\frac{2}{d}$.

6.A.1 limit of $q(t)$

We decompose our activation function as $\sigma(x) = \mu x + v\sigma^\perp(x)$ where $\sigma^\perp \in \text{Span}(H_{e_i})_{i \geq 2}$. In other words, we have $\mathbb{E}_G[\sigma^\perp(G)] = \mathbb{E}_G[\sigma^\perp(G)G] = 0$ and $\mathbb{E}_G[\sigma^\perp(G)^2] = 1$. Notice that conditional on $(\theta_i)_i$ sampled on the sphere of radius \sqrt{d} , we have for all $i \in \{1, \dots, N\}$ that $u_i \equiv \frac{\theta_i^T x_0}{\sqrt{d}} \sim x_0$

Chapter 6. The Random feature model

$\mathcal{N}(0, 1)$, and for all $j \in \{1, \dots, N\}$, we have $\text{Cov}(u_i, u_j) = \frac{\theta_i^T \theta_j}{d} = \left[\frac{\Theta \Theta^T}{d} \right]_{i,j}$. Similarly, for any $l \in \{1, \dots, d\}$ we have $\text{Cov}(u_j, [x_0]_l) = \frac{[\theta_j]_l}{\sqrt{d}}$. Now, using the Mehler-Kernel formula, we have

$$\mathbb{E}_{x_0} [[x_0]_l [z(x_0)]_j] = \sum_{k \geq 0} \frac{1}{k!} (\text{Cov}(u_j, [x_0]_l))^k \mathbb{E}_{x_0} [x_0 H_{e_k}(x_0)] \mathbb{E}_{u_j} [\sigma(u_j) H_{e_k}(u_j)] \quad (6.22)$$

which does not vanish only for $k = 1$ due to the first expectation on the RHS. Thus

$$\mathbb{E}_{x_0} [[x_0]_l [z(x_0)]_j] = \frac{[\theta_j]_l}{\sqrt{d}} \mu \quad (6.23)$$

and hence we find that $q(t) = \frac{\beta^{*T}}{\sqrt{d}} \mathbb{E}_{x_0} [x_0 z(x_0)^T] \frac{\beta_t}{\sqrt{N}} = \mu \frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \frac{\beta_t}{\sqrt{N}}$.

The result ought not be exact anymore when (θ_i) are sampled from a normal distribution, and we make the assumption that we can account for a correction term $o_d(1)$ which goes to 0 as d grows to infinity, hence $q(t) = \mu g(t) + o_d(1)$ in general.

6.A.2 limit of $p(t)$

Similarly for $p(t)$, we evaluate the kernel $U_{i,j} = \mathbb{E}_{x_0} [[z(x_0)]_i [z(x_0)]_j]$ for which the Mehler-Kernel formula provides

$$\begin{aligned} U_{i,j} &= \sum_{k \geq 0} \frac{1}{k!} (\text{Cov}(u_i, u_j))^k \mathbb{E}_{u_i} [\sigma(u_i) H_{e_k}(u_i)]^2 \\ &= \mu^2 \text{Cov}(u_i, u_j) + v^2 \sum_{k \geq 2} \frac{(\text{Cov}(u_i, u_j))^k}{k!} \mathbb{E}_{u_i} [\sigma^\perp(u_i) H_{e_k}(u_i)]^2. \end{aligned} \quad (6.24)$$

Intuitively, the terms $(\text{Cov}(u_i, u_j))^k$ for $k \geq 2$ are on a smaller order in d compared to $\text{Cov}(u_i, u_j)$ when $i \neq j$. We refer the reader to Lemma C.7 in Mei and Montanari (2019) where it is shown with some additional assumptions on σ (weakly differentiable with $\exists c_0, c_1, \forall x > 0, |\sigma(x)|, |\sigma'(x)| \leq c_0 e^{c_1 x}$) that:

$$\mathbb{E}_\Theta \left[\left\| U - \mu^2 \frac{\Theta \Theta^T}{d} - v^2 I_N \right\|_{\text{op}} \right] = o_d(1). \quad (6.25)$$

Therefore, we can bound:

$$\begin{aligned} |p(t) - \mu^2 h(t) - v^2 l(t)| &= \left| \left\langle \frac{\beta_t^T}{\sqrt{N}}, \left(U - \mu^2 \frac{\Theta \Theta^T}{d} - v^2 I_N \right) \frac{\beta_t^T}{\sqrt{N}} \right\rangle \right| \\ &\leq \left\| \frac{\beta_t}{\sqrt{N}} \right\| \cdot \left\| U - \mu^2 \frac{\Theta \Theta^T}{d} - v^2 I_N \right\|_{\text{op}} \cdot \left\| \frac{\beta_t}{\sqrt{N}} \right\| \\ &= l(t) \left\| U - \mu^2 \frac{\Theta \Theta^T}{d} - v^2 I_N \right\|_{\text{op}}. \end{aligned} \quad (6.26)$$

As per the general assumptions 6.1, $l(t)$ concentrates to a finite quantity $\bar{l}(t)$ at all times as d grows to infinity (that $\bar{l}(t)$ is finite is explicitly checked by the analytical computations of the generalization error). Thus by Markov's inequality we have at any fixed time t , $|p(t) - \mu^2 h(t) -$

$v^2 l(t) = o_d(1)$ with probability tending to one as $d \rightarrow +\infty$.

Notice also that we assume as before that $o_d(1)$ also contains the correction added when (θ_i) are sampled from a normal distribution.

6.B Cauchy's integral representation formula

In this section we complete the proof of propositions 6.1 and 6.2. We show how to derive the Cauchy integral representation of the two functions $l(t)$ and $h(t)$ by similar analysis of Sect. 6.4.1 for the representation of $g(t)$.

6.B.1 Representation formula for $l(t)$

We define the function $L_t(z) = \frac{\beta_t^T}{\sqrt{N}} R(z) \frac{\beta_t}{\sqrt{N}}$ and the auxiliary functions $U_t(z) = \frac{Y^T Z}{N} R(z) \frac{\beta_t}{\sqrt{N}}$ and $V(z) = \frac{Y^T Z}{N} R(z) \frac{Z^T Y}{N}$. We find a set of 2 integro-differential equations using the gradient flow equation for $\frac{d\beta_t}{dt}$ (as in the derivation of 6.14)

$$\begin{aligned} \frac{1}{2} \frac{\partial L_t(z)}{\partial t} &= U_t(z) - l(t) - (z + \delta) L_t(z) \\ \frac{\partial_t U_t(z)}{\partial t} &= V(z) - \mathcal{R}_z U_t - (z + \delta) U_t(z) \end{aligned} \quad (6.27)$$

Similarly $G_t(z)$ and $g(t)$, we also have that $l(t) = -\oint_{\Gamma} \frac{dz}{2i\pi} L_t(z) = \mathcal{R}_z L_t$. So we get a pair of integro-differential equations in this case (whereas for $G_t(z)$ we had only one such equation). However, we have one additional differential equation in this case. Pursuing with the Laplace transform operator¹ the equations (6.27) become

$$\begin{aligned} \mathcal{L}L_p(z) &= \frac{1}{\frac{1}{2}p + z + \delta} \left(\frac{1}{2}L_0(z) + \mathcal{L}U_p(z) - \mathcal{L}l(p) \right) \\ \mathcal{L}U_p(z) &= \frac{1}{p + z + \delta} \left(U_0(z) + \frac{V(z)}{p} - \mathcal{L}\mathcal{R}_z U_p \right) \end{aligned} \quad (6.28)$$

and re-injecting $\mathcal{L}U_p$ from the second equation into the first equation we find

$$\mathcal{L}L_p(z) = \frac{1}{\frac{1}{2}p + z + \delta} \left(\frac{L_0(z)}{2} - \mathcal{L}l(p) \right) + \frac{1}{(\frac{1}{2}p + z + \delta)(p + z + \delta)} \left(U_0(z) + \frac{V(z)}{p} - \mathcal{L}\mathcal{R}_z U_p \right). \quad (6.29)$$

With similar considerations as before, with p large enough to have $-\delta$ is outside the loop Γ , we see the terms $\mathcal{L}l(p)$ and $\mathcal{L}\mathcal{R}_z U_p$ don't contribute to the former equation when the operator \mathcal{R}_z is applied

$$\mathcal{R}_z \mathcal{L}L_p(z) = \mathcal{R}_z \left\{ \frac{1}{2} \frac{L_0(z)}{\frac{1}{2}p + z + \delta} + \frac{1}{(\frac{1}{2}p + z + \delta)(p + z + \delta)} \left(U_0(z) + \frac{V(z)}{p} \right) \right\}. \quad (6.30)$$

¹Defined as $(\mathcal{L}f)(p) = \int_0^{+\infty} dt e^{-pt} f(t)$ for $\text{Re } p$ large enough. We also use the notation $\mathcal{L}f_p$ to mean $(\mathcal{L}f)(p)$ specially when there are other variables involved. For example $\mathcal{L}L_p(z) = \int_0^{+\infty} dt e^{-pt} L_t(z)$.

Chapter 6. The Random feature model

Finally, there remains to use the commutativity of \mathcal{R}_z and \mathcal{L} (for $\text{Re } p$ large enough by Fubini's theorem) and compute the inverse Laplace transforms to find

$$l(t) = \mathcal{R}_z \left\{ e^{-2t(z+\delta)} \left[L_0(z) + 2 \frac{e^{t(\delta+z)} - 1}{\delta+z} U_0(z) + \left(\frac{e^{t(\delta+z)} - 1}{\delta+z} \right)^2 V(z) \right] \right\} \quad (6.31)$$

Expanding further the terms individually

$$l(t) = \mathcal{R}_z \left\{ e^{-2t(z+\delta)} L_0(z) + 2e^{-t(z+\delta)} \left(\frac{1 - e^{-t(z+\delta)}}{\delta+z} \right) U_0(z) + \left(\frac{1 - e^{-t(z+\delta)}}{\delta+z} \right)^2 V(z) \right\}. \quad (6.32)$$

We end-up (as for $g(t)$) with an expression where the time dependence is decoupled from random matrix expressions.

6.B.2 Representation formula for $h(t)$

The last term requires additional considerations. We will now use a double contour Γ_x, Γ_y enclosing the eigenvalues of $\frac{Z^T Z}{\sqrt{N}}$ and such that $\Gamma_x \cap \Gamma_y = \emptyset$. We consider the operators $\mathcal{R}_x, \mathcal{R}_y$ associated to each contour. Contrary to the previous two representations, when computing the multiple derivatives $h^{(k)}(t)$, due to the Θ matrix in $h(t)$, there appears pairs of matrices $\frac{Z^T Z}{\sqrt{N}}$. In terms of generating functions, this translates into a "2-variable resolvent" functions

$$H_t(x, y) = \frac{\beta_t^T}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{\beta_t}{\sqrt{N}}, \quad (6.33)$$

which has the property $h(t) = \mathcal{R}_{x,y} H_t$, and two auxiliary functions

$$Q_t(x, y) = \frac{\beta_t^T}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T Y}{N}, \quad \text{and} \quad W(x, y) = \frac{Y^T Z}{N} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T Y}{N}. \quad (6.34)$$

Using the former method for equation (6.27) leads to the following integro-differential equations:

$$\begin{aligned} \frac{\partial H_t(x, y)}{\partial t} &= Q_t(x, y) + Q_t(y, x) - \mathcal{R}_x H_t(y) - \mathcal{R}_y H_t(x) - (x + y + 2\delta) H_t(x, y) \\ \frac{\partial Q_t(x, y)}{\partial t} &= W(x, y) - \mathcal{R}_x Q_t(y) - (x + \delta) Q_t(x, y) \end{aligned} \quad (6.35)$$

Then the Laplace transform on the first equation reads

$$\mathcal{L} H_p(x, y) = \frac{1}{p + x + y + 2\delta} \left[H_0(x, y) + \mathcal{L} \{ Q_t(x, y) + Q_t(y, x) - \mathcal{R}_x H_t(y) - \mathcal{R}_y H_t(x) \} \right]. \quad (6.36)$$

Notice that \mathcal{R}_x and \mathcal{R}_y commute with each other as being integrals over a compact set Γ_x, Γ_y respectively. So by Fubini we can name indifferently $\mathcal{R}_{x,y} = \mathcal{R}_x \mathcal{R}_y = \mathcal{R}_y \mathcal{R}_x$. Notice also that $\mathcal{R}_x H_t(y)$ is not a function of x anymore, thus for p large enough to have $|2\delta + x + y| > 0$ for all

$(x, y) \in \Gamma_x \times \Gamma_y$, we find

$$\mathcal{R}_{x,y} \left\{ \frac{\mathcal{R}_x H_t(y)}{p+x+y+2\delta} \right\} = \mathcal{R}_y \left\{ \mathcal{R}_x \left\{ \frac{\mathcal{R}_x H_t(y)}{p+x+y+2\delta} \right\} \right\} = \mathcal{R}_y \{0\} = 0. \quad (6.37)$$

Symmetrically, the same statement can be made for $\mathcal{R}_y H_t(x)$, so applying the operator $\mathcal{R}_{x,y}$ and the result (6.37) to (6.36) we find

$$\mathcal{R}_{x,y} \mathcal{L} H_p(x, y) = \mathcal{R}_{x,y} \left\{ \frac{H_0(x, y) + \mathcal{L} Q_p(x, y) + \mathcal{L} Q_p(y, x)}{p+x+y+2\delta} \right\}. \quad (6.38)$$

Finally, we have $\mathcal{R}_{x,y} \mathcal{L} H_p(x, y) = \mathcal{L} \mathcal{R}_{x,y} H_p(x, y) = \mathcal{L} h(p)$. The Laplace transform of the second equation of (6.35) provides

$$\mathcal{L} Q_p(x, y) = \frac{1}{p+x+\delta} \left(Q_0(x, y) + \frac{W(x, y)}{p} - \mathcal{R}_x \mathcal{L} Q_p(y) \right). \quad (6.39)$$

Before injecting this equation into (6.38) (and its symmetrical result in x and y), notice that one term will not contribute under the operator $\mathcal{R}_{x,y}$

$$\mathcal{R}_{x,y} \left\{ \frac{\mathcal{R}_x \mathcal{L} Q_p(y)}{(p+x+y+2\delta)(p+x+\delta)} \right\} = \mathcal{R}_y \{0\} = 0 \quad (6.40)$$

and finally, using $W(x, y) = W(y, x)$, we obtain

$$\mathcal{L} h(p) = \mathcal{R}_{x,y} \left\{ \frac{1}{p+x+y+2\delta} \left(H_0(x, y) + \frac{Q_0(x, y) + \frac{W(x, y)}{p}}{p+x+\delta} + \frac{Q_0(y, x) + \frac{W(x, y)}{p}}{p+y+\delta} \right) \right\}. \quad (6.41)$$

Eventually, applying inverse Laplace transform we get the representation

$$\begin{aligned} h(t) &= \mathcal{R}_{x,y} \left\{ e^{-t(x+y+2\delta)} H_0(x, y) \right\} \\ &+ \mathcal{R}_{x,y} \left\{ e^{-t(2\delta+x+y)} \left(\frac{e^{t(\delta+y)} - 1}{\delta+y} Q_0(x, y) + \frac{e^{t(\delta+x)} - 1}{\delta+x} Q_0(y, x) \right) \right\} \\ &+ \mathcal{R}_{x,y} \left\{ \frac{1 - e^{-t(x+\delta)}}{x+\delta} \frac{1 - e^{-t(y+\delta)}}{y+\delta} W(x, y) \right\} \end{aligned} \quad (6.42)$$

6.B.3 Remark on the consistency with the minimum least squares estimator

It can be seen, at least formally, that the integral representation formula correctly retrieves the minimum least-squares estimator formulas in the limit $t \rightarrow \infty$. Indeed, commuting \lim_t and \mathcal{R}_z we find

$$\begin{aligned} \lim_{t \rightarrow \infty} g(t) &= \mathcal{R}_z \left\{ \frac{1}{z+\delta} K(z) \right\} = \sum_{i=1}^N \frac{\beta^{*T}}{\sqrt{d}} v_i \mathcal{R}_z \left\{ \frac{1}{(\lambda_i+z)(z+\delta)} \right\} v_i^T \frac{Z^T Y}{N} \\ &= \sum_{i=1}^N \frac{\beta^{*T}}{\sqrt{d}} \frac{v_i v_i^T}{(\lambda_i - \delta)} \frac{Z^T Y}{N} = K(-\delta). \end{aligned} \quad (6.43)$$

On the other hand, we expect

$$\lim_{t \rightarrow +\infty} g(t) = \lim_t \frac{\beta^{*T} \Theta^T}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \beta_t = \frac{\beta^{*T} \Theta^T}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \beta_\infty \quad (6.44)$$

with β_∞ defined as the minimum least-squares estimator. Thus, we clearly have:

$$\frac{\beta^{*T} \Theta^T}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \beta_\infty = \frac{\beta^{*T} \Theta^T}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} \left(\frac{Z^T Z}{N} + \delta I \right)^{-1} \frac{Z^T}{\sqrt{N}} \frac{Y}{\sqrt{N}} = K(-\delta) \quad (6.45)$$

The same calculations can be done on each term $h(t), l(t)$.

6.B.4 Representation formula for the training error

The derivation of $\mathcal{E}_{\text{train}}(t)$ is quite straightforward based on the previous terms derived for the test error. Firstly, expanding the expression of $\mathcal{E}_{\text{train}}(t)$ we get:

$$\mathcal{E}_{\text{train}}(t) = \frac{1}{n} \left\| Y - Z \frac{\beta_t}{\sqrt{N}} \right\|^2 + \lambda \left\| \frac{\beta_t}{\sqrt{N}} \right\|^2 = \frac{\|Y\|^2}{n} - \frac{2}{n} Y^T Z \frac{\beta_t}{\sqrt{N}} + \frac{1}{n} \left\| \frac{Z \beta_t}{\sqrt{N}} \right\|^2 + \frac{\delta}{c} \left\| \frac{\beta_t}{\sqrt{N}} \right\|^2 \quad (6.46)$$

Reusing the function $U_t(z)$ from Sect. 6.B.1, and defining $u(t) = \mathcal{R}_z U_t(z) = \frac{1}{N} Y^T \frac{Z \beta_t}{\sqrt{N}}$ and $\tilde{h}(t) = \frac{1}{N} \left\| \frac{Z \beta_t}{\sqrt{N}} \right\|^2$, we get:

$$\mathcal{E}_{\text{train}}(t) = \frac{\|Y\|^2}{n} + \frac{1}{c} \left(-2u(t) + \tilde{h}(t) + \delta l(t) \right) \quad (6.47)$$

Furthermore, reusing the differential equation found for $U_t(z)$, a simpler solution can be extracted for $u(t)$:

$$u(t) = \mathcal{R}_z \left\{ e^{-t(z+\delta)} U_0(z) + \frac{1 - e^{-t(z+\delta)}}{z + \delta} V(z) \right\} \quad (6.48)$$

The second term $\tilde{h}(t)$ can also be derived from the expression $L_t(z)$ which is also defined in appendix 6.B.1. We find $\tilde{h}(t) = \mathcal{R}_z \{ z L_t(z) \}$. Hence the terms $\delta l(t)$ and $\tilde{h}(t)$ can be grouped together with $\tilde{h}(t) + \delta l(t) = \mathcal{R}_z \{ (z + \delta) L_t(z) \}$. Expanding from the expression of $\mathcal{R}_z \mathcal{L} L_t(z)$ we find

$$\begin{aligned} (\tilde{h} + \delta l)(t) = \mathcal{R}_z \left\{ (z + \delta) e^{-2t(z+\delta)} L_0(z) + 2e^{-t(z+\delta)} \left(1 - e^{-t(z+\delta)} \right) U_0(z) \right. \\ \left. + \frac{(1 - e^{-t(\delta+z)})^2}{\delta + z} V(z) \right\}. \end{aligned} \quad (6.49)$$

Remarkably, all the terms can be summed together in (6.47) and we retrieve a simpler expres-

sion

$$\mathcal{E}_{\text{train}}(t) = \frac{\|Y\|^2}{n} + \frac{1}{c} \mathcal{R}_z \left\{ (z + \delta) e^{-2t(z+\delta)} L_0(z) - 2e^{-2t(z+\delta)} U_0(z) - \frac{1 - e^{-2t(\delta+z)}}{\delta+z} V(z) \right\}. \quad (6.50)$$

6.C High-dimensional limit

In this appendix we use assumption 6.1 in Section 6.2.3 to compute limiting expressions of traces.

As $d \rightarrow \infty$, the mean of β_0 or β^* converges to 0. Let's consider the auxiliary functions $U_0(z), G_0(z), Q_0(x, y)$. These three terms have only occurrence of β_0 and β^* on each side of the matrix-vector multiplication composition (notice β^* is also included in the term Y): they can be written in the form $F(H) = \frac{\beta_0^T}{\sqrt{N}} H \frac{\beta^*}{\sqrt{d}}$ where H is a random matrix independent of β_0, β^* . For instance we have $G_0(z) = F\left(R(z) \frac{\Theta}{\sqrt{d}}\right)$. As the mean of $F(H)$ is precisely 0, assuming concentration, we have that these terms go to 0 when $d \rightarrow \infty$. The same considerations can be applied to the term ξ from Y .

Besides, when a vector such as β_0 is expressed on both side of another expression such as $F(H) = \frac{\beta_0^T}{\sqrt{N}} H \frac{\beta_0}{\sqrt{N}}$, it can still be rewritten as the trace $F(H) = \text{Tr} \left[H \frac{\beta_0 \beta_0^T}{N} \right]$ so that we can effectively use the independence of H with β_0 and compute the expectation $\mathbb{E}_{\beta_0} [F(H)] = \frac{r^2}{N} \text{Tr} [H]$. Hence if $F(H)$ concentrates as $N \rightarrow \infty$, we can replace it by $\lim_N \frac{r^2}{N} \text{Tr} [H]$.

In the sequel we will adopt the following notation. For any sequence of matrices $(M_k) \in \mathbb{R}^{k \times k}$ we set $\text{Tr}_k [M_k] = \lim_{k \rightarrow \infty} \frac{1}{k} \text{Tr} [M_k]$.

Therefore, in general, applying the concentration arguments above, we can substitute the limiting expressions with the following terms

$$L_0(z) = \frac{\beta_0^T}{\sqrt{N}} R(z) \frac{\beta_0}{\sqrt{N}} \xrightarrow{d \rightarrow \infty} r^2 \text{Tr}_N [R(z)] \quad (6.51)$$

$$K(z) = \frac{\beta^{*T}}{\sqrt{d}} \frac{\Theta^T}{\sqrt{d}} R(z) \frac{Z^T Y}{N} \xrightarrow{d \rightarrow \infty} \text{Tr}_d \left[\frac{\Theta^T}{\sqrt{d}} R(z) \frac{Z^T}{\sqrt{N}} \frac{X}{\sqrt{N}} \right] \quad (6.52)$$

$$H_0(x, y) = \frac{\beta_0^T}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{\beta_0}{\sqrt{N}} \xrightarrow{d \rightarrow \infty} r^2 \text{Tr}_N \left[R(x) \frac{\Theta \Theta^T}{d} R(y) \right] \quad (6.53)$$

$$V(z) = \frac{Y^T Z}{N} R(z) \frac{Z^T Y}{N} \xrightarrow{d \rightarrow \infty} \text{Tr}_d \left[\frac{X^T}{\sqrt{N}} \frac{Z}{\sqrt{N}} R(z) \frac{Z^T}{\sqrt{N}} \frac{X}{\sqrt{N}} \right] + s^2 \text{Tr}_N \left[\frac{Z}{\sqrt{N}} R(z) \frac{Z^T}{\sqrt{N}} \right] \quad (6.54)$$

$$W(x, y) \xrightarrow{d \rightarrow \infty} \text{Tr}_d \left[\frac{X^T}{\sqrt{N}} \frac{Z}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T}{\sqrt{N}} \frac{X}{\sqrt{N}} \right] + s^2 \text{Tr}_N \left[\frac{Z}{\sqrt{N}} R(x) \frac{\Theta \Theta^T}{d} R(y) \frac{Z^T}{\sqrt{N}} \right] \quad (6.55)$$

As for the training error, all the required terms are given by $V(z), L_0(z), U_0(z)$, of which only $V(z), L_0(z)$ contributes to the result as $d \rightarrow \infty$

Finally, we apply the gaussian equivalence principle with the substitution described in 6.4.2 with the linearization $Z \rightarrow Z_{\text{lin}}$ with $Z_{\text{lin}} \equiv \frac{\mu}{\sqrt{d}} X \Theta^T + \nu \Omega$. This substitution is applied throughout all the occurrences of Z , including in the resolvents $z \rightarrow R(z)$.

6.D Linear Pencil

6.D.1 Main matrix

The main approach of the linear-pencil method is to design a block-matrix $M_{x,y} = \sum_{i,j} E_{i,j} \otimes M_{x,y}^{(i,j)}$ where the blocks $M_{x,y}^{(i,j)}$ are either a gaussian random matrix or a scalar matrix, and $E_{i,j}$ is the matrix with matrix elements $(E_{i,j})_{k,l} = \delta_{ki} \delta_{lj}$. The subscripts indicate explicitly the dependence on two complex variables $(x, y) \in \mathbb{C}^2$. Importantly, this matrix is inverted using block-inversion formula to have an expression of the form $M_{x,y}^{-1} = \sum_{i,j} E_{i,j} \otimes (M_{x,y}^{-1})^{(i,j)}$ such that some blocks $(M_{x,y}^{-1})^{(i,j)}$ match the different matrix terms in equations (6.51).

In order to define our main linear pencil matrix, we first need to introduce some additional upper-level blocks: $U^T = [\frac{X}{\sqrt{N}}, \nu \frac{\Omega}{\sqrt{N}}]$ and $V^T = [\mu \frac{\Theta}{\sqrt{d}}, I]$. In addition, in order to keep a consistent symmetry and structure to our block-matrix, we will use the following blocks in reverse order: $\tilde{U}^T = [\nu \frac{\Omega}{\sqrt{N}}, \frac{X}{\sqrt{N}}]$ and $\tilde{V}^T = [I, \mu \frac{\Theta}{\sqrt{d}}]$. Furthermore, we let $K_x = (-xI + \frac{Z_{\text{lin}}^T Z_{\text{lin}}}{\sqrt{N}})^{-1}$ and $L_x = (-xI + U U^T V V^T)^{-1}$ and $R_x = (-xI + V V^T U U^T)^{-1}$ and $\tilde{K}_x = (-xI + \frac{Z_{\text{lin}} Z_{\text{lin}}^T}{\sqrt{N}})^{-1}$. The following identities (which can be obtained with the push-through identity) provide additional relations which can be used later:

$$\frac{Z_{\text{lin}}}{\sqrt{N}} = U^T V \quad (6.56)$$

$$L_x U U^T = U \tilde{K}_x U^T \quad (6.57)$$

$$V V^T L_x = V K_x V^T \quad (6.58)$$

$$-x \tilde{K}_x = I - \left(-xI + \frac{Z_{\text{lin}} Z_{\text{lin}}^T}{N} \right)^{-1} \frac{Z_{\text{lin}} Z_{\text{lin}}^T}{N} = I - \frac{Z_{\text{lin}}}{\sqrt{N}} K_x \frac{Z_{\text{lin}}^T}{\sqrt{N}} \quad (6.59)$$

We define our main block-matrix consisting in 13×13 blocks where the upper-level blocks

U, V, \bar{U}, \bar{V} are to be considered as "flattened":

$$M_{x,y} = \left[\begin{array}{cccc|cccc|c} -xI & -V^T & 0 & 0 & \frac{\Theta}{\sqrt{d}} & 0 & 0 & 0 & 0 \\ 0 & I & U & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & U^T & 0 & 0 & 0 & 0 & 0 \\ V & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & \frac{\Theta^T}{\sqrt{d}} \\ \hline 0 & 0 & 0 & 0 & 0 & I & \bar{U} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & \bar{U}^T & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & -\bar{V} \\ \hline 0 & 0 & 0 & 0 & 0 & \bar{V}^T & 0 & 0 & -yI \end{array} \right] \quad (6.60)$$

This is precisely the block-matrix M given at the end of Sect. 6.4.

6.D.2 Linear-pencil inversion and relation to the matrix terms

The inverse of $M_{x,y}$ can be computed by splitting it into higher-level blocks. These blocks are highlighted with the lines and double-lines depicted in equation (6.60): the block-matrix is split into a 2×2 block-matrix recursively in order to apply the block-matrix inversion formula recursively. Starting with the higher level split:

$$M_{x,y} = \left[\begin{array}{c|c} M_{1,1} & M_{1,2} \\ \hline 0 & M_{2,2} \end{array} \right] \Rightarrow M_{x,y}^{-1} = \left[\begin{array}{c|c} M_{1,1}^{-1} & -M_{1,1}^{-1}M_{1,2}M_{2,2}^{-1} \\ \hline 0 & M_{2,2}^{-1} \end{array} \right] \quad (6.61)$$

It is now quite straightforward algebra to proceed with the remaining blocks. Starting with $M_{1,1}$:

$$M_{1,1}^{-1} = \left[\begin{array}{cccc} K_x & K_x V^T & -K_x \frac{Z_{\text{in}}^T}{\sqrt{N}} & K_x \frac{Z_{\text{in}}^T}{\sqrt{N}} U^T \\ -U \frac{Z_{\text{in}}}{\sqrt{N}} K_x & -xL_x & xL_x U & -xL_x U U^T \\ \frac{Z_{\text{in}}}{\sqrt{N}} K_x & \frac{Z_{\text{in}}}{\sqrt{N}} V^T L_x & -x\tilde{K}_x & x\tilde{K}_x U^T \\ -VK_x & -VV^T L_x & V \frac{Z_{\text{in}}^T}{\sqrt{N}} \tilde{K}_x & -xR_x \end{array} \right] \quad (6.62)$$

For $M_{2,2}$, with an additional split:

$$M_{2,2} = \left[\begin{array}{c|c} I & N_{1,2} \\ \hline 0 & N_{2,2} \end{array} \right] \Rightarrow M_{2,2}^{-1} = \left[\begin{array}{c|c} I & -N_{1,2}N_{2,2}^{-1} \\ \hline 0 & N_{2,2}^{-1} \end{array} \right] \quad (6.63)$$

Chapter 6. The Random feature model

A straightforward algebra calculation provides the result of $M_{2,2}^{-1}$:

$$M_{2,2}^{-1} = \begin{bmatrix} I & \frac{\Theta^T}{\sqrt{d}} K_y \bar{V}^T & -\frac{\Theta^T}{\sqrt{d}} K_y \frac{Z_{\text{lin}}^T}{\sqrt{N}} & \frac{\Theta^T}{\sqrt{d}} K_y \frac{Z_{\text{lin}}^T}{\sqrt{N}} \bar{U}^T & -\frac{\Theta^T}{\sqrt{d}} K_y \\ 0 & -y \bar{R}_y & y \bar{U} \bar{K}_y & -y \bar{U} \bar{U}^T \bar{L}_y & \bar{U} \frac{Z_{\text{lin}}}{\sqrt{N}} K_y \\ 0 & \bar{K}_y \frac{Z_{\text{lin}}}{\sqrt{N}} \bar{V}^T & -y \bar{K}_y & y \bar{U}^T \bar{L}_y & -\frac{Z_{\text{lin}}}{\sqrt{N}} K_y \\ 0 & -\bar{L}_y \bar{V} \bar{V}^T & \bar{L}_y \bar{V} \frac{Z_{\text{lin}}^T}{\sqrt{N}} & -y \bar{L}_y & \bar{V} K_y \\ 0 & -K_y \bar{V}^T & K_y \frac{Z_{\text{lin}}^T}{\sqrt{N}} & -K_y \frac{Z_{\text{lin}}^T}{\sqrt{N}} \bar{U}^T & K_y \end{bmatrix} \quad (6.64)$$

Finally, using $Q = K_x \frac{\Theta \Theta^T}{d} K_y$ we obtain the third block of $M_{x,y}$:

$$-M_{1,1}^{-1} M_{1,2} M_{2,2}^{-1} = \begin{bmatrix} -K_x \frac{\Theta}{\sqrt{d}} & -Q \bar{V}^T & Q \frac{Z_{\text{lin}}^T}{\sqrt{N}} & -Q \frac{Z_{\text{lin}}^T \bar{U}^T}{\sqrt{N}} & Q \\ \frac{U Z_{\text{lin}}}{\sqrt{N}} K_x \frac{\Theta}{\sqrt{d}} & \frac{U Z_{\text{lin}}}{\sqrt{N}} Q \bar{V}^T & -\frac{U Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T}{\sqrt{N}} & \frac{U Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T \bar{U}^T}{\sqrt{N}} & -\frac{U Z_{\text{lin}}}{\sqrt{N}} Q \\ -\frac{Z_{\text{lin}}}{\sqrt{N}} K_x \frac{\Theta}{\sqrt{d}} & -\frac{Z_{\text{lin}}}{\sqrt{N}} Q \bar{V}^T & \frac{Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T}{\sqrt{N}} & -\frac{Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T \bar{U}^T}{\sqrt{N}} & \frac{Z_{\text{lin}}}{\sqrt{N}} Q \\ V K_x \frac{\Theta}{\sqrt{d}} & V Q \bar{V}^T & -V Q \frac{Z_{\text{lin}}^T}{\sqrt{N}} & V Q \frac{Z_{\text{lin}}^T \bar{U}^T}{\sqrt{N}} & -V Q \end{bmatrix} \quad (6.65)$$

Notice now that all the matrix terms in equations (6.51) are actually contained in some of the blocks of our matrix (note that $\text{Tr}_d \left[\frac{X^T X}{n} \right] = 1$):

$$\bar{L}_0(y) = r^2 \text{Tr}_N [K_y] \quad (6.66)$$

$$\bar{K}(y) = \text{Tr}_d \left[\frac{\Theta^T}{\sqrt{d}} K_y \frac{Z_{\text{lin}}^T}{\sqrt{N}} \bar{U}^T \right]_{1,2} \quad (6.67)$$

$$\bar{H}_0(x, y) = r^2 \text{Tr}_N [Q] \quad (6.68)$$

$$\bar{W}(x, y) = s^2 \frac{\phi}{\psi} \text{Tr}_n \left[\frac{Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T}{\sqrt{N}} \right] + \text{Tr}_d \left[\frac{U Z_{\text{lin}}}{\sqrt{N}} Q \frac{Z_{\text{lin}}^T \bar{U}^T}{\sqrt{N}} \right]_{1,2} \quad (6.69)$$

$$\bar{V}(x) = s^2 \frac{\phi}{\psi} \text{Tr}_n [I_n + x \bar{K}_x] + \left(\text{Tr}_d [x L_x U U^T]_{1,1} + \text{Tr}_d \left[\frac{X^T X}{N} \right] \right) \quad (6.70)$$

Or equivalently, with the block coordinates of the inverse matrix $M_{x,y}^{-1}$:

$$\bar{L}_0(y) = r^2 \text{Tr}_N \left[(M_{x,y}^{-1})^{(13,13)} \right] \quad (6.71)$$

$$\bar{K}(y) = \text{Tr}_d \left[(M_{x,y}^{-1})^{(7,12)} \right] \quad (6.72)$$

$$\bar{H}_0(x, y) = r^2 \text{Tr}_N \left[(M_{x,y}^{-1})^{(1,13)} \right] \quad (6.73)$$

$$\bar{W}(x, y) = s^2 \frac{\phi}{\psi} \text{Tr}_n \left[(M_{x,y}^{-1})^{(4,10)} \right] + \text{Tr}_d \left[(M_{x,y}^{-1})^{(2,12)} \right] \quad (6.74)$$

$$\bar{V}(x) = s^2 \frac{\phi}{\psi} \left(1 - \text{Tr}_n \left[(M_{x,y}^{-1})^{(4,4)} \right] \right) + \left(-\text{Tr}_d \left[(M_{x,y}^{-1})^{(2,5)} \right] + \frac{\phi}{\psi} \right) \quad (6.75)$$

In the next section we show how to derive further each trace of the squared matrices from the block matrix $M_{x,y}$. In order to deal with self-adjoint matrices, we double the dimensions with

$\tilde{M}_{x,y}$:

$$\tilde{M}_{x,y} = \begin{bmatrix} \mathbf{0} & M_{x,y} \\ M_{x,y}^\dagger & \mathbf{0} \end{bmatrix} \quad (6.76)$$

and find the inverse:

$$\tilde{M}_{x,y}^{-1} = \begin{bmatrix} \mathbf{0} & (M_{x,y}^\dagger)^{-1} \\ M_{x,y}^{-1} & \mathbf{0} \end{bmatrix} \quad (6.77)$$

6.D.3 Structural terms of the limiting traces

The matrix $M_{x,y}$ is a block-matrix constituted with either gaussian random matrices, or constant matrices (proportional to I). More precisely, letting S be the matrix of the coefficients of the constant blocks of $M_{x,y}$ (and \tilde{S} for $\tilde{M}_{x,y}$), and A the random blocks part (\tilde{A} respectively) we write: $\tilde{M}_{x,y} = \sum_{i,j} E_{i,j} \otimes \tilde{M}_{x,y}^{(i,j)}$ where $\tilde{M}_{x,y}^{(i,j)} = \tilde{S}^{(i,j)} + \tilde{A}^{(i,j)}$ is the block of size (N_i, N_j) . Also notice that letting $\mathbb{L} = \{(i, j) \mid N_i = N_j\}$, the fact that the constant blocks are supposed to be proportional to an identity matrix implies that: $\forall (i, j) \notin \mathbb{L} \implies \tilde{S}^{(i,j)} = \mathbf{0} = z_{i,j} \mathbf{0}_{N_i, N_j}$ with $\mathbf{0}_{N_i, N_j}$ the zero-matrix of size $N_i \times N_j$ and otherwise $\forall (i, j) \in \mathbb{L} \implies \tilde{S}^{(i,j)} = z_{i,j} I_{N_i}$ with $\tilde{B} = (z_{i,j})$ the matrix of size 26×26 .

Now we want to find a matrix $\tilde{G} \in \mathbb{R}^{26 \times 26}$ such that

$$[\tilde{G}]_{i,j} = \text{Tr}_{N_i} \left[(\tilde{M}_{x,y}^{-1})^{(i,j)} \right], \quad \forall (i, j) \in \mathbb{L}, \quad (6.78)$$

An important theorem in Mingo and Speicher (2017) (chapter 9, equ. (9.5) and theorem 2), which we show again in the next section, states that there is a solution \tilde{G} of the equation

$$\tilde{B}\tilde{G} = I + \eta(\tilde{G})\tilde{G} \quad (6.79)$$

which satisfies (6.78). In this equation $\eta(\tilde{G})$ is the matrix mapping defined element-wise as:

$$[\eta(\tilde{G})]_{i,j} = \delta_{\mathbb{L}}(i, j) \cdot \sum_{k,l \in \mathbb{L}} \sigma(i, k; l, j) \cdot [\tilde{G}]_{k,l} \quad (6.80)$$

and where σ satisfies the relation for all (i, k, l, j) such that $N_i = N_j$ and $N_k = N_l$ (and keeping in mind that the N_k are growing with the dimension d):

$$\forall (r, s) \in \{1, \dots, N_i\} \times \{1, \dots, N_j\}, r \neq s \implies \sigma(i, k; l, j) = \lim_{d \rightarrow \infty} N_k \cdot \mathbb{E} \left[[\tilde{A}^{(i,k)}]_{r,s} [\tilde{A}^{(l,j)}]_{s,r} \right] \quad (6.81)$$

We remark that the setting here, and in particular equation (6.79), is in fact more general than in Mingo and Speicher (2017) (chapter 9, equ. (9.5)) and we provide an independent and self-contained (formal) derivation of (6.79) in Chapter 3 with various methods.

