The Matrix Dyson Equation for Machine Learning: Correlated Linearizations and the Test Error in Random Features Regression

Hugo Latourelle-Vigeant

Department of Mathematics and Statistics
McGill University, Montreal

April, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Contemporary machine learning models, particularly deep learning models, are frequently trained on large datasets within high-dimensional feature spaces, presenting challenges for traditional analytical approaches. Notably, the effective generalization of highly overparameterized models contradicts conventional statistical wisdom. Furthermore, the presence of non-linear activations in artificial neural networks adds complexity to their analysis. To simplify theoretical analysis, it is often assumed that training data is sampled from an unstructured distribution. While such analyses offer insights into certain aspects of machine learning, they fall short in elucidating how neural networks extract information from the structure of the data, crucial for their success in real-world applications.

Fortunately, random matrix theory has emerged as a valuable tool for theoretically understanding certain machine learning procedures. Various techniques have been employed to explore large random matrices through asymptotic deterministic equivalents. One such approach involves substituting the random resolvent associated with a large random matrix with the solution of a deterministic fixed-point equation known as the matrix Dyson equation. Another effective technique, known as the linearization trick, involves embedding a matrix expression into a larger random matrix, termed a linear matrix pencil, with a simplified correlation structure.

In this thesis, we extend the matrix Dyson equation framework to derive an anisotropic global law for a broad class of pseudo-resolvents with general correlation structures. This extension enables the analysis of spectral properties of a wide range of random matrices using a simpler and deterministic solution to the matrix Dyson equation. Through the development of this theory, we address critical aspects such as existence-uniqueness, spectral support bounds, and stability properties. These considerations are essential for constructing deterministic equivalents for pseudo-resolvents of a class of correlated linear pencils.

Leveraging this theoretical framework, we provide an asymptotically exact deterministic expression for the empirical test error of random features ridge regression. The random features model, characterized by its non-linear activation function and potential for overparameterization, emerges as a powerful model for studying phenomena observed in real-life machine learning models, such as multiple descent and implicit regularization. Our exact expression facilitates a precise characterization of the implicit regularization of the model and unveils connections between random features regression and closely related kernel methods. Since we make no particular assumptions about the distribution of the data and response variable, our work represents a significant step towards understanding how neural networks exploit specific data structures.

Abrégé

Les modèles d'apprentissage automatique contemporains, en particulier les modèles d'apprentissage profond, sont souvent entraînés sur de vastes ensembles de données à l'intérieur d'espaces de dimensions élevées, ce qui pose des défis pour les approches analytiques traditionnelles. Notamment, la généralisation adéquate de modèles surparamétrés contredit les conventions statistiques. De plus, la présence de fonctions d'activations non linéaires dans les réseaux neuronaux artificiels ajoute de la complexité à leur analyse. Pour simplifier cette analyse théorique, il est souvent supposé que les données d'entraînement sont échantillonnées à partir d'une distribution non structurée. Bien que de telles analyses permettent de mieux comprendre certains aspects de l'apprentissage automatique, elles ne parviennent pas à élucider comment les réseaux neuronaux extraient des informations de la structure des données, élément crucial de leur réussite dans les applications du monde réel.

La théorie des matrices aléatoires s'est révélée être un outil précieux pour la compréhension de certaines procédures d'apprentissage automatique. Diverses techniques ont été utilisées pour explorer les grandes matrices aléatoires par le biais d'équivalents déterministes asymptotiques. Une telle approche implique de substituer la résolvante associé à une grande matrice aléatoire par la solution d'une équation à point fixe déterministe appelée équation de Dyson matricielle. Une autre technique efficace, connue sous le nom de truc de linéarisation, consiste à intégrer une expression matricielle dans une plus grande matrice aléatoire, appelée linéarisation, avec une structure de corrélation simplifiée.

Dans cette thèse, nous étendons le cadre théorique de l'équation de Dyson matricielle pour dériver une loi globale anisotrope pour une large classe de pseudo-résolvantes avec des structures de corrélation générales. Cette extension permet l'analyse des propriétés spectrales d'une large gamme de matrices aléatoires à l'aide d'une solution plus simple et déterministe de l'équation de Dyson matricielle. À travers le développement de cette théorie, nous abordons des aspects critiques tels que l'existence d'une solution unique, les bornes de support spectral et les propriétés de stabilité. Ces considérations sont essentielles pour construire des équivalents déterministes pour les pseudo-résolvantes d'une classe de linéarisations corrélées.

En tirant parti de ce cadre théorique, nous fournissons une expression déterministe asymptotiquement exacte pour l'erreur de validation empirique de la régression ridge pour le modèle de caractéristiques aléatoires. Le modèle à caractéristiques aléatoires, caractérisé par sa fonction d'activation non linéaire et son potentiel de surparamétrage, émerge comme un modèle puissant pour étudier les phénomènes observés dans les modèles d'apprentissage automatique de la vie réelle, tels que la descente multiple et la régularisation implicite. Notre expression exacte facilite une caractérisation précise de la régularisation implicite du modèle et révèle des liens entre le modèle de caractéristiques aléatoires et les méthodes de noyau étroitement reliée. Comme nous ne faisons aucune hypothèse particulière par rapport à la distribution des données, notre travail représente une avancée significative vers la compréhension de la manière dont les réseaux neuronaux exploitent des structures de données spécifiques.

Acknowledgements

I wish to extend my deepest gratitude to Professors Courtney and Elliot Paquette, my supervisors, for their guidance and support throughout my academic journey. I consider myself incredibly fortunate to have had the opportunity to learn from them, and I am grateful for their constant availability, advice, and encouragement. Their mentorship has been instrumental in shaping me into the researcher I am today. They exemplify excellence in both research and mentorship, serving as role models for aspiring researchers like myself. I would also like to express my heartfelt thanks to Courtney for generously funding my studies and research. Courtney and Elliot, thank you for everything.

I am also grateful to Professor Tim Hoheisel for introducing me to the field of mathematical optimization and providing continuous guidance. Although optimization is underrepresented in this thesis, it has significantly shaped my academic journey and greatly influenced my research interests.

While I cannot name everyone, I would like to express my gratitude to the numerous brilliant students and faculty members who have made my time at McGill University memorable and enjoyable.

Last but not least, I extend my heartfelt thanks to my family, friends, and partner for their unwavering support, encouragement, and understanding throughout this journey.

Contribution

This thesis focuses on results that were established in [LP23] by the author and his co-advisor Elliot Paquette. We provide a brief summary of the contributions of the author to the original work presented in this thesis.

First, Chapter 3 develops the theory for the matrix Dyson equation for general linearizations. The result and the proof regarding the existence of a solution to the matrix Dyson equation is entirely novel. Furthermore, we establish stability properties of the MDE through the innovative use of the Carathedory-Riffen-Finsler pseudometric, a methodology that holds independent interest. More generally, the approach used to demonstrate that the unique solution to the matrix Dyson equation serves as an asymptotic deterministic equivalent for a pseudo-resolvent draws inspiration from the Gaussian concentration approach in [LLC18], making us the first, to our knowledge, to apply this concept in the context of the matrix Dyson equation.