For example, we have $M_{x,y}^{(5,1)} = \mu \frac{\Theta^T}{\sqrt{d}}$ of size $d \times N$ and $M_{x,y}^{(1,7)} = \frac{\Theta}{\sqrt{d}}$ of size $N \times d$. So this is $\tilde{M}_{x,y}^{(5,14)} = \mu \frac{\Theta^T}{\sqrt{d}}$ and $\tilde{M}_{x,y}^{(1,20)} = \frac{\Theta}{\sqrt{d}}$, with $N_5 = N_{20} = d$ and $N_{14} = N_1 = N$. For $r = 1, s = 2$ (or any

Chapter 6. The Random feature model

other suitable indices) we find:

$$\sigma(5, 14; 1, 20) = \lim_{d \rightarrow \infty} \mu \frac{N}{d} \mathbb{E} [[\Theta]_{1,2}^2] = \mu \psi$$

In fact, a careful inspection of all the blocks in row 5 and all the blocks in column 20 shows that we have $[\eta(\tilde{G})]_{5,20} = \mu \psi [\tilde{G}]_{14,1}$.

Calculating all the terms of $\eta(\tilde{G})$ is quite cumbersome, but it can be done automatically with the help of a computer algebra system. Still, this approach yields many equations for each 26×26 terms of \tilde{G} . However, some initial structure can also be provided for this matrix. Looking back at $\tilde{M}_{x,y}^{-1}$, it is clear that some blocks will have the same limiting traces (potentially seen using the aforementioned push-through identities). For instance, $(M_{1,1}^{-1})^{(1,1)} = K_x = -(M_{1,1}^{-1})^{(6,1)}$ (expanding the U, V blocks), so $(M_{x,y}^{-1})^{(1,1)} = -(M_{x,y}^{-1})^{(6,1)}$, in other words $(\tilde{M}_{x,y}^{-1})^{(14,1)} = -(\tilde{M}_{x,y}^{-1})^{(19,1)}$, and thus we expect $[\tilde{G}]_{14,1} = -[\tilde{G}]_{19,1}$. Non-squared blocks can also be mapped to 0 in \tilde{G} . In the end, taking every block into account, \tilde{G} is expected to be of the form:

$$\tilde{G} = \left[\begin{array}{c|c} 0 & G^\dagger \\ \hline G & 0 \end{array} \right] \quad (6.82)$$

with

$$G = \left[\begin{array}{c|c|c} G_{1,1} & G_{1,2} & G_{1,3} \\ \hline 0 & 1 & G_{2,3} \\ \hline 0 & 0 & G_{3,3} \end{array} \right] \quad (6.83)$$

(which has 13×13 scalar matrix elements) where:

$$G_{1,3} = \left[\begin{array}{c|c|c|c|c|c} -q_1 & 0 & 0 & -vq_6^{yx} & 0 & q_1 \\ \hline 0 & \mu q_7^{yx} & 0 & 0 & q_2 & 0 \\ \hline vq_6^{xy} & 0 & 0 & v^2 q_3 & 0 & -vq_6^{xy} \\ \hline 0 & 0 & q_4 & 0 & 0 & 0 \\ \hline 0 & \mu^2 q_5 & 0 & 0 & \mu q_7^{xy} & 0 \\ \hline q_1 & 0 & 0 & vq_6^{yx} & 0 & -q_1 \end{array} \right] \quad (6.84)$$

$$G_{1,1} = \left[\begin{array}{c|c|c|c|c|c} g_1^x & 0 & g_1^x & 0 & 0 & vg_2^x \\ \hline 0 & h_1^x & 0 & 0 & h_4^x & 0 \\ \hline -vg_2^x & 0 & h_2^x & 0 & 0 & v^2 h_5^x \\ \hline 0 & 0 & 0 & g_3^x & 0 & 0 \\ \hline 0 & -\mu^2 h_3^x & 0 & 0 & h_1^x & 0 \\ \hline -g_1^x & 0 & -g_1^x & 0 & 0 & h_2^x \end{array} \right] \quad (6.85)$$

$$G_{3,3} = \left[\begin{array}{cc|cc|c} h_2^y & 0 & 0 & v^2 h_5^y & 0 & v g_2^y \\ 0 & h_1^y & 0 & 0 & h_4^y & 0 \\ \hline 0 & 0 & g_3^y & 0 & 0 & 0 \\ -g_1^y & 0 & 0 & h_2^y & 0 & g_1^y \\ 0 & -\mu^2 h_3^y & 0 & 0 & h_1^y & 0 \\ \hline -g_1^y & 0 & 0 & -v g_2^y & 0 & g_1^y \end{array} \right] \quad (6.86)$$

$$G_{1,2} = \left[\begin{array}{c} 0 \\ \hline t_1^x \\ 0 \\ \hline 0 \\ \hline \mu h_3^x \\ 0 \end{array} \right] \quad G_{2,3} = \left[0 \quad \mu h_3^y \mid 0 \mid 0 \quad t_1^y \mid 0 \right] \quad (6.87)$$

All (non-vanishing) matrix elements depend on the complex variables x and y . This is indicated by the upper-script notation with x, y, xy, yx . Some quantities depend only on x , some only on y , and some on both x and y . Among the ones that depend on both variables the quantities $q_6^{xy}, q_6^{yx}, q_7^{xy}, q_7^{yx}$ are *non-symmetric*, while q_1, q_2, q_3, q_4, q_5 are *symmetric* (e.g., $q_1^{x,y} = q_1^{y,x}$). We choose not to use the upper-script notation for the symmetric quantities in order to distinguish them from the *non-symmetric* ones.

Eventually, with a careful mapping between $\tilde{M}_{x,y}^{-1}$ and \tilde{G} in equations (6.66), only $g_1^x, t_1^x, h_4^x, g_3^x$ and the symmetric terms q_1, q_2, q_4 are needed and equations (6.66) take the form:

$$\bar{L}_0(x) = r^2 g_1^x \quad (6.88)$$

$$\bar{K}(x) = t_1^x \quad (6.89)$$

$$\bar{H}_0(x, y) = r^2 q_1 \quad (6.90)$$

$$\bar{W}(x, y) = s^2 \frac{\phi}{\psi} q_4 + q_2 \quad (6.91)$$

$$\bar{V}(x) = s^2 \frac{\phi}{\psi} (1 - g_3^x) + \left(\frac{\phi}{\psi} - h_4^x \right) \quad (6.92)$$

6.D.4 Solution of the fixed point equation

The fixed-point equations as described in (6.79) for the given matrices $\tilde{S}, \eta(\tilde{G}), \tilde{G}$ is a priori a system of 26×26 algebraic equations. are computed using Sympy in python, a symbolic calculation tool. In effect this is really a fixed point equation for G a priori involving 13×13 algebraic equations. It turns out that many matrix elements vanish and (using the symbolic calculation tool Sympy in python) we can extract a system of 39 algebraic equations which are given in the following:

$$0 = g_1^x (-\mu^2 h_4^x + x) - g_2^x v + 1 \quad (6.93)$$

$$0 = g_1^x (-\mu^2 h_4^x + x) + h_2^x \quad (6.94)$$

$$0 = g_2^x v (-\mu^2 h_4^x + x) + h_5^x v^2 \quad (6.95)$$

$$0 = -g_1^y (\mu^2 q_2 - \mu t_1^x - \mu t_1^y + 1) + v q_6^{xy} - q_1 (-\mu^2 h_4^x + x) \quad (6.96)$$

$$0 = -g_2^y v (\mu^2 q_2 - \mu t_1^x - \mu t_1^y + 1) + v^2 q_3 - v q_6^{yx} (-\mu^2 h_4^x + x) \quad (6.97)$$

$$0 = g_1^y (\mu^2 q_2 - \mu t_1^x - \mu t_1^y + 1) - v q_6^{xy} + q_1 (-\mu^2 h_4^x + x) \quad (6.98)$$

$$0 = \frac{\mu^2 \phi g_3^x h_3^x}{\psi} - h_1^x + 1 \quad (6.99)$$

$$0 = \frac{\phi g_3^x h_1^x}{\psi} - h_4^x \quad (6.100)$$

$$0 = -\frac{\mu \phi g_3^x h_3^x}{\psi} - t_1^x \quad (6.101)$$

$$0 = \frac{\mu^2 \phi g_3^x q_5}{\psi} + \frac{\mu^2 \phi h_3^y q_4}{\psi} - \mu q_7^{yx} \quad (6.102)$$

$$0 = \frac{\mu \phi g_3^x q_7^{xy}}{\psi} + \frac{\phi h_1^y q_4}{\psi} - q_2 \quad (6.103)$$

$$0 = -\frac{\phi g_1^x g_3^x v^2}{\psi} + g_2^x v \quad (6.104)$$

$$0 = -\frac{\phi g_1^x g_3^x v^2}{\psi} - h_2^x + 1 \quad (6.105)$$

$$0 = \frac{\phi g_3^x h_2^x v^2}{\psi} - h_5^x v^2 \quad (6.106)$$

$$0 = -\frac{\phi g_1^y v^2 q_4}{\psi} + \frac{\phi g_3^x v^2 q_1}{\psi} - v q_6^{xy} \quad (6.107)$$

$$0 = \frac{\phi g_3^x v^3 q_6^{yx}}{\psi} + \frac{\phi h_2^y v^2 q_4}{\psi} - v^2 q_3 \quad (6.108)$$

$$0 = \frac{\phi g_1^y v^2 q_4}{\psi} - \frac{\phi g_3^x v^2 q_1}{\psi} + v q_6^{xy} \quad (6.109)$$

$$0 = g_3^x \left(\frac{\mu^2 h_3^x}{\psi} - g_1^x v^2 - 1 \right) + 1 \quad (6.110)$$

$$0 = g_3^y \left(\frac{\mu^2 q_5}{\psi} + v^2 q_1 \right) + q_4 \left(\frac{\mu^2 h_3^x}{\psi} - g_1^x v^2 - 1 \right) \quad (6.111)$$

$$0 = -\mu^2 \psi g_1^x h_1^x - \mu^2 h_3^x \quad (6.112)$$

$$0 = -\mu^2 \psi g_1^x h_4^x - h_1^x + 1 \quad (6.113)$$

$$0 = -\mu^2 \psi g_1^x t_1^x + \mu \psi g_1^x + \mu h_3^x \quad (6.114)$$

$$0 = -\mu^3 \psi g_1^x q_7^{yx} - \mu^2 \psi g_1^x h_3^y + \mu^2 \psi h_1^y q_1 - \mu^2 q_5 \quad (6.115)$$

$$0 = -\mu^2 \psi g_1^x q_2 + \mu^2 \psi h_4^y q_1 + \mu \psi g_1^x t_1^y - \mu q_7^{xy} \quad (6.116)$$

$$0 = -g_2^x v - h_2^x + 1 \quad (6.117)$$

$$0 = \mu \psi g_1^y h_1^y + \mu h_3^y \quad (6.118)$$

$$0 = \mu\psi g_1^y h_4^y - t_1^y \quad (6.119)$$

$$0 = -\frac{\phi g_1^y g_3^y v^2}{\psi} - h_2^y + 1 \quad (6.120)$$

$$0 = \frac{\phi g_3^y h_2^y v^2}{\psi} - h_5^y v^2 \quad (6.121)$$

$$0 = \frac{\phi g_1^y g_3^y v^2}{\psi} - g_2^y v \quad (6.122)$$

$$0 = \frac{\mu^2 \phi g_3^y h_3^y}{\psi} - h_1^y + 1 \quad (6.123)$$

$$0 = \frac{\phi g_3^y h_1^y}{\psi} - h_4^y \quad (6.124)$$

$$0 = g_3^y \left(\frac{\mu^2 h_3^y}{\psi} - g_1^y v^2 - 1 \right) + 1 \quad (6.125)$$

$$0 = -g_2^y v - h_2^y + 1 \quad (6.126)$$

$$0 = -\mu^2 \psi g_1^y h_1^y - \mu^2 h_3^y \quad (6.127)$$

$$0 = -\mu^2 \psi g_1^y h_4^y - h_1^y + 1 \quad (6.128)$$

$$0 = -g_1^y (-\mu^2 h_4^y + y) - h_2^y \quad (6.129)$$

$$0 = -g_2^y v (-\mu^2 h_4^y + y) - h_5^y v^2 \quad (6.130)$$

$$0 = g_1^y (-\mu^2 h_4^y + y) - g_2^y v + 1 \quad (6.131)$$

6.D.5 Reduction of the solutions

The previous system of equations can be reduced further by substitutions with a computer algebra system. We find the variables $g_3^x, t_1^x, h_4^x, g_1^x, h_1^x$ are linked through the algebraic system:

$$\begin{cases} 0 = 1 + g_1^x \left(-\mu^2 h_4^x - \frac{\phi}{\psi} g_3^x u^2 + x \right) \\ 0 = -h_4^x + g_3^x \left(-\mu^2 \phi g_1^x h_4^x + \frac{\phi}{\psi} \right) \\ 0 = \frac{\phi}{\psi} (1 - g_3^x) - g_1^x x - 1 \\ 0 = \mu\psi g_1^x h_4^x - t_1^x \\ 0 = 1 - h_1^x - \mu t_1^x \end{cases} \quad (6.132)$$

Notice this system can be shrunk further down to 3 equations to get to the main result in 6.1 using the substitution h_1^x with the 5th equation and g_3^x with the 3rd equation. Also, by symmetry we find the same equations for $g_3^y, t_1^y, h_4^y, g_1^y, h_1^y$.

For the other variables, a set of equations link q_1, q_2, q_4, q_5 . Notice there can many different representations depending on the reductions that are applied. Here we only show the example

which has been used throughout the computations:

$$\begin{cases} 0 = -\mu^2 g_1^y q_2 + \mu^2 h_4^x q_1 + \mu g_1^y t_1^x + \mu g_1^y t_1^y - \frac{\phi g_1^y q_4 v^2}{\psi} - g_1^y - q_1 x + \frac{q_1 v^2 (\phi - \psi g_1^x x - \psi)}{\psi} \\ 0 = \mu (\phi - \psi g_1^x x - \psi) (-\mu g_1^x q_2 + \mu h_4^y q_1 + g_1^x t_1^y) + \frac{\phi h_1^y q_4}{\psi} - q_2 \\ 0 = -\mu^2 g_1^x h_1^x q_4 + \frac{\mu^2 q_5 (\phi - \psi g_1^y y - \psi)}{\phi \psi} - g_1^x q_4 v^2 - q_4 + \frac{q_1 v^2 (\phi - \psi g_1^y y - \psi)}{\phi} \\ 0 = \mu^2 \phi g_1^x g_1^y h_1^y q_4 - \frac{\mu^2 g_1^x q_5 (\phi - \psi g_1^x x - \psi)}{\psi} + \psi g_1^x g_1^y h_1^y + h_1^y q_1 - \frac{q_5}{\psi} \end{cases} \quad (6.133)$$

In conclusion, we can obtain 3 systems with (4, 5, 5)-equations or 3 systems with (4, 3, 3)-equations (so a total of 10), as in the main result 6.1 (as discussed above these various systems are all equivalent and depend on the applied reductions).

The solutions are not necessarily unique and one has to choose the appropriate ones with care. In our experimental results using Matlab with the "vpasolve" function, conditioning on $\text{Im } g_1^x > 0$ and $\text{Im } g_3^x > 0$ provided a unique solution to (6.132) for $x \in \mathbb{R}_+$ (or $x \in \mathbb{R} \times i[0, \epsilon]$ for ϵ close to 0); while conditioning on $g_1^x, g_3^x \in \mathbb{R}_+$ provided a unique solution to (6.132) for $x \in \mathbb{R}_-$. We remind that we use $x \in \mathbb{R}_-$ exclusively in the time limit $t \rightarrow \infty$ in result 6.2 while we use $x \in \mathbb{R}_+$ in the situation of result 6.1. In addition, we found that selecting the appropriate solutions for x and y as just described for (6.132) also led to a unique solution for 6.133 in our experiments.

6.E Numerical results

All the experiments are run on a standard desktop configuration:

1. Matlab R2019b is used to generate the heatmaps or 3D landscapes. Most exemples can be generated in less than 12h on a standard machine.
2. The experimental comparisons run on a standard instance of a Google collaboratory notebook in less than a few hours.

6.E.1 Numerical computations

We take equation (6.4) as an example of how to proceed with the numerical experiments. Specifically we consider the second integral in the Cauchy integral representation of $g(t)$

$$g_2(t) = -\frac{1}{2i\pi} \oint_{\Gamma} dz \left\{ \frac{1 - e^{-t(z+\delta)}}{z + \delta} K(z) \right\}. \quad (6.134)$$

We choose a contour with $\lambda^* = \max \text{Sp} \frac{Z^T Z}{N}$ with two positive fixed constants ϵ, Δ :

$$\Gamma = \{\gamma \lambda^* \pm i\Delta | -\epsilon \leq \gamma \leq 1 + \epsilon\} \cup \{\epsilon \lambda^* + \gamma i\Delta | -1 \leq \gamma \leq 1\} \cup \{-\epsilon + \gamma i\Delta | -1 \leq \gamma \leq 1\} \quad (6.135)$$

Now, the integrand is continuous in $\lambda^* + \epsilon$ and $-\epsilon$ for ϵ small enough. So taking the limit $\epsilon \rightarrow 0$ and $\Delta \rightarrow 0$

$$g(t) = \lim_{\Delta \rightarrow 0} \frac{1}{2i\pi} \int_0^{\lambda^*} \left\{ \frac{1 - e^{-t(r+\delta+i\Delta)}}{r + \delta + i\Delta} K(r + i\Delta) - \frac{1 - e^{-t(r+\delta)-i\Delta}}{r + \delta - i\Delta} K(r - i\Delta) \right\} dr \quad (6.136)$$

which is simply

$$g(t) = \int_0^{\lambda^*} \frac{1 - e^{-t(r+\delta)}}{r + \delta} \lim_{\Delta \rightarrow 0} \frac{1}{2i\pi} \left\{ K(r + i\Delta) - K(r - i\Delta) \right\} dr \quad (6.137)$$

Obviously the inward term is also given by the limit $\lim_{\Delta \rightarrow 0} \frac{1}{\pi} \text{Im} K(r + i\Delta)$. So this all there is to compute from the former algebraic equations are appropriate imaginary parts. This can be done by taking a discretized interval $0 \leq r_1 \leq \dots \leq r_K \leq \lambda^*$, and solving the algebraic equations for the imaginary value $\text{Im} t_1^x$ for $x = r_i, i = 1, \dots, K$.

We proceed similarly with the terms containing two complex variables x and y (or two resolvents). For instance for $W(x, y)$ one uses the limit in $\Delta_x, \Delta_y \rightarrow 0$ of $\rho(x, y)$ where

$$\rho(x, y) = \lim_{\Delta_x \rightarrow 0} \lim_{\Delta_y \rightarrow 0} \left[\frac{-1}{4\pi^2} \left\{ W(r_x + i\Delta_x, r_y + i\Delta_y) - W(r_x + i\Delta_x, r_y - i\Delta_y) \right\} - \frac{-1}{4\pi^2} \left\{ W(r_x - i\Delta_x, r_y + i\Delta_y) - W(r_x - i\Delta_x, r_y - i\Delta_y) \right\} \right] \quad (6.138)$$

or equivalently

$$\rho(x, y) = \lim_{\Delta_x, \Delta_y \rightarrow 0} \frac{1}{2\pi^2} \text{Re} \left\{ W(r_x + i\Delta_x, r_y - i\Delta_y) - W(r_x + i\Delta_x, r_y + i\Delta_y) \right\} \quad (6.139)$$

6.E.2 Technical considerations

Dirac distributions with 1-variable functions: It happens that the limiting distribution $\frac{Z^T Z}{N}$ may contain a mixture of a Dirac peak at 0 and a continuous measure. For instance, $K(z)$ may contain a branch cut in the interval $\mathcal{C}^* = [\lambda_1, \lambda^*]$ with $\lambda_0 = 0 < \lambda_1 < \lambda^* < \infty$ along with an isolated pole in 0 with: $K(z) = \frac{\alpha}{0-z} + K_c(z)$ (where $K_c: \mathbb{C} \setminus \mathcal{C}^* \rightarrow \mathbb{C}$). For instance, equation (6.137) becomes:

$$g(t) = \alpha \frac{1 - e^{-t\delta}}{\delta} + \int_{\lambda_0}^{\lambda^*} dr \frac{1 - e^{-t(\delta+r)}}{r + \delta} \lim_{\Delta \rightarrow 0} \frac{1}{\pi} \text{Im} K_c(r + i\Delta) \quad (6.140)$$

The weight α can be retrieved by computing $\alpha = \lim_{\epsilon \rightarrow 0^+} (-i\epsilon) K(i\epsilon) = \lim_{\epsilon \rightarrow 0^+} \epsilon \text{Im} K(i\epsilon)$.

Dirac distributions with 2-variables functions: Similarly, we can have an isolated pole at 0 for x, y for $W(x, y)$. In that case, we can write down $W(x, y)$ as for instance:

$$W(x, y) = \frac{\alpha_{xy}}{(0-x)(0-y)} + \frac{\alpha_x}{0-x} W_y(y) + \frac{\alpha_y}{0-y} W_x(x) + W_{xy}(x, y) \quad (6.141)$$

where W_x, W_y are defined on $\mathbb{C} \setminus \mathcal{C}^* \rightarrow \mathbb{C}$ and $W_{xy} : (\mathbb{C} \setminus \mathcal{C}^*)^2 \rightarrow \mathbb{C}$. Firstly, We can easily find α_{xy} with:

$$\alpha_{xy} = \lim_{\epsilon \rightarrow 0^+} (-\epsilon^2) \operatorname{Re} W(i\epsilon, i\epsilon) \quad (6.142)$$

Secondly, all the considered 2-variables functions are symmetrical with respect to x and y : $W(x, y) = W(y, x)$ which implies that $\alpha_x = \alpha_y$ and $W_x(r) = W_y(r)$ for all $r \in \mathbb{C} \setminus \mathcal{C}^*$. Therefore, if we have $\gamma_t(z) = \frac{1-e^{-t(\delta+z)}}{z+\delta}$, we have to compute:

$$\begin{aligned} \mathcal{R}_{x,y} \{ \gamma_t(x) \gamma_t(y) W(x, y) \} &= \gamma_t(0)^2 \alpha_{xy} + \iint_{[\lambda_0, \lambda^*]^2} \gamma_t(u) \gamma_t(v) \rho(u, v) du dv \\ &+ 2\gamma_t(0) \int_{\lambda_0}^{\lambda^*} dr \gamma_t(r) \lim_{\Delta \rightarrow 0^+} \frac{\alpha_x}{2i\pi} \left\{ W_y(r+i\Delta) - W_y(r-i\Delta) \right\} \end{aligned} \quad (6.143)$$

But because we don't have access to α_x nor W_y directly, we can use the full form:

$$\begin{aligned} \mathcal{R}_{x,y} \{ \gamma_t(x) \gamma_t(y) W(x, y) \} &= \gamma_t(0)^2 \alpha_{xy} + \iint_{[\lambda_0, \lambda^*]^2} \gamma_t(u) \gamma_t(v) \rho(u, v) du dv \\ &+ 2\gamma_t(0) \int_{\lambda_0}^{\lambda^*} dr \gamma_t(r) \lim_{\Delta \rightarrow 0^+} \lim_{\epsilon \rightarrow 0^+} \frac{-i\epsilon}{2i\pi} \left\{ W(i\epsilon, r+i\Delta) - W(i\epsilon, r-i\Delta) \right\} \end{aligned} \quad (6.144)$$

This comes from the fact that for $\epsilon \rightarrow 0$ we have: $W(i\epsilon, r+i\Delta) \sim \frac{\alpha_x}{-i\epsilon} W_y(r+i\Delta)$. Because we expect a real result, we ought to have numerically:

$$\begin{aligned} \mathcal{R}_{x,y} \{ \gamma_t(x) \gamma_t(y) W(x, y) \} &= \gamma_t(0)^2 \alpha_{xy} + \iint_{[\lambda_0, \lambda^*]^2} \gamma_t(u) \gamma_t(v) \rho(u, v) du dv \\ &+ \gamma_t(0) \int_{\lambda_0}^{\lambda^*} dr \gamma_t(r) \lim_{\Delta \rightarrow 0^+} \lim_{\epsilon \rightarrow 0^+} \frac{\epsilon}{\pi} \operatorname{Re} \left\{ W(i\epsilon, r-i\Delta) - W(i\epsilon, r+i\Delta) \right\} \end{aligned} \quad (6.145)$$

1-variable distributions in 2-variables functions Finally, it can happen that the 2-variables functions $W(x, y)$ actually generates a distribution $\rho(u, v) = \rho_c(u, v) + \mu(u)\delta(v-u)$ which may be the sum of a continuous measure $\rho_c(u, v)$ as described above, and another measure $\mu(u)\delta(v-u) = \delta(u-v)\mu(v)$.

6.E.3 Additional heatmaps

We provide additional heatmaps that complement those of Sect. 6.3. Notice that all the heat-maps are always derived from a 3D mesh comprising 30×100 points as in Fig. 6.E.7.

Instead of fixing λ , we can rescale it and fix $\delta = c\lambda$. As we have seen, the λ parameter seems to

affect the length of the time scale on the first plateau. Rescaling it as seen in Fig. 6.E.1, the interpolation threshold time scale becomes constant in the over-parametrized regime at fixed δ , and the results are consistent with what is observed empirically in Nakkiran et al. (2020a).

We notice also that under the configuration in Fig. 6.E.2 where $r = 0$ (the noise of the second layer vanishes), the second plateau seems to vanish with the test error.

One of the effects of a large λ is that it removes the double descent on the test error, which is consistent with the description in Mei and Montanari (2019). Another effect is that it seems to add an additional "two-stage decrease" in the training error as can be seen in Fig. 6.E.4 and also in the experiments in Figs. 6.E.10, 6.E.11.

Note that the previous figures are performed for the activation function $\sigma(x) = \text{Relu}(x) - \frac{1}{\sqrt{2\pi}}$ while Figs. 6.E.5 and 6.E.6 are displayed other activation functions, $\sigma(x) = \tanh(x)$ and $\sigma(x) = \tanh(5x)$. We can see that the epoch-wise structures are more marked when the slope of the activation function is bigger in the second case.

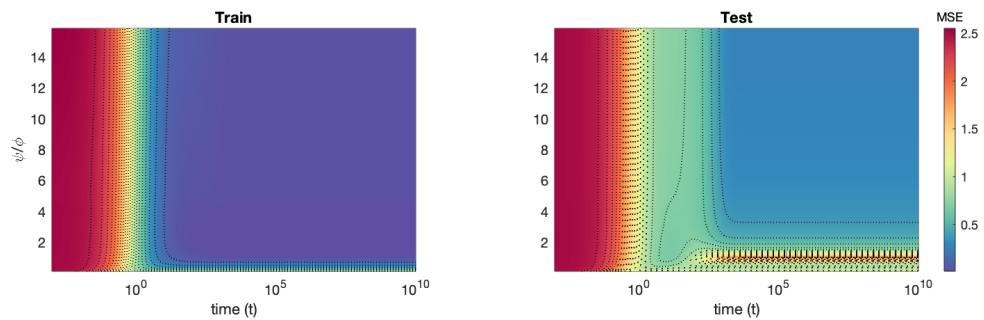


Figure 6.E.1: Analytical training error and test error evolution at fixed δ with parameters $(\mu, \nu, \phi, r, s, \delta) = (0.5, 0.3, 3, 2., 0.4, 0.001)$

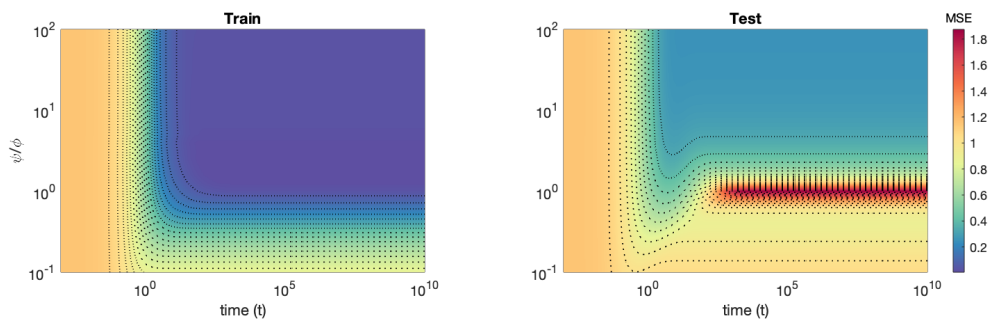


Figure 6.E.2: Analytical training error and test error evolution with parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 3, 0, 0.4, 0.001)$

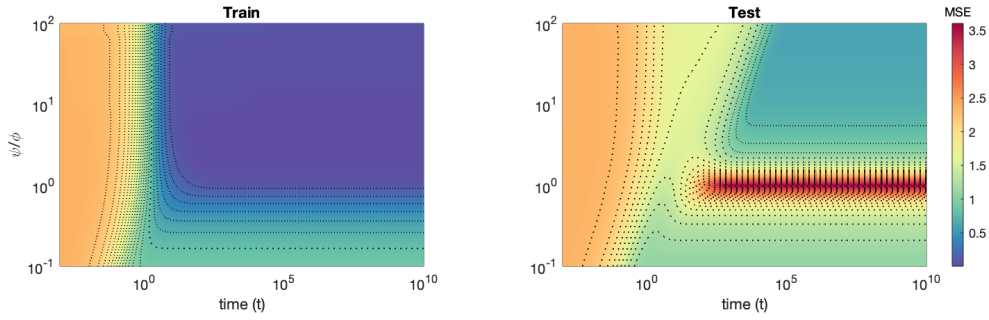


Figure 6.E.3: Analytical training error and test error evolution with parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 0.5, 2, 0.1, 0.003)$

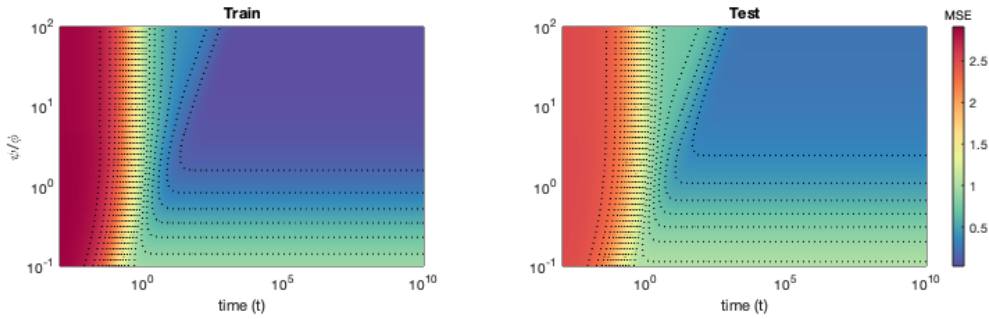


Figure 6.E.4: Analytical training error and test error evolution with parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 3, 0, 0.4, 0.1)$

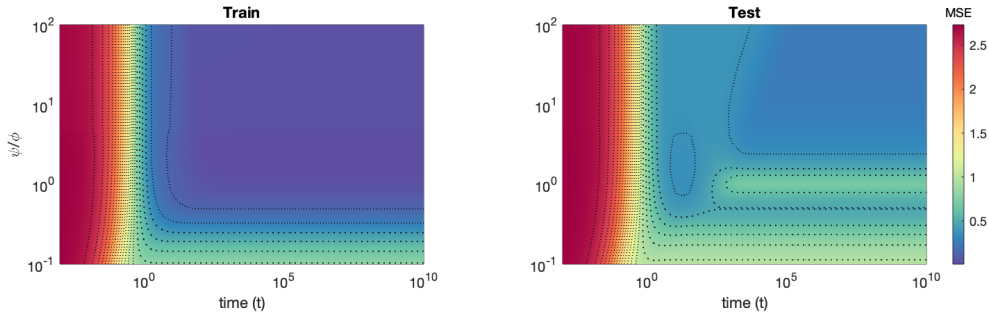


Figure 6.E.5: Analytical training error and test error evolution with parameters corresponding to $\sigma(x) = \tanh(x)$ with $(\mu, \nu, \phi, r, s, \lambda) = (0.61, 0.15, 3, 0, 0.4, 0.001)$

6.E.4 Comparison with experimental simulations

We have already shown on figure 6.3.1 in Sect. 6.3 that the analytical formulas for the training and generalization errors match the experimental curves in the limit of $t \rightarrow +\infty$. Here we provide additional evidence that this is also the case for the whole time-evolution in Figs. 6.E.8 and 6.E.9 as the dimension d increases.

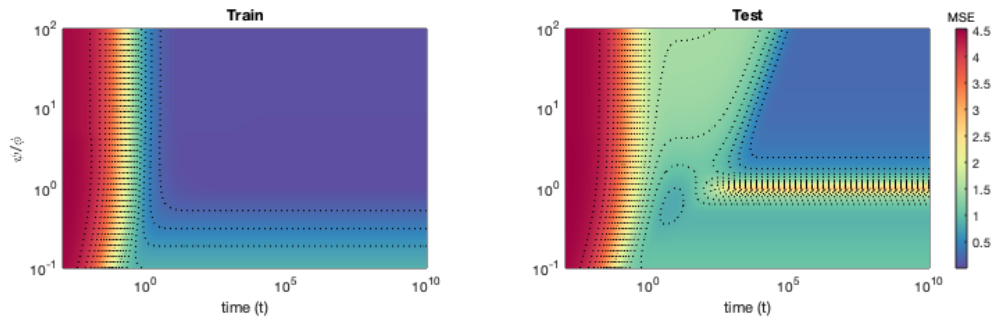


Figure 6.E.6: Analytical training error and test error evolution with parameters corresponding to $\sigma(x) = \tanh(5x)$ with $(\mu, \nu, \phi, r, s, \lambda) = (0.79, 0.47, 3, 2, 0.4, 0.001)$

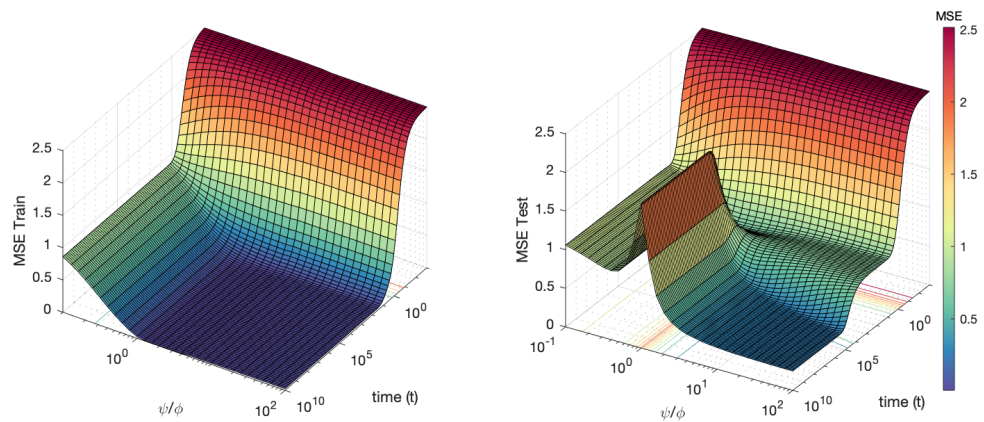


Figure 6.E.7: Analytical training error with parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 3, 2., 0.4, 0.001)$

In 6.E.10, 6.E.11, we can see that the epoch-wise descent structures of the training error and test error can be captured correctly experimentally for long time. Note that we have taken $d = 100$ small enough to be able to run these experiments for such a long timescale.

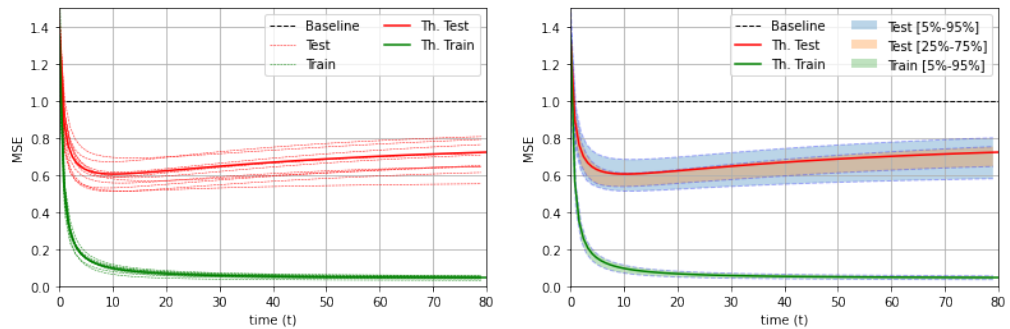


Figure 6.E.8: Analytical training error and test error profile with parameters $(\mu, \nu, \phi, \psi, r, s, \lambda) = (0.5, 0.3014, 1.4, 1.8, 1.0, 0, 0.01)$ compared to 10 experimental runs ($\sigma = \text{Relu} - \frac{1}{\sqrt{2\pi}}$) with $d = 200$ and $dt = 0.01$

Chapter 6. The Random feature model

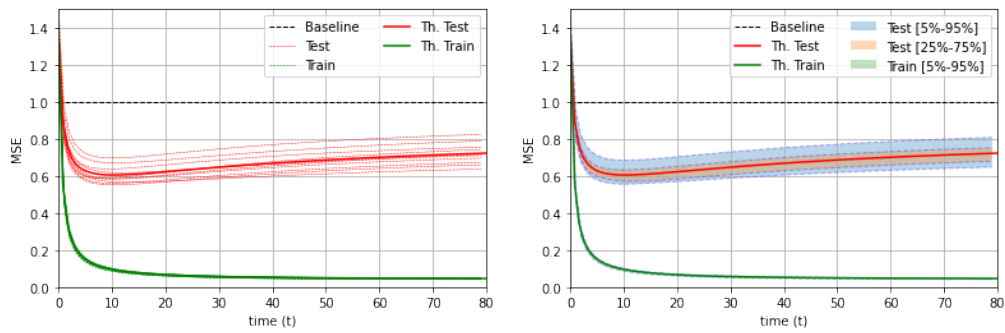


Figure 6.E.9: Analytical training error and test error profile with parameters $(\mu, \nu, \phi, \psi, r, s, \lambda) = (0.5, 0.3014, 1.4, 1.8, 1.0, 0, 0.01)$ compared to 10 experimental runs ($\sigma = \text{Relu} - \frac{1}{\sqrt{2\pi}}$) with $d = 1000$ and $dt = 0.01$

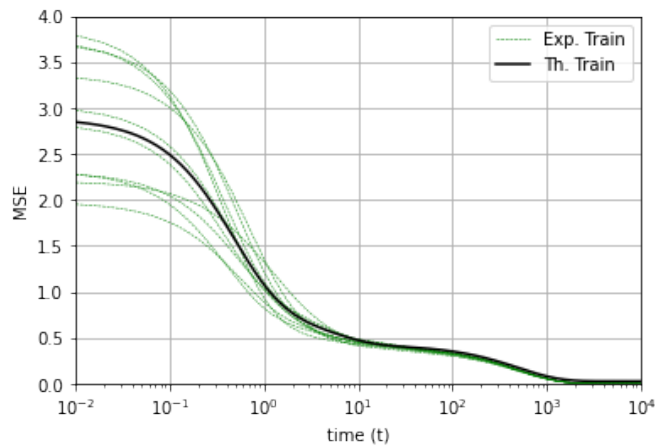


Figure 6.E.10: Analytical training error with parameters $(\mu, \nu, \phi, \psi, r, s, \lambda) = (0.5, 0.3, 300, 3, 2, 0.4, 0.1)$ compared to 10 experimental runs ($\sigma = \text{Relu} - \frac{1}{\sqrt{2\pi}}$) with $d = 100$ and $dt = 0.01$

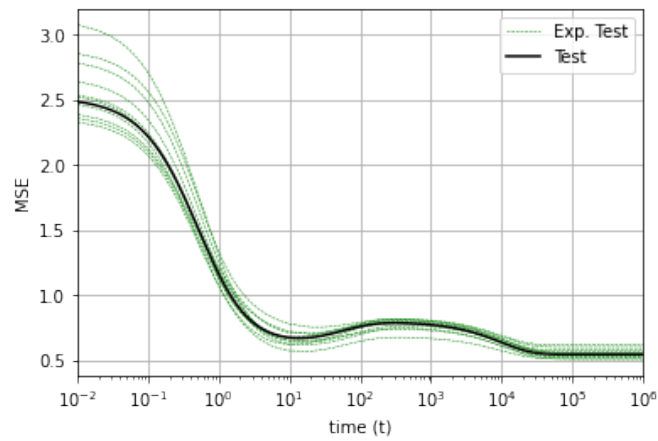


Figure 6.E.11: Analytical test error with parameters $(\mu, \nu, \phi, \psi, r, s, \lambda) = (0.5, 0.3, 6, 3, 2, 0.4, 0.0001)$ compared to 10 experimental runs with $d = 100$ and $dt = 0.01$ for $0 \leq t \leq 10^4$ and $dt = 0.1$ for $10^4 \leq t \leq 10^6$

Beyond the linear setting with matrix completion **Part III**

7 The rank-one model: a non-convex setting

This chapter is based on the work (Bodin and Macris, 2021b) which investigates model 1.4. More precisely, we consider a rank-one symmetric matrix corrupted by additive noise. The rank-one matrix is formed by an n -component unknown vector on the sphere of radius \sqrt{n} , and we consider the problem of estimating this vector from the corrupted matrix in the high dimensional limit of n large, by gradient descent for a quadratic cost function on the sphere. Explicit formulas for the whole time evolution of the overlap between the estimator and unknown vector, as well as the cost, are rigorously derived. In the long time limit we recover the well known spectral phase transition, as a function of the signal-to-noise ratio. The explicit formulas also allow to point out interesting transient features of the time evolution. Our analysis technique is based on recent progress in random matrix theory and uses *local versions* of the semi-circle law.