In Chapter 4, the theory developed in Chapter 3 is applied to provide a deterministic equivalent for the empirical test error of random features ridge regression. Both the result and its accompanying proof are original contributions developed by the author under the supervision and with support from co-advisor Elliot Paquette.

Throughout Chapter 3 and Chapter 4, the author played a lead role in formulating and proving the majority of the results, proposing propositions, developing novel proof concepts that led to the main results and conducting numerical simulations. The author benefited from the guidance and insights of co-supervisors Courtney and Elliot Paquette, whose input was instrumental in shaping the direction of the project. Konstantinos Christopher Tsiolis contributed to the initial development of numerical simulations results for Chapter 4, upon which the author expanded.

Contents

List of Figures

List of Abbreviations & Symbols

Abbreviations

CIFAR	Canadian Institute For Advanced Research
CRF	Carathéodory-Riffen-Finsler
GOE	Gaussian orthogonal ensemble
i.i.d.	Independent and identically distributed
MDE	Matrix Dyson equation
MNIST	Modified National Institute of Standards and Technology
RMDE	Regularized matrix Dyson equation
WLOG	Without loss of generality

Symbols

$0_{n\times d}$	$n \times d$ matrix of zeros
$\mathbb{B}$	Open complex unit ball
$\mathbb{C}$	Complex numbers
$\mathbb{E}$	Expectation
$\mathbb{H}$	Upper-half complex plane

$\mathbb{P}$	Probability
$\mathbb{R}$	Real numbers
$\mathscr{B}_\epsilon(x)$	Open ball of radius ϵ centered at x
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
diag	Diagonal or block-diagonal matrix
$\mathbb{I}$	Indicator function
$\lesssim$	Less than or equal to up to a constant
$\lVert \cdot \rVert$	Euclidean or operator norm
$\lVert \cdot \rVert_*$	Nuclear norm
$\lVert \cdot \rVert_F$	Frobenius norm
$\propto$	Proportional to
supp	Support
tr	Matrix trace
$\vee$	Maximum
$\wedge$	Minimum
I_n	$n \times n$ identity matrix
M^*	Conjugate transpose of M
M^T	Transpose of M
$M^{-1}_{j,k}$	j, k sub-block of the inverse of M
$o_n(f)$	Functions $g(n)$ such that $\lim_{n\to\infty} g(n)/f(n) = 0$

1

Introduction

The development of artificial intelligence is characterized by a transition from elementary rule-based systems to the dominant paradigm of data-driven methodologies, particularly in the realm of machine learning. Fundamentally, machine learning seeks to establish a statistical relationship based on a given set of examples, typically comprising samples and their associated response variables.

The trajectory of machine learning has undergone a profound transformation, fueled by significant advancements in processing power and the widespread accessibility of vast datasets. In stark contrast to contemporary practices, the machine learning landscape in 1959 is exemplified by the work of Samuel [Sam59]. This study utilized a rudimentary machine learning model with fewer than 50 parameters trained on 53 000 datapoints to play checkers. Notably, the author reported that the program was able to outperform the person who programmed it in a game of checkers. Nowadays, machine learning models are trained using large datasets within high-dimensional feature spaces. For instance, the GPT-3 model, a precursor to the widely-used conversational agent ChatGPT, stands out with an impressive 175 billion parameters and training on 374 billion data points [Bro+20]. Additionally, the image generation systems DALL-E and DALL-E 2, capable of producing photo-realistic images from text prompts, leverages a staggering 12 billion and 3.5 billion parameters respectively, trained on 250 million and 650 million data points respectively [Ram+21; Ram+22]. These examples exemplify an enduring trend of embracing progressively larger models in contemporary machine learning practices.

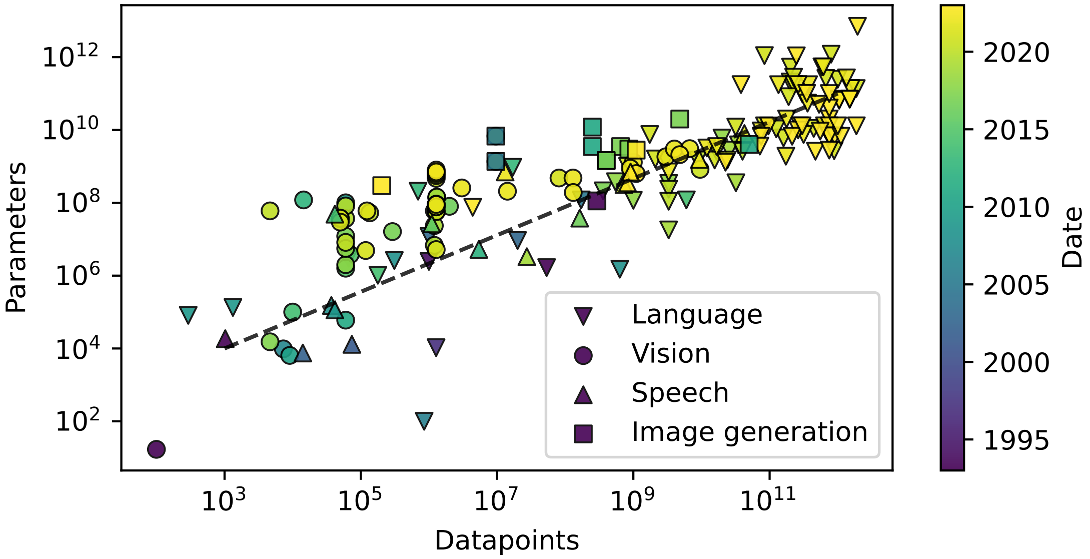

Figure 1.1: Relationship between the number of datapoints and the number of parameters in machine learning models, color-coded by date. The data, sourced from [Epo22], reveals a distinct trend of utilizing more datapoints with an increasing number of parameters over time.

This escalation in model complexity raises critical questions about the theoretical underpinnings of such expansive models. This inquiry is driven by the so-called “curse of dimensionality”, a term coined by Bellman in his exposition on dynamic programming [Bel10]. The concept is loosely used to convey the idea that low-dimensional intuition, as well as computational or theoretical approaches, may break down entirely in high-dimensional scenarios [CL22]. Manifestations of this curse are observable in various aspects in the context of machine learning. As a first example, traditional performance metrics for optimization techniques often rely on worst-case complexity. Yet, in real-world, high-dimensional scenarios, the likelihood of encountering a data point associated with the worst-case running time of a given algorithm may be exceedingly low. Therefore, while worst-case complexity offers assurances regarding running time, it may not accurately reflect the algorithm's expected performance in high dimensions. A second example is found in the analysis of artificial neural networks. When overparameterized, meaning the number of trainable parameters exceeds the number of data points, these models possess extraordinary capacity. In fact, they are capable of fitting given data perfectly, even when the labels are pure noise [Zha+21]. In this scenario, the fact that they still exhibit good generalization performance contravenes conventional statistical knowledge. Furthermore, the presence of non-linear activation functions adds complexity, making them challenging to analyze analytically.