7.1 Introduction

Gradient descent dynamic is at the root of machine learning methods, and in particular, its stochastic version augmented by various ad-hoc methods, has been very successful at finding "good" minima of cost functions Lecun et al. (1998). However, rigorous detailed results on the full time evolution of the dynamics are scarce even for simple models and usual gradient descent. In this contribution, we show how to completely solve for the whole time evolution for a simple paradigm of non-linear estimation; the problem of estimating a rank-one spike embedded in noise.

Let $\theta^* \in \mathbb{S}^{n-1}(\sqrt{n})$ a *hidden* vector on the $n-1$ dimensional sphere of radius \sqrt{n} , i.e., $\theta^* = (\theta_1^*, \dots, \theta_n^*)^T$ and $\|\theta^*\|_2^2 = n$. We consider the *data* matrix Y with elements $Y = \theta^* \theta^{*T} + \sqrt{\frac{n}{\lambda}} \xi$ where $\lambda > 0$ is the signal-to-noise parameter and $\xi = (\xi_{i,j})_{1 \leq i, j \leq n}$ a symmetric random noise matrix with i.i.d $\xi_{i,j}$ for $i \leq j$. The goal is to recover θ^* given that Y and λ are known. This model is usually considered for a gaussian noise symmetric matrix $\xi_{ij} \sim \mathcal{N}(0, 1)$, $i \leq j$, and is variously called the noisy rank-one matrix estimation problem or the spiked Wigner model. In this chapter, all the results hold under the general assumption that $\mathbb{E}\xi_{ij} = 0$, $\mathbb{E}\xi_{ij}^2 = 1 + O(\delta_{ij})$

and for all integers p we have $\mathbb{E}|\xi_{ij}|^p$ finite.¹

We consider the cost function ($\|\cdot\|_F$ the Frobenius norm)

$$\mathcal{H}(\theta) = \frac{1}{2n^2} \|Y - \theta\theta^T\|_F^2 - \frac{1}{2n^2} \|Y - \theta^*\theta^{*T}\|_F^2 \quad (7.1)$$

(normalized so that, $\mathcal{H}(\theta^*) = 0$, and at the same time, the limit $n \rightarrow +\infty$ is well defined) and want to characterize the time evolution of the estimator for θ^* provided by gradient descent dynamics on the sphere. In gradient descent, an initial (deterministic) vector $\theta_0 \in \mathbb{S}^{n-1}(\sqrt{n})$ is updated through the autonomous ordinary differential equation

$$\frac{d\theta_t}{dt} = -\eta(\nabla_{\theta}\mathcal{H}(\theta_t) - \frac{\theta_t}{n}\langle\theta_t, \nabla_{\theta}\mathcal{H}(\theta_t)\rangle) \quad (7.2)$$

where $\eta \in \mathbb{R}_+^*$ is a learning rate. The second term on the right hand side enforces the constraint $\theta_t \in \mathbb{S}^{n-1}(\sqrt{n})$ at all times (see Appendix 7.E). The main quantities of interest to be computed are the time evolutions of the cost $\mathcal{H}(\theta_t)$ and overlap $q(t) = n^{-1}\langle\theta^*, \theta_t\rangle$ in the high-dimensional limit $n \rightarrow +\infty$. We note that the overlap is equivalent to the mean-square-error $n^{-1}\|\theta_t - \theta^*\|_2^2 = 2(1 - \frac{\langle\theta^*, \theta_t\rangle}{n})$.

Contribution: We compute the full time evolution of the cost and overlap in the scaling limit $\lim_{n \rightarrow +\infty} \mathcal{H}(\theta_{t=\tau n/\eta})$ and $\lim_{n \rightarrow +\infty} \frac{\langle\theta^*, \theta_{t=\tau n/\eta}\rangle}{n}$ for all $\tau > 0$. Explicit formulas are expressed solely in terms of a modified Bessel function of first order in theorems 7.1 and 7.2 (Section 7.2). The formulas allow to explore the asymptotic behavior as $\tau \rightarrow +\infty$, as well as transient behavior by computing one and two dimensional integrals numerically (Section 7.2). In the long time limit we recover (analytically) as expected the phase transition at $\lambda = 1$ with a limiting value of the overlap equal to $\text{sign}(\langle\theta^*, \theta_0\rangle)\sqrt{1 - 1/\lambda} \mathbb{1}(\lambda > 1)$. This is the well known BBP-like phase transition found in the spectral method Pécché (2004); Féral and Pécché (2006); Baik et al. (2005b). The transient behavior also exhibits interesting features. For example, depending on the magnitude of the initial overlap $n^{-1}\langle\theta^*, \theta_0\rangle$ and $\lambda > 1$ for intermediate times we find that the overlap may display a maximum and then decrease to its limiting value. Such results may therefore give guidelines for applying early stopping during gradient descent to get a better estimate of the signal. We note that in the asymptotic limit of large n we require an initial overlap which is bounded away from zero uniformly in n . There are interesting situations where the signal θ^* has some structure and this is not an unnatural situation. These points are further discussed in Section 7.2.2.

On the technical side the analysis is based on a set of integro-differential equations (derived in Section 7.3) satisfied by matrix elements of the resolvent of the noise matrix $\langle\theta^*, (\frac{1}{\sqrt{n}}\xi - z)^{-1}\theta_t\rangle$ and $\langle\theta_t, (\frac{1}{\sqrt{n}}\xi - z)^{-1}\theta_t\rangle$, $z \in \mathbb{C} \setminus \mathbb{R}$. These quantities concentrate with respect to the probability law of the noise matrix as $n \rightarrow +\infty$ (for deterministic θ^* and θ_0). The main steps to prove

¹The notation $O(\delta_{ij})$ means that the second moment of off-diagonal elements is 1 but the variance of diagonal elements can be different. For example $\xi_{ij} \sim \mathcal{N}(0, 1)$, $i < j$, and $\xi_{ii} \sim \mathcal{N}(0, 2)$, corresponds to Wigner's Gaussian Orthogonal Ensemble. We refer to the general case as the generalized Wigner ensemble.

concentration are explained in Section 7.4. They combine concentration properties of the matrix elements of the resolvents with an adaptation of Gronwall type arguments to the integro-differential equations. Concentration of matrix elements of resolvents of random matrices amount to study the spectrum on a *local* scales. Such results are only a decade old in random matrix theory and go under the name of *local* semi-circle laws Erdős et al. (2008); Bloemendal et al. (2014); Benaych-Georges and Knowles (2016b). They have found many applications and here we provide one more. In Section 7.5 we present an exact analysis of the integro-differential equations and deduce the formulas for the time evolution of the overlap and cost.

Related Work: The statistical limits of the symmetric as well as non-symmetric spiked Wigner model have been elucidated in great detail in the Bayesian framework in a series of works Korada and Macris (2009); Barbier et al. (2016); Lelarge and Miolane (2018); Miolane (2017) where expressions for mutual information (in the form of low-dimensional variational problems) and minimum-mean-square-error are rigorously computed. This analysis has also been carried on for estimation of low-rank tensors corrupted by additive gaussian noise Lesieur et al. (2017b); Barbier et al. (2017); Perry et al. (2020). The dynamical behaviour under Approximate Message Passing (AMP) has also been investigated in detail and, depending on the exact model and prior, large computational-to-statistical gaps are found Barbier et al. (2016); Lesieur et al. (2017b). We note that these settings are different from the one of the present chapter in that θ^* as well as θ_0 are random. When the prior of the spike is unbiased with zero mean (for example uniform on the sphere or binary) an initial strictly positive overlap (uniformly in n), is necessary to start the AMP algorithm, much like gradient descent, and hence the initial condition cannot be chosen at random. In this connection, the behaviour of AMP under spectral initialization has been derived in the work Montanari and Venkataramanan (2021). We note that spectral initialization is not an option for us because it yields a stationary point of gradient flow (see appendix 7.F for a justification).

Starting with the early work of Burer and Monteiro (2005, 2003) the efficiency of gradient descent techniques has been uncovered in recent years for a host of low-rank matrix recovery modern problems, e.g., in PCA, low-rank matrix factorization, matrix completion, phase retrieval, phase synchronization, Ge et al. (2017a); Bhojanapalli et al. (2016); Ge et al. (2017b); De Sa et al. (2015); Park et al. (2017); Ling et al. (2019); Bandeira et al. (2016). We also refer to Chi et al. (2019a) for a general review and references. Underpinning the efficiency of gradient descent in such non-convex problems, is a high-level result Lee et al. (2016), stating that when the landscape satisfies a *strict saddle property* (i.e., critical points are strict saddles or minima) gradient descent with sufficiently small discrete step size and random initialization will converge almost surely to a minimum Lee et al. (2016). The spiked Wigner models falls in this category at least for n finite: critical points of the cost function on the sphere $\mathcal{S}^{n-1}(\sqrt{n})$ are the eigenvectors of Y and it is easy to show that almost surely (with respect to the noise matrix ξ) the largest eigenvector is a minimum while all the other ones are strict saddles. Therefore gradient descent will converge for small enough step size to the largest eigenvector and the spectral properties of Y imply that for $\lambda > 1$ with high probability this largest eigenvector has

an overlap with θ^* close to $\pm\sqrt{1-1/\lambda}$ (these known facts are briefly reviewed in Appendix 7.F).

While these approaches are able to provide guarantees and convergence rates of gradient descent and variants thereof, they do not provide the full time-evolution and do not say much about intermediate or transient times. This is what we achieve in this chapter for the admittedly simple Wigner spiked models. We believe that the techniques used here can be extended to other problems of interest in regression and learning. Recently, pure gradient descent was studied for the much harder optimization of the cost of a mixed matrix-tensor inference problem Sarao Mannelli et al. (2019); Mannelli et al. (2019) (see also Sarao Mannelli et al. (2020) for Langevin dynamics) and it was shown how the structure of saddles and minima determines the phase transition thresholds. This is based on a set of very sophisticated integro-differential CSHCK equations Crisanti et al. (1993); Cugliandolo and Kurchan (1993) with a long history in the framework of Langevin dynamics on spin-glass landscapes in statistical physics. While these derivation of the CSHCK equations for the inference problem are non-rigorous and their solution entirely numerical, they contain in principle the whole time evolution of the system (in the context of spin-glasses the formalism has been made rigorous Ben Arous et al. (2004)). The integro-differential equations and methods of the present chapter are entirely *different* (and involve different objects) even when specializing to the matrix case. We note that for the mixed matrix-tensor case the CSHCK formalism is quite intractable, but nevertheless in the pure matrix case it should be possible to retrieve our final analytical solution as (partly) done in Cugliandolo and Dean (1995) for the spherical spin-glass. We briefly comment on possible extensions of our formalism in the conclusion.

Organization of this chapter: The main theorems and illustrations of analytical formulas for the whole time-evolution of the overlap and cost are presented in Section 7.2. The heart of the method presented here is contained in sections 7.3 (derivation of integro-differential equations), 7.4 (local semi-circle laws and concentration of solutions), 7.5 (analytical solution of integro-differential equations). Appendices contain proofs, of intermediate results and technical material.

In the rest of the chapter, it is understood that the noise matrix ξ satisfies: (i) $\mathbb{E}\xi_{ij} = 0$, (ii) $\mathbb{E}\xi_{ij}^2 = 1 + O(\delta_{ij})$, (iii) $\mathbb{E}|\xi_{ij}|^p$ finite for all $p \in \mathbb{N}$. We use the notations $H = n^{-1/2}\xi$, \mathbb{P} for its probability law, and $X_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} X$ for convergence in probability, i.e., $\lim_{n \rightarrow +\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$ for any $\epsilon > 0$.

7.2 Analytical solutions and illustrations

We solve gradient descent dynamics (7.2) in the scaling limit $t = \tau n/\eta$, with fixed $\tau > 0$ and $n \rightarrow +\infty$. The main quantities that we determine in the scaling limit are the overlap $q(\tau) = \frac{1}{n} \langle \theta^*, \theta_{n\tau/\eta} \rangle$ and the cost $\mathcal{H}(\theta_{n\tau/\eta})$. We remark that the overlap is directly linked to the mean-square error $n^{-1} \|\theta^* - \theta_{n\tau/\eta}\|^2 = 2(1 - q(\tau))$.

The initial condition θ_0 is fixed such that $q(0) = \alpha$ where $\alpha \in [-1, 1]$ is independent of n . It will become clear that: (i) If θ_t is a solution with initial condition $q(0) = \alpha$ then $-\theta_t$ is a solution with $q(0) = -\alpha$; (ii) For $\alpha = 0$ the solution remains trivial $q(\tau) = 0$. Therefore the reader can keep in mind that $\alpha > 0$ (all the analysis is valid for any α though).

7.2.1 Main results

The solution of the gradient descent dynamics can be entirely expressed thanks to a scaled moment generating function of Wigner's semi-circle law $\mu_{\text{sc}}(s) = \frac{1}{2\pi} \sqrt{4-s^2} \chi_{[-2,2]}(s)$,

$$M_\lambda(\tau) = \int_{-\infty}^{\infty} ds \mu_{\text{sc}}(s) e^{s \frac{\tau}{\sqrt{\lambda}}} \quad (7.3)$$

Setting $s = 2 \cos \theta$ we have $M_\lambda(\tau) = 2 \int_0^\pi \frac{d\theta}{\pi} (\sin \theta)^2 e^{\frac{2\tau}{\sqrt{\lambda}} \cos \theta}$. Integration by parts then shows that $M_\lambda(\tau) = \frac{\sqrt{\lambda}}{\tau} I_1(\frac{2\tau}{\sqrt{\lambda}})$ where $I_1(x) = \int_0^\pi \frac{d\theta}{\pi} (\cos \theta) e^{x \cos \theta}$ is a modified Bessel function of the first kind.

Theorem 7.1 (Time evolution of the overlap). *Let $\theta_0 \in \mathbb{S}^{n-1}(\sqrt{n})$ an initial condition such that $q(0) = \frac{1}{n} \langle \theta^*, \theta_0 \rangle = \alpha$ for a fixed $\alpha \in [-1, +1]$. The overlap converges in probability to a deterministic limit:*

$$q(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \bar{q}(\tau) = \frac{\hat{q}(\tau)}{\sqrt{\hat{p}(\tau)}} \quad (7.4)$$

where

$$\hat{q}(\tau) = \alpha e^{(1+\frac{1}{\lambda})\tau} \left[1 - \frac{1}{\lambda} \int_0^\tau ds e^{-(1+\frac{1}{\lambda})s} M_\lambda(s) \right] \quad (7.5)$$

and

$$\hat{p}(\tau) = M_\lambda(2\tau) + 2\alpha \int_0^\tau ds \hat{q}(s) M_\lambda(2\tau - s) + \int_0^\tau \int_0^\tau dudv \hat{q}(u) \hat{q}(v) M_\lambda(2\tau - u - v). \quad (7.6)$$

Theorem 7.2 (Time evolution of the cost). *Under the same conditions as in theorem 7.1 the cost converges to a deterministic limit $\mathcal{H}(\theta_{\tau n/\eta}) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 1 - \frac{1}{2} \frac{d}{d\tau} \{ \ln \hat{p}(\tau) \}$.*

Using asymptotic properties of the Bessel function and the Laplace method it is possible to calculate the asymptotics of the integrals in (7.5) and (7.6) for $\tau \rightarrow +\infty$. We find for the overlap $\lim_{\tau \rightarrow \infty} \bar{q}(\tau) = \text{sign}(\alpha) \sqrt{1 - \lambda^{-1}} \mathbb{1}(\lambda \geq 1)$. The overlap displays the well known phase transition at $\lambda = 1$ also predicted by the spectral method. The asymptotic values can also be derived independently from theorem 7.1 by directly looking at the stationary equation $\nabla_\theta \mathcal{H}(\theta_\infty) - \frac{\theta_\infty}{n} \langle \theta_\infty, \nabla_\theta \mathcal{H}(\theta_\infty) \rangle = 0$. This is discussed in Appendix 7.G for completeness. It is also possible to go one step further in the asymptotics to argue that at the transition $\lambda = 1$ the power law behavior holds $\bar{q}(\tau) \sim (\frac{2}{\pi\tau})^{1/4}$ (see Appendix 7.I).

Besides the transition at $\lambda = 1$, for finite λ , a detailed analysis of the equations of theorem 7.1

which are described also in Appendix 7.I allows to derive the first order asymptotic behavior of \bar{q} for large τ . These tedious calculations are carried out analytically in detail and checked numerically. Specifically, in the regime $1 < \lambda < +\infty$ we find

$$\bar{q}(\tau) - \text{sign}(\alpha) \sqrt{1 - \frac{1}{\lambda}} \sim \frac{\text{sign}(\alpha)}{2\sqrt{\pi}\lambda^{\frac{1}{4}} \sqrt{1 - \frac{1}{\lambda}} \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2} \tau^{-\frac{3}{2}} e^{-(1 - \frac{1}{\sqrt{\lambda}})^2 \tau} \quad (7.7)$$

As for $0 < \lambda < 1$, we retrieve a power law behavior:

$$\bar{q}(\tau) \sim \frac{\alpha \left(\frac{2}{\pi}\right)^{\frac{1}{4}}}{\lambda^{\frac{5}{8}} \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \sqrt{1 - \alpha^2 + \frac{\alpha^2}{\lambda \left(\frac{1}{\sqrt{\lambda}} - 1\right)^2}} \tau^{-\frac{3}{4}} \quad (7.8)$$

The noise-less regime $\lambda = +\infty$ is an elementary case for which the overlap can be obtained very simply. Taking the inner product of (7.2) with θ^* we find the differential equation (for $t = \tau n/\eta$) $\frac{dq(\tau)}{d\tau} = q(\tau) - q(\tau)^3$, $q(0) = \alpha$, which has the solution $q(\tau) = \alpha(\alpha^2 + (1 - \alpha^2)e^{-2\tau})^{-1/2}$. As we will see, in the noisy case there is no closed form first order ODE for $q(\tau)$ and we must solve integro-differential equations for suitable generating functions (or an infinite hierarchy of coupled differential equations for generalized overlaps). As a sanity check, we can verify that theorem 7.1 leads to the same expression when $\lambda \rightarrow +\infty$. Explicitly, we find $\lim_{\lambda \rightarrow +\infty} \hat{q}(\tau) = \alpha e^\tau$ and $\lim_{\lambda \rightarrow +\infty} \hat{p}(\tau) = 1 - \alpha^2 + \alpha^2 e^{2\tau}$ which implies the noiseless expression for the overlap.

7.2.2 Discussion and numerical experiments

Theorems 7.1 and 7.2 provide theoretical predictions for the full time evolution of the overlap and risk in the high dimensional limit $n \rightarrow +\infty$. In this section (and Appendix 7.J) we briefly illustrate and discuss this time evolution. Moreover in Appendix 7.J we also compare the theoretical predictions with simulations of discrete step size gradient descent for runs over multiple samples of ξ .

Choice of the initial condition

Given $\theta^* \in \mathbb{S}^{n-1}(\sqrt{n})$ if we choose the initial condition θ_0 uniformly at random we expect α a random variable of zero mean and standard deviation $n^{-1/2}$. To analyze this case one should deal with finite n corrections to the dynamics which is beyond the scope of this thesis. Numerical plots of our formulas (fig 1a, 5a) show when $\alpha \rightarrow 0$ gradient flow kicks-off at larger and larger times; this suggests that if $\alpha \sim n^{-1/2}$ gradient flow kicks-off once a large enough time-scale has elapsed. Our analysis is presumably valid beyond this time-scale (we do not have a proof of this claim), but estimating this time-scale is open. As mentioned in the introduction AMP suffers from similar issues. However, there are many interesting situations where the signal has some *structure* which is partially known, and it is then very natural to have $\alpha > 0$ (uniformly in n). For example signals which may have a non-zero empirical expectation

$\rho > 0$, for instance with components distributed as $\text{Ber}(\frac{\rho+1}{2})$ in $\{-1, 1\}$. Then we can take the initial all-one vector $\theta_0 = \mathbb{1}_n$, and thus $\alpha = \rho > 0$ which naturally kicks-off the gradient flow, and our analysis applies.

Time evolution of the overlap

Figure 7.2.1 shows the theoretical overlap at all times $\tau \in \mathbb{R}^+$ for two initial conditions $\alpha = 0.1$ and $\alpha = 0.5$ and any signal-to-noise ratio λ . Let us say a few words about the *transient behaviors* that are observed.

On the one hand, the closer α gets to 0, the longer it takes for the gradient descent to "kick-in": the overlap stays longer close to 0 before reaching its asymptotic behavior. An additional example for $\alpha = 0.01$ illustrates this fact in Appendix 7.J. On the other hand, we clearly see that when the initial overlap α is not too close to 0, the time evolution is *not* monotone even for $\lambda > 1$, and a specific bump is reached at early times where the overlap reaches a maximum before dropping down to its limit. In fact this is clearly suggested by (7.7) for $\alpha < 1$. This can be seen in particular in the case $\alpha = 0.5$ in Figure 7.2.1 (b). This suggests that in practice, in such situations, it is worth using early-stopping techniques to optimize the estimation of the signal. The increase of the overlap above the spectral estimate for finite times is a consequence of the side information $\alpha > 0$ that standard PCA does not have. As a side note we mention that in the Bayesian setting with known prior the information theoretic overlap is at least as good or better than PCA (for a $\mathcal{N}(0, 1)$ prior they are equal).

In the case $\lambda = 1$ one can show that $\hat{q}(\tau) = \alpha (I_0(2\tau) + I_1(2\tau))$ (with modified Bessel functions of the first kind) and it is numerically much easier to evaluate the asymptotic behavior of $q(\tau)$. The calculation yields $q(\tau) \sim (\frac{2}{\pi\tau})^{\frac{1}{4}}$ (see Appendix 7.I). Furthermore plotting a family of curves with $\lambda = 1$ and $\alpha \in (0, 1)$ in Figure 7.2.2, it appears that this asymptote also seems to act as an *upper-bound*.

Time evolution of the cost

We also have predictions for the evolution of cost at any time for any values of (α, λ) . This is illustrated in Figure 7.2.3. As seen in the analysis of Section 7.A, Equ. (7.41) the cost has two additive contributions basically interpreted as $q(\tau)^2$ and $p_1(\tau) = n^{-1} \langle \theta_\tau, H\theta_\tau \rangle$. The second contribution equals $n^{-1} \text{Tr} H\theta_\tau \theta_\tau^T$ can be interpreted as a similarity measure of the reconstructed matrix $\theta_\tau \theta_\tau^T$ and the noise matrix H , and is thus a "proxy" for assessing over-fitting in this particular setting. Interestingly, in the depicted example where $\lambda = 2, \alpha = 0.1$, $p_1(\tau)$ is shown to decrease the risk at early stages at a fast rate, until it slightly "heals" for $\tau \geq 3$. Conversely, when $\alpha = 0.5$, we see $p_1(\tau)$ does not decrease as much in early stages, and the healing phenomenon does not occur. At the same time, as observed on 7.2.1 (b) $q(\tau)$ is not monotonous: it increases at early stages and decreases down to its limiting value later.

Chapter 7. The rank-one model: a non-convex setting

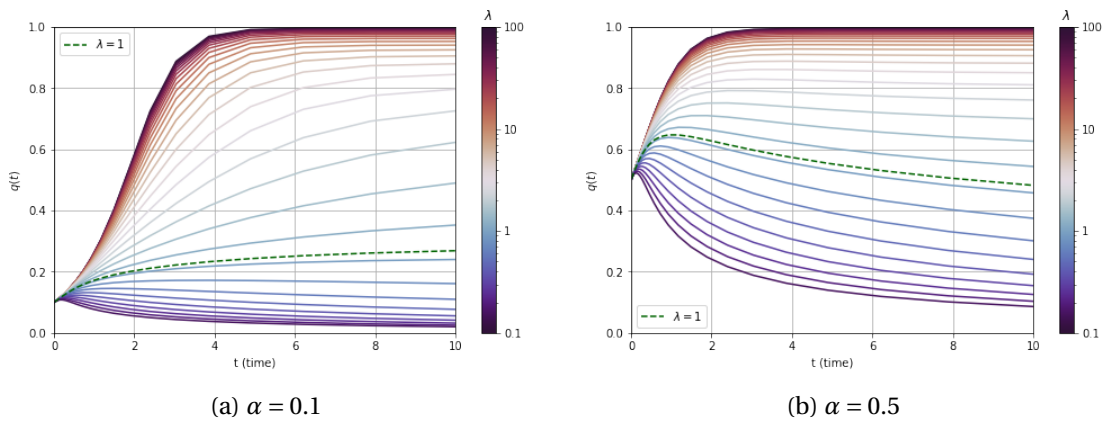


Figure 7.2.1: Overlap as a function of time according to theorem 7.1 for two initial conditions and different signal-to-noise ratios. Thick dotted line corresponds to $\lambda = 1$ and tends to zero slowly as $(2/\pi\tau)^{1/4}$. For $\lambda < 1$ the curves tend to zero and for $\lambda > 1$ they tend to $\sqrt{1 - 1/\lambda}$.

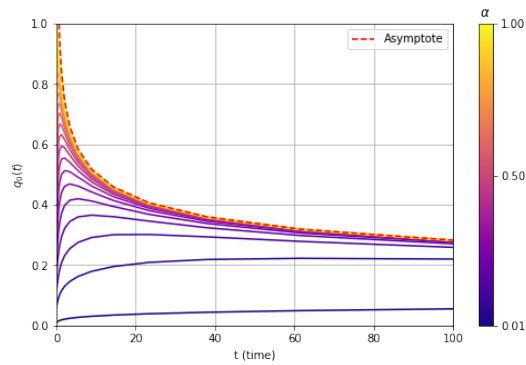


Figure 7.2.2: Overlap comparison for $\lambda = 1$ with a range of values for α

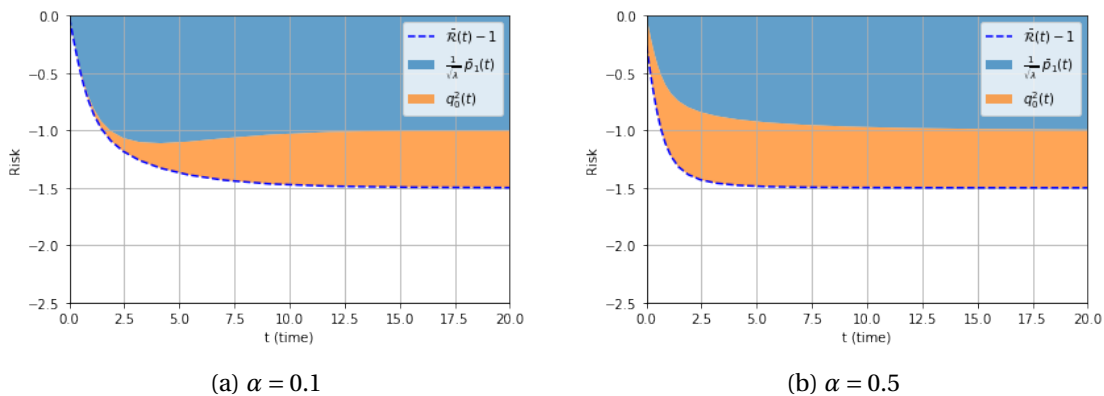


Figure 7.2.3: Cost evolution for $\lambda = 5$

7.3 Integro-differential equations

We study gradient descent in a regime where $t = \tau n/\eta$, $n \rightarrow +\infty$, with τ fixed. Abusing slightly notation we set $\theta_{\tau n/\eta} \rightarrow \theta_\tau$ so that equation (7.2) reads

$$\frac{d\theta_\tau}{d\tau} = -n\nabla_{\theta} \mathcal{H}(\theta_\tau) + \theta_\tau \langle \theta_\tau, \nabla_{\theta} \mathcal{H}(\theta_\tau) \rangle = \frac{1}{n^2} Y \theta_\tau - \frac{1}{n^2} \langle \theta_\tau, Y \theta_\tau \rangle \theta_\tau \quad (7.9)$$

We define $H = n^{-1/2} \xi$ the suitably normalized noise matrix. Besides the basic overlap $q(\tau) = \frac{1}{n} \langle \theta^*, \theta_\tau \rangle$, another one also plays an important role, namely $p_1(\tau) = \frac{1}{n} \langle \theta_\tau, H \theta_\tau \rangle$.

Using $Y = \theta^* \theta^{*T} + \frac{n}{\sqrt{\lambda}} H$ we find

$$\frac{d\theta_\tau}{d\tau} = q(\tau) \theta^* + \frac{1}{\sqrt{\lambda}} H \theta_\tau - \left(q(\tau)^2 + \frac{p_1(\tau)}{\sqrt{\lambda}} \right) \theta_\tau \quad (7.10)$$

It is not possible to write down a closed set of equations that involve only $q(\tau)$ and $p_1(\tau)$, but only for a hierarchy of such objects, or for their generating functions. We now introduce these generating functions and then give the closed set of equations which they satisfy.

The $n \times n$ matrix $H = n^{-1/2} \xi$ is drawn with the probability law \mathbb{P} . Fix any small $\delta > 0$ and let \mathcal{S}_δ^n the set of realizations of H such that all eigenvalues fall in an interval $I_\delta = [-2 - \delta, 2 + \delta]$. Then $\mathbb{P}(\mathcal{S}_\delta^n) \rightarrow 1$ as $n \rightarrow +\infty$ (see for example Erdős (2011)). In the rest of this section it is understood that $H \in \mathcal{S}_\delta^n$. In particular the resolvent matrix² $\mathcal{R}(z) = (H - zI)^{-1}$ is well defined for $z \in \mathbb{C} \setminus I_\delta$ if $H \in \mathcal{S}_\delta^n$.

For any contour $\mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$ with $\rho > 2 + \delta$ we can define three generating functions

$$Q_\tau(z) = \frac{1}{n} \langle \theta_\tau, \mathcal{R}(z) \theta^* \rangle, \quad P_\tau(z) = \frac{1}{n} \langle \theta_\tau, \mathcal{R}(z) \theta_\tau \rangle, \quad R(z) = \frac{1}{n} \langle \theta^*, \mathcal{R}(z) \theta^* \rangle. \quad (7.11)$$

From standard holomorphic functional calculus for matrices (see for example Dunford and Schwartz (1988)) we have

$$q(\tau) = - \oint_{\mathcal{C}} \frac{dz}{2\pi i} Q_\tau(z), \quad p_1(\tau) = - \oint_{\mathcal{C}} \frac{dz}{2\pi i} z P_\tau(z). \quad (7.12)$$

Note that these two overlaps are part of a hierarchy of overlaps $q_k(\tau) \equiv \frac{\langle \theta^*, H^k \theta_\tau \rangle}{n} = - \oint_{\mathcal{C}} \frac{dz}{2\pi i} z^k Q_\tau(z)$ and $p_k(\tau) \equiv \frac{\langle \theta_\tau, H^k \theta_\tau \rangle}{n} = - \oint_{\mathcal{C}} \frac{dz}{2\pi i} z^k P_\tau(z)$, $k \geq 1$, which can all be calculated by the methods of this chapter (note $q(\tau)$ corresponds to $k = 0$).

Proposition 7.1. *For any realization $H \in \mathcal{S}_\delta$ and any $z \in \mathbb{C} \setminus I_\delta$ the generating functions (7.11)*

²Here I is the identity $n \times n$ matrix and we will slightly abuse notation by omitting it and simply write $(H - z)^{-1}$.

satisfy the integro-differential equation

$$\begin{cases} \frac{d}{d\tau} Q_\tau(z) = q(\tau)R(z) + \frac{1}{\sqrt{\lambda}}(zQ_\tau(z) + q(\tau)) - \left(q^2(\tau) + \frac{1}{\sqrt{\lambda}}p_1(\tau)\right)Q_\tau(z) \\ \frac{1}{2}\frac{d}{d\tau} P_\tau(z) = q(\tau)Q_\tau(z) + \frac{1}{\sqrt{\lambda}}(zP_\tau(z) + 1) - \left(q^2(\tau) + \frac{1}{\sqrt{\lambda}}p_1(\tau)\right)P_\tau(z) \end{cases} \quad (7.13)$$

where $q(\tau) = -\oint_{\mathcal{C}} \frac{dz}{2\pi i} Q_\tau(z)$ and $p_1(\tau) = -\oint_{\mathcal{C}} \frac{dz}{2\pi i} zP_\tau(z)$.

Proof. Let us derive the first equation. Using (7.10)

$$\begin{aligned} \frac{d}{d\tau} Q_\tau(z) &= \frac{1}{n} \langle \theta^*, (H-z)^{-1} \frac{d\theta_\tau}{d\tau} \rangle = \frac{q(\tau)}{n} \langle \theta^*, (H-z)^{-1} \theta^* \rangle + \frac{1}{n\sqrt{\lambda}} \langle \theta^*, (H-z)^{-1} H \theta_\tau \rangle \\ &\quad - \left(q(\tau)^2 + \frac{p_1(\tau)}{\sqrt{\lambda}} \right) \frac{1}{n} \langle \theta^*, (H-z)^{-1} \theta_\tau \rangle \end{aligned} \quad (7.14)$$

Using $(H-z)^{-1}H = I + z(H-z)^{-1}$ in the second term in the right hand side, we immediately get the first equation of (7.13). Let us now derive the second equation. Again using (7.10) and since $(H-zI)$ is a symmetric matrix

$$\begin{aligned} \frac{d}{d\tau} P_\tau(z) &= \frac{2}{n} \langle \theta_\tau, (H-z)^{-1} \frac{d\theta_\tau}{d\tau} \rangle = 2 \frac{q(\tau)}{n} \langle \theta_\tau, (H-z)^{-1} \theta^* \rangle + \frac{2}{n\sqrt{\lambda}} \langle \theta_\tau, (H-z)^{-1} H \theta_\tau \rangle \\ &\quad - 2 \left(q(\tau)^2 + \frac{p_1(\tau)}{\sqrt{\lambda}} \right) \frac{1}{n} \langle \theta_\tau, (H-z)^{-1} \theta_\tau \rangle \end{aligned} \quad (7.15)$$

Thus using again $(H-z)^{-1}H = I + z(H-z)^{-1}$ and $\langle \theta_\tau, \theta_\tau \rangle = 1$ we get the second equation of (7.13). \square

7.4 Concentration results

We introduce the Stieltjes transform of the semi-circle law $\mu_{\text{sc}}(s) = \frac{1}{2\pi} \sqrt{4-s^2} \chi_{[-2,2]}(s)$,

$$G_{\text{sc}}(z) = \int_{\mathbb{R}} ds \frac{\mu_{\text{sc}}(s)}{s-z} = \frac{1}{2}(-z + \sqrt{z^2-4}), \quad z \in \mathbb{C} \setminus [-2,2]. \quad (7.16)$$

It is a classical result of random matrix theory Erdős (2011) that, for any $z \in \mathbb{C} \setminus [-2,2]$,

$$\frac{1}{n} \text{Tr} \mathcal{R}(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} G_{\text{sc}}(z)$$

. However here we will need convergence in probability of *matrix elements* of the resolvent (for given z and also uniformly in z). This tool is provided by recent results in random matrix theory that go under the name of *local semi-circle laws* Bloemendal et al. (2014).

Recall that \mathcal{S}_δ^n is the set of realizations of $H = \frac{1}{\sqrt{n}}\xi$ with eigenvalues in $I_\delta = [-2-\delta, 2+\delta]$, $\delta > 0$, and that $\lim_{n \rightarrow +\infty} \mathbb{P}(\mathcal{S}_\delta^n) = 1$. It will be convenient to use the notation \mathbb{P}_δ for the conditional probability law of H conditioned on the event $H \in \mathcal{S}_\delta^n$.

7.4.1 Initial condition analysis

We first derive natural initial conditions for the integro-differential equations (7.13) when $\frac{1}{n}\langle\theta_0, \theta^*\rangle = q(0) = \alpha \in [-1, +1]$. We claim (corollary 7.1 below) that the initial conditions $Q_0(z), P_0(z)$ as well as $R(z)$ concentrate on explicit functions $\bar{Q}_0(z), \bar{P}_0(z), \bar{R}(z)$. The main tool is the following proposition which we prove in Section 7.B (based on a theorem in Bloemendal et al. (2014)):

Proposition 7.2. *Fix $\delta > 0, \epsilon > 0$. For any fixed $z \in \mathbb{C} \setminus I_\delta$ and any deterministic sequence of unit vectors $u^{(n)}, v^{(n)} \in \mathbb{S}^{n-1}(1)$ the n -sphere of unit radius, we have*

$$\lim_{n \rightarrow \infty} \mathbb{P}_\delta(|\langle u^{(n)}, \mathcal{R}(z)v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{sc}(z)| > \epsilon) = 0. \quad (7.17)$$

Applying this proposition to the three pairs of unit vectors $(u^{(n)}, v^{(n)}) = (\frac{\theta_0}{\sqrt{n}}, \frac{\theta^*}{\sqrt{n}}), (\frac{\theta_0}{\sqrt{n}}, \frac{\theta_0}{\sqrt{n}})$, and $(\frac{\theta^*}{\sqrt{n}}, \frac{\theta^*}{\sqrt{n}})$ we directly obtain

Corollary 7.1. *Fix $\alpha \in [-1, +1]$ and θ_0 such that $\frac{1}{n}\langle\theta_0, \theta^*\rangle = \alpha$. For $z \in \mathbb{C} \setminus I_\delta$ we have convergence in probability of $Q_0(z), P_0(z), R(z)$ to the Stieljes transform of the semi-circle law:*

$$Q_0(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\delta} \bar{Q}_0(z) = \alpha G_{sc}(z), \quad P_0(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\delta} \bar{P}_0(z) = G_{sc}(z), \quad R(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\delta} \bar{R}(z) = G_{sc}(z). \quad (7.18)$$

7.4.2 Concentration of the overlap for finite times

We consider the integro-differential equations (7.13) for the limiting initial conditions $(\bar{Q}_0(z), \bar{P}_0(z)) = (\alpha G_{sc}(z), G_{sc}(z))$ and limiting $\bar{R}(r) = G_{sc}(z)$. More explicitly we define $\bar{Q}_\tau(z), \bar{P}_\tau(z)$ as the (holomorphic over $z \in \mathbb{C} \setminus I_\delta$) solutions of

$$\begin{cases} \frac{d}{d\tau} \bar{Q}_\tau(z) = \bar{q}(\tau) \bar{R}(z) + \frac{1}{\sqrt{\lambda}}(z \bar{Q}_\tau(z) + \bar{q}(\tau)) - \left(\bar{q}^2(\tau) + \frac{1}{\sqrt{\lambda}} \bar{p}_1(\tau) \right) \bar{Q}_\tau(z) \\ \frac{1}{2} \frac{d}{d\tau} \bar{P}_\tau(z) = \bar{q}(\tau) \bar{Q}_\tau(z) + \frac{1}{\sqrt{\lambda}}(z \bar{P}_\tau(z) + 1) - \left(\bar{q}^2(\tau) + \frac{1}{\sqrt{\lambda}} \bar{p}_1(\tau) \right) \bar{P}_\tau(z) \end{cases} \quad (7.19)$$

where by definition $\bar{q}(\tau) = -\oint_{\mathcal{C}} \frac{dz}{2\pi i} \bar{Q}_\tau(z)$ and $\bar{p}_1(\tau) = -\oint_{\mathcal{C}} \frac{dz}{2\pi i} z \bar{P}_\tau(z)$, and the initial conditions are $\bar{Q}_0(z) = \alpha G_{sc}(z), \bar{P}_0(z) = G_{sc}(z)$. The explicit calculation of the solutions $\bar{Q}_\tau(z), \bar{P}_\tau(z)$ in Section 7.5 shows that they exist and they are holomorphic for $z \in \mathbb{C} \setminus I_\delta$.