Fortunately, the scale of contemporary models not only brings about the curse of dimensionality but also offers a blessing. Conceptualizing the problem as random, we may leverage the fact that scalar observations of large random systems often follow a law of large numbers effect and concentrate around a deterministic quantity. If this quantity accurately characterizes observed behavior in practice, one can argue that the chosen model is satisfactory. To describe this deterministic limit, we can turn to tools from random matrix theory. The origins of random matrix theory can be traced back to the work of Wishart, who investigated the eigenvalues of large sample covariance matrices [Wis28]. The theory gained significant momentum after the contributions of Wigner, who explored the spacing between the eigenvalues of a symmetric random matrix, using it as a model for the spacings between the lines in the spectrum of heavy atomic nuclei [Wig55].

In the analysis of machine learning, random matrix theory appears because one chooses to simplify the model by assuming that a component of it is random. To illustrate this concept, consider a traditional supervised problem setting. In a typical supervised problem scenario, we are provided with a labeled dataset $\mathscr{D} = \{(x_j, y_j)\}_{j=1}^{n_{\text{train}}}$, where each sample $x_j \in \mathbb{R}^{n_0}$ is associated with a label $y_j \in \mathbb{R}$ for $j = 1, 2, \ldots, n_{\text{train}}$. For conciseness, we organize the data into a matrix $X \in \mathbb{R}^{n_{\text{train}} \times n_0}$, where the jth row of X represents x_j^T, and a vector $y \in \mathbb{R}^{n_{\text{train}}}$ representing the labels. The objective is to establish a relationship between the inputs x_j and the corresponding outputs y_j. To achieve this, we confine our focus to a class of parametric functions. One of the simplest models is the linear model, which represents the relationship using the mapping $x \mapsto x^T w$ for some weights $w \in \mathbb{R}^{n_0}$. Once we have selected a model, we use the dataset to find optimal weights by minimizing a loss function. A fundamental loss is obtained by taking the squared norm of the residuals and leads the optimization problem $\min_{w \in \mathbb{R}^d} \|y - Xw\|^2$. This optimization problem is called linear regression. It serves as a simple tractable model to study the behavior of iterative minimization procedures used to train more complex models. Using gradient descent with a constant step size, a straightforward iterative optimization algorithm, generates a sequence of iterates $\{w_k\}_{k=0}^{\infty}$, where $w_0 \in \mathbb{R}^{n_0}$ is arbitrary and $w_{k+1} = w_k - \gamma X^T(Xw_k - y)$ for $k = 1, 2, \ldots$. Assuming that the linear model is correct, that is there exists a ground truth vector $w_* \in \mathbb{R}^{n_0}$ such that $y = Xw_*$, we can unfold the iterative procedure and express $w_k - w_* = (I_{n_0} - \gamma X^T X)^k (w_0 - w_*)$. Substituting this expression into the loss function, we can represent the loss function after k iterations of gradient descent as $\|y - Xw_k\|^2 = \|X(I_{n_0} - \gamma X^T X)^k (w_0 - w_*)\|^2 = (w_0 - w_*)^T X^T X (I_{n_0} - \gamma X^T X)^{2k} (w_0 - w_*)$. Although this expression represents the training loss of the simplest model trained using the simplest optimization algorithm, the behavior of the loss during training depends non-trivially on the data matrix through $X^T X (I_{n_0} - \gamma X^T X)^{2k}$. However, in machine learning, it is common to assume that the data is sampled from a certain distribution. Therefore, to analyze the training loss of a linear regression model at the iterates of gradient descent, it is reasonable to assume that the entries of X are random variables. Under some additional statistical assumptions, tools from random matrix theory can be employed to study the polynomial of random matrices $X^T X (I_{n_0} - \gamma X^T X)^{2k}$ for large n_{train} and n_0. More precisely, it can be shown that as both n_{train} and n_0 grow to infinity proportionally, the loss function at the iterates of gradient descent approaches a deterministic limit, dependent only on the distribution of the entries of X through their first two moments [Paq+23]. This type of analysis can be extended to other loss functions and variants of stochastic gradient descent [Paq+23; Col+23; Lee+22; PP21; Paq+21]. Such analyzes provide realistic expectations for training time and aid in selecting hyperparameters without conducting extensive grid searches.

Another related area where random matrix theory is beneficial is in analyzing the loss landscape associated with the training of machine learning models. These landscapes, which are typically non-convex and often non-smooth, present challenges in optimization. Nonetheless, empirical models often achieve remarkable performance. By examining the eigenvalues and alignment of eigenvectors of the Hessian matrix, which captures second-order information about the loss function, we can gain valuable insights into the structure and properties of the landscape [LM21]. A better understanding of the loss landscape can help explain the successes and limitations of various optimization algorithms.

In practice, the linear model often proves insufficiently expressive for representing complex functions. Hence, a more general class of parametric function is given by fully connected 2-layer feed-forward neural networks. In this model, the relationship between samples and labels is modeled using the function $x \mapsto \sigma(x^T W)w$, where $W \in \mathbb{R}^{n_0 \times d}$ and $w \in \mathbb{R}^d$ are respectively the weight matrix and vector, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ denotes an activation function applied entrywise. The activation function introduces non-linearity into the model, and allows the representation of more complex functions. Common examples of activation functions include the rectified linear unit (ReLU), hyperbolic tangent function, and sigmoid function. Although we omitted it for the sake of discussion, we note that the output of the network is often scaled by a normalizing constant to maintain stability as dimensions increase proportionally. Using a regularized version of the norm squared loss, we consider the minimization problem

$$\min_{W \in \mathbb{R}^{n_0 \times d},\, w \in \mathbb{R}^d} \|y - Aw\|^2 + \delta \|w\|^2$$

where we defined $A = d^{-1/2}\sigma(XW)$ for notational convenience. The addition of the regularization term makes the optimization problem strongly convex. In fact, the minimization problem admits the closed-form solution $w_{\text{ridge}} = A^T(AA^T + \delta I_{n_{\text{train}}})^{-1} y$, which is called the ridge estimator. To quantify how well our model performs on the training dataset, we can look at the squared norm of the residuals $\|y - AA^T(AA^T + \delta I_{n_{\text{train}}})^{-1} y\|^2$. Expanding the squared norm, we get a quantity that depends on bilinear forms of rational expressions in AA^T. Motivated similarly to the linear model case, we may assume that the dataset is sampled from a distribution. Alternatively, a common theoretical exploration technique involves assuming randomness in the weight matrix. Given that neural networks are often initialized with random weights, this corresponds to constructing a two-layer neural network, fixing the first-layer weights at random initialization, and training only the second layer. This approach gives rise to the popular random features model introduced in [RR07]. Unlike the linear model, the random features model incorporates a non-linear activation function and has the capacity for overparameterization. However, it remains tractable, serving as a suitable tool for studying phenomena observed in real-life machine learning models, such as multiple descent [MM22; AP20b; Bel+19] and implicit regularization [Cho22; Jac+20]. Further details on the random features model and a result about the empirical test error will be discussed in Chapter 4.

A more practical application of random matrix theory in machine learning lies in selecting suitable estimators for generalization error [WHS22]. While generalization error serves as a crucial metric in machine learning, accurately estimating it poses a challenge. Random matrix theory also finds application to study scaling laws [Bah+21], which delineate power-law scaling relationships among various dimensions of system size and computational resources. Understanding these scaling laws offers insights into the performance of machine learning models as the size of the training dataset increases. Such insights are indispensable for devising more efficient and scalable machine learning algorithms.

In the examples provided and in many other problems, theoretical analysis often hinges on comprehending the behavior of bilinear forms or traces of rational expressions of random matrices. Crucially, the behavior of the object in such cases is not contingent upon any finite number of individual entries; rather, it necessitates an understanding of the collective behavior arising from global interactions between these entries. An important consideration in this approach is ensuring that the distribution from which random variables are drawn accurately reflects the statistical properties of real-world scenarios. For instance, in the context of the random features model, where random weights are employed, the model effectively represents the neural network at its early training stages and serves as a valuable theoretical tool. However, it falls short in capturing the full generality of a 2-layer neural network. Moreover, while isotropic distributions often suffice to provide interesting theoretical results, they may not consistently mirror the statistical properties of real-world datasets. For example, the Modified National Institute of Standards and Technology (MNIST) database [Den12], which contains handwritten digits from 0 to 9, cannot be accurately characterized by assuming that samples are drawn from a multivariate standard normal distribution alone, as illustrated in Figure 1.2. Instead, a more realistic model might entail a spiked Gaussian mixture distribution. On a related note, empirical evidence suggests that part of the efficacy of neural networks stems from their capacity to leverage the inherent structure within data to derive suitable representations, a concept commonly referred to as feature learning [Ba+22]. Isotropic distributions, by their very nature, fall short in capturing such structural nuances. To attain a more comprehensive understanding, it becomes imperative to incorporate more intricate correlations within the random matrix framework.

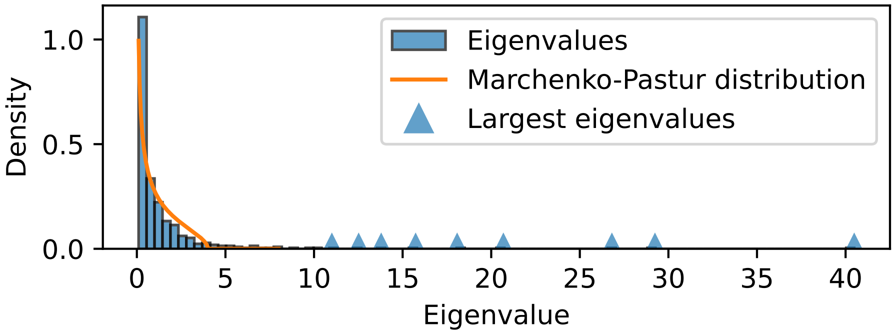

Figure 1.2: The spectrum of the sample covariance matrix associated with the MNIST dataset can be decomposed into a noise component, depicted by the overlapping Marchenko-Pastur distribution with shape parameter 1, and a low-rank signal component characterized by outlier eigenvalues.

In this thesis, which builds upon the work of a previous article co-authored by the author and his co-supervisor Elliot Paquette [LP23], we commence by delving into foundational concepts that serve as essential background material for our subsequent discussions. Following this, we introduce a comprehensive framework tailored to analyze rational expressions of random matrices characterized by general correlation structures. Central to our framework is the utilization of the matrix Dyson equation, a deterministic fixed-point equation, which offers a means to study scalar observations of random matrices. We contribute to the existing literature on the matrix Dyson equation by expanding its application to derive deterministic equivalents for general pseudo-resolvents. These pseudo-resolvents naturally emerge from our adoption of the linearization trick, also known as the linear pencil method. This approach involves representing rational functions of random matrices as blocks of inverses of larger random matrices that linearly depend on their random matrix inputs. Such linearizations possess simpler correlation structures, rendering them more amenable to certain types of analysis. This notably explains why they have been successfully utilized in conjunction with tools from operator-valued free probability to analyze simple machine learning models. After reviewing relevant literature and laying the groundwork for our study, we systematically develop our framework. This involves establishing the existence of a unique solution to the matrix Dyson equation, demonstrating the stability of the equation in a suitable context, and leveraging this stability to assert that the solution serves as a surrogate for studying random matrices.

Given the preceding discussion, we have tailored our framework with a specific focus on its relevance to machine learning applications. In order to illustrate the practical utility of our framework, we apply it to analyze the empirical test error of random features ridge regression. This application enables us to precisely quantify the implicit regularization inherent in the model. Furthermore, it allows us to draw interesting connections to a closely related kernel method. Importantly, our framework enables the consideration of anisotropic random features models, yielding results consistent with empirical observations on real-world datasets.

2

Preliminaries

Before delving into the core of this thesis, it is essential to establish a foundational understanding of key concepts. In this preliminary chapter, we introduce fundamental topics in complex analysis and random matrix theory. These concepts serve as essential tools for comprehending the discussions that follow. Instead of presenting a comprehensive introduction to these topics, we will focus on concepts directly relevant to the objectives of this thesis and refer the reader to appropriate references for more information.

2.1 Complex Analysis

Since our focus lies in understanding the behavior of rational expressions of matrices, we will frequently encounter matrix-valued expressions. Certain concepts from complex analysis, typically introduced for functions of complex variables, can be naturally extended to more general complex normed vector spaces. We will review these concepts as they are crucial for our purposes. Additionally, we will define the Carathéodory-Riffen-Finsler-pseudometric and discuss its properties. Utilizing this pseudometric, we will introduce the Earle-Hamilton fixed-point theorem, which will play a vital role in constructing our theoretical framework.

2.1.1 Holomorphic Functions

The content presented in this section is standard and can be found in various references, such as [Hil48]. Throughout this discussion, we consider $\mathscr{X}$ and $\mathscr{Y}$ as normed vector spaces over the complex numbers, with $\mathscr{D}$ being an open subset of $\mathscr{X}$.