One can show that the concentration result of corollary 7.1 extends to all finite times. This can be done by a Grönwall stability type argument. A difficulty with respect to the standard argument is that here we deal with an integro-differential equation instead of purely ordinary differential equation. For this reason we need a *uniform* (over z) concentration result which strengthens proposition 7.2. The following is proved in Section 7.B.

Proposition 7.3. *Fix $\delta > 0, \epsilon > 0$. Recall $\mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$ for $\rho \geq 2 + \delta$. For any deterministic sequence of unit vectors $u^{(n)}, v^{(n)} \in \mathbb{S}^{n-1}(1)$ the n -sphere of unit radius, we have*

$$\lim_{n \rightarrow \infty} \mathbb{P}_\delta\left(\sup_{z \in \mathcal{C}} |\langle u^{(n)}, \mathcal{R}(z)v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{sc}(z)| > \epsilon\right) = 0. \quad (7.20)$$

Applying this proposition to appropriate pairs of unit vectors as previously we get directly:

Corollary 7.2. *Fix $\alpha \in [-1, +1]$ and θ_0 such that $\frac{1}{n}\langle \theta_0, \theta^* \rangle = \alpha$. Let $\mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$ for some $\rho \geq 2 + \delta$. Recall $\bar{Q}_0(z) = \alpha G_{sc}(z)$, $\bar{P}_0(z) = G_{sc}(z)$, $\bar{R}(z) = G_{sc}(z)$. Then $\sup_{z \in \mathcal{C}} |Q_0(z) - \bar{Q}_0(z)|$, $\sup_{z \in \mathcal{C}} |P_0(z) - \bar{P}_0(z)|$, $\sup_{z \in \mathcal{C}} |R(z) - \bar{R}(z)|$ all converge in \mathbb{P}_δ -probability to zero.*

In Section 7.C this corollary is used to prove:

Proposition 7.4. *Fix $\alpha \in [-1, +1]$ and θ_0 such that $\frac{1}{n}\langle \theta_0, \theta^* \rangle = \alpha$. Fix any $T > 0$. We have convergences at any $\tau \in [0, T]$ of the following overlaps to the deterministic limits $q(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}}$ $\bar{q}(\tau)$, $p_1(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}}$ $\bar{p}_1(\tau)$ where here convergence is with respect to the probability law \mathbb{P} of the generalized Wigner ensemble.*

Remark 7.1. *With a bit more work the proof of this corollary can be strengthened to also show that for any $z \in \mathbb{C} \setminus I_\delta$ and $\tau \in [0, T]$ we have convergence in probability of $Q_\tau(z)$, $P_\tau(z)$ to the deterministic solutions of the integro-differential equations (7.19), i.e., $Q_\tau(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\delta} \bar{Q}_\tau(z)$, $P_\tau(z) \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\delta} \bar{P}_\tau(z)$, as well as convergence of all overlaps $q_k(\tau)$, $p_k(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}}$ $\bar{q}_k(\tau)$, $\bar{p}_k(\tau)$ ($k \geq 1$). Since we will not need these results we omit their proof.*

7.5 Solution of integro-differential equations and overlap

In this section we analyze (7.19) for $z \in \mathbb{C} \setminus I_\delta$ with the initial conditions $(\bar{Q}_0(z), \bar{P}_0(z), \bar{R}(z)) = (\alpha G_{sc}(z), G_{sc}(z), G_{sc}(z))$. In the process we obtain $\bar{q}(\tau) \equiv -\int_{\mathbb{C}} \frac{dz}{2\pi i} \bar{Q}_\tau(z)$.

Proof of formulas (7.5) and (7.6) in theorem 7.1. We use a change of variable $\hat{Q}_\tau(z) = e^{F(\tau)} \bar{Q}_\tau(z)$ and $\hat{P}_\tau(z) = e^{2F(\tau)} \bar{P}_\tau(z)$ with $F(\tau) = \int_0^\tau ds \left(\bar{q}^2(s) + \frac{1}{\sqrt{\lambda}} \bar{p}_1(s) \right)$. Similarly, we define also $\hat{q}(\tau) = e^{F(\tau)} \bar{q}(\tau)$, $\hat{p}(\tau) = e^{2F(\tau)}$. We have $\bar{q}(\tau) = \hat{q}(\tau) / \sqrt{\hat{p}(\tau)}$, and therefore in order to determine the overlap it suffices to determine $\hat{q}(\tau)$ and $\hat{p}(\tau)$. With the change of variables equations (7.13) become

$$\begin{cases} \frac{d}{d\tau} \hat{Q}_\tau(z) = \hat{q}(\tau) \left(\bar{R}(z) + \frac{1}{\sqrt{\lambda}} \right) + \frac{z}{\sqrt{\lambda}} \hat{Q}_\tau(z) \\ \frac{1}{2} \frac{d}{d\tau} \hat{P}_\tau(z) = \hat{q}(\tau) \hat{Q}_\tau(z) + \frac{1}{\sqrt{\lambda}} \hat{p}(\tau) + \frac{z}{\sqrt{\lambda}} \hat{P}_\tau(z) \end{cases} \quad (7.21)$$

We analyze these equations in the Laplace domain. Recall the Laplace transformation $\mathcal{L}f(p) = \int_0^{+\infty} d\tau e^{-p\tau} f(\tau)$, $\text{Re } p > a \in \mathbb{R}_+$, which is well defined as long as $|f(\tau)| \leq e^{a\tau}$. All functions involved below in Laplace transforms satisfy this requirement for some $a \in \mathbb{R}_+$ large enough independent of n . It will often be convenient to use the notations $\mathcal{L}(f(t))(p) = \int_0^{+\infty} d\tau e^{-p\tau} f(\tau)$, $\mathcal{L}Q_p(z) = \int_0^{+\infty} d\tau e^{-p\tau} Q_\tau(z)$, $\mathcal{L}P_p(z) = \int_0^{+\infty} d\tau e^{-p\tau} P_\tau(z)$.

A) Derivation of (7.5) for $\hat{q}(\tau)$. Taking the Laplace transform of the first equation in (7.21)

$$p \mathcal{L} \hat{Q}_p(z) - \hat{Q}_0(z) = \mathcal{L} \hat{q}(p) \left(\bar{R}(z) + \frac{1}{\sqrt{\lambda}} \right) + \frac{z}{\sqrt{\lambda}} \mathcal{L} \hat{Q}_p(z) \quad (7.22)$$

7.5 Solution of integro-differential equations and overlap

Notice that $\hat{Q}_0(z) = e^{F(0)} \bar{Q}_0(z) = \alpha G_{\text{sc}}(z)$ and $\bar{R}(z) = G_{\text{sc}}(z)$, and hence we can re-arrange the terms,

$$\mathcal{L}\hat{Q}_p(z) = \alpha \frac{\sqrt{\lambda} G_{\text{sc}}(z)}{p\sqrt{\lambda} - z} + \mathcal{L}\hat{q}(p) \frac{\sqrt{\lambda} G_{\text{sc}}(z) + 1}{p\sqrt{\lambda} - z}. \quad (7.23)$$

Now, assuming $\text{Re } p > \frac{2+\delta}{\sqrt{\lambda}}$ (recall $\text{Re } p > 0$) leaves the point $p\sqrt{\lambda}$ outside the contour \mathcal{C} . Using Fubini first and the definition of \hat{q} secondly

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} \int_0^{+\infty} d\tau e^{-p\tau} \hat{Q}_\tau(z) = \int_0^{+\infty} d\tau e^{-p\tau} \oint_{\mathcal{C}} \frac{dz}{2\pi i} \hat{Q}_\tau(z) = - \int_0^{+\infty} d\tau e^{-p\tau} \hat{q}(\tau) \quad (7.24)$$

Thus we have $\oint_{\mathcal{C}} \frac{dz}{2\pi i} \mathcal{L}\hat{Q}_p(z) = -\mathcal{L}\hat{q}(p)$ on the left side of (7.23) while a straightforward calculation using Fubini on compact sets shows

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} \frac{G_{\text{sc}}(z)}{p\sqrt{\lambda} - z} = \frac{1}{2\pi i} \oint_{\mathcal{C}} \int_{-2}^2 \frac{\mu_{\text{sc}}(l) dl dz}{(l-z)(p\sqrt{\lambda} - z)} = \int_{-2}^2 \frac{\mu_{\text{sc}}(l) dl}{l - p\sqrt{\lambda}} = G_{\text{sc}}(p\sqrt{\lambda}) \quad (7.25)$$

So taking Cauchy integration formula on both sides of (7.23) we get

$$-\mathcal{L}\hat{q}(p) = \alpha \sqrt{\lambda} G_{\text{sc}}(p\sqrt{\lambda}) + \mathcal{L}\hat{q}(p) \sqrt{\lambda} G_{\text{sc}}(p\sqrt{\lambda}) \quad (7.26)$$

Thus we find:

$$\mathcal{L}\hat{q}(p) = -\frac{\alpha G_{\text{sc}}(p\sqrt{\lambda})}{\frac{1}{\sqrt{\lambda}} + G_{\text{sc}}(p\sqrt{\lambda})} = \alpha \frac{1 + \frac{1}{\sqrt{\lambda}} G_{\text{sc}}(p\sqrt{\lambda})}{p - (1 + \frac{1}{\lambda})} \quad (7.27)$$

where the last equality can be checked from the explicit expression (7.16) of $G_{\text{sc}}(z)$. It remains to invert this equation in the time domain. To do so we first notice that

$$G_{\text{sc}}(p\sqrt{\lambda}) = -\frac{1}{\sqrt{\lambda}} \int_{-2}^2 ds \mu_{\text{sc}}(s) \int_0^{+\infty} d\tau e^{(\frac{s}{\sqrt{\lambda}} - p)\tau} = -\frac{1}{\sqrt{\lambda}} \int_0^{+\infty} d\tau e^{-p\tau} M_\lambda(\tau) \quad (7.28)$$

where we recall that $M_\lambda(\tau)$ is the scaled moment generating function of the semi-circle law (7.3). The interchange of integrals in the third equality is justified by Fubini. Using $\mathcal{L}(e^{(1+\frac{1}{\lambda})\tau})(p) = (p - (1 + \frac{1}{\lambda}))^{-1}$, equation (7.27) becomes

$$\mathcal{L}\hat{q}(p) = \alpha \mathcal{L}(e^{(1+\frac{1}{\lambda})t})(p) - \frac{\alpha}{\lambda} \mathcal{L}(e^{(1+\frac{1}{\lambda})t}) \mathcal{L}M_\lambda(p) \quad (7.29)$$

This is easily transformed back in the time-domain using standard properties of the Laplace transform to get (7.5).

B) A useful identity. For the derivation of $\hat{p}(\tau)$ we will need the following identity derived in Appendix 7.H

$$-\oint_{\mathcal{C}} \frac{dz}{2\pi i} \hat{Q}_u(z) e^{\frac{z(\tau-u)}{\sqrt{\lambda}}} = \alpha M_\lambda(2\tau - u) + \int_0^u ds \hat{q}(s) M_\lambda(2\tau - u - s) \quad (7.30)$$

where we recall \mathcal{C} is the circle with center the origin and radius $\rho > 2 + \delta$.

C) *Derivation of $\hat{p}(\tau)$.* Taking the Laplace transform of the second equation in (7.21) we find

$$\frac{1}{2}(p\mathcal{L}\hat{P}_p(z) - \hat{P}_0(z)) = \mathcal{L}(\hat{q}(\tau)\hat{Q}_\tau(z))(p) + \frac{1}{\sqrt{\lambda}}\mathcal{L}\hat{p}(p) + \frac{z}{\sqrt{\lambda}}\mathcal{L}\hat{P}_p(z) \quad (7.31)$$

and using $\hat{P}_0(z) = e^{F(0)}\bar{P}_0(z) = G_{\text{sc}}(z)$ we can rearrange the terms to get

$$\mathcal{L}\hat{P}_p(z) = \frac{1}{p - \frac{z\sqrt{\lambda}}{2}} \left(G_{\text{sc}}(z) + 2\sqrt{\lambda}\mathcal{L}(\hat{q}(\tau)\hat{Q}_\tau(z))(p) + \frac{2}{\sqrt{\lambda}}\mathcal{L}\hat{p}(p) \right). \quad (7.32)$$

Then using $(p - \frac{2z}{\sqrt{\lambda}})^{-1} = \mathcal{L}(e^{\frac{2zt}{\sqrt{\lambda}}})(p)$ and

$$2\mathcal{L}(e^{\frac{2zt}{\sqrt{\lambda}}})\mathcal{L}(\hat{q}(t)\hat{Q}_t(z)) = \mathcal{L}\left(2\int_0^t \hat{q}(u)\hat{Q}_u(z)e^{\frac{2z(t-u)}{\sqrt{\lambda}}} du\right), \quad (7.33)$$

and replacing in (7.32) we get

$$\mathcal{L}\hat{P}_p(z) = \mathcal{L}\left(e^{\frac{2zt}{\sqrt{\lambda}}}\right)(p)G_{\text{sc}}(z) + 2\mathcal{L}\left(\int_0^\tau \hat{q}(u)\hat{Q}_u(z)e^{\frac{2z(\tau-u)}{\sqrt{\lambda}}} du\right)(p) + \frac{2}{\sqrt{\lambda}}\mathcal{L}\left(e^{\frac{2zt}{\sqrt{\lambda}}}\right)(p)\mathcal{L}\hat{p}(p). \quad (7.34)$$

Now we take $\text{Re } p > 4/\sqrt{\lambda}$ and choose the contour \mathcal{C} , encircling the interval I_δ , but such that it does not encircle the point $z = \frac{1}{2}p\sqrt{\lambda}$, and integrate each term of (7.34) along this contour. First note that the contribution of the last term vanishes since $\mathcal{L}\left(e^{\frac{2zt}{\sqrt{\lambda}}}\right)(p) = (p - \frac{2z}{\sqrt{\lambda}})^{-1}$ and the pole $z = \frac{1}{2}p\sqrt{\lambda}$ lies in the exterior of \mathcal{C} . Then there remains

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} \mathcal{L}\hat{P}_p(z) = \oint_{\mathcal{C}} \frac{dz}{2\pi i} \mathcal{L}\left(e^{\frac{2zt}{\sqrt{\lambda}}}\right)(p)G_{\text{sc}}(z) + 2\oint_{\mathcal{C}} \frac{dz}{2\pi i} \mathcal{L}\left(\int_0^\tau \hat{q}(u)\hat{Q}_u(z)e^{\frac{2z(\tau-u)}{\sqrt{\lambda}}} du\right)(p). \quad (7.35)$$

For the left hand side we have

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} \int_0^{+\infty} d\tau e^{-p\tau} \hat{P}_\tau(z) = \int_0^{+\infty} d\tau e^{-p\tau} \oint_{\mathcal{C}} \frac{dz}{2\pi i} \hat{P}_\tau(z) = - \int_0^{+\infty} d\tau e^{-p\tau} \hat{p}(\tau) \quad (7.36)$$

where the first equality follows from Fubini and the second by functional calculus Dunford and Schwartz (1988). For the first term on the right hand side of (7.35) we find (see Appendix 7.H for details)

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) \int_0^{+\infty} d\tau e^{-p\tau} e^{\frac{z\tau}{\sqrt{\lambda}}} = - \int_0^{+\infty} d\tau e^{-p\tau} M_\lambda(\tau). \quad (7.37)$$

Finally it remains to treat the last contour integral in (7.35). Using again Fubini and (7.30) we

find

$$\begin{aligned}
 & \oint_{\mathcal{C}} \frac{dz}{2\pi i} \int_0^{+\infty} d\tau e^{-p\tau} \int_0^\tau \hat{q}(u) \hat{Q}_u(z) e^{\frac{2z(\tau-u)}{\sqrt{\lambda}}} du = \int_0^{+\infty} d\tau e^{-p\tau} \int_0^\tau \hat{q}(u) \oint_{\mathcal{C}} \frac{dz}{2\pi i} \hat{Q}_u(z) e^{\frac{2z(\tau-u)}{\sqrt{\lambda}}} du \\
 & = - \int_0^{+\infty} d\tau e^{-p\tau} \int_0^\tau du \hat{q}(u) \left[\alpha M_\lambda(2\tau - u) + \int_0^u ds \hat{q}(s) M_\lambda(2\tau - u - s) \right] \\
 & = - \int_0^{+\infty} d\tau e^{-p\tau} \left[\alpha \int_0^\tau du \hat{q}(u) M_\lambda(2\tau - u) + \frac{1}{2} \int_0^\tau \int_0^\tau du ds \hat{q}(s) M_\lambda(2\tau - u - s) \right] \quad (7.38)
 \end{aligned}$$

Putting together (7.35), (7.36), (7.37), (7.38) we obtain (7.6) in the Laplace domain. Going back to the time domain we obtain (7.6). \square

7.6 Conclusion and future work

Tracking gradient descent dynamics and their variants for different scores and loss functions can be used to provide meaningful insights on a learning algorithm and for example, help monitor its progress and avoid over-fitting. As computational capabilities increase with distributed systems allowing for bigger datasets and larger systems to be treated, a good understanding of the dynamics can help account for computational cost.

We have seen in this work that for the rank-one matrix recovery problem in the regime of large dimensions, probabilistic concentrations naturally occur that can be captured by the local semi-circle laws in random matrix theory obtained in the last decade. In particular, suitable generating functions constructed out of the resolvent of the noise matrix concentrate around the solutions of a set of deterministic integro-differential equations. We have been able to completely solve these equations thereby tracking the dynamics for all times. It is also observed that the analytical solution provides a good approximation for the expected behavior of the learning algorithm, even for dimensions as low as $n < 100$.

The method and integro-differential equations derived here can be generalized to different models. For instance, we will show in forthcoming work how it is possible to apply it to certain neural-network architectures, and in particular the random feature models. This allows us to better understand the dynamical emergence of interesting behaviors such as the double descent phenomenon. The generalisation is possible, in essence, when the dynamics can be captured by spectral properties of some "resolvent matrix". Depending on the system though, performing random matrix averages can be arbitrarily complicated. For problems where the dynamics is not captured by some resolvent matrix, such as a genuine tensor problem (with tensor of order greater equal than three) it is not so clear how to proceed since there are no obvious spectral notions for tensors. One option would be to approach the problem by looking at the dynamics of the alternating least square method.

Appendix

7.A Analysis of the cost

Proof of theorem 7.2. Expanding the Frobenius norm in the cost and using $\|\theta_\tau\|^2 = \|\theta^*\|^2 = n$ we find

$$\begin{aligned}\mathcal{H}(\theta_\tau) &= \frac{1}{2n^2} \{-2\text{Tr}Y\theta_\tau\theta_\tau^T + \text{Tr}(\theta_\tau\theta_\tau^T\theta_\tau\theta_\tau^T)\} - \frac{1}{2n^2} \{-2\text{Tr}Y\theta^*\theta^{*T} + \text{Tr}(\theta^*\theta^{*T}\theta^*\theta^{*T})\} \\ &= \frac{1}{n^2} \langle \theta^*, Y\theta^* \rangle - \frac{1}{n^2} \langle \theta_\tau, Y\theta_\tau \rangle.\end{aligned}\quad (7.39)$$

Using that $Y = \theta^*\theta^{*T} + \frac{n}{\sqrt{\lambda}}H$ (recall $H = \frac{1}{\sqrt{n}}\xi$) we get

$$\begin{aligned}\mathcal{H}(\theta_\tau) &= \left(1 + \frac{1}{\sqrt{\lambda}} \frac{\langle \theta^*, H\theta^* \rangle}{n}\right) - \left(\frac{\langle \theta_\tau, \theta^* \rangle^2}{n^2} + \frac{1}{\sqrt{\lambda}} \frac{\langle \theta_\tau, H\theta_\tau \rangle}{n}\right) \\ &= \left(1 + \frac{1}{\sqrt{\lambda}} \frac{\langle \theta^*, H\theta^* \rangle}{n}\right) - \left(q(\tau)^2 + \frac{p_1(\tau)}{\sqrt{\lambda}}\right).\end{aligned}\quad (7.40)$$

By the law of large numbers $\frac{\langle \theta^*, H\theta^* \rangle}{n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0$ and since $q(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \bar{q}(\tau)$ and $p_1(\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \bar{p}_1(\tau)$ we have

$$\mathcal{H}(\theta_\tau) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 1 - \left(\bar{q}(\tau)^2 + \frac{\bar{p}_1(\tau)}{\sqrt{\lambda}}\right).\quad (7.41)$$

Now it remains to recall the definition of $F(\tau)$ and $\hat{p}_0(\tau) = e^{2F(\tau)}$, to see that

$$\bar{q}(\tau)^2 + \frac{\bar{p}_1(\tau)}{\sqrt{\lambda}} = \frac{dF(\tau)}{d\tau} = \frac{1}{2} \frac{d}{d\tau} \ln \hat{p}(\tau).\quad (7.42)$$

The result of the theorem follows from (7.41) and (7.42). \square

7.B Proof of propositions 7.2 and 7.3

The proof is based the following *local* semi-circle law (theorem 2.12 in Bloemendal et al. (2014)):

Theorem 7.3 (isotropic local semi-circle law Bloemendal et al. (2014)). *For any $\omega \in (0, 1)$*

Chapter 7. The rank-one model: a non-convex setting

consider the following domain in the upper half-plane

$$S(\omega, n) = \left\{ z \in \mathbb{C} \mid |\operatorname{Re}(z)| \leq \frac{1}{\omega}, \frac{1}{n^{1-\omega}} \leq \operatorname{Im}(z) \leq \frac{1}{\omega} \right\}.$$

Then for all $\delta, D > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$, and any unit vectors $u, v \in \mathbb{S}_n(1)$:

$$\sup_{z \in S(\omega, n)} \mathbb{P} \left(|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > n^\delta \left[\sqrt{\frac{\operatorname{Im} G_{\text{sc}}(z)}{n \operatorname{Im} z}} + \frac{1}{n \operatorname{Im} z} \right] \right) < \frac{1}{n^D} \quad (7.43)$$

where \mathbb{P} is the probability law on the generalized Wigner matrix.

Proof of proposition 7.2. First we note that for $\operatorname{Im} z \neq 0$ since $\lim_{n \rightarrow +\infty} \mathbb{P}(\mathcal{S}_\delta^n) = 1$ we have

$$\lim_{n \rightarrow +\infty} \mathbb{P}(|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > \epsilon) = \lim_{n \rightarrow +\infty} \mathbb{P}_\delta(|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > \epsilon). \quad (7.44)$$

We consider first the cases $\operatorname{Im} z$ strictly positive, negative, and then give the extra argument needed for $\operatorname{Im} z = 0$.

First we take $\operatorname{Im} z > 0$. We can find $n_1 \in \mathbb{N}, \omega \in (0, 1)$ such that $z \in S(\omega, n_1)$ and henceforth, for all $n \geq n_1$, $z \in S(\omega, n)$. Taking $\delta = \frac{1}{4}, D = 1$ and applying theorem 7.3 yields the existence of n_0 such that for all $n \geq \max(n_0, n_1)$:

$$\mathbb{P} \left(|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > n^{\frac{1}{4}} \left[\sqrt{\frac{\operatorname{Im} G_{\text{sc}}(z)}{n \operatorname{Im} z}} + \frac{1}{n \operatorname{Im} z} \right] \right) < \frac{1}{n}. \quad (7.45)$$

Set $l(n, z) = n^{\frac{1}{4}} \left[\sqrt{\frac{\operatorname{Im} G_{\text{sc}}(z)}{n \operatorname{Im} z}} + \frac{1}{n \operatorname{Im} z} \right]$. Since $\lim_{n \rightarrow \infty} l(n, z) = 0$, we can find n_2 such that for all $n \geq n_2$ we have $l(n, z) < \epsilon$. Thus for all $n \geq \max(n_0, n_1, n_2)$ we have the set inclusion in the generalized Wigner ensemble

$$\{H : |\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > \epsilon\} \subset \{H : |\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > l(n, z)\} \quad (7.46)$$

and therefore

$$\mathbb{P}(|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| > \epsilon) < \frac{1}{n}. \quad (7.47)$$

Applying this inequality to a deterministic sequence $(u^{(n)}, v^{(n)})$ on the unit sphere and taking the limit $n \rightarrow \infty$ concludes the proof for $\operatorname{Im} z > 0$.

To deal with $\operatorname{Im} z < 0$ it suffices to remark that $|\langle u, \mathcal{R}(z)v \rangle - \langle u, v \rangle G_{\text{sc}}(z)| = |\langle u, \mathcal{R}(\bar{z})v \rangle - \langle u, v \rangle G_{\text{sc}}(\bar{z})|$. Alternatively one could use a version of theorem 7.3 for the lower half-plane.

Consider now $z = x$ with $x \in \mathbb{R} \setminus I_\delta$ and $H \in \mathcal{S}_\delta^n$. Take a complex number $x + iy$, $0 < y \leq$

$\frac{\epsilon}{2}|x - (2 + \delta)|^2$. From the mean value theorem we have

$$\begin{aligned} & |(\langle u, \mathcal{R}(x)v \rangle - \langle u, v \rangle G_{\text{sc}}(x)) - (\langle u, \mathcal{R}(x + iy)v \rangle - \langle u, v \rangle G_{\text{sc}}(x + iy_k))| \\ & \leq |y| \sup_{y>0} \left| \frac{d}{dy} \langle u, \mathcal{R}(x + iy)v \rangle \right|. \end{aligned} \quad (7.48)$$

Since for $H \in \mathcal{S}_\delta^n$

$$\left| \frac{d}{dy} \langle u, \mathcal{R}(x + iy)v \rangle \right| = |\langle u, (x + iy - H)^{-2}v \rangle| \leq \frac{1}{(x - (2 + \delta))^2 + y^2} \quad (7.49)$$

we deduce from (7.48) and the triangle inequality

$$\begin{aligned} |\langle u, \mathcal{R}(x)v \rangle - \langle u, v \rangle G_{\text{sc}}(x)| & \leq |\langle u, \mathcal{R}(x + iy)v \rangle - \langle u, v \rangle G_{\text{sc}}(x + iy_k)| + \frac{y}{|x - (2 + \delta)|^2} \\ & \leq |\langle u, \mathcal{R}(x + iy)v \rangle - \langle u, v \rangle G_{\text{sc}}(x + iy_k)| + \frac{\epsilon}{2}. \end{aligned} \quad (7.50)$$

Thus for realizations $H \in \mathcal{S}_\delta^n$, the event $|\langle u, \mathcal{R}(x)v \rangle - \langle u, v \rangle G_{\text{sc}}(x)| > \epsilon$ implies the event $|\langle u, \mathcal{R}(x + iy)v \rangle - \langle u, v \rangle G_{\text{sc}}(x + iy)| \geq \frac{\epsilon}{2}$ for any $0 < y \leq \frac{\epsilon}{2}|x - (2 + \delta)|^2$. In other words

$$\mathbb{P}_\delta(|\langle u, \mathcal{R}(x)v \rangle - \langle u, v \rangle G_{\text{sc}}(x)| > \epsilon) \leq \mathbb{P}_\delta(|\langle u, \mathcal{R}(x + iy)v \rangle - \langle u, v \rangle G_{\text{sc}}(x + iy)| \geq \frac{\epsilon}{2}). \quad (7.51)$$

By the previous results for $\text{Im } z > 0$ we conclude that these probabilities tend to zero as $n \rightarrow +\infty$. □

Proof of proposition 7.3. The proof uses a discretization argument together with the union bound. Consider the discrete set of N points on the contour \mathcal{C} , $z_k = \rho e^{i\theta_k}$, $\theta_k = \frac{2\pi k}{N}$, $k = 0, \dots, N - 1$. First, Observe that from the union bound

$$\begin{aligned} & \mathbb{P}\left(\max_{k=0, \dots, N} |\langle u^{(n)}, \mathcal{R}(z_k)v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{\text{sc}}(z_k)| > \epsilon\right) \\ & \leq \sum_{k=0}^N \mathbb{P}\left(|\langle u^{(n)}, \mathcal{R}(z_k)v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{\text{sc}}(z_k)| > \epsilon\right) \end{aligned} \quad (7.52)$$

thus from proposition (7.2)

$$\lim_{n \rightarrow +\infty} \mathbb{P}\left(\max_{k=0, \dots, N} |\langle u^{(n)}, \mathcal{R}(z_k)v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{\text{sc}}(z_k)| > \epsilon\right) = 0 \quad (7.53)$$

Second, for any $z = \rho e^{i\theta} \in \mathcal{C}$ there exist a θ_k such that $|\theta - \theta_k| \leq \frac{1}{N}$. Applying the triangle inequality $|b| \leq |a| + |b - a|$ for $a = \langle u^{(n)}, \mathcal{R}(z)v^{(n)} \rangle - \langle u^{(n)}, G_{\text{sc}}(z)v^{(n)} \rangle$ and $b = \langle u^{(n)}, \mathcal{R}(z_k)v^{(n)} \rangle -$

$\langle u^{(n)}, G_{\text{sc}}(z_k) v^{(n)} \rangle$, and the mean value theorem, we get

$$\begin{aligned} |\langle u^{(n)}, \mathcal{R}(z_k) v^{(n)} \rangle - \langle u^{(n)}, G_{\text{sc}}(z_k) v^{(n)} \rangle| &\leq |\langle u^{(n)}, \mathcal{R}(z) v^{(n)} \rangle - \langle u^{(n)}, \mathcal{R}_H(z) v^{(n)} \rangle| \\ &+ \frac{1}{N} \sup_{\theta \in [0, 2\pi]} \left| \frac{d}{d\theta} \langle u^{(n)}, \mathcal{R}(\rho e^{i\theta}) v^{(n)} \rangle \right| \end{aligned} \quad (7.54)$$

We can take the supremum of the right hand side over $z \in \mathcal{C}$ and then the maximum of the right hand side over $k = 0, \dots, N-1$ to deduce

$$\begin{aligned} \max_{k=0, \dots, N} |\langle u^{(n)}, \mathcal{R}(z_k) v^{(n)} \rangle - \langle u^{(n)}, G_{\text{sc}}(z_k) v^{(n)} \rangle| &\leq \sup_{z \in \mathcal{C}} |\langle u^{(n)}, \mathcal{R}(z_k) v^{(n)} \rangle - \langle u^{(n)}, \mathcal{R}(z_k) v^{(n)} \rangle| \\ &+ \frac{1}{N} \sup_{\theta \in [0, 2\pi]} \left| \frac{d}{d\theta} \langle u^{(n)}, \mathcal{R}(\rho e^{i\theta}) v^{(n)} \rangle \right| \end{aligned} \quad (7.55)$$

Since

$$\frac{d}{d\theta} \langle u^{(n)}, \mathcal{R}(\rho e^{i\theta}) v^{(n)} \rangle = i \rho e^{i\theta} \langle u^{(n)}, (\rho e^{i\theta} - H)^{-2} v^{(n)} \rangle \quad (7.56)$$

we deduce from Cauchy-Schwarz, that with probability tending to one as $n \rightarrow +\infty$

$$\frac{1}{N} \sup_{\theta \in [0, 2\pi]} \left| \frac{d}{d\theta} \langle u^{(n)}, \mathcal{R}(\rho e^{i\theta}) v^{(n)} \rangle \right| \leq \frac{\rho}{N(\rho-2)^2} \quad (7.57)$$

Therefore taking $N > \frac{2\rho}{\epsilon(\rho-2)^2}$ we find from (7.53), (7.57) and (7.55)

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left(\sup_{z \in \mathcal{C}} |\langle u^{(n)}, \mathcal{R}(z) v^{(n)} \rangle - \langle u^{(n)}, v^{(n)} \rangle G_{\text{sc}}(z)| \geq \frac{\epsilon}{2} \right) = 0 \quad (7.58)$$

for any $\epsilon > 0$. This concludes the proof. \square

7.C Proof of proposition 7.4

We assume the condition $H \in \mathcal{S}_\delta^n$ so that $\|\mathcal{R}(z)\|_{\text{op}} \leq (\rho-2)^{-1}$ for all $z \in \mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$ and $\rho > 2 + \delta$. The condition is relaxed at the very end.

The proof of proposition 7.4 is based on a Gronwall type argument. As explained in Section 7.4 the difficulty here is that we have an integro-differential equation instead of a plain ordinary differential equation and the usual Lipschitz condition is not a priori satisfied. For this reason, given that $H \in \mathcal{S}_\delta^n$, we need preliminary bounds on $\sup_{z \in \mathcal{C}} |Q_\tau(z)|$, $\sup_{z \in \mathcal{C}} |P_\tau(z)|$, $\sup_{z \in \mathcal{C}} |R(z)|$, $\sup_{z \in \mathcal{C}} |\bar{R}(z)|$ and on $\sup_{z \in \mathcal{C}} |\bar{Q}_\tau(z)|$, $\sup_{z \in \mathcal{C}} |\bar{P}_\tau(z)|$, for $\tau \in [0, T]$. Here we do not seek the best possible bounds but rather we just need that all quantities are bounded (with high probability for the first three).

For the first *four* quantities the bound easily follows from their definition (7.11). By Cauchy-Schwartz we obtain that $\sup_{z \in \mathcal{C}} |Q_\tau(z)|$, $\sup_{z \in \mathcal{C}} |P_\tau(z)|$ and $\sup_{z \in \mathcal{C}} |R(z)|$ are upper bounded

by $(\rho - 2)^{-1}$. For $\sup_{z \in \mathcal{C}} |\bar{R}(z)|$ we can use the integral representation to get the same (loose) bound.

The remaining *two* quantities are here defined through the solution of the integro-differential equation (7.19) which we take as a starting point to prove a bound. In Section 7.5 we compute exactly the combination $\bar{q}(\tau)^2 + \frac{1}{\lambda} \bar{p}_1(\tau) \equiv \frac{1}{2} \ln \hat{p}(\tau)$ and find $\hat{p}(\tau)$ given by formula (7.6). It can be checked that this is a continuous function for any compact time interval, so $\sup_{\tau \in [0, T]} |\bar{q}(\tau)^2 + \frac{1}{\sqrt{\lambda}} \bar{p}_1(\tau)| \leq L_*(T) < +\infty$ for any $T > 0$ (in fact one can even take L_* independent of T but we will not need this information). Then, integrating the first equation in (7.19) over $[0, \tau]$, using the triangle inequality, and then taking suprema, we deduce

$$\sup_{z \in \mathcal{C}} |\bar{Q}_\tau(z)| \leq \sup_{z \in \mathcal{C}} |\bar{Q}_0(z)| + \left(\frac{\rho}{\rho - 2} + \frac{2\rho}{\sqrt{\lambda}} + \rho^2 L_*(T) \right) \int_0^\tau ds \sup_{z \in \mathcal{C}} |\bar{Q}_s(z)| \quad (7.59)$$

Iterating this inequality a standard calculation yields any $\tau \in [0, T]$

$$\sup_{z \in \mathcal{C}} |\bar{Q}_\tau(z)| \leq \sup_{z \in \mathcal{C}} |\bar{Q}_0(z)| e^{T \left(\frac{\rho}{\rho - 2} + \frac{2\rho}{\sqrt{\lambda}} + \rho^2 L_*(T) \right)} \leq \frac{\alpha}{\rho - 2} e^{T \left(\frac{\rho}{\rho - 2} + \frac{2\rho}{\sqrt{\lambda}} + \rho^2 L_*(T) \right)} \quad (7.60)$$

where we used $\bar{Q}_0(z) = \alpha G_{\text{sc}}(z)$, and for $|G_{\text{sc}}(z)| \leq \frac{1}{\rho - 2}$ for $z \in \mathcal{C}$. The definition of $\bar{q}(\tau)$ in terms of a contour integral implies immediately $\sup_{\tau \in [0, T]} |\bar{q}(\tau)| \leq L(T)$ where $L(T)$ is the right hand side of (7.60) multiplied by ρ . Now, integrating the second equation in (7.19) over $[0, \tau]$, using the triangle inequality, and then taking suprema again, we deduce

$$\begin{aligned} \frac{1}{2} \sup_{z \in \mathcal{C}} |\bar{P}_\tau(z)| &\leq \frac{1}{2} \sup_{z \in \mathcal{C}} |\bar{P}_0(z)| + \frac{\alpha^2 \rho \tau}{(\rho - 2)^2} e^{2T \left(\frac{\rho}{\rho - 2} + \frac{2\rho}{\sqrt{\lambda}} + \rho^2 L_*(T) \right)} + \frac{\tau}{\sqrt{\lambda}} \\ &\quad + \left(\frac{\rho}{\sqrt{\lambda}} + L_*(T) \right) \int_0^\tau \sup_{z \in \mathcal{C}} |\bar{P}_s(z)| \end{aligned} \quad (7.61)$$

Again a standard calculation yields (using the initial condition $\bar{P}_0(z) = G_{\text{sc}}(z)$)

$$\sup_{z \in \mathcal{C}} |\bar{P}_\tau(z)| \leq \left(\frac{1}{\rho - 2} + \frac{2\alpha^2 \rho T}{(\rho - 2)^2} e^{2T \left(\frac{\rho}{\rho - 2} + \frac{2\rho}{\sqrt{\lambda}} + \rho^2 L_*(T) \right)} + \frac{2\tau}{\sqrt{\lambda}} \right) e^{T \left(\frac{\rho}{\sqrt{\lambda}} + L_*(T) \right)} \quad (7.62)$$

Note that this implies the bound $\sup_{\tau \in [0, T]} |\bar{p}_1(\tau)| \leq L_1(T)$ where $L_1(T)$ is the right hand side of (7.62) multiplied by ρ^2 .

We now have all the elements to adapt a Gronwall type argument.