A function $f : \mathscr{D} \mapsto \mathscr{Y}$ is termed *Fréchet differentiable* at $x \in \mathscr{D}$ if there exists a bounded linear operator $\mathrm{D}f(x) : \mathscr{X} \mapsto \mathscr{Y}$ satisfying

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - \mathrm{D}f(x)h\|}{\|h\|} = 0.$$

When this condition holds, $\mathrm{D}f(x)$ is referred to as the Fréchet derivative of f at x. A function f is deemed *holomorphic* in $\mathscr{D}$ if the Fréchet derivative of f exists as a bounded complex linear map from $\mathscr{X}$ onto $\mathscr{Y}$ for every $x \in \mathscr{D}$. We will denote by $\mathrm{Hol}(\mathscr{D}, \mathscr{Y})$ the set of all holomorphic functions from $\mathscr{D}$ to $\mathscr{Y}$. Fréchet differentiability extends the concept of complex differentiability. Importantly, when f is holomorphic in $\mathscr{D}$, the nth order Fréchet derivative of f at x, denoted $D^n f(x)$, exists as a symmetric multilinear mapping from $\mathscr{X}^n = \mathscr{X} \times \cdots \times \mathscr{X}$ to the completion of $\mathscr{Y}$ as a Banach space. This connection allows us to establish a link between holomorphicity and analyticity. Let $\mathscr{B}_r(x_0)$ denote the open ball of radius r centered at x_0 in $\mathscr{X}$. If $f : \mathscr{B}_r(x_0) \subseteq \mathscr{X} \mapsto \mathscr{Y}$ is a bounded holomorphic function, then

$$f(x) = \sum_{n=0}^{\infty} \frac{\mathrm{D}^n f(x_0)}{n!}(x - x_0)^n$$

for all $x \in \mathscr{B}_r(x_0)$. Moreover, this series converges uniformly on $\mathscr{B}_s(x_0)$ for every $s \in (0, r)$ [Hil48, Theorem 3.17.1].

2.1.2 Carathéodory-Riffen-Finsler Pseudometric

Let $\mathscr{D}$ be a domain in a normed linear space over the complex numbers $\mathscr{X}$. We define the *infinitesimal Carathéodory-Riffen-Finsler (CRF)-pseudometric* as

$$\alpha : (x, v) \in \mathscr{D} \times \mathscr{X} \mapsto \sup\{\|\mathrm{D}f(x)v\| \, : \, f \in \mathrm{Hol}(\mathscr{D}, \mathbb{B})\} \in \mathbb{R},$$

where $\mathbb{B}$ denotes the open complex unit ball of unit radius [Har03; Har79]. The pseudometric α is an *infinitesimal Finsler pseudometric* on $\mathscr{D}$, meaning that it is non-negative, lower semicontinuous, locally bounded, and satisfies $\alpha(x, tv) = |t|\alpha(x, v)$ for every $(x, v) \in \mathscr{D} \times \mathscr{X}$ and $t \in \mathbb{R}$.

Let Γ be the set of all curves in $\mathscr{D}$ with piecewise continuous derivatives, referred to as *admissible* curves, and define

$$\mathscr{L} : \gamma \in \Gamma \mapsto \int_0^1 \alpha(\gamma(y), \gamma'(t))\mathrm{d}t \in \mathbb{R}.$$

The pseudometric α is a seminorm at each point in $\mathscr{D}$, and we interpret $\mathscr{L}(\gamma)$ as the length of the curve γ measured with respect to α [Har03]. Then, the *CRF-pseudometric* ρ of $\mathscr{D}$ is defined as

$$\rho : (x, y) \in \mathscr{D}^2 \mapsto \inf\{\mathscr{L}(\gamma) \,:\, \gamma \in \Gamma,\, \gamma(0) = x,\, \gamma(1) = y\} \in \mathbb{R}_{\geq 0}.$$

As the name suggests, ρ is a pseudometric.

Our main tool for extrapolating results about norms is the Schwarz-Pick inequality, which we state here for completeness.

Proposition 2.1.1 ([Har79, Proposition 3])**.** *Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be domains in complex normed vector spaces, and let ρ_1 and ρ_2 be the associated CRF-pseudometrics. If $f : \mathscr{D}_1 \mapsto \mathscr{D}_2$ is holomorphic, then $\rho_2(f(x), f(y)) \leq \rho_1(x, y)$ for all $x, y \in \mathscr{D}_1$.*

In fact, the inequality in Proposition 2.1.1 can be replaced by an equality when the function f is a biholomorphic mapping. This means that the CRF-pseudometric is biholomorphically invariant [HFS07]. In some sense, Proposition 2.1.1 indicates that the CRF-pseudometric is non-expansive on the space of holomorphic functions mapping a domain onto itself.

If we denote by $\rho_{\mathbb{B}}$ the CRF-pseudometric on the complex open unit disk $\mathbb{B}$, then Proposition 2.1.1 becomes particularly useful because $\rho_{\mathbb{B}}$, also known as the *Poincaré metric*, admits the closed form expression

$$\rho_\Delta(z_1, z_2) = \operatorname{arctanh}\left|\frac{z_1 - z_2}{1 - \bar{z}_1 z_2}\right|. \tag{2.1}$$

For a derivation of (2.1), refer to [Har79, Example 2].

2.1.3 Holomorphic Fixed-Point Theorem

One of the most well-known fixed-point results is the Banach fixed-point theorem, also known as the Banach contraction mapping theorem, which asserts that contractive self-maps on Banach spaces have unique fixed points. The *Earle-Hamilton fixed-point theorem* can be viewed as an extension of the Banach fixed-point theorem to holomorphic functions. Essentially, it asserts that every strictly holomorphic function defined on a domain of a complex Banach space has a unique fixed point. Here, by strictly holomorphic function, we refer to a holomorphic function $f : \mathscr{D} \mapsto \mathscr{D}$ such that there exists $\epsilon \in \mathbb{R}_{>0}$ such that $y \in f(\mathscr{D})$ for every $x \in f(\mathscr{D})$ and $y \in \mathscr{D}$ satisfying $\|x - y\| < \epsilon$. This theorem leverages the CRF-pseudometric, in conjunction with Proposition 2.1.1, to reduce strict holomorphicity to contractiveness. Subsequently, it employs the Banach fixed-point theorem to establish its conclusion. Given the significance of the Earle-Hamilton fixed-point theorem to subsequent sections, we provide its statement and proof.

Theorem 2.1.1 (Earle-Hamilton Fixed-Point [EH70])**.** *Let $\mathscr{D}$ be a non-empty domain in a complex Banach space $\mathscr{X}$. If $f : \mathscr{D} \mapsto \mathscr{D}$ is a bounded strictly holomorphic function, then f has a unique fixed point in $\mathscr{D}$. Furthermore, for any $x_0 \in \mathscr{D}$, the sequence $x_{k+1} = f(x_k)$ converges, in norm, to the unique fixed point of f.*

Proof. We follow the proof techniques outlined in [Har03, Theorem 3.1] and [Har79, Theorem 4]. By assumption, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathscr{B}_\epsilon(f(x)) \subseteq \mathscr{D}$ for every $x \in \mathscr{D}$. To clarify, $\mathscr{B}_\epsilon(x)$ denotes the open ball of radius ϵ centered at x. Since f is bounded, we can assume without loss of generality that $\mathscr{D}$ is bounded by setting $\mathscr{D} = \bigcup_{x\in\mathscr{D}} \mathscr{B}_\epsilon(f(x))$.

Let $\delta = \epsilon(\sup_{x,y\in\mathscr{D}} \|x - y\|)^{-1}$ and $x \in \mathscr{D}$. Define the function

$$h : y \in \mathscr{D} \mapsto f(y) + \delta\left(f(y) - f(x)\right) \in \mathscr{D}.$$

Then, $\delta\|f(y)-f(x)\| \leq \epsilon$ and h is a holomorphic mapping of $\mathscr{D}$ onto itself. For every $v \in \mathscr{X}$, we have $\mathrm{D}h(x)v = (1+\delta)\mathrm{D}f(x)v$. Therefore, for $g \in \mathrm{Hol}(\mathscr{D},\mathbb{B})$, it follows from the definition of the infinitesimal CRF-pseudometric that

$$(1+\delta)\|\mathrm{D}g(h(x))\mathrm{D}f(x)v\| = \|\mathrm{D}g(h(x))\mathrm{D}h(x)v\| = \|\mathrm{D}(g \circ h)(x)v\| \leq \alpha(x,v),$$

where α denotes the infinitesimal CRF-pseudometric on $\mathscr{D}$. Since $g \circ h = g \circ (y \mapsto y + \delta(y - f(x))) \circ f \in \mathrm{Hol}(\mathscr{D},\mathbb{B})$, this implies that $\alpha(f(x),\mathrm{D}f(x)v) \leq (1+\delta)^{-1}\alpha(x,v)$. According to [Har79, Lemma 1], we get $\rho(f(x),f(y)) \leq (1+\delta)^{-1}\rho(x,y)$ for every $x,y \in \mathscr{D}$.

Now, let x_0 be arbitrary, and consider the sequence $\{x_k\}_{k=0}^{\infty}$ with $x_{k+1} = f(x_k)$ for every $k = 1,2,\ldots$. It follows easily from the above that the sequence is Cauchy with respect to the CRF-pseudometric on $\mathscr{D}$. Since the underlying Banach space is complete with respect to the norm $\|\cdot\|$, it only remains to show that the CRF-pseudometric majorizes the norm, i.e. $\rho(x,y) \geq c\|x-y\|$ for some $c \in \mathbb{R}_{>0}$.

Consider $x,y \in \mathscr{D}$. By the Hahn-Banach theorem, there exists a linear operator $L \in \mathscr{X}^*$ with $\|L\| = 1$ such that $L(x-y) = \|x-y\|$. Define the holomorphic function

$$h : w \in \mathscr{D} \mapsto \frac{L(u-y)}{\sup_{x,y\in\mathscr{D}}\|x-y\|} \in \mathbb{B}.$$

Then, if we let ρ be the CRF-pseudometric on $\mathscr{D}$ and $\rho_{\mathbb{B}}$ be the CRF-pseudometric associated with $\mathbb{B}$, it follows from Proposition 2.1.1 and (2.1) that

$$\operatorname{arctanh}\left(\frac{\|x-y\|}{\sup_{x,y\in\mathscr{D}}\|x-y\|}\right) = \rho_{\mathbb{B}}\left(h(x),h(y)\right) \leq \rho(x,y).$$

In particular, this implies that the sequence $\{x_k\}_{k=0}^{\infty}$ is Cauchy with respect to the norm on $\mathscr{X}$. Since $(\mathscr{X},\|\cdot\|)$ is complete, there exists a limit point $x_\infty \in \mathscr{D}$. This limit point must be a fixed point of f.

To demonstrate uniqueness, assume that $x,y \in \mathscr{D}$ are two fixed points for f. Then, we must have $\rho(x,y) = \rho(f(x),f(y)) \leq (1+\delta)^{-1}\rho(x,y)$, implying $\rho(x,y) = 0$. As the CRF-pseudometric majorizes the norm in $\mathscr{D}$, the fixed point is unique. □

2.2 Matrix Identities

In this section, we introduce various matrix identities, including the block inversion lemma and the Herglotz-Nevanlinna representation theorem, providing essential tools in random matrix theory.

2.2.1 General Matrix Identities

We begin by presenting some standard, general matrix identities.

Lemma 2.2.1. *If $M_1, M_2 \in \mathbb{C}^{n\times n}$ are non-singular, then $M_1^{-1} - M_2^{-1} = M_1^{-1}(M_2 - M_1)M_2^{-1}$.*

Proof. Multiply on the left by M_1 and on the right by M_2. □

Lemma 2.2.2. *Let $z \in \mathbb{C}$, $M \in \mathbb{C}^{n\times n}$ and assume that $\|M\| \le a < b \le |z|$ from some $a, b \in \mathbb{R}_{\ge 0}$. Then, $M - zI_n$ is non-singular and $\|(M - zI_n)^{-1}\| \le (b-a)^{-1}$.*

Proof. For every $v \in \mathbb{C}^n$, we have $\|(M - zI_n)v\| \ge \|zv\| - \|Mv\| \ge (|z| - \|M\|)\|v\|$. This implies that $M - zI_n$ is non-singular. Choosing $v = (M - zI_n)^{-1}u$ for some unit vector u, we have $\|(M - zI_n)^{-1}u\| \le (|z| - \|M\|)^{-1}$. Taking the supremum over unit vectors u and using the definition of spectral norm, we obtain the desired result. □

Lemma 2.2.3. *For every $M_1, M_2^T \in \mathbb{C}^{n\times d}$ and $z \in \mathbb{C}$ such that both $M_1M_2 - zI_n$ and $M_2M_1^T - zI_d$ are non-singular, we have $M_1(M_2M_1 - zI_d)^{-1} = (M_1M_2 - zI_n)^{-1}M_1$.*

Proof. Left-multiply the equation on both sides by $M_1M_2 - zI_n$ and right-multiply by $M_2M_1 - zI_d$. □

2.2.2 Real and Imaginary Parts of Matrices

Apart from general matrix identities, we will also need to consider the real and imaginary parts of matrices. Just like complex numbers, we can decompose a complex matrix $M \in \mathbb{C}^{n\times n}$ as $M = \Re[M] + i\Im[M]$ where $2\Re[M] = M + M^*$ and $2i\Im[M] = M - M^*$. The real and imaginary parts of M are Hermitian. The following lemma states that the norm of the real and imaginary parts of a matrix are bounded by the norm of the matrix itself.

Lemma 2.2.4. *For every $M \in \mathbb{C}^{n\times n}$, $\|\Re[M]\| \vee \|\Im[M]\| \le \|M\|$.*

Proof. Let $v \in \mathbb{C}^n$ be a complex unitary vector. By Cauchy-Schwarz's inequality, $\|Mv\|^2 = \|Mv\|^2\|v\|^2 \geq |v^*Mv|^2 = |v^*\Re[M]v + iv^*\Im[M]v|^2$. Since both $\Re[M]$ and $\Im[M]$ are Hermitian, the quadratic forms $v^*\Re[M]v$ and $v^*\Im[M]v$ are real. Hence, $|v^*\Re[M]v + iv^*\Im[M]v|^2 = (v^*\Re[M]v)^2 + (v^*\Im[M]v)^2$. Taking the supremum over all unitary vectors v, we obtain the desired result. □

The proof of Lemma 2.2.4 is as crucial as the statement itself, if not more so. For instance, the following lemma follows directly from this argument.