Proof of proposition 7.4. We start by deriving preliminary bounds We set $Q_\tau(z) - \bar{Q}_\tau(z) = \Delta_\tau^Q(z)$, $P_\tau(z) - \bar{P}_\tau(z) = \Delta_\tau^P(z)$, $R(z) - \bar{R}(z) = \Delta^R(z)$, $q(\tau) - \bar{q}(\tau) = \delta^q(\tau)$, $p_1(\tau) - \bar{p}_1(\tau) = \delta^{p_1}(\tau)$. Note for later use that all the $\sup_{z \in \mathcal{C}} |\cdot|$ of these differences are bounded by some finite positive constant depending only on ρ, α, λ, T . Taking the difference of (7.19) and (7.13) we find after a

bit of algebra

$$\begin{aligned}
 \frac{d}{d\tau} \Delta_\tau^Q(z) &= \delta^q(\tau) \Delta^R(z) + \delta^q(\tau) \bar{R}(z) + \bar{q}(\tau) \Delta^R(z) + \frac{1}{\sqrt{\lambda}} (z \Delta_\tau^Q(z) + \delta^q(\tau)) \\
 &\quad - (q(\tau) + \bar{q}(\tau)) \delta^q(\tau) \Delta_\tau^Q(z) - (q(\tau) + \bar{q}(\tau)) \delta^q(\tau) \bar{Q}_\tau(z) - \bar{q}(\tau)^2 \Delta_\tau^Q(z) \\
 &\quad - \frac{1}{\sqrt{\lambda}} (\delta^{p_1}(\tau) \Delta_\tau^Q(z) - \delta^{p_1}(\tau) \bar{Q}_\tau(z) - \bar{p}_1(\tau) \Delta_\tau^Q(z))
 \end{aligned} \tag{7.63}$$

and

$$\begin{aligned}
 \frac{d}{d\tau} \Delta_\tau^P(z) &= \delta^q(\tau) \Delta_\tau^Q(z) + \delta^q(\tau) \bar{Q}_\tau(z) + \bar{q}(\tau) \Delta_\tau^Q(z) + \frac{1}{\sqrt{\lambda}} z \Delta_\tau^P(z) \\
 &\quad - (q(\tau) + \bar{q}(\tau)) \delta^q(\tau) \Delta_\tau^P(z) - (q(\tau) + \bar{q}(\tau)) \delta^q(\tau) \bar{P}_\tau(z) - \bar{q}(\tau)^2 \Delta_\tau^P(z) \\
 &\quad - \frac{1}{\sqrt{\lambda}} (\delta^{p_1}(\tau) \Delta_\tau^P(z) - \delta^{p_1}(\tau) \bar{P}_\tau(z) - \bar{p}_1(\tau) \Delta_\tau^P(z))
 \end{aligned} \tag{7.64}$$

After integrating the above equations over the interval $[0, \tau]$, using the triangle inequality, and the inequalities $|\delta^q(\tau)| \leq \rho \sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)|$, $|\delta^{p_1}(\tau)| \leq \rho^2 \sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)|$, $|q(\tau)| \leq 1$, $\sup_{\tau \in [0, T]} |\bar{q}(\tau)| < L(T)$, $\sup_{\tau \in [0, T]} |\bar{p}_1(\tau)| < L_1(T)$, we deduce (with $L = \max(L(T), L_1(T))$)

$$\begin{aligned}
 \sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)| &\leq \sup_{z \in \mathcal{C}} |\Delta_0^Q(z)| + \rho \sup_{z \in \mathcal{C}} |\Delta^R(z)| \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + \rho \sup_{z \in \mathcal{C}} |\bar{R}(z)| \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| \\
 &\quad + L\tau \sup_{z \in \mathcal{C}} |\Delta^R(z)| + \frac{2\rho}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + (1+L)\rho \int_0^\tau ds (\sup_{z \in \mathcal{C}} |\Delta_s^Q(z)|)^2 \\
 &\quad + (1+L)\rho \int_0^\tau ds (\sup_{z \in \mathcal{C}} |\Delta_s^Q(z)|)^2 \sup_{z \in \mathcal{C}} |\bar{Q}_s(z)| + L^2 \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| \\
 &\quad + \frac{\rho^2}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + \frac{\rho^2}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \sup_{z \in \mathcal{C}} |\bar{Q}_s(z)| \\
 &\quad + \frac{L}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)|
 \end{aligned} \tag{7.65}$$

and

$$\begin{aligned}
 \sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)| &\leq \sup_{z \in \mathcal{C}} |\Delta_0^P(z)| + \rho \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)|^2 + \rho \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| \sup_{z \in \mathcal{C}} |\bar{Q}_s(z)| \\
 &\quad + L \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + \frac{\rho}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| + (1+L)\rho \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \\
 &\quad + (1+L)\rho \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| \sup_{z \in \mathcal{C}} |\bar{P}_s(z)| + L^2 \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \\
 &\quad + \frac{\rho^2}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)|^2 + \frac{\rho^2}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \sup_{z \in \mathcal{C}} |\bar{P}_s(z)| + \frac{L}{\sqrt{\lambda}} \int_0^\tau ds \sup_{z \in \mathcal{C}} |\Delta_s^P(z)|
 \end{aligned} \tag{7.66}$$

Now, using (7.60) and (7.62) we can "linearize" the right hand side to obtain two inequalities

of the form (where $C(\rho, \alpha, \lambda, T)$ is a suitable constant)

$$\sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)| \leq \sup_{z \in \mathcal{C}} |\Delta_0^Q(z)| + L\tau \sup_{z \in \mathcal{C}} \Delta^R(z) + C(\rho, \alpha, \lambda, T) \int_0^\tau ds \{ \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \} \quad (7.67)$$

and

$$\sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)| \leq \sup_{z \in \mathcal{C}} |\Delta_0^P(z)| + C(\rho, \alpha, \lambda, T) \int_0^\tau ds \{ \sup_{z \in \mathcal{C}} |\Delta_s^Q(z)| + \sup_{z \in \mathcal{C}} |\Delta_s^P(z)| \} \quad (7.68)$$

Summing (7.67) and (7.68) and iterating the resulting integral inequality we deduce

$$\sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)| + \sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)| \leq \{ \sup_{z \in \mathcal{C}} |\Delta_0^Q(z)| + \sup_{z \in \mathcal{C}} |\Delta_0^P(z)| + LT \sup_{z \in \mathcal{C}} \Delta^R(z) \} e^{2TC(\rho, \alpha, \lambda, T)} \quad (7.69)$$

By corollary 7.2 we conclude that for $\tau \in [0, T]$ $\sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)|$ and $\sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)|$ converge in \mathbb{P}_δ -probability to zero.

Finally, we can look at the overlaps. Observe that $|q(\tau) - \bar{q}(\tau)| = |\int_{\mathcal{C}} \frac{dz}{2\pi i} \Delta_\tau^Q(z)|$ so $|q(\tau) - \bar{q}(\tau)| \leq \rho \sup_{z \in \mathcal{C}} |\Delta_\tau^Q(z)|$ and $|p_1(\tau) - \bar{p}_1(\tau)| = |\int_{\mathcal{C}} \frac{dz}{2\pi i} z \Delta_\tau^P(z)| \leq \rho^2 \sup_{z \in \mathcal{C}} |\Delta_\tau^P(z)|$. Therefore $|q(\tau) - \bar{q}(\tau)|$ and $|p_1(\tau) - \bar{p}_1(\tau)|$ converge with \mathbb{P}_δ -probability to 0. But since $\lim_{n \rightarrow +\infty} \mathbb{P}(H \in \mathcal{S}_\delta^n) = 1$ it is easy to see (by the law of total probability) that $|q(\tau) - \bar{q}(\tau)|$ and $|p_1(\tau) - \bar{p}_1(\tau)|$ also converge with \mathbb{P} -probability to 0. □

7.D Laplace Transform applicability

Laplace transform can be applied appropriately with the condition of deriving a bound of the form $e^{a\tau}$ with $a > 0$ for the terms $\hat{q}(\tau)$, $\hat{p}(\tau)$ first, and $\hat{Q}_\tau(z)$, $\hat{P}_\tau(z)$ secondly. For $\hat{q}(\tau)$ because $M_\lambda(s)$ is positive on $[0, \tau]$ and $\alpha \hat{q}(\tau)$ remains positive at all time, we derive the bound

$$0 \leq |\hat{q}(\tau)| \leq e^{(1+\frac{1}{\lambda})\tau} \quad (7.70)$$

Next we find a bound for $M_\lambda(\tau)$ with the definition (7.3)

$$|M_\lambda(\tau)| \leq 2 \int_0^\pi \frac{d\theta}{\pi} |\sin(\theta)|^2 |e^{\cos(\theta)\frac{2\tau}{\sqrt{\lambda}}}| \leq 2e^{\frac{2\tau}{\sqrt{\lambda}}} \quad (7.71)$$

For $\hat{p}(\tau)$ using the previous bound and $|\alpha| \leq 1$

$$|\hat{p}(\tau)| \leq |M_\lambda(2\tau)| + 2 \int_0^\tau |\hat{q}(s)| |M_\lambda(2\tau - s)| ds + \int_0^\tau \int_0^\tau |\hat{q}(u) \hat{q}(v)| |M_\lambda(2\tau - u - v)| dudv \quad (7.72)$$

$$\leq 2e^{\frac{4\tau}{\sqrt{\lambda}}} + 4 \int_0^\tau e^{(1+\frac{1}{\lambda})s + \frac{2}{\sqrt{\lambda}}(2\tau-s)} ds + 2 \int_0^\tau \int_0^\tau e^{(1+\frac{1}{\lambda})(u+v) + \frac{2}{\sqrt{\lambda}}(2\tau-u-v)} dudv \quad (7.73)$$

Hence

$$\frac{1}{2}|\hat{p}(\tau)|e^{-\frac{4\tau}{\sqrt{\lambda}}} \leq 1 + 2 \int_0^\tau e^{(1-\frac{1}{\sqrt{\lambda}})^2 s} ds + \int_0^\tau \int_0^\tau e^{(1-\frac{1}{\sqrt{\lambda}})^2 u} e^{(1-\frac{1}{\sqrt{\lambda}})^2 v} dudv \quad (7.74)$$

$$\leq \left(1 + \int_0^\tau e^{(1-\frac{1}{\sqrt{\lambda}})^2 s} ds\right)^2 \quad (7.75)$$

$$\leq \left(1 + \frac{e^{(1-\frac{1}{\sqrt{\lambda}})^2 \tau} - 1}{(1-\frac{1}{\sqrt{\lambda}})^2}\right)^2 \quad (7.76)$$

$$\leq e^{2(1-\frac{1}{\sqrt{\lambda}})^2 \tau} \left(e^{-(1-\frac{1}{\sqrt{\lambda}})^2 \tau} + \frac{1 - e^{-(1-\frac{1}{\sqrt{\lambda}})^2 \tau}}{(1-\frac{1}{\sqrt{\lambda}})^2} \right)^2 \quad (7.77)$$

$$\leq e^{2(1-\frac{1}{\sqrt{\lambda}})^2 \tau} \left(1 + \frac{1}{(1-\frac{1}{\sqrt{\lambda}})^2} \right)^2 \quad (7.78)$$

Hence with $C_\lambda = 2 \left(1 + \frac{1}{(1-\frac{1}{\sqrt{\lambda}})^2}\right)^2$ we have the exponential bound $|\hat{p}(\tau)| \leq C_\lambda e^{2(1+\frac{1}{\lambda})\tau}$.

Now going back to equations (7.21) we have the system

$$\begin{cases} \frac{d}{d\tau} e^{-\frac{z\tau}{\sqrt{\lambda}}} \hat{Q}_\tau(z) = e^{-\frac{z\tau}{\sqrt{\lambda}}} \hat{q}(\tau) \left(\bar{R}(z) + \frac{1}{\sqrt{\lambda}} \right) \\ \frac{1}{2} \frac{d}{d\tau} e^{-\frac{2z\tau}{\sqrt{\lambda}}} \hat{P}_\tau(z) = e^{-\frac{2z\tau}{\sqrt{\lambda}}} \hat{q}(\tau) \hat{Q}_\tau(z) + \frac{1}{\sqrt{\lambda}} \hat{p}(\tau) e^{-\frac{2z\tau}{\sqrt{\lambda}}} \end{cases} \quad (7.79)$$

Hence integrating over $[0, \tau]$ provides

$$\begin{cases} \hat{Q}_\tau(z) = \bar{Q}_0(z) e^{\frac{z\tau}{\sqrt{\lambda}}} + \left(\bar{R}(z) + \frac{1}{\sqrt{\lambda}} \right) \int_0^\tau ds e^{\frac{z(\tau-s)}{\sqrt{\lambda}}} \hat{q}(s) \\ \hat{P}_\tau(z) = e^{\frac{2z\tau}{\sqrt{\lambda}}} \bar{P}_0(z) + 2 \int_0^\tau ds e^{\frac{2z(\tau-s)}{\sqrt{\lambda}}} \hat{q}(s) \hat{Q}_s(z) + \frac{2}{\sqrt{\lambda}} \int_0^\tau ds \hat{p}(s) e^{\frac{2z(\tau-s)}{\sqrt{\lambda}}} \end{cases} \quad (7.80)$$

Notice again that we have $|G_{sc}(z)| \leq \frac{1}{\rho-2}$ for $z \in \mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$ where $\rho > 2$.

For $\hat{Q}_\tau(z)$ we find

$$|\hat{Q}_\tau(z)| \leq \frac{|\alpha|}{\rho-2} e^{\frac{\operatorname{Re}(z)\tau}{\sqrt{\lambda}}} + \left(\frac{1}{\rho-2} + \frac{1}{\sqrt{\lambda}} \right) \int_0^\tau ds e^{\frac{\operatorname{Re}(z)(\tau-s)}{\sqrt{\lambda}}} e^{(1+\frac{1}{\lambda})s} \quad (7.81)$$

$$\leq e^{\frac{\rho\tau}{\sqrt{\lambda}}} \left(\frac{1}{\rho-2} + \left(\frac{1}{\rho-2} + \frac{1}{\sqrt{\lambda}} \right) \int_0^\tau ds e^{(\frac{1}{\lambda} - \frac{\rho}{\sqrt{\lambda}})s} \right) \quad (7.82)$$

$$\leq e^{\frac{\rho\tau}{\sqrt{\lambda}}} \left(\frac{1}{\rho-2} + \left(\frac{1}{\rho-2} + \frac{1}{\sqrt{\lambda}} \right) e^{\frac{\tau}{\lambda}} \int_0^\tau ds e^{-s} \right) \quad (7.83)$$

$$\leq e^{\frac{\rho\tau}{\sqrt{\lambda}}} \left(\frac{1}{\rho-2} + \left(\frac{1}{\rho-2} + \frac{1}{\sqrt{\lambda}} \right) e^{(\frac{1}{\lambda}+1)\tau} (1 - e^{-\tau}) \right) \quad (7.84)$$

$$\leq e^{(1+\frac{\rho}{\sqrt{\lambda}}+\frac{1}{\lambda})\tau} \left(\frac{2}{\rho-2} + \frac{1}{\sqrt{\lambda}} \right) \quad (7.85)$$

7.E Enforcing the spherical constraint in gradient dynamics

With $C'_{\rho,\lambda} = \frac{2}{\rho-2} + \frac{1}{\sqrt{\lambda}}$ we thus have $|\hat{Q}_\tau(z)| \leq C'_{\rho,\lambda} e^{(1+\frac{\rho}{\sqrt{\lambda}}+\frac{1}{\lambda})\tau}$ for any $z \in \mathcal{C}$. Similarly for $\hat{P}_\tau(z)$:

$$|\hat{P}_\tau(z)| \leq e^{\frac{2\rho\tau}{\sqrt{\lambda}}} \left(\frac{1}{\rho-2} + 2C'_{\rho,\lambda} \int_0^\tau e^{\frac{-2\rho s}{\sqrt{\lambda}} + (1+\frac{1}{\lambda})s + (1+\frac{\rho}{\sqrt{\lambda}}+\frac{1}{\lambda})s} ds + \frac{2C_\lambda}{\sqrt{\lambda}} \int_0^\tau e^{\frac{-2\rho s}{\sqrt{\lambda}} + 2(1-\frac{1}{\sqrt{\lambda}})^2 s} ds \right) \quad (7.86)$$

$$\leq e^{\frac{2\rho\tau}{\sqrt{\lambda}}} \left(\frac{1}{\rho-2} + 2C'_{\rho,\lambda} e^{2(1+\frac{1}{\lambda})\tau} + \frac{2}{\sqrt{\lambda}} e^{2(1+\frac{1}{\lambda})\tau} \right) \quad (7.87)$$

$$\leq e^{2(1+\frac{\rho}{\sqrt{\lambda}}+\frac{1}{\lambda})\tau} \left(\frac{1}{\rho-2} + 2C'_{\rho,\lambda} + \frac{2}{\sqrt{\lambda}} C_\lambda \right) \quad (7.88)$$

Hence with $C''_{\lambda,\rho} = \frac{1}{\rho-2} + 2C'_{\rho,\lambda} + \frac{2}{\sqrt{\lambda}} C_\lambda$ we find $|\hat{P}_\tau(z)| \leq C''_{\lambda,\rho} e^{2(1+\frac{\rho}{\sqrt{\lambda}}+\frac{1}{\lambda})\tau}$ for any $z \in \mathcal{C}$.

7.E Enforcing the spherical constraint in gradient dynamics

The second term in equation (7.2) enforces the spherical constraint $\theta_t \in \mathbb{S}^{n-1}(\sqrt{n})$ at all times. This is well known but we briefly recall how to derive it for completeness. Since the n dimensional sphere is embedded in \mathbb{R}^n the covariant gradient D_θ can be obtained by projecting the usual gradient ∇_θ on a tangent plane. This projection is obtained by subtracting the component along a radius of the sphere, i.e., $\frac{\theta}{\sqrt{n}} \langle \frac{\theta}{\sqrt{n}}, \nabla_\theta \mathcal{H}(\theta) \rangle$. Therefore gradient descent reads

$$\frac{d\theta_t}{dt} = \eta D_\theta \mathcal{H}(\theta_t) = \eta \left(\nabla_\theta \mathcal{H}(\theta_t) - \frac{\theta_t}{n} \langle \theta_t, \nabla_\theta \mathcal{H}(\theta_t) \rangle \right). \quad (7.89)$$

It is easily checked that $\frac{d\|\theta_t\|_2^2}{dt} = 0$ and since $\theta_0 \in \mathbb{S}^{n-1}(\sqrt{n})$ we have $\theta_t \in \mathbb{S}^{n-1}(\sqrt{n})$ for all times. Indeed

$$\frac{d\|\theta_t\|_2^2}{dt} = 2 \langle \theta_t, \frac{d\theta_t}{dt} \rangle = 2\eta \left(\langle \theta_t, \nabla_\theta \mathcal{H}(\theta_t) \rangle - \frac{\langle \theta_t, \theta_t \rangle}{n} \langle \theta_t, \nabla_\theta \mathcal{H}(\theta_t) \rangle \right) = 0. \quad (7.90)$$

7.F Strict saddle property

We say that the strict saddle property is satisfied if the critical points of the cost are *strict* saddles or minima (a strict saddle has by definition at least one strictly negative eigenvalue of the Hessian). It is known from Lee et al. (2016) that for a cost satisfying the strict saddle property, gradient descent with small enough discrete time steps converges to a minimum, almost surely with respect to the initial condition. In the present context (as shown below) the critical points are given by the eigenvectors of $A \equiv \frac{\sqrt{\lambda}}{n} Y = \frac{\sqrt{\lambda}}{n} \theta^* \theta^{*T} + \frac{1}{\sqrt{n}} \xi$ - call them $v_i \in \mathcal{S}^{n-1}(\sqrt{n})$, $i = 1, \dots, n$ - and the Hessian at v_i is proportional to $\alpha_i I - A$ where α_i is the corresponding eigenvalue. For a random $n \times n$ matrix and *fixed* λ the spectrum is almost surely non-degenerate,³ i.e., $\alpha_1 < \alpha_2 < \dots < \alpha_n$, so the strict saddle property is almost surely satisfied. Moreover the top eigenvector v_n has positive definite Hessian and is a minimum,

³However for a fixed realization when λ varies we can have eigenvalue crossings.

Chapter 7. The rank-one model: a non-convex setting

while for the other ones are strict saddles with non-zero positive and negative eigenvalues. Now, for $\lambda > 1$ we know, that for n large enough with high probability, $\{\alpha_1 < \dots < \alpha_{n-1}\} \subset [-2, 2]$, $\alpha_n \approx \sqrt{\lambda} + 1/\sqrt{\lambda} > 2$ and $n^{-1}|\langle \theta_*, v_n \rangle| \approx \sqrt{1 - 1/\lambda}$ (where $a \approx b$ means $|a - b| = o_n(1)$) P     (2004); F  ral and P     (2006). This explains that for $\lambda > 1$ gradient descent with a small enough discrete time steps will converge to v_n and the overlap approach $\pm\sqrt{1 - 1/\lambda}$.

The critical points on the sphere $\mathcal{S}^{n-1}(\sqrt{n})$ satisfy $D_\theta \mathcal{H}(\theta) = 0$ where $D_\theta = (1 - \frac{1}{n}\theta\theta^T)\nabla_\theta$ is the covariant derivative. We have

$$D_\theta \mathcal{H}(\theta) \propto \frac{1}{n} \langle \theta, A\theta \rangle \theta - A\theta = 0 \quad (7.91)$$

and has n solutions $\theta = v_i$, $i = 1, \dots, n$. The Hessian matrix on the sphere is (up to a positive prefactor)

$$D_\theta D_\theta^T \mathcal{H}(\theta) \propto (1 - \frac{1}{n}\theta\theta^T) \left(\frac{1}{n} \langle \theta, A\theta \rangle I - A \right) \quad (7.92)$$

and for each critical point $\theta = v_i$ we find $D_\theta D_\theta^T \mathcal{H}(v_i) \propto \frac{1}{n^2}(\alpha_i I - A)$. This has $n-1$ eigenvectors v_j , $j \neq i$ (perpendicular to v_i and tangent to the sphere) with eigenvalues $\alpha_i - \alpha_j$, $j \neq i$, and one eigenvector v_i with 0 eigenvalue. For fixed λ there is no degeneracy $\alpha_1 < \alpha_2 < \dots < \alpha_n$, almost surely and v_n is a minimum while v_j , $j \neq n$ are strict saddles.

7.G Analysis of the stationary equation

The stationary equations corresponding to (7.13) are given by setting the time derivatives on the left hand side to zero.

$$\begin{cases} \bar{q}^\infty \left(\bar{R}(z) + \frac{1}{\sqrt{\lambda}} \right) + \left(\frac{z}{\sqrt{\lambda}} - (\bar{q}^\infty)^2 - \frac{1}{\sqrt{\lambda}} \bar{p}_1^\infty \right) \bar{Q}_\infty(z) = 0 \\ \bar{q}^\infty \bar{Q}_\infty(z) + \frac{1}{\sqrt{\lambda}} + \left(\frac{z}{\sqrt{\lambda}} - (\bar{q}^\infty)^2 - \frac{1}{\sqrt{\lambda}} \bar{p}_1^\infty \right) \bar{P}_\infty(z) = 0 \end{cases} \quad (7.93)$$

where $\bar{q}^\infty \equiv -\int_{\mathcal{C}} \frac{dz}{2\pi i} \bar{Q}_\infty(z)$, $\bar{p}_1^\infty \equiv -\int_{\mathcal{C}} \frac{dz}{2\pi i} z \bar{P}_\infty(z)$, $\bar{R}(z) = G_{\text{sc}}(z)$, and $\mathcal{C} = \{z \in \mathbb{C} \mid z = \rho e^{i\theta}, \theta \in [0, 2\pi]\}$, $\rho > 2$. Here we show how to derive all possible solutions of these equations. One expects that the set of solutions contains the limiting solution for $\tau \rightarrow +\infty$ and we check that this is indeed the case.

From (7.93) we get

$$\begin{cases} \bar{Q}_\infty(z) = \bar{q}^\infty \frac{\sqrt{\lambda} G_{\text{sc}}(z) + 1}{\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty - z} \\ \bar{P}_\infty(z) = (\bar{q}^\infty)^2 \sqrt{\lambda} \frac{\sqrt{\lambda} G_{\text{sc}}(z) + 1}{(\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty - z)^2} + \frac{1}{\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty - z} \end{cases} \quad (7.94)$$

Let us first assume that $|\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty| \leq 2$. We integrate the second equation over the contour \mathcal{C} . One can show that integral of the first term on the right hand side vanishes. Thus we find the condition by $\bar{p}_1^\infty = \sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty$ which implies $\bar{q}^\infty = 0$. This implies in turn that

$Q_\infty(z) = 0$, $P_\infty(z) = (\bar{p}_1^\infty - z)^{-1}$ and $|\bar{p}_1^\infty| \leq 2$.

Now assume that $|\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty| > 2$. Integrating the first equation of (7.94) over \mathcal{C} we find

$$\bar{q}^\infty = \sqrt{\lambda} \bar{q}^\infty \int_{z \in \mathcal{C}} \frac{dz}{2\pi i} \frac{G_{\text{sc}}(z)}{z - (\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty)} \quad (7.95)$$

The solution $\bar{q}^\infty = 0$ is again a possibility $Q_\infty(z) = 0$, $P_\infty(z) = (\bar{p}_1^\infty - z)^{-1}$ and $|\bar{p}_1^\infty| > 2$.

Now assume that $\bar{q}^\infty \neq 0$ (and still $|\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty| > 2$). Computing the contour integral we find the equation $1 = -\sqrt{\lambda} G_{\text{sc}}(\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty)$ which provides a solution and a condition

$$\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty = \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda} \quad (7.96)$$

$$\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty \geq \frac{2}{\sqrt{\lambda}} \quad (7.97)$$

Notice that the initial condition $\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty > 2$ is satisfied for all $\lambda \neq 1$, while $\frac{1}{\sqrt{\lambda}} + \sqrt{\lambda} \geq \frac{2}{\sqrt{\lambda}}$ is equivalent to $\lambda \geq 1$. So a solution can only exist when $\lambda > 1$. Integrating the second equation in (7.94) over \mathcal{C} we find

$$-\frac{1}{\lambda} = (\bar{q}^\infty)^2 \int_{z \in \mathcal{C}} \frac{dz}{2\pi i} \frac{G_{\text{sc}}(z)}{\left(z - (\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty)\right)^2} = -(\bar{q}^\infty)^2 \frac{dG_{\text{sc}}(z)}{dz} \Big|_{\sqrt{\lambda}(\bar{q}^\infty)^2 + \bar{p}_1^\infty} \quad (7.98)$$

Then using the explicit expression of $G_{\text{sc}}(z)$ we find that $(\bar{q}^\infty)^2 = 1 - \frac{1}{\lambda}$, with $\lambda > 1$. Furthermore we have from (7.94) and (7.96)

$$\begin{cases} \bar{Q}_\infty(z) = \left(1 - \frac{1}{\lambda}\right) \frac{\sqrt{\lambda} G_{\text{sc}}(z) + 1}{\sqrt{\lambda + \frac{1}{\lambda}} - z} \\ \bar{P}_\infty(z) = \left(1 - \frac{1}{\lambda}\right) \sqrt{\lambda} \frac{\sqrt{\lambda} G_{\text{sc}}(z) + 1}{\left(\sqrt{\lambda + \frac{1}{\lambda}} - z\right)^2} + \frac{1}{\sqrt{\lambda + \frac{1}{\lambda}} - z} \end{cases} \quad (7.99)$$

Note that multiplying the second equation in (7.99) by z and integrating over \mathcal{C} yields $\bar{p}_1^\infty = \frac{2}{\sqrt{\lambda}}$. This is consistent with (7.96).

We conclude by noting that the solutions that are attainable from the time evolution when $\lambda > 1$ are $\{\bar{q}^\infty = 0, |\bar{p}_1^\infty| \leq 2\}$ and $\{\bar{q}^\infty = \pm \sqrt{1 - \frac{1}{\lambda}}, \bar{p}_1^\infty = \frac{2}{\lambda}\}$. The first one is "attained" from an initial condition with $\alpha = \frac{1}{n} \langle \theta^*, \theta_0 \rangle = 0$. In this case gradient descent "does not start" and $\bar{q}^\infty = \bar{q}(0) = 0$, $\bar{p}_1^\infty = \bar{p}_1(0) = \frac{1}{n} \langle \theta_0, H\theta_0 \rangle$ and $\bar{p}_1(0) \leq 2$ with high probability. The other two solutions correspond to the initial conditions $\alpha = \frac{1}{n} \langle \theta^*, \theta_0 \rangle$ with $\alpha > 0$ and $\alpha < 0$. When $\lambda \leq 1$, there is only one possible solution $\{\bar{q}^\infty = 0, |\bar{p}_1^\infty| \leq 2\}$.

7.H Intermediate identities

We derive a number of identities requiring interchange of integrals.

Chapter 7. The rank-one model: a non-convex setting

A) *Derivation of (7.30).* To prove (7.30) we start with (7.23) in the form

$$\mathcal{L}\hat{Q}_p(z) = \alpha G_{\text{sc}}(z) \mathcal{L}(e^{\frac{zI}{\sqrt{\lambda}}})(p) + \mathcal{L}\hat{q}(p) \mathcal{L}(e^{\frac{zI}{\sqrt{\lambda}}})(p) \left(G_{\text{sc}}(z) + \frac{1}{\sqrt{\lambda}} \right) \quad (7.100)$$

and invert it back to the time domain

$$\hat{Q}_\tau(z) = \alpha G_{\text{sc}}(z) e^{\frac{z\tau}{\sqrt{\lambda}}} + G_{\text{sc}}(z) \int_0^\tau ds \hat{q}(s) e^{\frac{z(\tau-s)}{\sqrt{\lambda}}} + \frac{1}{\sqrt{\lambda}} \int_0^\tau ds \hat{q}(s) e^{\frac{z(\tau-s)}{\sqrt{\lambda}}}. \quad (7.101)$$

So this generating function is entirely known. Now we multiply this equation by $e^{\frac{z(\tau-u)}{\sqrt{\lambda}}}$ and integrate along \mathcal{C} . It is easy to see that, by Fubini's theorem, for the last term on the right hand side, the contour integral and the s -integral can be exchanged. Therefore the contour integral of the last term on the right hand side vanishes because $e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}}$ is holomorphic in the whole complex plane. For the other two terms on the right hand side we use the semi-circle law representation of $G_{\text{sc}}(z)$ to obtain (see below for details) to obtain

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) e^{\frac{z(2\tau-u)}{\sqrt{\lambda}}} = -M_\lambda(2\tau - u) \quad (7.102)$$

and

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) \int_0^\tau ds \hat{q}(s) e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}} = - \int_0^\tau ds \hat{q}(s) M_\lambda(2\tau - s - u). \quad (7.103)$$

Putting together (7.102), (7.103) and (7.101) we obtain the claimed identity (7.30).

B) *Derivation of (7.102).* From the semi-circle law representation of G_{sc}

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) e^{\frac{z(2\tau-u)}{\sqrt{\lambda}}} = \oint_{\mathcal{C}} \frac{dz}{2\pi i} \int_{-2}^2 ds \frac{\mu_{\text{sc}}(s)}{s-z} e^{\frac{z(2\tau-u)}{\sqrt{\lambda}}} \quad (7.104)$$

It is easy to see that Fubini's theorem can be applied to interchange the integrals. Indeed the contour integral over \mathcal{C} can be parametrized so that we then have two integrals with bounded functions over bounded intervals. So

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) e^{\frac{z(2\tau-u)}{\sqrt{\lambda}}} = \int_{-2}^2 ds \mu_{\text{sc}}(s) \oint_{\mathcal{C}} \frac{dz}{2\pi i} \frac{e^{\frac{z(2\tau-u)}{\sqrt{\lambda}}}}{s-z} = - \int_{-2}^2 ds \mu_{\text{sc}}(s) e^{\frac{s(2\tau-u)}{\sqrt{\lambda}}} = -M_\lambda(2\tau - u) \quad (7.105)$$

C) *Derivation of (7.103).* We proceed similarly. First,

$$\oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) \int_0^\tau ds \hat{q}(s) e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}} = \oint_{\mathcal{C}} \frac{dz}{2\pi i} \int_{-2}^2 dx \int_0^\tau ds \mu_{\text{sc}}(x) \hat{q}(s) \frac{e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}}}{x-z} \quad (7.106)$$

Again, it is clear that the contour integral can be parametrized so that we all integrals are over bounded intervals and all functions are bounded, so that Fubini's theorem applies. Thus

$$\begin{aligned} \oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) \int_0^\tau ds \hat{q}(s) e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}} &= \int_0^\tau ds \hat{q}(s) \int_{-2}^2 dx \mu_{\text{sc}}(x) \oint_{\mathcal{C}} \frac{dz}{2\pi i} \frac{e^{\frac{z(2\tau-s-u)}{\sqrt{\lambda}}}}{x-z} \\ &= - \int_0^\tau ds \hat{q}(s) M_\lambda(2\tau-s-u) \end{aligned} \quad (7.107)$$

D) *Derivation of (7.37)*. Again, using Fubini and then Cauchy's theorem,

$$\begin{aligned} \oint_{\mathcal{C}} \frac{dz}{2\pi i} G_{\text{sc}}(z) \int_0^{+\infty} d\tau e^{-p\tau} e^{\frac{z\tau}{\sqrt{\lambda}}} &= \int_0^{+\infty} d\tau e^{-p\tau} \oint_{\Gamma'} \frac{dz}{2\pi i} G_{\text{sc}}(z) e^{\frac{z\tau}{\sqrt{\lambda}}} \\ &= \int_0^{+\infty} d\tau e^{-p\tau} \oint_{\mathcal{C}} \frac{dz}{2\pi i} e^{\frac{z\tau}{\sqrt{\lambda}}} \int_{-2}^2 ds \frac{\mu_{\text{sc}}(s)}{s-x} \\ &= \int_0^{+\infty} d\tau e^{-p\tau} \int_{-2}^2 ds \mu_{\text{sc}}(s) \oint_{\mathcal{C}} \frac{dz}{2\pi i} \frac{e^{\frac{z\tau}{\sqrt{\lambda}}}}{s-z} \\ &= - \int_0^{+\infty} d\tau e^{-p\tau} \int_{-2}^2 ds \mu_{\text{sc}}(s) e^{\frac{s\tau}{\sqrt{\lambda}}} \\ &= - \int_0^{+\infty} d\tau e^{-p\tau} M_\lambda(\tau) \end{aligned} \quad (7.108)$$

7.I Asymptotic analysis of \bar{q}

7.I.1 limit when $\lambda > 1$

We deduce the limiting behavior for $\lambda > 1$. The next order correction is given in 7.I.3. Rewriting the first term from theorem 7.1, we have for $\tau \in \mathbb{R}^+$

$$e^{-(1+\frac{1}{\lambda})\tau} \hat{q}(\tau) = \alpha \left[1 - \frac{1}{\lambda} \int_0^\tau e^{-(1+\frac{1}{\lambda})s} M_\lambda(s) ds \right]. \quad (7.109)$$

We notice that in the limit $\tau \rightarrow \infty$, the right hand side of the integral is the laplace transform

$$\int_0^\infty e^{-(1+\frac{1}{\lambda})s} M_\lambda(s) ds = \mathcal{L} M_\lambda \left(1 + \frac{1}{\lambda} \right) \quad (7.110)$$

and we have seen the connection with resolvent in (7.28)

$$\mathcal{L} M_\lambda \left(1 + \frac{1}{\lambda} \right) = -\sqrt{\lambda} G_{\text{sc}} \left(\left(1 + \frac{1}{\lambda} \right) \sqrt{\lambda} \right). \quad (7.111)$$

Chapter 7. The rank-one model: a non-convex setting

But $X^2 + (1 + \frac{1}{\lambda})\sqrt{\lambda}X + 1 = 0$ has two roots: $\{-\sqrt{\lambda}; \frac{-1}{\sqrt{\lambda}}\}$. To ensure $G_{\text{sc}}(z) \in \mathbb{C}_+$ when $z \in \mathbb{C}_+$, we have $-\sqrt{\lambda}$ for $\lambda < 1$ and $\frac{-1}{\sqrt{\lambda}}$ for $\lambda > 1$. Thus we conclude

$$\lim_{\tau \rightarrow \infty} e^{-(1+\frac{1}{\lambda})\tau} \hat{q}(\tau) = \begin{cases} 0 & (\lambda < 1) \\ \alpha(1 - \frac{1}{\lambda}) & (\lambda > 1) \end{cases} \quad (7.112)$$

Therefore, in the regime $\lambda > 1$, we find the asymptotic behavior for $\tau \rightarrow \infty$

$$\hat{q}(\tau) \sim \alpha e^{(1+\frac{1}{\lambda})\tau} (1 - \frac{1}{\lambda}). \quad (7.113)$$

A careful analysis of the terms entering $\hat{p}(\tau)$ shows the main contribution stems from the last term, on the square $\mathcal{C} = [\sqrt{\tau}, \tau]^2$ (as the integral can be neglected on $[0, \tau]^2 \setminus \mathcal{C}$):

$$\hat{p}(\tau) \simeq \int_{\sqrt{\tau}}^{\tau} \int_{\sqrt{\tau}}^{\tau} \hat{q}(u) q(v) M_{\lambda}(2\tau - u - v) du dv. \quad (7.114)$$

Using the approximation of $\hat{q}(t)$ in (7.113) for large $t \in \mathcal{C}$, we can further approximate

$$\hat{p}(\tau) \simeq \alpha^2 \left(1 - \frac{1}{\lambda}\right)^2 \int_{\sqrt{\tau}}^{\tau} \int_{\sqrt{\tau}}^{\tau} e^{(1+\frac{1}{\lambda})(u+v)\tau} M_{\lambda}(2\tau - u - v) du dv \quad (7.115)$$

and a change of variables $u = \tau - x, v = \tau - y$ provides

$$\hat{p}(\tau) \simeq \alpha^2 e^{2(1+\frac{1}{\lambda})\tau} \left(1 - \frac{1}{\lambda}\right)^2 \iint_{[0, \tau(1-\frac{1}{\lambda})]^2} e^{-(1+\frac{1}{\lambda})(x+y)} M_{\lambda}(x+y) dy dx. \quad (7.116)$$

Now, notice the integral converges towards a non-zero value K_{λ} when $\tau \rightarrow \infty$

$$K_{\lambda} = \iint_{[0, \infty]^2} e^{-(1+\frac{1}{\lambda})(x+y)} M_{\lambda}(x+y) dy dx. \quad (7.117)$$

Using a further change of variable $s = x + y$ we find

$$K_{\lambda} = \int_{x=0}^{\infty} \int_{s=x}^{\infty} e^{-(1+\frac{1}{\lambda})s} M_{\lambda}(s) ds dx = \int_{s=0}^{\infty} \int_{x=0}^s e^{-(1+\frac{1}{\lambda})s} M_{\lambda}(s) dx ds. \quad (7.118)$$

Hence again, we find a connection with a Laplace transform (with a derivative from the additional s term inside the integral)

$$K_{\lambda} = \int_{s=0}^{\infty} e^{-(1+\frac{1}{\lambda})s} s M_{\lambda}(s) ds = -(\mathcal{L} M_{\lambda})' \left(1 + \frac{1}{\lambda}\right) \quad (7.119)$$

As $\mathcal{L} M_{\lambda}(p) = -\sqrt{\lambda} G_{\text{sc}}(p\sqrt{\lambda})$, and considering that $G'_{\text{sc}}(z) = -\frac{G_{\text{sc}}(z)}{2G_{\text{sc}}(z)+z}$, and that $G_{\text{sc}}((1 +$

$\frac{1}{\lambda})\sqrt{\lambda}) = -\frac{1}{\sqrt{\lambda}}$ in the case when $\lambda > 1$, we conclude

$$K_\lambda = \lambda \frac{\frac{1}{\sqrt{\lambda}}}{-2\frac{1}{\sqrt{\lambda}} + (1 + \frac{1}{\lambda})\sqrt{\lambda}} = \frac{1}{1 - \frac{1}{\lambda}}. \quad (7.120)$$

Finally, with (7.120) and (7.116) we find

$$\hat{p}(\tau) \sim \alpha^2 \left(1 - \frac{1}{\lambda}\right) e^{2(1 + \frac{1}{\lambda})\tau} \quad (7.121)$$

and for $\alpha > 0$, we can conclude $\lim_{\tau \rightarrow \infty} \bar{q}(\tau) = \sqrt{1 - \frac{1}{\lambda}}$.

7.1.2 Asymptotic analysis of $\lambda < 1$

The case $\lambda < 1$ is computationally more involved as $\hat{q}(\tau)$ converges to 0, and hence we need to find the rate of convergence towards 0 of this term and that of $\hat{p}(\tau)$ in order to deduce the one from $\bar{q}(\tau)$. Though it is not the main topic of the chapter, we provide some calculus elements to achieve this. We start with a lemma to find a suitable expression for $\hat{q}(\tau)$. Most of the calculations has been checked with Mathematica (a notebook is provided in the supplementary material).