Lemma 2.2.5. *Let $M \in \mathbb{C}^{n\times n}$. If there exists $a \in \mathbb{R}_{>0}$ such that $\Re[M] \succeq aI_n$, $\Re[M] \preceq -aI_n$, $\Im[M] \succeq aI_n$ or $\Im[M] \preceq -aI_n$, then M is non-singular and $\|M^{-1}\| \leq a^{-1}$.*

Proof. Assume that $\Re[M] \succeq aI_n$ or $\Re[M] \preceq -aI_n$. By the proof of Lemma 2.2.4, we have $\|Mv\| \geq |v^*\Re[M]v| \geq a$ for every unitary $v \in \mathbb{C}^n$. Hence, M is non-singular. Taking $v = \frac{M^{-1}u}{\|M^{-1}u\|}$ for some unitary $u \in \mathbb{C}^n \setminus \{0\}$, we have $\|M^{-1}u\| \leq a^{-1}$. Taking the supremum over all unitary u gives the result first half of the result. The second half follows similarly. □

Another important lemma relates the real and imaginary parts of an inverse matrix to those of the original matrix.

Lemma 2.2.6. *Let $M \in \mathbb{C}^{n\times n}$ be invertible. Then, $\Re[M^{-1}] = M^{-1}\Re[M]M^{-*}$ and $\Im[M^{-1}] = -M^{-1}\Im[M]M^{-*}$.*

Proof. Write $M = \Re[M] + i\Im[M]$. Since the matrix $\Re[M]$ is Hermitian and $i\Im[M]$ is skew-Hermitian, we have $M^* = (\Re[M]+i\Im[M])^* = \Re[M]-i\Im[M]$. By the definition of matrix real and imaginary parts as well as Lemma 2.2.1, $2\Re[M^{-1}] = M^{-1}+M^{-*} = M^{-1}(M^*+M)M^{-*} = 2M^{-1}\Re[M]M^{-*}$. The proof for the imaginary part is similar. □

2.2.3 Matrix Norms

Throughout this thesis, we will utilize several useful inequalities involving norms. The first one pertains to the Frobenius norm of a product.

Lemma 2.2.7. *Let $M_1 \in \mathbb{C}^{n\times d}$ and $M_2 \in \mathbb{C}^{d\times m}$ be arbitrary matrices. Then, $\|M_1M_2\|_F \leq \|M_1\|\|M_2\|_F$ and $\|M_1M_2\|_F \leq \|M_1\|_F\|M_2\|$.*

Proof. By definition, $\|M_1M_2\|_F^2 = \operatorname{tr}(M_2^*M_1^*M_1M_2)$. Using the cyclic property of the trace, $\operatorname{tr}(M_2^*M_1^*M_1M_2) = \operatorname{tr}(M_2M_2^*M_1^*M_1)$. Let $M_2M_2^* = U\Lambda U^*$ for some unitary $U \in \mathbb{C}^{d\times d}$ and real positive semidefinite diagonal Λ. With $W = U^*M_1^*M_1U$, we have $\operatorname{tr}(M_2^*M_1^*M_1M_2) = \operatorname{tr}(M_2M_2^*M_1^*M_1) = \sum_{j=1}^d \Lambda_{j,j}W_{j,j}$. Indeed, as $\|M_2\|^2 = \|M_2M_2^*\| = \max_j \Lambda_{j,j}^2$, we have $\sum_{j=1}^d \Lambda_{j,j}W_{j,j} \leq \|M_2\|^2 \operatorname{tr}(U^*M_1^*M_1U) = \|M_2\|^2\|M_1\|_F^2$. A similar argument can be made for the second inequality. □

The second inequality related the trace to the spectral norm and the nuclear norm.

Lemma 2.2.8. *Let $M, U \in \mathbb{C}^{n\times n}$. Then, $|\operatorname{tr}(UM)| \leq \|U\|_*\|M\|$.*

Proof. Let $\{u_j\}_{j=1}^n$ and $\{m_j\}_{j=1}^n$ be a non-increasing enumeration of the singular values of U and M, respectively. By Von Neumann's trace inequality, $|\operatorname{tr}(UM)| \leq \sum_{j=1}^n u_jm_j \leq m_1\sum_{j=1}^n u_j = \|U\|_*\|M\|$. □

2.2.4 Matrix-Valued Herglotz Function

In what follows, we let $\mathbb{H} := \{z \in \mathbb{C} : \Im[z] > 0\}$ denote the complex upper half-plane. A matrix-valued analytic function $M : \mathbb{H} \mapsto \mathbb{C}^{n\times n}$ satisfying $\Im[M(z)] \succeq 0$ for all $z \in \mathbb{H}$ is termed *Herglotz* [GT00]. The Nevanlinna, or Riesz-Herglotz, representation theorem is a fundamental result in complex analysis. In this thesis, we will require a matrix version of this theorem, which we state here for completeness.

Theorem 2.2.1 (Nevanlinna representation [GT00, Theorem 5.4(iv) and Theorem 2.3(iii)])**.** *Let $M : \mathbb{H} \mapsto \mathbb{C}^{n\times n}$ be a matrix-valued Herglotz function. Then, there exists a matrix-valued measure Ω on the bounded Borel subsets of $\mathbb{R}$ satisfying $\int_{\mathbb{R}} \frac{v^*\Omega(\mathrm{d}\lambda)v}{1+\lambda^2} < \infty$ for all $v \in \mathbb{C}^n$ such that*

$$M(z) = C + Dz + \int_{\mathbb{R}} \left(\frac{1}{\lambda - z} - \frac{\lambda}{1+\lambda^2}\right)\Omega(\mathrm{d}\lambda)$$

for every $z \in \mathbb{H}$ with $C = \Re[M(i)]$ and $D = \lim_{\eta\to\infty}(i\eta)^{-1}M(i\eta) \geq 0$. Furthermore, if $\lim_{\eta\to\infty} -i\eta M(i\eta) = E \in \mathbb{R}^{n\times n}$, then $M(z) = \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{\lambda - z}$ and $\int_{\mathbb{R}} \Omega(\mathrm{d}\lambda) = E$.

The Nevanlinna representation theorem, which we adapted from [GT00, Theorem 5.4(iv) and Theorem 2.3(iii)], is a powerful tool, especially due to the Stieltjes inversion lemma for the matrix-valued measure.

Lemma 2.2.9 (Stieltjes inversion formula [GT00, Theorem 5.4(v) and Theorem 5.4(vi)])**.** *Let M be a matrix-valued Herglotz function and Ω be the associated matrix-valued measure in the Nevanlinna representation theorem. Then, for $\lambda_1, \lambda_2 \in \mathbb{R}$ with $\lambda_1 \leq \lambda_2$,*

$$2^{-1}\Omega(\{\lambda_1\}) + 2^{-1}\Omega(\{\lambda_2\}) + \Omega((\lambda_1, \lambda_2)) = \pi^{-1} \lim_{\eta \downarrow 0} \int_{\lambda_1}^{\lambda_2} \Im[M(\lambda + i\eta)] \mathrm{d}\lambda.$$

Additionally, the absolutely continuous part Ω_{ac} of Ω is given by $\Omega_{\mathrm{ac}}(\mathrm{d}\lambda) = \lim_{\eta \downarrow 0} \pi^{-1} \Im[M(\lambda + i\eta)] \mathrm{d}\lambda$.