Lemma 7.1. $\hat{q}(\tau)$ has the following equivalent form:

$$\hat{q}(\tau) = \alpha \left(1 - \frac{1}{\lambda}\right) e^{(1 + \frac{1}{\lambda})\tau} \mathbb{1}_{(1, +\infty)}(\lambda) + \frac{2\alpha}{\pi\lambda} e^{\frac{2}{\sqrt{\lambda}}\tau} \int_0^\pi e^{\frac{2}{\sqrt{\lambda}}(\cos(\theta) - 1)\tau} \frac{\sin(\theta)^2}{(1 + \frac{1}{\lambda}) - \frac{2}{\sqrt{\lambda}} \cos(\theta)} d\theta \quad (7.122)$$

Proof. Starting with $\hat{q}(\tau)$ from (7.1), one can use a similar expression of M_λ

$$\frac{e^{-(1 + \frac{1}{\lambda})\tau}}{\alpha} \hat{q}(\tau) = 1 - \frac{2}{\pi\lambda} \int_0^\pi \int_0^\tau e^{\left(\frac{2}{\sqrt{\lambda}} \cos(\theta) - (1 + \frac{1}{\lambda})\right)s} \sin(\theta)^2 ds d\theta \quad (7.123)$$

The inward integral can further be integrated (notice the constant term in the exponent is non-zero)

$$\frac{e^{-(1 + \frac{1}{\lambda})\tau}}{\alpha} \hat{q}(\tau) = 1 - \frac{2}{\pi\lambda} \int_0^\pi \left(e^{\left(\frac{2}{\sqrt{\lambda}} \cos(\theta) - (1 + \frac{1}{\lambda})\right)\tau} - 1 \right) \frac{\sin(\theta)^2}{\frac{2}{\sqrt{\lambda}} \cos(\theta) - (1 + \frac{1}{\lambda})} d\theta \quad (7.124)$$

Using proposition 7.5 with the constant $a = \frac{1 + \frac{1}{\lambda}}{\frac{2}{\sqrt{\lambda}}} > 1$, one can simplify

$$a - \sqrt{a^2 - 1} = \sqrt{\lambda} \frac{1 + \frac{1}{\lambda} - |1 - \frac{1}{\lambda}|}{2} = \begin{cases} \frac{1}{\sqrt{\lambda}} & (\lambda > 1) \\ \sqrt{\lambda} & (\lambda < 1) \end{cases} \quad (7.125)$$

and we finally find

$$\frac{2}{\pi\lambda} \int_0^\pi \frac{\sin(\theta)^2}{\frac{2}{\sqrt{\lambda}} \cos(\theta) - (1 + \frac{1}{\lambda})} d\theta = \frac{1}{\pi\sqrt{\lambda}} \int_0^\pi \frac{\sin(\theta)^2}{\cos(\theta) - a} d\theta = \begin{cases} -\frac{1}{\lambda} & (\lambda > 1) \\ -1 & (\lambda < 1) \end{cases} \quad (7.126)$$

using the solution (7.126) in (7.124) concludes the proof. \square

Proposition 7.5. *For any $a > 1$, we have:*

$$\int_0^\pi \frac{\sin(\theta)^2}{\cos(\theta) - a} d\theta = \pi(\sqrt{a^2 - 1} - a) \quad (7.127)$$

Proof. Bioche's rules suggest a change of variable $u = \tan(\frac{\theta}{2})$, we find on the left-hand side

$$\int_0^\pi \frac{\sin(\theta)^2}{\cos(\theta) - a} d\theta = \int_0^\infty \frac{(\frac{2u}{u^2+1})^2}{\frac{1-u^2}{1+u^2} - a} \frac{2du}{1+u^2} = \int_0^\infty \frac{8u^2}{[(1-a) - (1+a)u^2](1+u^2)^2} du \quad (7.128)$$

Using the constant $K = \frac{a-1}{a+1}$ (or equivalently $a = \frac{1+K}{1-K}$) we can rewrite

$$\int_0^\pi \frac{\sin(\theta)^2}{\cos(\theta) - a} d\theta = -4(1-K) \int_0^\infty \frac{u^2}{(K+u^2)(1+u^2)^2} du \quad (7.129)$$

and make a classical partial fraction decomposition of the inward term of the integral

$$\frac{u^2}{(K+u^2)(1+u^2)^2} = \frac{1}{(1-K)^2} \left(\frac{u^2}{K+u^2} - \frac{u^2}{1+u^2} \right) - \frac{1}{1-K} \frac{u^2}{(1+u^2)^2} \quad (7.130)$$

$$= \frac{1}{(1-K)^2} \left(\frac{1}{1+u^2} - \frac{K}{K+u^2} \right) - \frac{1}{1-K} \left[\frac{1}{1+u^2} - \frac{1}{(1+u^2)^2} \right] \quad (7.131)$$

$$= \frac{K}{(1-K)^2} \left(\frac{1}{1+u^2} - \frac{1}{K+u^2} \right) + \frac{1}{1-K} \frac{1}{(1+u^2)^2} \quad (7.132)$$

Then on the one hand, with change of variable $u = \tan(x)$ we have:

$$\int_0^\infty \frac{du}{(1+u^2)^2} = \int_0^{\frac{\pi}{2}} \frac{dx}{1+\tan^2(x)} = \int_0^{\frac{\pi}{2}} \cos^2(x) dx = \frac{\pi}{4} \quad (7.133)$$

On the other hand, with change of variable $u = \sqrt{K} \tan(x)$ we have:

$$\int_0^\infty \frac{du}{K+u^2} = \int_0^{\frac{\pi}{2}} dx = \frac{\pi}{2} \frac{1}{\sqrt{K}} \quad (7.134)$$

Thus:

$$-4(1-K) \int_0^\infty \frac{u^2 du}{(K+u^2)(1+u^2)^2} = -\pi \left[\frac{2K}{1-K} \left(1 - \frac{1}{\sqrt{K}} \right) + 1 \right] \quad (7.135)$$

and:

$$\frac{2K}{1-K} \left(1 - \frac{1}{\sqrt{K}} \right) + 1 = a - \sqrt{a^2 - 1} \quad (7.136)$$

□

Going back to the case $\lambda < 1$, we can simplify the expression from equation (7.122)

$$\hat{q}(\tau) = \frac{2\alpha}{\pi\lambda} e^{\frac{2}{\sqrt{\lambda}}\tau} \int_0^\pi e^{\frac{2}{\sqrt{\lambda}}(\cos(\theta)-1)\tau} \frac{\sin(\theta)^2}{(1 + \frac{1}{\lambda}) - \frac{2}{\sqrt{\lambda}}\cos(\theta)} d\theta \quad (7.137)$$

Further, with $u = \frac{2}{\sqrt{\lambda}}(1 - \cos(\theta))$ we rewrite (7.137) to apply Watson's lemma

$$\hat{q}(\tau) = \frac{2\alpha}{\pi\lambda} e^{\frac{2}{\sqrt{\lambda}}\tau} \left(\frac{\sqrt{\lambda}}{2}\right)^{\frac{1}{2}} \int_0^{\sqrt{\lambda}} e^{-u\tau} \frac{(u(2 - \frac{\sqrt{\lambda}}{2}u))^{\frac{1}{2}}}{(1 + \frac{1}{\lambda}) - \frac{2}{\sqrt{\lambda}}(1 - \frac{\sqrt{\lambda}}{2}u)} du \quad (7.138)$$

Therefore, Watson's lemma provides the asymptotic equivalence

$$\hat{q}(\tau) \sim \frac{2\alpha}{\pi\lambda} e^{\frac{2}{\sqrt{\lambda}}\tau} \left(\frac{\sqrt{\lambda}}{2}\right)^{\frac{3}{2}} \frac{2^{\frac{1}{2}}\Gamma(\frac{3}{2})\tau^{-\frac{3}{2}}}{(1 + \frac{1}{\lambda}) - \frac{2}{\sqrt{\lambda}}} \quad (7.139)$$

With $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$ we have therefore

$$\hat{q}(\tau) \sim \frac{\alpha\tau^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}\lambda^{\frac{1}{4}}\left(1 - \frac{1}{\sqrt{\lambda}}\right)^2} \quad (7.140)$$

The remaining term $\hat{p}(\tau)$ can further be analyzed by splitting each integral from theorem 7.1 and analyzing the terms with the asymptotic form $e^{\frac{4\tau}{\sqrt{\lambda}}}\tau^{-\frac{3}{2}}$. For instance, we get easily the first term for which we have

$$M_\lambda(2\tau) = \frac{\sqrt{\lambda}}{2\tau} I_1\left(\frac{4\tau}{\sqrt{\lambda}}\right) \sim \frac{\sqrt{\lambda}e^{\frac{4\tau}{\sqrt{\lambda}}}}{2t\sqrt{2\pi}\frac{4\tau}{\sqrt{\lambda}}} \sim \frac{\lambda^{\frac{3}{4}}e^{\frac{4\tau}{\sqrt{\lambda}}}}{2^{\frac{5}{2}}\sqrt{\pi}\tau^{\frac{3}{2}}} \quad (7.141)$$

The other terms require more technical considerations. We will use both former approximations from the equivalence relations (7.140) and (7.141). However, these approximations are only valid for large τ while the integral for the second term is applied on the whole range $[0, \tau]$. Therefore, we split the integration intervals into two segments, say $[0, \sqrt{\tau}]$ and $[\sqrt{\tau}, \tau]$, and apply the approximations in the domains where it is valid.

Starting with the second term, as $2\tau - s > \tau$ for all $s \in [0, \tau]$, we can already apply the relation (7.141) and split further the integrals:

$$\int_0^\tau \hat{q}(s)M_\lambda(2\tau - s)ds \simeq \frac{\lambda^{\frac{3}{4}}e^{\frac{4}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}} \left[\int_0^{\sqrt{\tau}} \hat{q}(s) \frac{e^{-\frac{2}{\sqrt{\lambda}}s}}{(2\tau - s)^{\frac{3}{2}}} ds + \int_{\sqrt{\tau}}^\tau \hat{q}(s) \frac{e^{-\frac{2}{\sqrt{\lambda}}s}}{(2\tau - s)^{\frac{3}{2}}} ds \right] \quad (7.142)$$

Chapter 7. The rank-one model: a non-convex setting

Then the integrand on the first segment of (7.142) is further approximated using $\frac{1}{(2\tau-s)^{\frac{3}{2}}} = \frac{1}{(2\tau)^{\frac{3}{2}}}$. Indeed, as $s \leq \sqrt{\tau}$ we have $s = o(\tau)$. In the end we retrieve the laplace transform of \hat{q} :

$$\int_0^{\sqrt{\tau}} \hat{q}(s) \frac{e^{-\frac{2}{\sqrt{\lambda}}s}}{(2\tau-s)^{\frac{3}{2}}} ds \simeq \frac{1}{(2\tau)^{\frac{3}{2}}} \int_0^{\sqrt{\tau}} \hat{q}(s) e^{-\frac{2}{\sqrt{\lambda}}s} ds \simeq \frac{1}{(2\tau)^{\frac{3}{2}}} \mathcal{L}\hat{q}\left(\frac{2}{\sqrt{\lambda}}\right) \quad (7.143)$$

From (7.27) which remains valid at $z = 2$ with $G_{\text{sc}}(2) = -1$, we can even derive further the constant term

$$\mathcal{L}\hat{q}\left(\frac{2}{\sqrt{\lambda}}\right) = \frac{-\alpha G_{\text{sc}}(2)}{\frac{1}{\sqrt{\lambda}} + G_{\text{sc}}(2)} = \frac{\alpha}{\frac{1}{\sqrt{\lambda}} - 1} \quad (7.144)$$

In the second segment of the integral in (7.142), we use the approximation from (7.140) and use change of variable $r = \frac{s}{\tau}$

$$\int_{\sqrt{\tau}}^{\tau} \hat{q}(s) \frac{e^{-\frac{2}{\sqrt{\lambda}}s}}{(2\tau-s)^{\frac{3}{2}}} ds \simeq \frac{\alpha}{2\sqrt{\pi}\lambda^{\frac{1}{4}} \left[\left(1 + \frac{1}{\lambda}\right) - \frac{2}{\sqrt{\lambda}} \right]} \int_{\frac{1}{\sqrt{\tau}}}^1 \frac{1}{r^{\frac{3}{2}}(2-r)^{\frac{3}{2}}} \frac{\tau}{\tau^{\frac{3}{2}+\frac{3}{2}}} dr \quad (7.145)$$

The integral from the right side can be solved:

$$\int_{\frac{1}{\sqrt{\tau}}}^1 \frac{dr}{r^{\frac{3}{2}}(2-r)^{\frac{3}{2}}} = \frac{1 - \frac{1}{\sqrt{\tau}}}{\sqrt{\frac{1}{\sqrt{\tau}}\left(2 - \frac{1}{\sqrt{\tau}}\right)}} \sim \frac{\tau^{\frac{1}{4}}}{\sqrt{2}} \quad (7.146)$$

Putting things together with (7.146) in (7.145) we get

$$\int_{\sqrt{\tau}}^{\tau} \hat{q}(s) \frac{e^{-\frac{2}{\sqrt{\lambda}}s}}{(2\tau-s)^{\frac{3}{2}}} ds \simeq \frac{\alpha\tau^{-\frac{7}{4}}}{2\sqrt{2\pi}\lambda^{\frac{1}{4}} \left[\left(1 + \frac{1}{\lambda}\right) - \frac{2}{\sqrt{\lambda}} \right]} \quad (7.147)$$

So, the main contribution comes from the first integral of equation (7.142) with the coefficient $\tau^{-\frac{3}{2}}$

$$2\alpha \int_0^{\sqrt{\tau}} \hat{q}(s) M_{\lambda}(2\tau-s) ds \sim \frac{\alpha^2 \lambda^{\frac{3}{4}} e^{\frac{4}{\sqrt{\lambda}}\tau} \tau^{-\frac{3}{2}}}{2^{\frac{3}{2}} \sqrt{\pi} \left(\frac{1}{\sqrt{\lambda}} - 1 \right)} \quad (7.148)$$

The third term with the double-integral requires extending the previous calculation idea on each rectangle: $I_1 = [0, \sqrt{\tau}]^2$, $I_2 = [0, \sqrt{\tau}] \times [\sqrt{\tau}, \tau]$, $I_2' = [\sqrt{\tau}, \tau] \times [0, \sqrt{\tau}]$ and $I_3 = [\sqrt{\tau}, \tau]^2$. As we will see, only the integral on I_1 brings a contribution of order $\tau^{-\frac{3}{2}}$ and the others can be neglected.

Interval $I_1 = [0, \sqrt{\tau}]^2$ On this interval, $2\tau - u - v \gg 1$ so we consider

$$\iint_{I_1} \hat{q}(u) \hat{q}(v) M_{\lambda}(2\tau - u - v) dudv \simeq \frac{\lambda^{\frac{3}{4}} e^{\frac{4}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}} \iint_{I_1} \hat{q}(u) \hat{q}(v) \frac{e^{-\frac{2}{\sqrt{\lambda}}(u+v)}}{(2\tau - u - v)^{\frac{3}{2}}} dudv \quad (7.149)$$

also, on I_1 we have $\frac{1}{(2\tau-u-v)^{\frac{3}{2}}} \simeq \frac{1}{(2\tau)^{\frac{3}{2}}}$, thus we are left to consider:

$$\iint_{I_1} \hat{q}(u)\hat{q}(v)e^{-\frac{2}{\sqrt{\lambda}}(u+v)}dudv \simeq \left[\mathcal{L}\hat{q}\left(\frac{2}{\sqrt{\lambda}}\right) \right]^2 \quad (7.150)$$

hence with (7.124) we find

$$\iint_{I_1} \hat{q}(u)\hat{q}(v)M_\lambda(2\tau-u-v)dudv \simeq \frac{\alpha^2\lambda^{\frac{3}{4}}e^{\frac{4}{\sqrt{\lambda}}\tau}\tau^{-\frac{3}{2}}}{2^{\frac{5}{2}}\sqrt{\pi}\left(\frac{1}{\sqrt{\lambda}}-1\right)^2} \quad (7.151)$$

Interval $I_2 = [0, \sqrt{\tau}] \times [\sqrt{\tau}, \tau]$ here we still have $2\tau - u - v \gg 1$ but also $v \geq \sqrt{\tau} \gg 1$ so with (7.140) we first get

$$\iint_{I_2} \hat{q}(u)\hat{q}(v)M_\lambda(2\tau-u-v)dudv \simeq \frac{\alpha \iint_{I_2} \hat{q}(u)v^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}v}M_\lambda(2\tau-u-v)dudv}{2\sqrt{\pi}\lambda^{\frac{1}{4}}\left[\left(1+\frac{1}{\lambda}\right)-\frac{2}{\sqrt{\lambda}}\right]} \quad (7.152)$$

and then:

$$\iint_{I_2} \hat{q}(u)v^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}v}M_\lambda(2\tau-u-v)dudv \simeq \frac{\lambda^{\frac{3}{4}}e^{\frac{4}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}} \iint_{I_2} \hat{q}(u)v^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}v} \frac{e^{-\frac{2}{\sqrt{\lambda}}(u+v)}}{(2\tau-u-v)^{\frac{3}{2}}}dudv \quad (7.153)$$

so

$$\iint_{I_2} \hat{q}(u)v^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}v}M_\lambda(2\tau-u-v)dudv \simeq \frac{\lambda^{\frac{3}{4}}e^{\frac{4}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}} \int_0^{\sqrt{\tau}} \hat{q}(u)e^{-\frac{2}{\sqrt{\lambda}}u} \int_{\sqrt{\tau}}^{\tau} \frac{1}{v^{\frac{3}{2}}(2\tau-u-v)^{\frac{3}{2}}}dvdu \quad (7.154)$$

At fixed $u \in [0, \sqrt{\tau}]$ With change of variable $s = \frac{v}{\tau}$ we find

$$\int_{\sqrt{\tau}}^{\tau} \frac{1}{v^{\frac{3}{2}}(2\tau-u-v)^{\frac{3}{2}}}dv = \int_{\frac{1}{\sqrt{\tau}}}^1 \frac{1}{\tau^{\frac{3}{2}}s^{\frac{3}{2}}(\tau(2-s)-u)^{\frac{3}{2}}}\tau ds = \frac{1}{\tau^2} \int_{\frac{1}{\sqrt{\tau}}}^1 \frac{1}{s^{\frac{3}{2}}\left((2-s)-\frac{u}{\tau}\right)^{\frac{3}{2}}}ds \quad (7.155)$$

Because $u \leq \sqrt{\tau}$ we have $\frac{u}{\tau} = o(1)$. Notice we have

$$\int_{\frac{1}{\sqrt{\tau}}}^1 \frac{ds}{s^{\frac{3}{2}}\left((2-s)-\frac{u}{\tau}\right)^{\frac{3}{2}}} = \left[\frac{2\left(\frac{u}{\tau}+2s-2\right)}{\left(\frac{u}{\tau}-2\right)^2\sqrt{s\left(2-s-\frac{u}{\tau}\right)}} \right]_{\frac{1}{\sqrt{\tau}}}^1 \quad (7.156)$$

$$= \frac{2}{\left(\frac{u}{\tau}-2\right)^2} \left[\frac{u}{\tau\sqrt{1-\frac{u}{\tau}}} - \frac{\frac{u}{\tau}+\frac{2}{\sqrt{\tau}}-2}{\sqrt{\frac{1}{\sqrt{\tau}}\left(2-\frac{1}{\sqrt{\tau}}-\frac{u}{\tau}\right)}} \right] \sim \frac{\tau^{\frac{1}{4}}}{\sqrt{2}} \quad (7.157)$$

Chapter 7. The rank-one model: a non-convex setting

Hence we have a term in $\tau^{-\frac{7}{4}}$ so the term on I_2 can be neglected compared to I_1 :

$$\iint_{I_2} \hat{q}(u)v^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}v}M_\lambda(2\tau-u-v)dudv \sim \frac{\lambda^{\frac{3}{4}}e^{\frac{4}{\sqrt{\lambda}}\tau}\tau^{-\frac{7}{4}}}{2\sqrt{2\pi}}\mathcal{L}\hat{q}\left(\frac{2}{\sqrt{\lambda}}\right) \quad (7.158)$$

Notice finally that the interval $I'_2 = [\sqrt{\tau}, \tau] \times [0, \sqrt{\tau}]$ is similar as the integrand is symmetric in its arguments.

Interval $I_3 = [\sqrt{\tau}, \tau]^2$ we can approximate both $\hat{q}(u), \hat{q}(v)$

$$\iint_{I_3} \hat{q}(u)\hat{q}(v)M_\lambda(2\tau-u-v)dudv \simeq \frac{\alpha^2 \iint_{I_3} (uv)^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}(u+v)}M_\lambda(2\tau-u-v)dudv}{4\pi\lambda^{\frac{1}{2}}\left[\left(1+\frac{1}{\lambda}\right)-\frac{2}{\sqrt{\lambda}}\right]^2} \quad (7.159)$$

Let's focus on the right hand side integral

$$f(\tau) = \iint_{I_3} (uv)^{-\frac{3}{2}}e^{\frac{2}{\sqrt{\lambda}}(u+v)}M_\lambda(2\tau-u-v)dudv \quad (7.160)$$

Now, using change of variable $u = \tau - x, v = \tau - y$ we have

$$f(\tau) = e^{\frac{4}{\sqrt{\lambda}}\tau} \iint_{[0, \tau(1-\frac{1}{\sqrt{\tau}})]^2} (\tau-x)^{-\frac{3}{2}}(\tau-y)^{-\frac{3}{2}}e^{\frac{-2}{\sqrt{\lambda}}(x+y)}M_\lambda(x+y)dx dy \quad (7.161)$$

with $s = x + y$

$$e^{\frac{4}{\sqrt{\lambda}}\tau}f(\tau) = \int_0^{\tau(1-\frac{1}{\sqrt{\tau}})} \int_x^{x+\tau(1-\frac{1}{\sqrt{\tau}})} (\tau-x)^{-\frac{3}{2}}(\tau-s+x)^{-\frac{3}{2}}e^{\frac{-2}{\sqrt{\lambda}}s}M_\lambda(s)ds dx \quad (7.162)$$

$$= \int_0^{2\tau(1-\frac{1}{\sqrt{\tau}})} \int_{\max(s-\tau(1-\frac{1}{\sqrt{\tau}}), 0)}^{\min(\tau(1-\frac{1}{\sqrt{\tau}}), s)} (\tau-x)^{-\frac{3}{2}}(\tau-s+x)^{-\frac{3}{2}}dx e^{\frac{-2}{\sqrt{\lambda}}s}M_\lambda(s)ds \quad (7.163)$$

$$= \int_0^{\tau(1-\frac{1}{\sqrt{\tau}})} \int_0^s (\tau-x)^{-\frac{3}{2}}(\tau-s+x)^{-\frac{3}{2}}dx e^{\frac{-2}{\sqrt{\lambda}}s}M_\lambda(s)ds \quad (7.164)$$

$$+ \int_{\tau(1-\frac{1}{\sqrt{\tau}})}^{2\tau(1-\frac{1}{\sqrt{\tau}})} \int_{s-\tau(1-\frac{1}{\sqrt{\tau}})}^{\tau(1-\frac{1}{\sqrt{\tau}})} (\tau-x)^{-\frac{3}{2}}(\tau-s+x)^{-\frac{3}{2}}dx e^{\frac{-2}{\sqrt{\lambda}}s}M_\lambda(s)ds \quad (7.165)$$

On the first integral, we find

$$\int_0^s (\tau-x)^{-\frac{3}{2}}(\tau-s+x)^{-\frac{3}{2}}dx = \left[\frac{2(2x-s)}{(2\tau-s)^2\sqrt{(\tau-x)(\tau+x-s)}} \right]_0^s = \frac{4s}{(2\tau-s)^2\sqrt{(\tau-s)\tau}} \quad (7.166)$$

However, $s \leq \tau - \sqrt{\tau}$ so $\sqrt{\tau} \leq \tau - s$ and $\tau + \sqrt{\tau} \leq 2\tau - s$ so:

$$\frac{4s}{(2\tau-s)^2\sqrt{(\tau-s)\tau}} \leq \frac{4\tau(1-\frac{1}{\sqrt{\tau}})}{\tau^2(1+\frac{1}{\sqrt{\tau}})^2\tau^{\frac{1}{2}}\tau^{\frac{1}{4}}} = 4\tau^{-\frac{7}{4}}(1+o_\tau(1)) \quad (7.167)$$

Therefore, we find:

$$\int_0^{\tau(1-\frac{1}{\sqrt{\tau}})} \int_0^s (\tau-x)^{-\frac{3}{2}} (\tau-s+x)^{-\frac{3}{2}} dx e^{\frac{-2}{\sqrt{\lambda}}s} M_\lambda(s) ds \leq 4\tau^{-\frac{7}{4}} (1 + o_\tau(1)) \mathcal{L}M_\lambda\left(\frac{2}{\sqrt{\lambda}}\right) \quad (7.168)$$

Noticeably, $\mathcal{L}M_\lambda\left(\frac{2}{\sqrt{\lambda}}\right) = -\sqrt{\lambda}G_{sc}(2) = \sqrt{\lambda}$. In the asymptotic limit, this term can be neglected due to $\tau^{-\frac{7}{4}}$ compared to $\tau^{-\frac{3}{2}}$.

Similarly, we find

$$\int_{s-\tau(1-\frac{1}{\sqrt{\tau}})}^{\tau(1-\frac{1}{\sqrt{\tau}})} (\tau-x)^{-\frac{3}{2}} (\tau-s+x)^{-\frac{3}{2}} dx = \frac{4(2\tau-s-2\sqrt{\tau})}{(2\tau-s)^2 \tau^{\frac{1}{4}} \sqrt{2\tau-\sqrt{\tau}-s}} \quad (7.169)$$

then, in $[\tau-\sqrt{\tau}, 2(\tau-\sqrt{\tau})]$ we approximate M_λ with its asymptotic expression. So we are left to evaluate

$$K(\tau) = \int_{\tau(1-\frac{1}{\sqrt{\tau}})}^{2\tau(1-\frac{1}{\sqrt{\tau}})} \frac{4(2\tau-s-2\sqrt{\tau})}{s^{\frac{3}{2}} (2\tau-s)^2 \tau^{\frac{1}{4}} \sqrt{2\tau-\sqrt{\tau}-s}} ds \quad (7.170)$$

Notice that $2\tau-s-2\sqrt{\tau} \leq \tau(1-\frac{1}{\sqrt{\tau}})$, and $2\tau^{\frac{1}{2}} \leq 2\tau-s$ and $\tau^{\frac{1}{2}} \leq 2\tau-\sqrt{\tau}-s$, hence

$$0 \leq K(\tau) \leq \frac{4\tau(1-\frac{1}{\sqrt{\tau}})}{(2\tau^{\frac{1}{2}})^2 \tau^{\frac{1}{4}} \sqrt{\tau^{\frac{1}{2}}}} \int_{\tau(1-\frac{1}{\sqrt{\tau}})}^{2\tau(1-\frac{1}{\sqrt{\tau}})} \frac{ds}{s^{\frac{3}{2}}} \quad (7.171)$$

So

$$0 \leq K(\tau) \leq \frac{(1-\frac{1}{\sqrt{\tau}})}{\tau^{\frac{1}{2}}} \left[-\frac{2}{s^{\frac{1}{2}}} \right]_{\tau(1-\frac{1}{\sqrt{\tau}})}^{2\tau(1-\frac{1}{\sqrt{\tau}})} \quad (7.172)$$

with a change of variable $u = s - (\tau - \sqrt{\tau})$ we find

$$K(\tau) = \frac{1}{\tau^{\frac{1}{4}}} \int_0^{\tau(1-\frac{1}{\sqrt{\tau}})} \frac{4(\tau(1-\frac{1}{\sqrt{\tau}}) - u)}{(\tau(1-\frac{1}{\sqrt{\tau}}) + u)^{\frac{3}{2}} (\tau(1+\frac{1}{\sqrt{\tau}}) - u)^2 \sqrt{\tau-u}} du \quad (7.173)$$

with another change of variable $u = \tau r$ we find:

$$K(\tau) = \frac{4}{\tau^{\frac{9}{4}}} \int_0^{1-\frac{1}{\sqrt{\tau}}} \frac{(1-\frac{1}{\sqrt{\tau}}-r)}{(1-\frac{1}{\sqrt{\tau}}+r)^{\frac{3}{2}} (1+\frac{1}{\sqrt{\tau}}-r)^2 \sqrt{1-r}} dr \quad (7.174)$$

Though this integral can be completely solved, we are only interested in bounding it. In particular, we find:

$$K(\tau) \leq \frac{4}{\tau^{\frac{9}{4}}} \int_0^{1-\frac{1}{\sqrt{\tau}}} \frac{(1-r)}{(1-\frac{1}{\sqrt{\tau}})^{\frac{3}{2}} (1-r)^2 \sqrt{1-r}} dr = \frac{4}{\tau^{\frac{9}{4}} (1-\frac{1}{\sqrt{\tau}})^{\frac{3}{2}}} \int_0^{1-\frac{1}{\sqrt{\tau}}} \frac{dr}{(1-r)^{\frac{3}{2}}} \quad (7.175)$$

So

$$K(\tau) \leq \frac{4}{\tau^{\frac{9}{4}}(1-\frac{1}{\sqrt{\tau}})^{\frac{3}{2}}} \left[\frac{2}{(1-r)^{\frac{1}{2}}} \right]_0^{1-\frac{1}{\sqrt{\tau}}} = \frac{8(1-\frac{1}{\sqrt{\tau}})}{\tau^{\frac{8}{4}}(1-\frac{1}{\sqrt{\tau}})^{\frac{3}{2}}} = 8\tau^{-\frac{8}{4}}(1+o(1)) \quad (7.176)$$

In the end, the integral on I_3 can also be neglected.

conclusion summing up all the main contributions from (7.141), (7.148) and (7.151) we find

$$\lim_{\tau \rightarrow \infty} \tau^{\frac{3}{2}} e^{-\frac{4\tau}{\sqrt{\lambda}}} \hat{p}(\tau) = \frac{\lambda^{\frac{3}{4}}}{2^{\frac{5}{2}}\sqrt{\pi}} + \frac{\alpha^2 \lambda^{\frac{3}{4}}}{2^{\frac{3}{2}}\sqrt{\pi}(\frac{1}{\sqrt{\lambda}}-1)} + \frac{\alpha^2 \lambda^{\frac{3}{4}}}{2^{\frac{5}{2}}\sqrt{\pi}(\frac{1}{\sqrt{\lambda}}-1)^2} \quad (7.177)$$

$$= \frac{\lambda^{\frac{3}{4}}}{2^{\frac{5}{2}}\sqrt{\pi}} \left[1 + \alpha^2 \left(\frac{2}{\frac{1}{\sqrt{\lambda}}-1} + \frac{1}{(\frac{1}{\sqrt{\lambda}}-1)^2} \right) \right] \quad (7.178)$$

and thus:

$$\frac{1}{\sqrt{\hat{p}(\tau)}} \sim \frac{2^{\frac{5}{4}}\pi^{\frac{1}{4}}}{\lambda^{\frac{3}{8}}} \left[1 - \alpha^2 + \frac{\alpha^2}{\lambda(\frac{1}{\sqrt{\lambda}}-1)^2} \right]^{-\frac{1}{2}} \tau^{\frac{3}{4}} e^{-\frac{2}{\sqrt{\lambda}}\tau} \quad (7.179)$$

Using back (7.140) we find

$$\bar{q}(\tau) \sim \frac{\alpha \left(\frac{2}{\pi}\right)^{\frac{1}{4}}}{\lambda^{\frac{5}{8}} \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2 \sqrt{1 - \alpha^2 + \frac{\alpha^2}{\lambda(\frac{1}{\sqrt{\lambda}}-1)^2}}} \tau^{-\frac{3}{4}} \quad (7.180)$$

Numerical evaluations from the functions of theorem 7.1 match correctly this expression for different values of (α, λ) , see Figure 7.I.1 (a) for instance.

7.I.3 Asymptotic analysis of $\lambda > 1$

Using the previous analysis for $\hat{q}(\tau)$ ((7.122) and (7.140)), we have an additional term:

$$\hat{q}(\tau) = \alpha \left(1 - \frac{1}{\lambda}\right) e^{(1+\frac{1}{\lambda})\tau} + \frac{\alpha \tau^{-\frac{3}{2}} e^{\frac{2}{\sqrt{\lambda}}\tau}}{2\sqrt{\pi}\lambda^{\frac{1}{4}} \left(1 - \frac{1}{\sqrt{\lambda}}\right)^2} + o(\tau^{-\frac{3}{2}} e^{\frac{2}{\sqrt{\lambda}}\tau}) \quad (7.181)$$

Now, for $\hat{p}(\tau)$, we have already seen the leading asymptotics in equation (7.121). For the next correction, we postulate through computer analysis that there exists a non-null constant $C \in \mathbb{R}_+^*$ such that it takes the form:

$$\hat{p}(\tau) = \alpha^2 \left(1 - \frac{1}{\lambda}\right) e^{2(1+\frac{1}{\lambda})\tau} \left[1 - 2\tau^{-\frac{3}{2}} e^{-2(1-\frac{1}{\sqrt{\lambda}})^2\tau} (C + o(1)) \right] \quad (7.182)$$

Hence the expression:

$$\frac{1}{\sqrt{\hat{p}(\tau)}} = \frac{e^{-(1+\frac{1}{\lambda})\tau}}{|\alpha|\sqrt{1-\frac{1}{\lambda}}} \left[1 + \tau^{-\frac{3}{2}} e^{-2(1-\frac{1}{\sqrt{\lambda}})^2\tau} (C + o(1)) \right] \quad (7.183)$$

Putting things together, we find:

$$\bar{q}(\tau) = \text{sign}(\alpha) \sqrt{1-\frac{1}{\lambda}} \left(1 + \frac{\tau^{-\frac{3}{2}} e^{-(1-\frac{1}{\sqrt{\lambda}})^2\tau} (1 + o(1))}{2(1-\frac{1}{\lambda})\sqrt{\pi}\lambda^{\frac{1}{4}} \left(1-\frac{1}{\sqrt{\lambda}}\right)^2} \right) \left(1 + \tau^{-\frac{3}{2}} e^{-2(1-\frac{1}{\sqrt{\lambda}})^2\tau} (C + o(1)) \right) \quad (7.184)$$

Hence the exponential term in the expression of \hat{q} dominates the one in the expression of \hat{p} . Therefore, expanding the asymptotic expansion provides the result:

$$\bar{q}(\tau) - \text{sign}(\alpha) \sqrt{1-\frac{1}{\lambda}} \sim \frac{\text{sign}(\alpha)}{2\sqrt{\pi}\lambda^{\frac{1}{4}} \sqrt{1-\frac{1}{\lambda}} \left(1-\frac{1}{\sqrt{\lambda}}\right)^2} \tau^{-\frac{3}{2}} e^{-(1-\frac{1}{\sqrt{\lambda}})^2\tau} \quad (7.185)$$

More specifically, equation (7.184) shows that the second order term of \hat{q} dominates the one of $\frac{1}{\sqrt{\hat{p}}}$ when we compute the final contribution in equation (7.185). Therefore, this fact can be emphasized with the equivalent limiting behavior:

$$\bar{q}(\tau) - \text{sign}(\alpha) \sqrt{1-\frac{1}{\lambda}} \sim \frac{1}{|\alpha|\sqrt{1-\frac{1}{\lambda}}} \left(\hat{q}(\tau) e^{-(1+\frac{1}{\lambda})\tau} - \alpha \left(1-\frac{1}{\lambda}\right) \right) \quad (7.186)$$

This form is actually more convenient because a numerical evaluation $\bar{q}(\tau)$ for large τ requires extra precision and computational resources due to the double-integral within the $\hat{p}(\tau)$ term. Therefore, it appears to be easier to observe the equivalent behavior in (7.186) rather than in (7.185). To illustrate this phenomenon, one can evaluate:

$$\psi(\tau) = |\alpha| \sqrt{1-\frac{1}{\lambda}} \left(\bar{q}(\tau) - \text{sign}(\alpha) \sqrt{1-\frac{1}{\lambda}} \right) e^{(1-\frac{1}{\sqrt{\lambda}})^2\tau} \quad (7.187)$$

$$\phi(\tau) = \left(\hat{q}(\tau) e^{-(1+\frac{1}{\lambda})\tau} - \alpha \left(1-\frac{1}{\lambda}\right) \right) e^{(1-\frac{1}{\sqrt{\lambda}})^2\tau} \quad (7.188)$$

$$\mathcal{A}(\tau) = \frac{\alpha}{2\sqrt{\pi}\lambda^{\frac{1}{4}} \left(1-\frac{1}{\sqrt{\lambda}}\right)^2} \tau^{-\frac{3}{2}} \quad (7.189)$$

and expect to observe $\psi(\tau) \sim \phi(\tau) \sim \mathcal{A}(\tau)$ when $\tau \rightarrow \infty$ for any $\lambda > 1$ and $\alpha \neq 0$. See Figure 7.1.1 (b) as an example where the computation of $\psi(\tau)$ had to be stopped earlier in time to cope with computational limits of the math library Scipy.

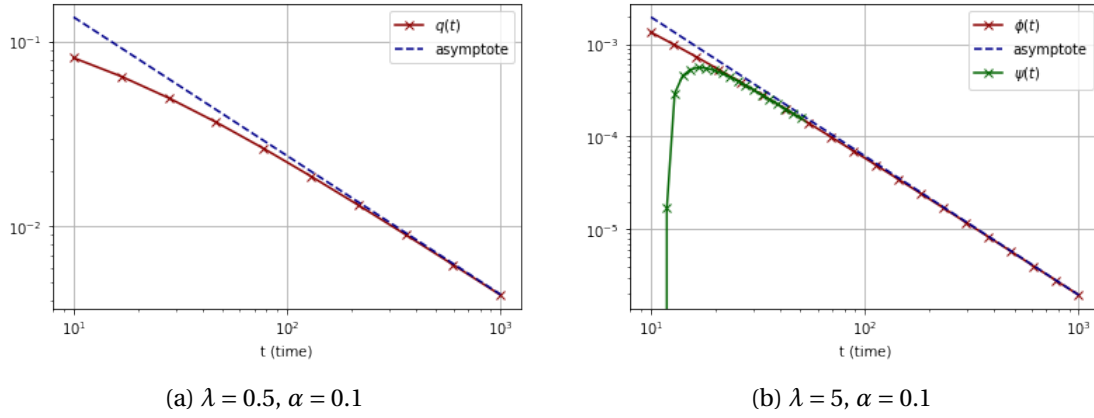


Figure 7.1.1: Example of a numerical evaluation of theorem 7.1 and comparisons with their respective asymptotes in log-scale for $\lambda < 1$ in (a) and $\lambda > 1$ in (b).

7.1.4 Asymptotic analysis for $\lambda = 1$

In the special case $\lambda = 1$ where the regime changes, one can write explicitly:

$$\hat{q}(\tau) = \alpha e^{2\tau} (1 - 1 + e^{-2\tau} (I_0(2\tau) + I_1(2\tau))) = \alpha [I_0(2\tau) + I_1(2\tau)] \quad (7.190)$$

and we find the first term of the asymptotic expansion in $\tau \rightarrow \infty$:

$$\hat{q}(\tau) \sim \alpha \frac{e^{2\tau}}{\sqrt{\pi\tau}}. \quad (7.191)$$

Some further analysis lead us to a similar estimate for $\hat{p}(\tau)$

$$\sqrt{\hat{p}(\tau)} \sim |\alpha| \frac{e^{2t}}{(2\pi\tau)^{\frac{1}{4}}} \quad (7.192)$$

and thus to conclude using (7.191) (for $\alpha > 0$):

$$\bar{q}(\tau) \sim \left(\frac{2}{\pi\tau}\right)^{\frac{1}{4}} \quad (7.193)$$

Using similar arguments as the case $\lambda < 1$ (see Section 7.1.2), we can check that the main asymptotic contribution in $\tau^{-\frac{1}{2}}$ comes from the third term of \hat{p} on the interval I_3 . Indeed, the first term in $M_1(2\tau)$ is obviously in $\tau^{-\frac{3}{2}}$. The second term can also be neglected, notice that we have:

$$\int_{\sqrt{\tau}}^{\tau} \hat{q}(s) \frac{e^{-2s}}{(2\tau - s)^{\frac{3}{2}}} ds \sim \frac{\alpha}{\sqrt{\pi}} \int_{\sqrt{\tau}}^{\tau} \frac{ds}{\sqrt{s} \sqrt{(2\tau - s)^{\frac{3}{2}}}} \sim \frac{\alpha}{\sqrt{\pi\tau}} \quad (7.194)$$

Also we don't have a constant term with the laplace transform of \hat{q} . Instead for any $t > 0$

$$\int_0^t \hat{q}(s) e^{-2s} ds = \frac{\alpha}{2} (e^{-2t} (1 + 4t) I_0(2t) + 4t e^{-2t} I_1(2t) - 1) \quad (7.195)$$

In particular when $t = \sqrt{\tau}$ and $\tau \rightarrow \infty$:

$$\int_0^{\sqrt{\tau}} \hat{q}(s) e^{-2s} ds \sim \frac{4\alpha\sqrt{\tau}}{\sqrt{4\pi\sqrt{\tau}}} \sim \frac{2\alpha\tau^{\frac{1}{4}}}{\sqrt{\pi}} \quad (7.196)$$

Hence with the additional term in $\tau^{-\frac{3}{2}}$ this gives a term in $\tau^{-\frac{5}{4}}$. We proceed similarly for the third term with the 4 segments I_1, I_2, I'_2, I_3 .

Interval $I_1 = [0, \sqrt{\tau}]^2$ Similar considerations using the result (7.196) lead to the asymptotics:

$$\iint_{I_1} \hat{q}(u)\hat{q}(v)e^{-2(u+v)} dudv \sim \frac{4\alpha^2\tau^{\frac{1}{2}}}{\pi} \quad (7.197)$$

Hence with the additional term in $\tau^{-\frac{3}{2}}$ this gives a term in τ^{-1} .

Interval $I_2 = [0, \sqrt{\tau}] \times [\sqrt{\tau}, \tau]$ We get:

$$\iint_{I_2} \hat{q}(u)\hat{q}(v)M_1(2\tau - u - v)dudv \simeq \frac{\alpha}{2\pi} \iint_{I_2} \hat{q}(u) \frac{e^{2v}}{\sqrt{v}} \frac{e^{2(2\tau - u - v)}}{(2\tau - u - v)^{\frac{3}{2}}} dudv \quad (7.198)$$

We can compute further the integral considering $u = o(\tau)$:

$$\begin{aligned} \int_v \frac{1}{\sqrt{v}(2\tau - u - v)^{\frac{3}{2}}} dv &= \frac{2}{2\tau - u} \left[\sqrt{\frac{v}{2\tau - u - v}} \right]_{\sqrt{\tau}}^{\tau} \\ &= \frac{2}{2\tau - u} \left[\sqrt{\frac{\tau}{2\tau - u}} - \sqrt{\frac{\sqrt{\tau}}{2\tau - u - \sqrt{\tau}}} \right] \\ &\sim \tau^{-1} \end{aligned} \quad (7.199)$$

Finally, using (7.196) gives:

$$\iint_{I_2} \hat{q}(u)\hat{q}(v)M_1(2\tau - u - v)dudv \sim \frac{\alpha^2}{\pi^{\frac{3}{2}}} e^{4\tau} \tau^{-\frac{3}{4}} \quad (7.200)$$

Interval $I_3 = [\sqrt{\tau}, \tau]^2$ On this interval we have:

$$\iint_{I_3} \hat{q}(u)\hat{q}(v)M_1(2\tau - u - v)dudv \simeq \frac{\alpha^2}{\pi} \iint_{I_3} e^{2(u+v)} \frac{I_1(2(2\tau - u - v))}{(2\tau - u - v)\sqrt{uv}} dudv \quad (7.201)$$

Let's focus on the right hand side integral:

$$f(\tau) = \iint_{I_3} e^{2(u+v)} \frac{I_1(2(2\tau - u - v))}{(2\tau - u - v)\sqrt{uv}} dudv \quad (7.202)$$

With $x = \tau - u$, $y = \tau - v$ we find:

$$e^{-4\tau} f(\tau) = \iint_{[0, \tau - \sqrt{\tau}]^2} e^{-2(x+y)} \frac{I_1(2(x+y))}{(x+y)\sqrt{(\tau-x)(\tau-y)}} dx dy \quad (7.203)$$

Now, consider further the change of variable: $x = (\tau - \sqrt{\tau})r$ and $y = (\tau - \sqrt{\tau})s$. we have:

$$\sqrt{\tau} e^{-4\tau} f(\tau) = \iint_{[0,1]^2} \sqrt{\tau} \frac{e^{-2(\tau-\sqrt{\tau})(r+s)} I_1(2(\tau-\sqrt{\tau})(r+s))}{(r+s)\sqrt{\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - r\right)\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - s\right)}} dr ds \quad (7.204)$$

Now, for all $r, s \in [0, 1]^2 \setminus \{(0,0)\}$, we have:

$$\lim_{\tau \rightarrow \infty} \sqrt{\tau} \frac{e^{-2(\tau-\sqrt{\tau})(r+s)} I_1(2(\tau-\sqrt{\tau})(r+s))}{(r+s)\sqrt{\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - r\right)\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - s\right)}} = \frac{1}{\sqrt{4\pi}(r+s)^{\frac{3}{2}}\sqrt{(1-r)(1-s)}} \quad (7.205)$$

and it can be shown that this function is integrable:

$$\iint_{[0,1]^2} \frac{dr ds}{\sqrt{4\pi}(r+s)^{\frac{3}{2}}\sqrt{(1-r)(1-s)}} = \sqrt{\frac{\pi}{2}} \quad (7.206)$$

Further, for all $r, s \in [0, 1]^2 \setminus \{(0,0)\}$ and for instance $\tau \geq 4$:

$$\sqrt{\tau} I_1(2(\tau-\sqrt{\tau})(r+s)) \leq \frac{1}{\sqrt{4\pi}\left(1-\frac{1}{\sqrt{\tau}}\right)(r+s)} \leq \frac{\sqrt{2}}{\sqrt{4\pi}(r+s)} \quad (7.207)$$

and

$$\frac{1}{\sqrt{\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - r\right)\left(\frac{1}{1-\frac{1}{\sqrt{\tau}}} - s\right)}} \leq \frac{1}{\sqrt{(1-r)(1-s)}} \quad (7.208)$$

Hence for all $\tau \geq 4$, the integrand is dominated by its limit times $\sqrt{2}$.

In conclusion, we have the main contribution term

$$\iint_{I_3} \hat{q}(u)\hat{q}(v)M_1(2\tau-u-v)dudv \sim \frac{\alpha^2}{\sqrt{2\pi\tau}} \quad (7.209)$$

7.1.5 Asymptotic analysis conclusion

We have seen the case $\lambda < 1$ in (7.180) and $\lambda > 1$ in (7.185). So compared to the first case $\lambda < 1$, the convergence towards the limit is reached with an exponential term $\exp\{-(1-\frac{1}{\sqrt{\lambda}})^2\tau\}$ in the asymptotic limit for $\lambda > 1$. It confirms the result that the convergence happens faster as λ grows to infinity, and that the exponential term vanishes as λ gets close to 1 - with an additional singularity in the denominator.

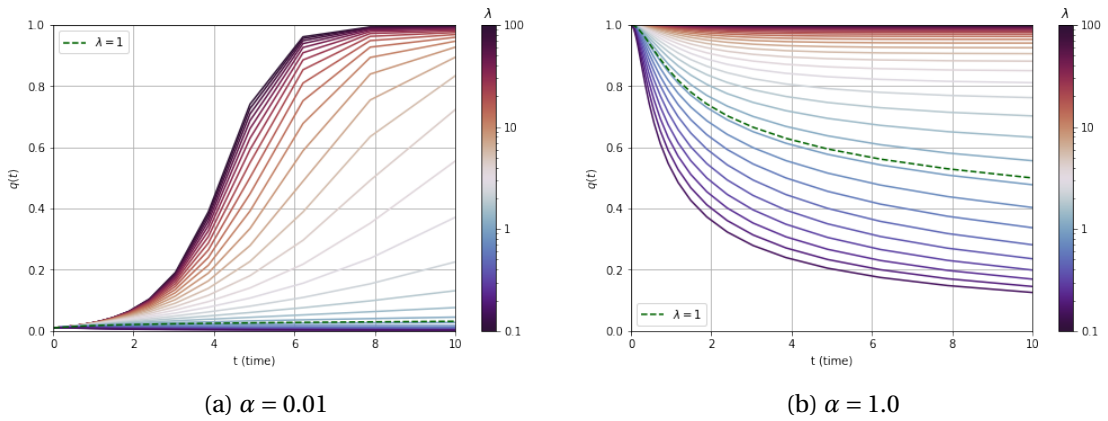


Figure 7.J.1: Comparison of the overlap over time with different configurations of λ parameter, and between two different values of α .

7.J Additional experiments

A python notebook is available in the supplementary material to reproduce all the examples.

7.J.1 Limiting gradient descent

We illustrate the predicted time evolution for cases α very close to 0 and α very close to 1 in Figure 7.J.1. Since $\alpha = 0$ leads to a null overlap evolution, a slight non-zero initial value of α is required to initiate the learning algorithm. The smaller the α the more is the asymptotic regime delayed. The opposite case $\alpha = 1$ brings another insight, namely when $\theta_0 = \pm\theta^*$ the effect of the noise inexorably disturbs the signal towards a lower limiting overlap (for $\lambda < \infty$).

7.J.2 Comparison with experimental gradient descent algorithm

The theoretical gradient descent prediction is compared with the experimental values when taking the data dimension n sufficiently large over multiple runs with new samples of the noise matrix. Discrete step size gradient descent is performed while keeping θ_t on $\mathcal{S}_d(\sqrt{n})$. We choose a $\delta_t > 0$ sufficiently small and consider discrete times $t_k = k\delta_t$ for $k \in \mathbb{N}$. We update θ_{t_k} in two steps: first with the gradient descent $\theta_{t_k + \frac{\delta_t}{2}} = \theta_{t_k} - \eta\delta_t \nabla \mathcal{H}(\theta_{t_k})$, and secondly projecting back on the sphere $\theta_{t_{k+1}} = \sqrt{n}\theta_{t_k + \frac{\delta_t}{2}} \|\theta_{t_k + \frac{\delta_t}{2}}\|^{-1}$. These steps are implemented using Tensorflow in Python and run seamlessly on a standard single computer configuration. The initial vectors θ_0 and θ^* are chosen deterministically as $\sqrt{n}\theta_0 = \alpha e_1 + \sqrt{1 - \alpha^2}e_2$ and $\sqrt{n}\theta^* = e_1$ with $(e_i)_{1 \leq i \leq n}$ the canonical basis of \mathbb{R}^n , while the noise matrix H is generated randomly. To account for the randomness of H at each execution, we perform 100 runs and give the quantiles for quantities of interest.

As shown in Figure 7.J.2, the learning curve matches the theoretical limiting curve with some fluctuations. As illustrated below, these fluctuations diminish as n is increased. Noticeably,

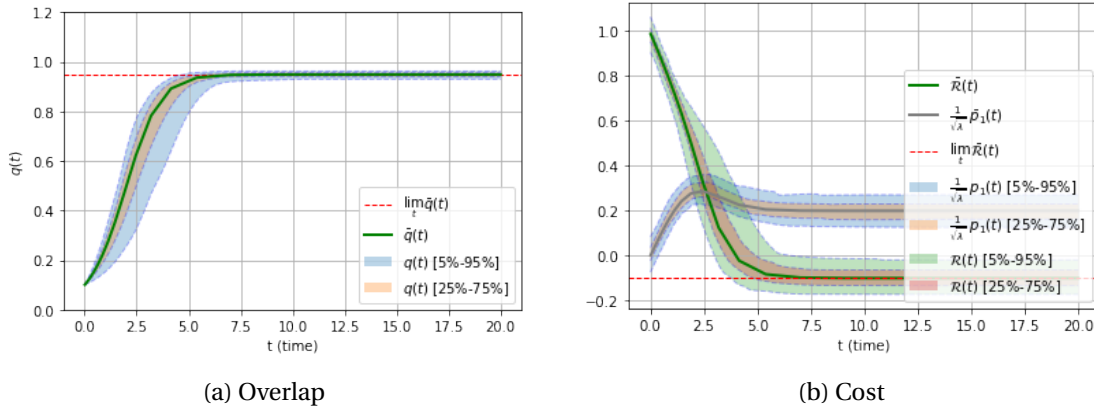


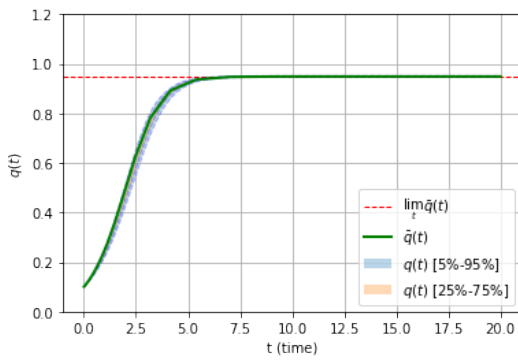
Figure 7.J.2: $\lambda = 10$, $n = 70$, $\alpha = 0.1$, $\delta_t = 0.1$

in the regime where $\lambda > 1$, smaller values of λ require higher values of n to keep the same concentration. Therefore, the formula from theorem 7.1 provides a good theoretical framework to predict the behavior of the experimental learning algorithm. Such formulas potentially allow to benchmark the time-evolution of gradient descent techniques and provide guidelines for early-stopping commonly used in machine learning.

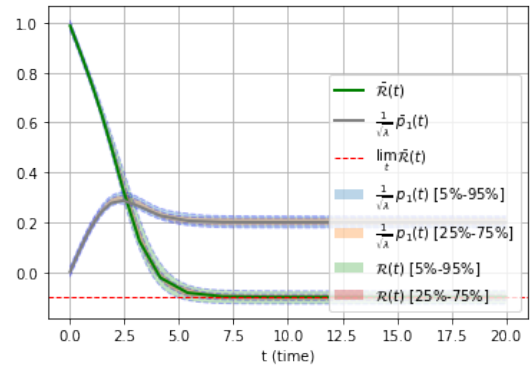
We provide a range of further different experiments for different values of λ , α , n .

Let us first comment the regime $\lambda > 1$ illustrated on Figures 7.J.3, 7.J.4, 7.J.5. Figure 7.J.3 clearly shows that increasing n up to 1000 concentrates the experimental curves around the expected limiting overlap and cost \bar{q} , $\bar{\mathcal{R}}$. We also see even more clearly the characteristic change of p_1 with a "self-healing" process at some specific point in the dynamics of the learning algorithm (recall that p_1 is a similarity measure between the reconstructed matrix $\theta_t \theta_t^T$ and the noise matrix H). This is also seen in Figures 7.J.4 and 7.J.5 for different values of λ and α . Figures 7.J.3 and 7.J.4 only differ in the value λ : we observe that decreasing this parameter closer to 1 not only decreases the overlap, but also increases the deviation from the limiting theoretical overlap \bar{q} - and thus as λ decreases higher values of n would thus be needed to match closely \bar{q} .

Finally, in the regime $\lambda < 1$, we observe on Figure 7.J.6 that similarity measure p_1 explodes and overtakes the risk.

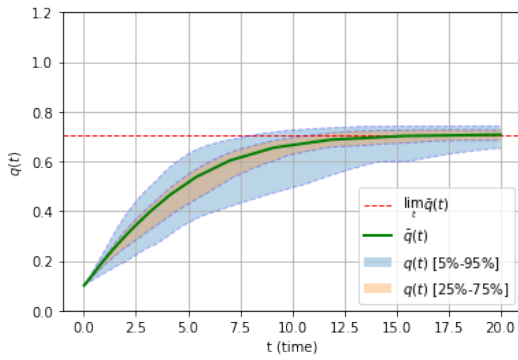


(a) Overlap

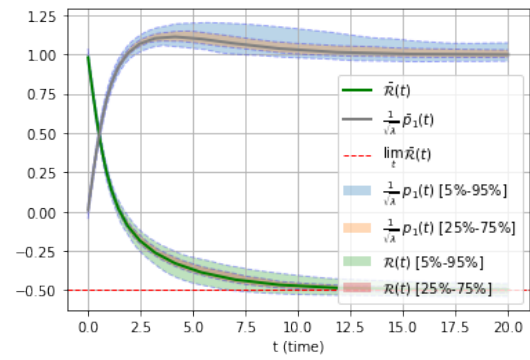


(b) Cost

Figure 7.J.3: $\lambda = 10, n = 1000, \alpha = 0.1, \delta_t = 0.1$

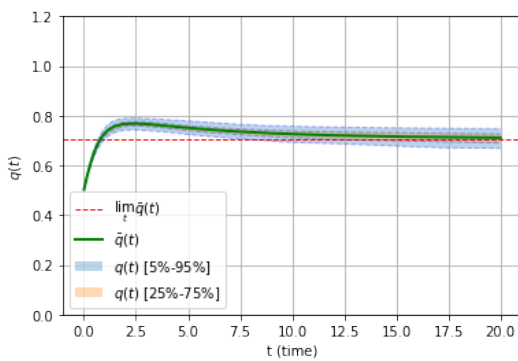


(a) Overlap

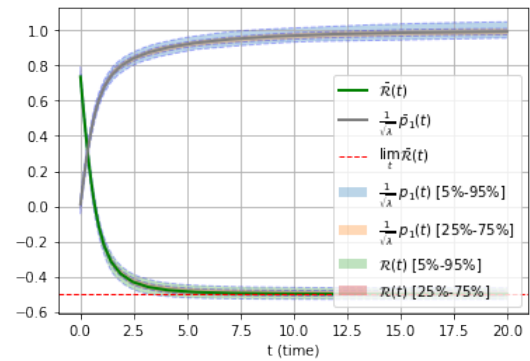


(b) Cost

Figure 7.J.4: $\lambda = 2, n = 1000, \alpha = 0.1, \delta_t = 0.1$



(a) Overlap



(b) Cost

Figure 7.J.5: $\lambda = 2, n = 1000, \alpha = 0.5, \delta_t = 0.1$

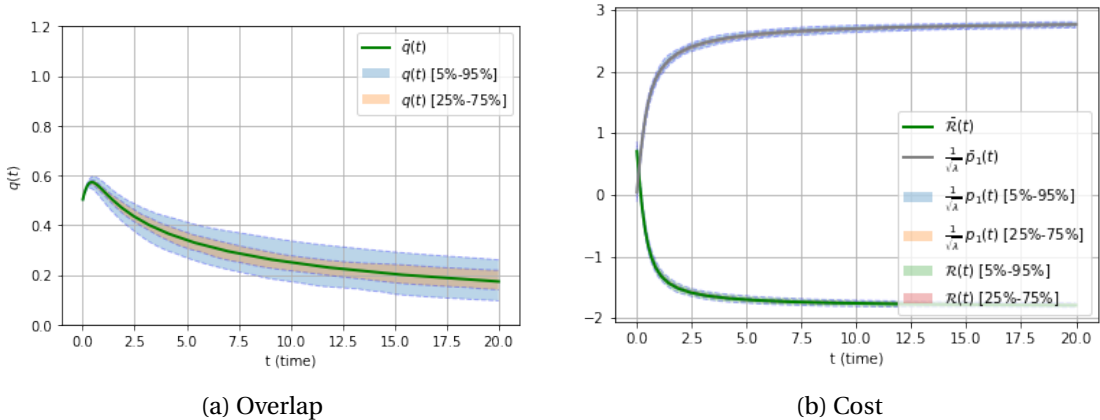


Figure 7.J.6: $\lambda = 0.5, n = 1000, \alpha = 0.5, \delta_t = 0.1$. Note the different scale for the cost.

8 Matrix denoising: an extensive rank model

This work is based on the contribution (Bodin and Macris, 2023) and investigates model 1.5 described in the introduction. We present a new approach to analyze the gradient flow for a positive semi-definite matrix denoising problem in an extensive-rank and high-dimensional regime. We use recent linear pencil techniques of random matrix theory to derive fixed point equations which track the complete time evolution of the matrix-mean-square-error of the problem. The predictions of the resulting fixed point equations are validated by numerical experiments. In this short note we briefly illustrate a few predictions of our formalism by way of examples, and in particular we uncover continuous phase transitions in the extensive-rank and high-dimensional regime, which connect to the classical phase transitions of the low-rank problem in the appropriate limit. The formalism has much wider applicability than shown in this communication.

8.1 Introduction

Matrix denoising and factorization play a crucial role in a variety of data science tasks such as matrix sensing, phase retrieval or synchronisation, or matrix completion. The problem consists in reducing the amount of noise or irrelevant information present in a dataset, allowing for more accurate analysis and interpretation of the data, as well as better computational efficiency and modeling by way of dimensionality reduction. The literature on the subject is immense and we refer to (Chen and Chi, 2018; Chi et al., 2019b) for recent overviews of applications and theory in various settings and formulations.

In this contribution we focus on the study of gradient-flow for the following statistical formulation for positive definite matrix denoising. We consider a "ground truth" signal $X^* \in \mathbb{R}^{n \times d}$ with randomly sampled independent entries $X_{ij}^* \sim \mathcal{N}(0, \frac{1}{n})$ where the dimensions n, d are such that $\phi = \frac{d}{n}$ is fixed. Then we define the corrupted data matrix $Y \in \mathbb{R}^{n \times n}$

$$Y = X^* X^{*T} + \frac{1}{\sqrt{\lambda}} \xi \tag{8.1}$$

Chapter 8. Matrix denoising: an extensive rank model

where ξ is an additive symmetric random noise with $\xi_{ij} = \xi_{ji} \sim \mathcal{N}(0, \frac{1}{n})$ and λ is (proportional to) the signal-to-noise ratio. The objective is to estimate the ground truth positive semi-definite matrix $X^* X^{*T}$ from the corrupted data matrix Y with a matrix XX^T such that $X \in \mathbb{R}^{n \times m}$ where m is set from the fixed ratio $\psi = \frac{m}{n}$. Note that we allow d and m to be different. The estimator studied in this contribution is given by the gradient flow $X(t)$ (t is time) for an objective function with regularization parameter μ , defined as

$$\mathcal{H}(X) = \frac{1}{4d} \|Y - XX^T\|_F^2 + \frac{\mu}{2d} \|X\|_F^2 \quad (8.2)$$

where $\|\cdot\|_F$ is the Frobenius norm. The initialization of gradient flow is $X(0) = X_0 \in \mathbb{R}^{n \times m}$ random with i.i.d matrix elements from $\mathcal{N}(0, \frac{1}{n})$. As a measure of performance we adopt the expected matrix-mean-square-error

$$\mathbb{E}\mathcal{E} = \frac{1}{d} \mathbb{E} \|X^* X^{*T} - XX^T\|_F^2 \quad (8.3)$$

where the expectation is over ξ , X^* , X_0 . Note that the objective function and performance measure are not the same and can be thought of as "training" and "generalization" errors in the language of machine learning.

Summary of main contributions:

- We derive a set of analytical fixed point equations whose solutions allow to compute the full performance curve $t \rightarrow \mathbb{E}\mathcal{E}_t$ for the extensive-rank and high-dimensional regime where m, d, n all tend to infinity while ϕ, ψ are kept fixed (results 1 and 2 in Sec. 8.2). Continuous time average behaviour of gradient flow is a proxy for the usual discrete gradient descent algorithm, and has the advantage that it is more amenable to analytical study. The numerical experiments confirm that (a) \mathcal{E}_t concentrates over its expectation; (b) theoretical predictions of gradient flow agree with gradient descent. See Fig. 8.2.1.
- We further push the analysis of these equations in the time limit $t = +\infty$ and display specific examples where a critical value λ_c can be calculated such that: (a) for $\lambda \leq \lambda_c$ the performance error of gradient flow is no better than the one of the null-estimator $X = 0$; (b) for $\lambda > \lambda_c$ better estimation is possible; (c) the phase transition between the two regimes is a continuous type phase transition. These results are displayed on Fig. 8.2.2.
- We analyze the limit $\phi = \psi \rightarrow 0$ (after n, m, d have been sent to infinity) and derive a connection with the low-rank setting. It turns out that the matrix-mean-square-error curve (when $t \rightarrow +\infty$) tends to the one of the rank-one problem and the phase transition reduces to the well known BBP transition at $\lambda_c = 1$.

We use tools based on modern results in random matrix theory. Central to our derivations, is the formalism of *linear-pencils*, that initially appeared in (Rashidi Far et al., 2006; Mingo and Speicher, 2017) and has been further improved recently in the context of neural networks

(Adlam and Pennington, 2020a; Bodin and Macris, 2021a, 2022). In particular we make use of extensions provided in (Bodin and Macris, 2022) to derive closed-form expressions of non-trivial averages over ξ , X^* , X_0 appearing in traces of complicated "rational" expressions of these random matrices. Although these techniques have not yet always been rigorously proven they have been used successfully in different applications, and the predictions are confirmed by numerical experiments. In addition, we use holomorphic functional calculus for matrices (Dunford and Schwartz, 1988).

Brief review of literature: The full time-evolution of gradient flow for the rank-one problem (the so-called spiked Wigner model with $d = m = 1$) has been solved and rigorously analyzed in much the same spirit as the present work in (Bodin and Macris, 2021b) with the difference that the spike is constrained to lie on a sphere all along the evolution. For the present extensive-rank setting rigorous or even analytical results on the whole time-evolution are scarce. Closely connected to our work is the recent paper (Tarmoun et al., 2021). An essential difference however is that in (Tarmoun et al., 2021) the initialization $X(0) = X_0$ is taken to have eigenvectors aligned with those of Y (this pre-processing can be implemented empirically in practice). Moreover the authors do not carry out the random matrix averages fully analytically. Gradient flow has been studied in a variety of settings more or less related to the present one, see (Gunasekar et al., 2017; Chou et al., 2020; Saxe et al., 2013; Mannelli et al., 2019; Arous et al., 2022; Liang et al., 2022).

Bayesian approaches are quite well understood for the low-rank problem (mainly rank-one). This context is quite different from the present one. To begin with it is not dynamical. One studies the Minimum-Mean-Square-Estimator (MMSE) computed as the conditional expectation of the signal with respect to the Bayesian posterior probability distribution (Montanari and Richard, 2014; Lelarge and Miolane, 2019; Luneau et al., 2020; Barbier and Macris, 2019; Miolane, 2017; Pourkamali and Macris, 2022b,a; Camilli et al., 2022; Barbier et al., 2022). Bayesian-optimal as well as mismatched estimation settings have been well studied with rigorous results on the mutual information, the MMSE, the cross-entropy, and the problem displays a rich phenomenology of first and higher order phase transitions depending on the nature of the priors. Related dynamics of the Approximate Message Passing (AMP) algorithms is also well understood for these problems (Lesieur et al., 2017b,a; Montanari and Venkataramanan, 2017). The realm of extensive-rank within such Bayesian and AMP approaches is quite open and very timely (Kabashima et al., 2016; Barbier and Macris, 2022; Maillard et al., 2022; Troiani et al., 2022; Camilli and Mézard, 2022).

Finally other types of non-dynamical approach belong to the class of spectral methods like Principal Component Analysis (PCA). The low rank case is covered by (Baik et al., 2005a; Péché, 2004; Benaych-Georges and Nadakuditi, 2011). For the extensive-rank setting the results are scarce and little is known except for ensembles of rotation invariant signals for which an interesting class of Rotation Invariant Estimators (RIE) has been proposed (Bun et al., 2017).

8.2 Results

8.2.1 Preliminaries

We simplify the notations by introducing the variables $Z = XX^T$ and $Z^* = X^*X^{*T}$ and the order parameters p and q such that $\mathcal{E} = r - 2q + p$ with:

$$q = \frac{1}{d} \text{Tr}[Z^* Z] \quad p = \frac{1}{d} \text{Tr}[Z^2] \quad r = \frac{1}{d} \text{Tr}[(Z^*)^2] \quad (8.4)$$

In the rank-one setting, p can be seen as a norm of the estimator while q represents the angle with the ground-truth. We consider the gradient flow

$$\frac{dX_t}{dt} = -\phi \nabla \mathcal{H}(X_t) \quad (8.5)$$

and track the evolution of the matrix mean-square error \mathcal{E}_t through the quantities q_t and p_t . The factor ϕ amounts to a rescaling of time which leads to more convenient expressions. With the additional notation $H = Y - \mu I_n$, expanding the gradient provides: $\frac{dX_t}{dt} = (H - Z_t)X_t$, which in turns provides the matrix Riccati differential equation:

$$\frac{dZ_t}{dt} = HZ_t + Z_tH - 2Z_t^2 \quad (8.6)$$

A general solution of this matrix differential equation is (see e.g., (Tarmoun et al., 2021)):

$$Z_t = e^{tH} X_0 \left(I_m + 2X_0^T \int_0^t e^{2sH} ds X_0 \right)^{-1} X_0^T e^{tH} \quad (8.7)$$

This formula is valid regardless of the dimensions n, m, d . In particular, when $m = d = 1$ this is the solution of the rank-1 gradient flow. In the high-rank case, it is not straightforward a priori how to track the evolution of the matrix Z_t as firstly the rank of $X_0X_0^T$ and X^*X^{*T} (or Y or H) are not necessarily equal when $d \neq m$, and secondly because the eigenvectors of the two matrices are not aligned at the initialization.

In the following, we will consider the high-dimensional limit $n, m, d \rightarrow \infty$ with d/n and m/n fixed and make the following assumptions:

- The limits of traces $p_t = \frac{1}{d} \text{Tr}[Z_t^2]$, $q_t = \frac{1}{d} \text{Tr}[Z^* Z_t]$ (and \mathcal{E}_t) concentrate on their expectation, as well as related traces used in the *linear-pencils* method in Sec. 8.3.
- We assume that H has a limiting spectral distribution whose support can be enlaced in a finite contour $\Gamma \subset \mathbb{C}$.

To keep notations lighter we shall abusively denote by p_t, q_t, \mathcal{E}_t their limiting deterministic values.

8.2.2 Main results

The MSE \mathcal{E}_t of the problem is completely given by q_t , p_t and the constant r which in the high-dimensional limit is found to be $r = 1 + \phi$ from the second moment of the Marchenko-Pastur law (Marchenko and Pastur, 1967). The main contribution of this chapter is the self-consistent set of equations that fully track q_t and p_t :

(Result 1) In the high dimensional limit, the overlap q_t evolves according the integral:

$$q_t = \int_{\mathbb{R}} \frac{z \rho_Q(z) dz}{1 - e^{-2tz} + z \tilde{q}_t e^{-2tz}} \quad (8.8)$$

with the auxiliary function \tilde{q}_t solution of the fixed-point equation:

$$\psi \tilde{q}_t = 1 + \int_{\mathbb{R}} \frac{(1 - e^{-2tz}) \rho_P(z) dz}{\frac{1}{\tilde{q}_t} (1 - e^{-2tz}) + z e^{-2tz}} \quad (8.9)$$

and ρ_P, ρ_Q are given by their inverse Stieltjes transforms $P(z), Q(z)$. These are the analytic solutions of the degree 3 polynomials such that $-zP(z) \rightarrow 1$ when $|z| \rightarrow \infty$ and $-zQ(z) \rightarrow 1$ when $|z| \rightarrow \infty$ where:

$$\begin{aligned} P^3 + P^2 (\lambda(\mu + z) + 1) + P\lambda(\mu + z - \phi + 1) + \lambda &= 0 \\ Q^3 \phi + Q^2 \left(\mu + z - 2\phi - 1 - \frac{1}{\lambda} \right) - Q(\mu + z - \phi - 2) &= 1 \end{aligned} \quad (8.10)$$

(Result 2) In the high-dimensional limit, the eigenvalue distribution of Z_t is found by the inverse Stieltjes-Transform of $h_t(z)$ where:

$$h_t(z) = \frac{-1}{2\pi i} \oint_{\Gamma} - \frac{(1 + e^{-2tx} (\frac{x}{\tilde{h}_t(z)} - 1)) P(x) dx}{x + z + z e^{-2tx} (\frac{x}{\tilde{h}_t(z)} - 1)} \quad (8.11)$$

$$\tilde{h}_t(z) = 1 + \frac{1}{\psi} \frac{-1}{2\pi i} \oint_{\Gamma} - \frac{(x + z - z e^{2tx}) P(x) dx}{x + z + z e^{-2tx} (\frac{x}{\tilde{h}_t(z)} - 1)} \quad (8.12)$$

in particular, we find:

$$p_t = -\frac{1}{2\phi} \frac{\partial^{(2)}}{\partial z^2} \left(\frac{1}{z} h_t \left(\frac{1}{z} \right) \right) \Big|_{z=0} \quad (8.13)$$

Note that a similar system of equations as (8.8) can be derived by calculating the first and second derivatives in z as given by (8.13) and using the integrands in (8.11). However the resulting formulas are too cumbersome to be presented here.

8.2.3 Discussions and experiments

Figure 8.2.1 provides an example of the calculation of q_t through time compared with experimental runs: we see a good agreement between the curves and the prediction.

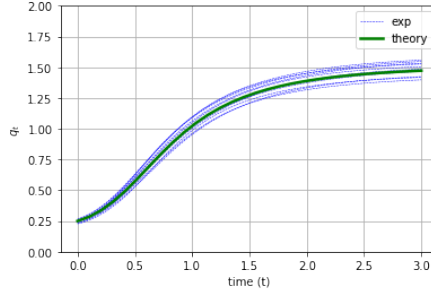


Figure 8.2.1: Comparison of q_t evolution with 10 runs of a gradient descent with $n = 100$, $m = 25$, $d = 75$ and $\lambda = 10^4$ and $\mu = 0$.

Asymptotic Limit $t \rightarrow \infty$: An interesting question is to study the asymptotics of q_t when $t \rightarrow \infty$. We take the Ansatz that $\tilde{q}_t \sim \gamma e^{2\alpha t}$ in this limit with $\alpha > 0$ and $\gamma > 0$ another constant and plug this in equation (8.9):

$$\psi = \frac{1}{\tilde{q}_t} + \int_{\mathbb{R}} \frac{(1 - e^{-2tz})\rho_P(z)dz}{1 - e^{-2tz} + z\tilde{q}_t e^{-2tz}} \quad (8.14)$$

$$\simeq \frac{e^{-2\alpha t}}{\gamma} + \int_{\mathbb{R}} \frac{(1 - e^{-2tz})\rho_P(z)dz}{1 - e^{-2tz} + z\gamma e^{-2(z-\alpha)t}} \quad (8.15)$$

$$\simeq \int_{\alpha}^{\infty} \rho_P(z)dz = 1 - F_P(\alpha) \quad (8.16)$$

With F_P the CDF of P . Such a solution exists when we can find α such that $F_P(\alpha) = 1 - \psi$, effectively selecting the proportion ψ of the eigenvalues of H in the interval $(\alpha, +\infty)$. Due to the assumption $\alpha > 0$, a further condition for the existence of such an α is $F_P(0) < 1 - \psi \leq 1$ or: $0 \leq \psi < 1 - F_P(0) \leq 1$. This implies that the Ansatz is valid in the *under-parameterized* regime ($m < n$). The asymptotic limit is thus given by $q_{\infty} = \lim_{t \rightarrow \infty} q_t = \int_{\alpha}^{\infty} z\rho_Q(z)dz$. Note that the alternative Ansatz that \tilde{q}_t converges towards a finite limit leads to a similar solution as but with $\alpha = 0$.

A similar line of reasoning lead us to consider the term $p_{\infty} = \frac{1}{\phi} \int_{\alpha}^{\infty} z^2 \rho_P(z)dz$ and thus a asymptotic mean square error:

$$\mathcal{E}_{\infty} = r - \int_{\alpha}^{\infty} \left(2z\rho_Q(z) - \frac{1}{\phi} z^2 \rho_P(z) \right) dz \quad (8.17)$$

As an example, for $\phi = \psi = 1$ and $\mu = \frac{1}{\lambda}$, and $\alpha = 0$ we expect from formula (8.17) that $\mathcal{E}_{\infty} = r$ when the support of ρ_P and ρ_Q is located below 0. This can be found by studying the discriminant $\Delta_P(\lambda, z)$ of the polynomial solved by P : because it is a order 3 polynomial with coefficients in \mathbb{R} when $z \in \mathbb{R}$, either the solutions are all real ($\Delta_P > 0$) implying $\rho_P(z) = 0$, or one is real and two are complex conjugate ($\Delta_P = 0$) implying $\rho_P(z) > 0$. At a specific λ , the support of ρ_P is located below 0 and touches $z = 0$. This λ_c is solution of $\Delta_P(\lambda_c, 0) = 0$ which provides the solution $\lambda_c = \frac{4}{27}$. The whole error curve at $t = +\infty$ is shown in Figure 8.2.2. The choice $\mu = \frac{1}{\lambda}$ is natural from a Bayesian point-of-view because it would correspond to the situation

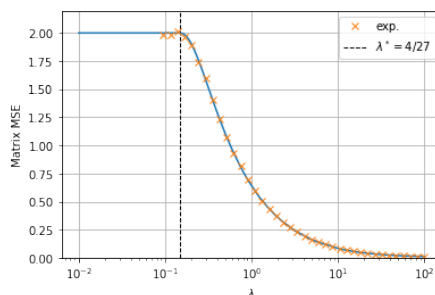


Figure 8.2.2: Experimental and theoretical \mathcal{E}_∞ with $\phi = \psi = 1, \mu = \frac{1}{\lambda}$

where the statistician matches its prior to the ground-truth when $\psi = \phi$.

Low rank limit when $\phi = \psi \rightarrow 0$: We bring to the reader’s attention that the objective function \mathcal{H} when $d = 1$ and $n \rightarrow \infty$ with $\mu = \frac{1}{\lambda}$ corresponds precisely to the spiked-Wigner problem. This suggest to look at the limit $\phi = \psi \rightarrow 0$. In this situation, we expect α should be close to the maximum eigenvalue of the bulk of ρ_Q . We make the following observation in Figure 8.2.3: as ϕ decreases, ρ_P in blue has two bulks of eigenvalues, one of which disappears as ϕ grows. On the other hand, ρ_Q in orange displays also two bulks at the same locations but the second bulk develops a mass as $\phi \rightarrow 0$. Therefore, we expect that α adjusts itself to the maximum eigenvalue of the first bulk of ρ_P . Furthermore, interestingly we see that these two bulks are getting closer when λ is closer to 1 as seen in Figure 8.2.4.

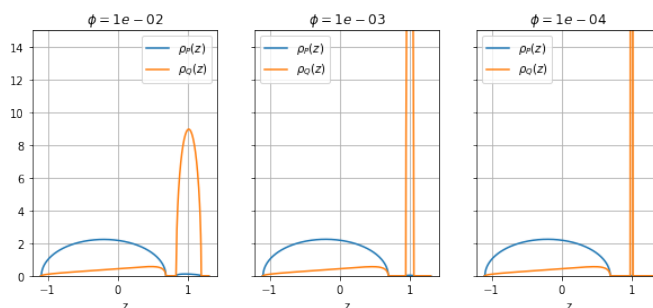


Figure 8.2.3: Bulk of eigenvalues for $\lambda = 5$ and different values of ϕ

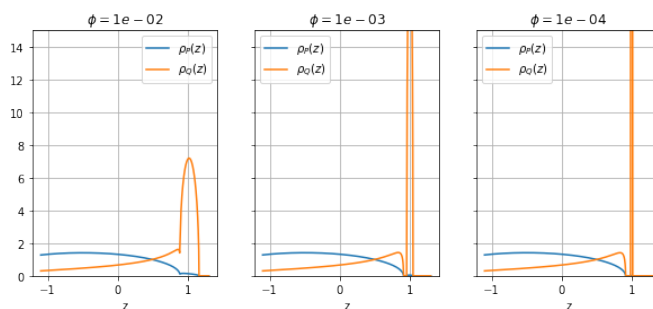


Figure 8.2.4: Bulk of eigenvalues for $\lambda = 2$ and different values of ϕ

Chapter 8. Matrix denoising: an extensive rank model

With these observations, we expect that $Q(z)$ has a pole $z = 1$ in the limit $\phi \rightarrow 0$. Let's consider a polynomial equation of \hat{Q} solving the reduced polynomial equation of Q with $\phi = 0$:

$$\hat{Q}^2(z-1) - \hat{Q}(z-2 + \frac{1}{\lambda}) - 1 = 0. \quad (8.18)$$

In order to find a potential pole, we consider $\mathcal{Q}(z) = (1-z)\hat{Q}(z)$ and check for potential limits of \mathcal{Q} when $z \rightarrow 1$. First of all, injecting \mathcal{Q} in the former polynomial equation, we find:

$$(\mathcal{Q}^2 \lambda + \mathcal{Q}(\lambda z - 2\lambda + 1) - \lambda z + \lambda) / (z-1) = 0. \quad (8.19)$$

Therefore, on the upper-complex plane we find the numerator equals 0, and by analytic continuation, the limit $z \rightarrow 1$ follows: $\mathcal{Q}(1)(\mathcal{Q}(1)\lambda - \lambda + 1) = 0$ so $\mathcal{Q}(1) \in \{0, 1 - \frac{1}{\lambda}\}$. It is interesting to notice the connection with the usual Bayesian overlap of the spiked Wigner model - since $\mathcal{Q}(1)$ represents the squared overlap q_∞ in the limit $\phi \rightarrow 0$. Pushing further this analysis for $P(z)$ allows to eventually get p_∞ and \mathcal{E}_∞ in the limit $\phi \rightarrow 0$ and check the connection with the Bayesian MMSE of the spiked Wigner model.

8.3 Sketch of Proof

Our method relies on considering the interaction of the random matrices X_0, X^*, ξ . We treat each term q_t and p_t separately with the linear-pencil technique. In both cases, we first factor out the X_0 matrix, then decouple the time dependency from the remaining random matrix expressions, and finally factor-out X^*, ξ .

Our results are derived in the limit $n, m, d \rightarrow +\infty$. For a sequence of matrices $A_N \in \mathbb{R}^{N \times N}$ we use the notation $\text{Tr}_N[A_N] = \lim_{N \rightarrow \infty} \frac{1}{N} \text{Tr}[A_N]$. As stated in Sec. 8.2 we assume that the limiting traces involved in the linear pencil method concentrate.

8.3.1 Tracking the angle q_t

The term $q_t = \text{Tr}_d[Z^* Z_t]$ can be completely recovered from a sub-block of the following linear-pencil M_q :

$$M_q = \left(\begin{array}{c|c|c|c|c|c|c} 0 & I_d & 0 & 0 & 0 & 0 & 0 \\ \hline I_d & 0 & 0 & 0 & 0 & 0 & W_t \\ \hline 0 & 0 & 0 & X_0 & 0 & 0 & I_n \\ \hline 0 & 0 & X_0^T & I_m & 0 & X_0^T & 0 \\ \hline 0 & 0 & 0 & 0 & L_t & I_n & 0 \\ \hline 0 & 0 & 0 & X_0 & I_n & 0 & 0 \\ \hline 0 & W_t^T & I_n & 0 & 0 & 0 & 0 \end{array} \right) \quad (8.20)$$

Where $W_t = X^{*T} e^{tH}$ and $L_t = 2 \int_0^t e^{2sH} ds$. A recursive application of the Schur-complement to compute M_q^{-1} shows that the block $(M_q^{-1})^{(1,1)}$ is the random matrix $X^{*T} Z_t X^*$. So in fact: $q_t = \text{Tr}_d \left[(M_q^{-1})^{(1,1)} \right]$.

The random matrices X_0, X^*, ξ are all independent and X_0 is not part of the terms W_t, L_t . Therefore, we can apply the linear-pencil theory on M_q over the random-matrix X_0 while considering the other random matrices as fixed. To this end, we note the constant part $C_q = \mathbb{E}_{X_0} [M_q]$, and consider matrix of sub-traces $g \in \mathbb{R}^{7 \times 7}$ such that for squared-blocks ij , $g_{ij} = \text{Tr}_{N_i} \left[(M_q^{-1})^{(i,j)} \right]$ where N_i is the size of the block ij in M_q^{-1} . Then we apply the fixed-point equation described in Appendix D of (Bodin and Macris, 2022) with $g_{ij} = \frac{1}{N_i} \text{Tr} \left[((C_q - \eta(g) \otimes I)^{-1})^{(ij)} \right]$ where $\eta(g)$ is the matrix defined by:

$$\eta(g) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi g_{44} & 0 & 0 & \psi g_{44} & 0 \\ 0 & 0 & 0 & g_{33} + g_{36} + g_{63} + g_{66} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi g_{44} & 0 & 0 & \psi g_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (8.21)$$

Further inversion of $C_q - \eta(g) \otimes I$ leads to:

$$g_{11} = \text{Tr}_d \left[g_{44} \psi W_t (g_{44} \psi L_t + I_n)^{-1} W_t^T \right] \quad (8.22)$$

$$g_{44} = \frac{1}{1 - g_{66}} \quad (8.23)$$

$$g_{66} = -\text{Tr}_n \left[L_t (g_{44} \psi L_t + I_n)^{-1} \right] \quad (8.24)$$

Let $\Gamma \subset \mathbb{C}$ be a contour enclosing the eigenvalues of H , we use the fact that for any functional f which applies on the eigenvalues of a matrix we have $f(H) = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) (H - zI_n)^{-1} dz$ to obtain:

$$g_{11} = \frac{-1}{2\pi i} \oint_{\Gamma} \frac{g_{44} \psi e^{2zt}}{1 + g_{44} \psi \int_0^t 2e^{2sz} ds} \text{Tr}_d \left[(H - zI_n)^{-1} Z^* \right] dz$$

which leads with $Q(z) = \text{Tr}_d \left[X^{*T} (H - zI_n)^{-1} X^* \right]$ to:

$$g_{11} = \frac{-1}{2\pi i} \oint_{\Gamma} \frac{g_{44} \psi z}{\psi g_{44} (1 - e^{-2tz}) + z e^{-2tz}} Q(z) dz \quad (8.25)$$

Similarly with $P(z) = \text{Tr}_n \left[(H - zI_n)^{-1} \right]$

$$g_{66} \psi = \frac{-1}{2\pi i} \oint_{\Gamma} \frac{1 - e^{-2tz}}{\psi g_{44} (1 - e^{-2tz}) + z e^{-2tz}} P(z) dz \quad (8.26)$$

We find the equations from the main results with $\tilde{q}_t = \frac{1}{\psi g_{44}}$.

Tracking the norm p_t

The term $p_t = \text{Tr}_d [Z_t^2]$ can also be recovered from a similar calculation but would lead to design a much larger linear-pencil. Another method is to track directly the eigenvalues of Z_t with the trace of the resolvent: $h_{11} = \text{Tr}_n [(Z_t - zI_n)^{-1}]$ with h the solution of the fixed point equation (Appendix D in (Bodin and Macris, 2022)) stemming from the following linear-pencil:

$$M_p = \left(\begin{array}{c|ccc|c} -zI_n & 0 & 0 & 0 & e^{tH} \\ \hline 0 & 0 & X_0 & 0 & I_n \\ \hline 0 & X_0^T & I_m & 0 & X_0^T & 0 \\ 0 & 0 & 0 & L_t & I_n & 0 \\ 0 & 0 & X_0 & I_n & 0 & 0 \\ \hline e^{tH} & I_n & 0 & 0 & 0 & 0 \end{array} \right) \quad (8.27)$$

Which yields the set of equations:

$$\begin{aligned} h_{11} &= -\text{Tr}_n \left[\left(L_t + \frac{1}{h_{33}} I_n \right) \left(e^{2tH} + zL_t + \frac{z}{h_{33}} I_n \right)^{-1} \right] \\ h_{33} &= 1 - \frac{1}{\psi} \text{Tr}_n \left[(zL_t + e^{2tH}) \left(e^{2tH} + zL_t + \frac{z}{h_{33}} I_n \right)^{-1} \right] \end{aligned}$$

Using the contour integration technique, we obtain:

$$h_{11} = \frac{-1}{2\pi i} \oint_{\Gamma} -\frac{\frac{1}{h_{33}} + \int_0^t 2e^{2sx} ds}{\frac{z}{h_{33}} + e^{2tx} + z \int_0^t 2e^{2sx} ds} P(x) dx \quad (8.28)$$

which is reduced to:

$$h_{11}(z) = \frac{-1}{2\pi i} \oint_{\Gamma} -\frac{1 + e^{-2tx} \left(\frac{x}{h_{33}} - 1 \right)}{x + z + ze^{-2tx} \left(\frac{x}{h_{33}} - 1 \right)} P(x) dx \quad (8.29)$$

Similarly for h_{33} :

$$h_{33}(z) = 1 + \frac{1}{\psi} \frac{-1}{2\pi i} \oint_{\Gamma} -\frac{(x + z - ze^{2tx}) P(x) dx}{x + z + ze^{-2tx} \left(\frac{x}{h_{33}} - 1 \right)} \quad (8.30)$$

Two possible ways to retrieve p_t from h_{11} and h_{33} : either with $\phi p_t = \frac{-1}{2\pi i} \oint_{\Gamma} z^2 h_{11}(z) dz$, or $\phi p_t = -\frac{1}{2} \frac{\partial^{(2)}}{\partial z^2} \left(\frac{1}{z} h_{11} \left(\frac{1}{z} \right) \right) |_{z=0}$. In both cases, there is an additional level of complexity in terms of calculation as it either requires a double-contour integration, or computing derivative and second derivative of the given functions yielding further new equations.

Quantities $Q(z), P(z)$

There remains to calculate the terms $Q(z), P(z)$ which depends only on the random matrices X^*, ξ and can be done altogether with the linear-pencil:

$$M_z = \begin{pmatrix} I_n & X^* & 0 & 0 \\ 0 & I_d & X^{*T} & 0 \\ 0 & 0 & (z + \mu)I_n - \frac{1}{\sqrt{\lambda}}\xi & X^* \\ 0 & 0 & X^{*T} & I_d \end{pmatrix} \quad (8.31)$$

Using the kernel $K = (H - zI_n)^{-1}$, we can calculate the inverse:

$$M_z^{-1} = \begin{pmatrix} I_n & -X^* & -Z^*K & Z^*KX^* \\ 0 & I_d & X^{*T}K & -X^{*T}KX^* \\ 0 & 0 & -K & KX^* \\ 0 & 0 & -X^{*T}K & I_d - X^{*T}KX^* \end{pmatrix} \quad (8.32)$$

So that $Q(z) = -f_{13}$ and $P(z) = f_{33}$ where we f is the analog of g and h with the former linear-pencils. In particular we expect the following structure:

$$f = \begin{pmatrix} 1 & 0 & -\phi Q(z) & 0 \\ 0 & 1 & 0 & -Q(z) \\ 0 & 0 & -P(z) & 0 \\ 0 & 0 & 0 & 1 - Q(z) \end{pmatrix} \quad (8.33)$$

We can further compute the fixed point equation with:

$$\eta(f) = \begin{pmatrix} 0 & 0 & f_{22}\phi + f_{24}\phi & 0 \\ 0 & f_{31} & 0 & f_{33} \\ 0 & 0 & \frac{f_{33}}{\lambda} + f_{42}\phi + f_{44}\phi & 0 \\ 0 & f_{31} & 0 & f_{33} \end{pmatrix} \quad (8.34)$$

After some algebraic reductions, we obtain the degree 3 polynomials given in equation (8.10). In general, these equations have multiple solutions but only one corresponds to the analytic solution associated to the appropriate trace of resolvent.

8.4 Conclusion

Our work primarily shows how we can take advantage of random matrix techniques to derive fixed-point equations solving the time evolution of the matrix-mean-square-error in the high-dimensional limit. Although we choose a specific data model, as future considerations, the matrix H can be generalized to other structures for which the same methods would apply. In particular, if only the noise structure changes, then only ρ_Q and ρ_P are changed. We will come back to these issues in a more extensive and detailed contribution.

9 Conclusion and future research directions

Throughout this thesis, we have explored a class of large learning models in a statistical limit that enables tracking the learning progression through the gradient-flow algorithm. More precisely, we have examined the Gaussian covariate model discussed in Chapter 5, which also captures a simplified representation of a 2-layer neural network as presented in Chapter 6. In this context, we have investigated the potential to predict the evolution of the gradient-descent algorithm for real-world datasets under specific conditions. In essence, this allows us to forecast the future values of the training error and test-error for such models and offers insights into their prospective performances. It also provides valuable time and resource savings compared to the computationally intensive weight-updating process, or the exploration of the hyper-parameter space.

This research can open the door to future possibilities for gaining deeper understanding into the learning dynamics of large models and comprehending the scaling laws governing the test-error and training error with various hyper-parameters. Specifically, these scaling laws aim to establish the optimal achievable loss, denoted as L , within the constraints of computational resources C , the model's size N , and the available data D . As previously cited in (Kaplan et al., 2020), there is a growing body of empirical evidence supporting the existence of such scaling laws for a variety of models, including large transformer language models. As mentioned in this same paper, *"the training curves follow predictable power-laws [...]"* and *"performance improves predictably as long as we scale N and D in tandem [...]"*. Notably, it has been stated that the architectural details such as network width and depth seem to have little impact in their investigations. So determining these laws is of utmost importance for budgeting the training of large models. However, even with these empirical findings, there is still no definitive consensus on the precise form they should have (Caballero et al., 2023). For instance, (Hoffmann et al., 2022) provide an empirical law that exemplifies this concept for some coefficients A, B, L_0, C_0 :

$$L = AN^{-\alpha} + BD^{-\beta} + L_0 \tag{9.1}$$

$$C = C_0ND \tag{9.2}$$

When the value of C is held constant and L is minimized, a precise power-law relationship emerges for N and D in the form of $N(C) = G \left(\frac{C}{C_0} \right)^a$ and $D(C) = G^{-1} \left(\frac{C}{C_0} \right)^b$ where the constants a, b and G depend on the other previous coefficients A and B . This opens-up a potential avenue and an interesting endeavour for applying the theory developed for the random-feature model and investigating the existence of similar laws in this context. However, a significant challenge lies in establishing a consistent definition of the computational budget C as no definitive formulation readily stands out. In our experimental framework, given that the first layer remains fixed, the primary computation cost arises from the update of $\beta_t \in \mathbb{R}^N$ which at most necessitates a fixed matrix-vector multiplication operation. Consequently, a viable computation cost may take the form of $C = \left(\frac{T}{\Delta t} \right) C_0 N^2$ for a learning step of size Δt and a time horizon T . Another critical consideration concerns the choice of this learning step Δt . For instance, it can be selected in relation to the largest eigenvalue of the Hessian of the loss function to minimize the number of iterations required to reach any specified level of accuracy while preserving the convergence towards β_∞ , as discussed in the introduction. Furthermore, other parameters that must also be considered include the regularization parameter λ which is also another constant of our framework. As an example, λ may be chosen to minimize the test-error while keeping the other parameters fixed. In Summary, it is not evident a priori whether scaling laws exist with the random feature model of Chapter 6, and if they do, what form they might take and whether they align with empirical observations from existing literature. Nevertheless, this model offers a potential path for investigating the emergence of such laws based on first principles.

While the Gaussian covariate model offers a theoretical foundation for analyzing supervised learning tasks, it is worth noting that similar random matrix methods can be effectively applied to address a broader spectrum of learning problems, such as the gradient-flow dynamics in matrix completion problems. In this thesis, we have presented a formula for analyzing the dynamics of the loss in the rank-one estimation problem with a non-trivial objective function, as discussed in Chapter 7. A typical next step for this model is the investigation of the non-symmetric case with $Y = uv^T + \sqrt{\frac{n}{\lambda}} \xi$. Here, $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$ are two vectors of size n and d which are not necessarily equal, but still considered in the limit of large values with the fixed ratio $\phi = \frac{n}{d}$ (Miolane, 2017). Another direction for future exploration is the analysis of alternative prior distributions on θ^* , such as the Laplace prior, to assess their impact on the learning curves. The objective function also needs to be adapted to account for the change of prior distribution. It remains unclear to what extent the technical results developed in Chapter 7 can be extended to this case. Alternatively, the addition of a noise term in the gradient-flow equation may be a viable path to explore further aspects of the optimization algorithm.

Additionally, we explored the extensive-rank model characterized by a quadratic number of parameters. The dynamics involve the alignment of the eigenvectors of the estimator with the data matrix, as detailed in Chapter 8. This model holds particular significance in the context of sample covariance matrix estimation. It can find some practical applications for the realistic datasets cases, thus presenting a potential avenue for improving the Gaussian covariate model

analysis and addressing the sub-sampling requirement for the training and test set. However, it is important to note that this model employs a gradient-flow method that may not achieve the same level of performance of the rotationally invariant estimator described in (Potters and Bouchaud, 2020). To improve our approach, one potential strategy involves adjusting the regularization term $\mu \|X\|_F^2$ as outlined in equation (8.2). For example, it can be refined into a more general form as $\|A^{\frac{1}{2}} X\|_F^2$, for an appropriate matrix A (with $A = \mu I$ in the previous case). Another line of research is the investigation of the non-symmetric instance. As for the rank-one case, we consider the data-matrix $Y = UV^T + \frac{1}{\sqrt{\lambda}}\xi$, featuring two random matrices U and V . In this situation, an algorithm of interest is the gradient-flow method applied to both estimators, U_t and V_t , simultaneously. It remains uncertain at this stage whether a closed system, comprised of a finite number of equations, can be derived. Nevertheless, it is worth noting that some progress appears feasible, albeit potentially leading to more intricate coupled matrix Riccati differential equations.

Finally, in a similar vein, other models are concerned with tensor estimation problems where Y is a tensor of order 3 or higher (Barbier et al., 2017), or even delve into the intricacies of a mixture of spiked-matrix tensor models (Mannelli et al., 2019). While these models are generally more complex, certain random-matrix methods may offer promising avenues to further extend our approaches to address some of these more intricate scenarios (for instance de Morais Goulart et al. (2022)).

At the heart of these models, we employ statistical methods of large-dimensional matrices with random entries, which have been described thoroughly in Chapter 2 and 3. This mathematical field, known as the random matrix theory, has been under development for over half a century and has found successful applications across various scientific disciplines. Consequently, it comes as no surprise that its relevance has been steadily increasing in the analysis of large-scale learning models, making it a promising way for understanding the learning dynamics of large neural networks. As described, the derivation of the equations governing the random feature model makes extensive use of the tools and techniques from random matrix theory, involving intricate operations with multiple matrices. However, there are still several hurdles to tackle more general models, such as a random feature model where we relax the frozen weight assumption. Such a model would offer a more realistic framework to understand full-sized neural networks. A notable contribution in this direction can be found in the research paper (Ba et al., 2022). While tracking the full gradient-flow of the weight matrix Θ (in Chapter 6) may prove challenging in the near future, it is in the realm of possibilities to explore a simpler but quite intricate scenario where a single gradient-step is applied to the loss with respect to Θ to update this weight matrix. This results in a new matrix, denoted as Θ_1 , which can subsequently be integrated within the regular random feature model and lead to new insightful analytical results. In particular, depending on the magnitude with respect to n of this initial gradient step, the weight matrix Θ_1 can exhibit for instance a similar structure as Θ with the addition of rank-one matrix, or even a new random matrix with multiple bulks. Besides, as in (Adlam and Pennington, 2020a), these studies often involve more general operations on large

Chapter 9. Conclusion and future research directions

random matrices, including some Hadamard products. While it is possible to derive these operations in certain specific cases, this is not yet always achievable in a fully general context. A parallel concern is the extension of our methods for generic loss and regularization functions as exemplified in Loureiro et al. (2021). As demonstrated in Chapter 5, in the asymptotic limit $t \rightarrow +\infty$, our methods align with these results when using the mean-squared-loss. However, it is not yet clear whether the dynamics for other types of objective functions can be derived as easily.

In conclusion, the future of random matrix theory in the context of large neural networks holds great promise, and it calls for further developments, perhaps in the form of a novel high-dimensional statistical framework tailored for large learning models.

Bibliography

- Adlam, B., Levinson, J., and Pennington, J. (2019). A Random Matrix Perspective on Mixtures of Nonlinearities for Deep Learning. *arXiv e-prints*, page arXiv:1912.00827.
- Adlam, B. and Pennington, J. (2020a). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 74–84. PMLR.
- Adlam, B. and Pennington, J. (2020b). Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032.
- Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020a). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.
- Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020b). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.
- Agoritsas, E., Biroli, G., Urbani, P., and Zamponi, F. (2018). Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002.
- Arous, G. B., Gheissari, R., and Jagannath, A. (2022). High-dimensional limit theorems for sgd: Effective dynamics and critical scaling.
- Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946. Curran Associates, Inc.
- Bai, Z. D. (1997). Circular law. *The Annals of Probability*, 25(1):494–529.
- Baik, J., Arous, G. B., and Péché, S. (2005a). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, page 1643.

Bibliography

- Baik, J., Ben Arous, G., and Péché, S. (2005b). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697.
- Bandeira, A. S., Boumal, N., and Voroninski, V. (2016). On the low-rank approach for semidefinite programs arising in synchronization and community detection. volume 49 of *Proceedings of Machine Learning Research*, pages 361–382, Columbia University, New York, New York, USA. PMLR.
- barbier, j., Dia, M., Macris, N., Krzakala, F., Lesieur, T., and Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Barbier, J., Hou, T., Mondelli, M., and Sáenz, M. (2022). The price of ignorance: how much does it cost to forget noise structure in low-rank matrix estimation? *arXiv preprint arXiv:2205.10009*.
- Barbier, J. and Macris, N. (2019). The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *Probability theory and related fields*, 174(3):1133–1185.
- Barbier, J. and Macris, N. (2022). Statistical limits of dictionary learning: random matrix theory and the spectral replica method. *Physical Review E*, 106(2):024136.
- Barbier, J., Macris, N., and Miolane, L. (2017). The Layered Structure of Tensor Estimation and its Mutual Information. In *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.
- Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019a). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019b). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116:201903070.
- Belkin, M., Hsu, D., and Xu, J. (2020a). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.
- Belkin, M., Hsu, D., and Xu, J. (2020b). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.
- Belkin, M., Ma, S., and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International*

-
- Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 541–549. PMLR.
- Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019c). Does data interpolation contradict statistical optimality? In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR.
- Ben Arous, G., Dembo, A., and Guionnet, A. (2004). Cugliandolo-kurchan equations for dynamics of spin-glasses. *Probability Theory and Related Fields*, 136.
- Benaych-Georges, F. and Knowles, A. (2016a). Lectures on the local semicircle law for wigner matrices. *arXiv preprint arXiv:1601.04055*.
- Benaych-Georges, F. and Knowles, A. (2016b). Lectures on the local semicircle law for Wigner matrices. working paper or preprint.
- Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521.
- Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, pages 3880–3888, Red Hook, NY, USA. Curran Associates Inc.
- Bloemendal, A., Erdős, L., Knowles, A., Yau, H.-T., and Yin, J. (2014). Isotropic local laws for sample covariance and generalized wigner matrices. *Electron. J. Probab.*, 19:53 pp.
- Bodin, A. and Macris, N. (2021a). Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. *Advances in Neural Information Processing Systems*, 34.
- Bodin, A. and Macris, N. (2021b). Rank-one matrix estimation: analytic time evolution of gradient descent dynamics. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 635–678. PMLR.
- Bodin, A. and Macris, N. (2022). Gradient flow in the gaussian covariate model: exact solution of learning curves and multiple descent structures.
- Bodin, A. and Macris, N. (2023). Gradient flow on extensive-rank positive semi-definite matrix denoising. In *2023 IEEE Information Theory Workshop (ITW)*, pages 365–370.
- Bordelon, B., Canatar, A., and Pehlevan, C. (2020). Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR.

Bibliography

- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Buchberger, B. (1965). Ein algorithmus zum auffinden der basiselemente des restklassenringes nach einem nulldimensionalen polynomideal. *Ph. D. Thesis, Math. Inst., University of Innsbruck*.
- Bun, J., Bouchaud, J.-P., and Potters, M. (2017). Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109.
- Burer, S. and Monteiro, R. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 95:329–357.
- Burer, S. and Monteiro, R. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103:427–444.
- Caballero, E., Gupta, K., Rish, I., and Krueger, D. (2023). Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*.
- Camilli, F., Contucci, P., and Mingione, E. (2022). An inference problem in a mismatched setting: a spin-glass model with Mattis interaction. *SciPost Phys.*, 12:125.
- Camilli, F. and Mézard, M. (2022). Matrix factorization with neural networks. *arXiv preprint arXiv:2212.02105*.
- Chen, L., Min, Y., Belkin, M., and Karbasi, A. (2021). Multiple descent: Design your own generalization curve. *Advances in Neural Information Processing Systems*, 34.
- Chen, Y. and Chi, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31.
- Chi, Y., Lu, Y. M., and Chen, Y. (2019a). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.
- Chi, Y., Lu, Y. M., and Chen, Y. (2019b). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.
- Chou, H.-H., Gieshoff, C., Maly, J., and Rauhut, H. (2020). Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*.

- Cox, D. A., Little, J., and O’Shea, D. (2007). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3/e (Undergraduate Texts in Mathematics)*. Springer-Verlag, Berlin, Heidelberg.
- Crisanti, A., Horner, H., and Sommers, H. (1993). The sphericalp-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92:257–271.
- Crisanti, A. and Sompolinsky, H. (2018). Path integral approach to random neural networks. *Phys. Rev. E*, 98:062120.
- Cugliandolo, L. F. and Dean, D. S. (1995). Full dynamical solution for a spherical spin-glass model. *Journal of Physics A: Mathematical and General*, 28(15):4213–4234.
- Cugliandolo, L. F. and Kurchan, J. (1993). Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Phys. Rev. Lett.*, 71:173–176.
- Cui, W., Rocks, J. W., and Mehta, P. (2020). The perturbative resolvent method: Spectral densities of random matrix ensembles via perturbation theory. *arXiv preprint arXiv:2012.00663*.
- D’Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. (2020). Double trouble in double descent: Bias and variance(s) in the lazy regime. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2280–2290. PMLR.
- d’Ascoli, S., Sagun, L., and Biroli, G. (2020). Triple descent and the two kinds of overfitting: where and why do they appear? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3058–3069. Curran Associates, Inc.
- de Morais Goulart, J. H., Couillet, R., and Comon, P. (2022). A random matrix perspective on random tensors. *The Journal of Machine Learning Research*, 23(1):12110–12145.
- De Sa, C., Olukotun, K., and Ré, C. (2015). Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, pages 2332–2341. JMLR.org.
- Deng, Z., Kammoun, A., and Thrampoulidis, C. (2021a). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495.
- Deng, Z., Kammoun, A., and Thrampoulidis, C. (2021b). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*. iaab002.
- Derezinski, M., Liang, F. T., and Mahoney, M. W. (2020). Exact expressions for double descent and implicit regularization via surrogate random design. In Larochelle, H., Ranzato, M.,

Bibliography

- Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5152–5164. Curran Associates, Inc.
- Dhifallah, O. and Lu, Y. M. (2020). A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*.
- Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.
- Dunford, N. and Schwartz, J. T. (1988). *Linear Operators*. Wiley Classics Library.
- Dyson, F. J. (1962). A brownian-motion model for the eigenvalues of a random matrix. *Journal of Mathematical Physics*, 3(6):1191–1198.
- d’Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. (2020). Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR.
- Edwards, S. F. and Jones, R. C. (1976). The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595.
- Engel, A. and Van den Broeck, C. (2001a). *Statistical mechanics of learning*. Cambridge University Press.
- Engel, A. and Van den Broeck, C. (2001b). *Statistical Mechanics of Learning*. Cambridge University Press.
- Erdős, L. (2011). Universality of wigner random matrices: a survey of recent results. *Russian Mathematical Surveys*, 66(3):507–626.
- Erdős, L., Schlein, B., and Yau, H.-T. (2008). Local semicircle law and complete delocalization for wigner random matrices. *Communications in Mathematical Physics*, 287(2):641–655.
- Féral, D. and Pécché, S. (2006). The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in Mathematical Physics*, 272.
- Ge, R., Jin, C., and Zheng, Y. (2017a). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. volume 70 of *Proceedings of Machine Learning Research*, pages 1233–1242, International Convention Centre, Sydney, Australia. PMLR.
- Ge, R., Jin, C., and Zheng, Y. (2017b). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning, ICML’17*, pages 1233–1242. JMLR.org.
- Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d’Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. (2019). Scaling description of generalization with number of parameters in deep learning. *CoRR*, abs/1901.01608.

- Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. (2020). Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401.
- Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. (2020a). Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR.
- Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. (2020b). Generalisation error in learning with random features and the hidden manifold model. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR.
- Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. (2022). The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR.
- Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30.
- Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints*, page arXiv:1903.08560.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
- Helton, J. W., Far, R. R., and Speicher, R. (2007). Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints. *International Mathematics Research Notices*, 2007(9):rnm086–rnm086.
- Helton, J. W., Mai, T., and Speicher, R. (2018). Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models.
- Hu, H. and Lu, Y. M. (2022). Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv preprint arXiv:2205.06798*.
- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.
- Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020a). Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR.

Bibliography

- Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. (2020b). Implicit regularization of random feature models. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4631–4640. PMLR.
- James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Kabashima, Y., Krzakala, F., Mezard, M., Sakata, A., and Zdeborova, L. (2016). Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv*, page 2001.08361v1.
- Kardar, M., Parisi, G., and Zhang, Y.-C. (1986). Dynamic scaling of growing interfaces. *Phys. Rev. Lett.*, 56:889–892.
- Kini, G. R. and Thrampoulidis, C. (2020). Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532.
- Korada, S. B. and Macris, N. (2009). Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.
- Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA. PMLR.
- Lelarge, M. and Miolane, L. (2018). Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3-4):859–929.
- Lelarge, M. and Miolane, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929.
- Lesieur, T., Krzakala, F., and Zdeborová, L. (2015). Phase transitions in sparse pca. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1635–1639. IEEE.
- Lesieur, T., Krzakala, F., and Zdeborová, L. (2017a). Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403.

- Lesieur, T., Miolane, L., Lelarge, M., Krzakala, F., and Zdeborová, L. (2017b). Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory, ISIT 2017, Aachen, Germany, June 25-30, 2017*, pages 511–515. IEEE.
- Liang, T., Sen, S., and Sur, P. (2022). High-dimensional asymptotics of langevin dynamics in spiked matrix models.
- Liao, Z. and Couillet, R. (2018). The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, pages 3072–3081. PMLR.
- Liao, Z., Couillet, R., and Mahoney, M. W. (2020). A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.
- Lin, L. and Dobriban, E. (2021). What causes the test error? going beyond bias-variance via anova. *J. Mach. Learn. Res.*, 22:155–1.
- Ling, S., Xu, R., and Bandeira, A. S. (2019). On the landscape of synchronization networks: A perspective from nonconvex optimization. *arXiv:1809.11083*.
- Loog, M., Viering, T., Mey, A., Krijthe, J. H., and Tax, D. M. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626.
- Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151.
- Lu, Y. M. and Yau, H.-T. (2022). An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv preprint arXiv:2205.06308*.
- Luneau, C., Macris, N., and Barbier, J. (2020). High-dimensional rank-one nonsymmetric matrix decomposition: the spherical case. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2646–2651. IEEE.
- Maillard, A., Krzakala, F., Mézard, M., and Zdeborová, L. (2022). Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083301.
- Mannelli, S. S., Krzakala, F., Urbani, P., and Zdeborova, L. (2019). Passed and spurious: Descent algorithms and local minima in spiked matrix-tensor models. volume 97 of *Proceedings of Machine Learning Research*, pages 4333–4342, Long Beach, California, USA. PMLR.
- Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

Bibliography

- Marcus, A. W., Spielman, D. A., and Srivastava, N. (2022). Finite free convolutions of polynomials. *Probability Theory and Related Fields*, 182(3-4):807–848.
- Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.
- Mehta, M. L. (2004). *Random matrices*. Elsevier.
- Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, page arXiv:1908.05355.
- Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.
- Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., Rathnayake, T., Vig, S., Granger, B. E., Muller, R. P., Bonazzi, F., Gupta, H., Vats, S., Johansson, F., Pedregosa, F., Curry, M. J., Terrel, A. R., Roučka, v., Saboo, A., Fernando, I., Kulal, S., Cimrman, R., and Scopatz, A. (2017). Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103.
- Mignacco, F., Krzakala, F., Urbani, P., and Zdeborová, L. (2020). Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550. Curran Associates, Inc.
- Mignacco, F., Urbani, P., and Zdeborová, L. (2021). Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*.
- Mingo, J. A. and Speicher, R. (2017). *Free probability and random matrices*, volume 35. Springer.
- Miolane, L. (2017). Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*.
- Misiakiewicz, T. (2022). Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*.
- Montanari, A. and Richard, E. (2014). A statistical model for tensor pca. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS 2014, pages 2897–2905, Cambridge, MA, USA. MIT Press.
- Montanari, A. and Venkataramanan, R. (2017). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*.
- Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321 – 345.

- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020a). Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*.
- Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. (2020b). Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*.
- Opper, M. (1998). *Statistical Mechanics of Generalization*, pages 922–925. MIT Press, Cambridge, MA, USA.
- Parisi, G., Urbani, P., and Zamponi, F. (2020). *Theory of Simple Glasses: Exact Solutions in Infinite Dimensions*. Cambridge University Press.
- Park, D., Kyriallidis, A., Carmanis, C., and Sanghavi, S. (2017). Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. volume 54 of *Proceedings of Machine Learning Research*, pages 65–74, Fort Lauderdale, FL, USA. PMLR.
- Péché, S. (2004). The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134:127–173.
- Pennington, J. and Worah, P. (2017). Nonlinear random matrix theory for deep learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2637–2646.
- Perry, A., Wein, A. S., and Bandeira, A. S. (2020). Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(1):230 – 264.
- Potters, M. and Bouchaud, J.-P. (2020). *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press.
- Pourkamali, F. and Macris, N. (2022a). Mismatched estimation of non-symmetric rank-one matrices under Gaussian noise. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1288–1293. IEEE.
- Pourkamali, F. and Macris, N. (2022b). Mismatched estimation of symmetric rank-one matrices under gaussian noise. In *International Zurich Seminar on Information and Communication (IZS 2022). Proceedings*, pages 84–88. ETH Zurich.
- Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. (2021). Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR MATH-AI Workshop*.

Bibliography

- Péché, S. (2019). A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24(none):1 – 7.
- Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.
- Rashidi Far, R., Oraby, T., Bryc, W., and Speicher, R. (2006). Spectra of large block matrices. *arXiv e-prints*, page cs/0610045.
- Richards, D., Mourtada, J., and Rosasco, L. (2021). Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR.
- Rubio, F. and Mestre, X. (2011). Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602.
- Sarao Mannelli, S., Biroli, G., Cammarota, C., Krzakala, F., Urbani, P., and Zdeborová, L. (2020). Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1).
- Sarao Mannelli, S., Biroli, G., Cammarota, C., Krzakala, F., and Zdeborová, L. (2019). Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8679–8689. Curran Associates, Inc.
- Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.
- Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262.
- Spigler, S., Geiger, M., d’Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. (2019a). A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001.
- Spigler, S., Geiger, M., d’Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. (2019b). A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001.
- Tao, T. (2012). *Topics in random matrix theory*, volume 132. American Mathematical Soc.
- Tao, T. and Vu, V. (2008). Random matrices: the circular law. *Communications in Contemporary Mathematics*, 10(02):261–307.

- Tarmoun, S., Franca, G., Haeffele, B. D., and Vidal, R. (2021). Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pages 10153–10161. PMLR.
- Troiani, E., Erba, V., Krzakala, F., Maillard, A., and Zdeborová, L. (2022). Optimal denoising of rotationally invariant rectangular matrices. *arXiv preprint arXiv:2203.07752*.
- Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Wang, C., Mattingly, J., and Lu, Y. M. (2017). Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and pca. *arXiv preprint arXiv:1712.04332*.
- Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327.
- Wu, D. and Xu, J. (2020). On the optimal weighted l2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123.
- Xiao, L. and Pennington, J. (2022). Precise learning curves and higher-order scaling limits for dot product kernel regression. *arXiv preprint arXiv:2205.14846*.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *ICLR 2017*, arXiv abs/1611.03530.

Antoine Bodin

Machine learning researcher with strong background and industrial experience in mathematical modeling, software engineering, and financial engineering



@ antoine.bodin.ch@gmail.com

Switzerland

antoine-bodin.org

EXPERIENCE

EPFL PhD Researcher in Computer Science

Swiss Federal Institute of Technology of Lausanne (EPFL)

Sept. 2019 – Dec. 2023 Lausanne, Switzerland

Research in machine learning theory and optimization with high-dimensional statistical methods.


 Visiting Researcher in Machine Learning

Harvard University

May 2023 – July 2023 Cambridge MA, USA

Conducted research on theoretical aspects of gradient descent optimization in high-dimensional spaces.

Maths Modeling ML PyTorch Jupyter Azure

 Data Scientist & Software Engineer

Credit Suisse, Compliance

Feb. 2016 - July 2019 Lausanne, Switzerland

Analysis and modeling with large datasets for world-wide transaction surveillance and anti-money laundering.

- Designed and developed a modern AML modeling framework
- Transformed and unified a large-scale data-processing pipeline
- Organized and led the data-science challenge for the participants of LauzHack (Lausanne Hackathon)

Modeling Python SparkML Data Pipelines

 Software Engineer Intern

Google, News Team

July 2015 - Sept. 2015 Mountain View CA, USA

Implemented modern headlines scoring methods for the spotlight section using NLP and various signals.

C/C++ Python BigTable Recommender Systems

INTERESTS AND AWARDS

Sport: climbing, biking, skiing

Hobbies: baking, guitar, history and game of chess and go

Google Hashcode 2015: finalist, 8# over 230

SKILLS

Programming

Python C/C++ OCaml

Others

PyTorch Tensorflow Git
Scikit-Learn Pandas Spark
Linux Docker Azure GCP
OpenCL OpenGL SQL

Mathematics

Stochastic Calculus Derivatives
Statistics Random Matrix Theory

EDUCATION

Master of Science (GPA 5.57/6)

Computer Science, Minor in Fin. Engineering

École Polytechnique Fédérale de Lausanne

Sept 2014 – August 2016

Master of Engineering

Applied Mathematics

École Centrale Paris (Grande École)

Sept 2012 – August 2016

Preparatory Classes

(Maths/Physics courses for Grandes Écoles)

Sept 2010 – August 2012

REFEREES

Prof. Nicolas Macris

@ EPFL

✉ nicolas.macris@epfl.ch

PUBLICATIONS

Conference Proceedings

- **Bodin, Antoine**, & Macris, N. (2023). Gradient flow on extensive-rank positive semi-definite matrix denoising. In *2023 IEEE Information Theory Workshop (ITW 2023)* (pp. 365–370). doi:10.1109/ITW55543.2023.10161669
- **Bodin, Antoine**, & Macris, N. (2021a). Model, sample, and epoch-wise descents: Exact solution of gradient flow in the random feature model. In *35th conference on neural information processing systems (NeurIPS 2021)*.
- **Bodin, Antoine**, & Macris, N. (2021b, 15–19 Aug). Rank-one matrix estimation: Analytic time evolution of gradient descent dynamics. In M. Belkin & S. Kpotufe (Eds.), *Proceedings of thirty fourth conference on learning theory (COLT 2021)* (Vol. 134, pp. 635–678). PMLR.

Preprints

- **Bodin, Antoine**, & Macris, N. (2022). Gradient flow in the gaussian covariate model: exact solution of learning curves and multiple descent structures. *arXiv e-prints*, arXiv:2212.06757. doi:10.48550/arXiv.2212.06757. arXiv: 2212.06757 [stat.ML]

Patent

- Daines, S., Voigt, B., Grazioli, C., Agrafiotis, V., **Bodin, Antoine**, Sidlauskas, D., ... McPherson, A. (2019, May). Systems and methods for integration of disparate data feeds for unified data monitoring. (US11227288B1).