While the matrix Herglotz function is defined on the complex upper half-plane, we can analytically extend it to the complex lower half-plane through an open interval on which the matrix-valued measure is not supported.

Lemma 2.2.10 (Analytic continuation [GT00, Lemma 5.6])**.** *Let M be a matrix-valued Herglotz function with the associated matrix-valued measure Ω in the Nevanlinna representation. Then, M can be analytically continued through the open interval (λ_1, λ_2) by reflexion if and only if Ω is supported on $\mathbb{R} \setminus (\lambda_1, \lambda_2)$.*

2.2.5 Block Matrices

The main portion of this thesis will focus on block matrices. An important aspect will be determining when block matrices are non-singular and being able to express their inverses using rational expressions of the blocks. The block inversion lemma utilizes the Schur complement to express the inverse of a block matrix in terms of the inverses of its blocks.

Lemma 2.2.11 (Block matrix inversion lemma [LS02, Theorem 2.1])**.** *Let*

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

be a block matrix with $A \in \mathbb{C}^{n \times n}$, $B, C^T \in \mathbb{C}^{n \times d}$ and $D \in \mathbb{C}^{d \times d}$. If A is non-singular, then M is non-singular if and only if $D - CA^{-1}B$ is non-singular. In this case, the inverse of M

is given by

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1}. \end{bmatrix}.$$

Alternatively, if D is non-singular, then M is non-singular if and only if $A - BD^{-1}C$ is non-singular. In this case, the inverse of M is given by

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}. \end{bmatrix}$$

Proof. The first part of the lemma follows from the decomposition

$$M = \begin{bmatrix} I_n & 0_{n\times d} \\ CA^{-1} & I_d \end{bmatrix} \begin{bmatrix} A & 0_{n\times d} \\ 0_{d\times n} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_n & A^{-1}B \\ 0_{d\times n} & I_d \end{bmatrix}$$

and the fact that the triangular block matrices are non-singular if and only if their diagonal blocks are non-singular. The second part follows from a similar decomposition. □

2.3 Random Matrix Theory

In this section, we begin by introducing the notion of the resolvent and its significant connection to eigenvalues and eigenvectors. We will then delve into the concept of a deterministic equivalent, followed by a concise overview of fundamental methods in random matrix theory, particularly the matrix Dyson equation and the linearization tricks. Those methods will form the foundation of the core thesis discussions.

2.3.1 Resolvent

When we aim to understand the behavior of large random matrices, it is common practice to examine scalar observations derived from the matrix itself. A pertinent quantity of interest in this context is the empirical spectral distribution $\mu_H = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(H)}$ associated with a Hermitian matrix $H \in \mathbb{C}^{n\times n}$. This distribution represents a probability measure on the real line, where δ_λ signifies the Dirac measure at λ and $\lambda_j(H)$ corresponds to the jth ordered eigenvalue of H. Just as probability distributions can be investigated through characteristic functions, the empirical spectral distribution is analyzed using the Stieltjes transform. This transform, denoted $m_\mu(z)$, is defined for a real probability measure μ with support $\mathrm{supp}(\mu)$ as $m_\mu(z) = \int \frac{\mu(\mathrm{d}\lambda)}{\lambda - z}$ for each $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$. The Stieltjes transform boasts attractive properties, such as being holomorphic within its domain, bounded, and positivity preserving. As its name implies, the Stieltjes transform offers an inverse formula to reconstruct the underlying measure. In the field of random matrix theory, the focus has traditionally been on the distribution of eigenvalues. Foundational outcomes concerning the weak convergence of the empirical spectral distribution of Wigner or Wishart matrices to continuous measures often begin by demonstrating the convergence of the Stieltjes transform, followed by the application of the inverse formula to retrieve the measure [Wig55; Wis28].

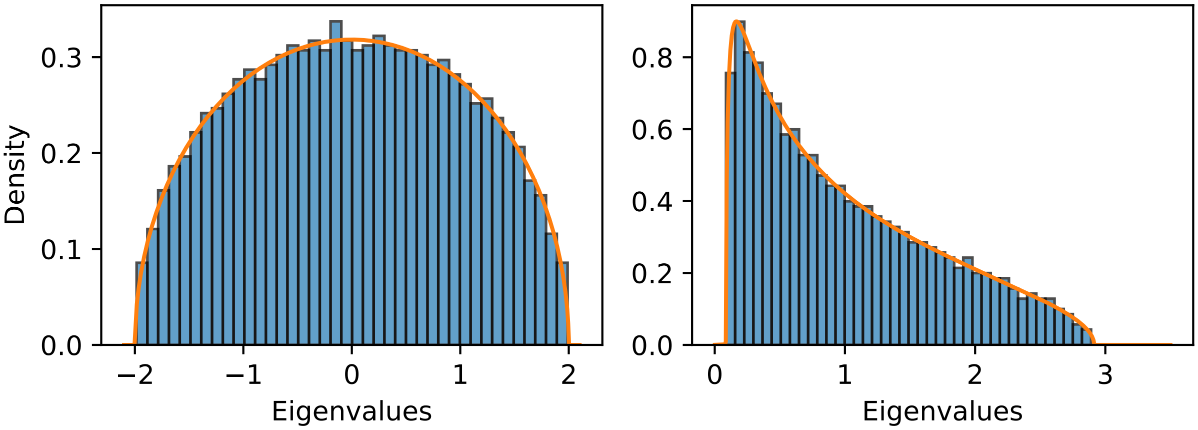

Figure 2.1: Eigenvalue densities for two different types of matrices. Left: eigenvalue density of a normalized GOE matrix, which converges to the Wigner semicircular distribution $\frac{\sqrt{\max(4-x^2,0)}}{2\pi}$; Right: eigenvalue density of a normalized Wishart matrix, which converges to the Marchenko-Pastur distribution, characterized by $\mu_{\mathrm{MP}}(\mathrm{d}x) = \max\{(1-r^{-1}), 0\}\delta_0(x) + \frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi r x}\mathrm{d}x$, where $\lambda_\pm = (1 \pm \sqrt{\rho})^2\sigma^2$ and $r = 1/2$ is the *shape parameter*.

The primary limitation of the Stieltjes transform approach is its inability to provide insight into the eigenvectors of a matrix. To address this, one can employ the *resolvent*, also known as *Green's function*, which is defined as $\mathcal{R}_H(z) = (H - zI_n)^{-1}$. The resolvent can be seen as a generalization of the Stieltjes transform for empirical spectral measures since

$$\frac{1}{n} \operatorname{tr} \mathcal{R}_H(z) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\lambda_j(H) - z} = \int \frac{\mu_H(\mathrm{d}\lambda)}{\lambda - z} = m_{\mu_H}(z).$$

The resolvent emerges as a central object in this thesis, as it captures nearly all spectral information of H while being more analytically tractable than analyzing the matrix directly. We compile some of the resolvent's fundamental properties in the following lemma.

Lemma 2.3.1. *Let $H \in \mathbb{C}^{n \times n}$ be a Hermitian matrix. Then, the resolvent $\mathcal{R}_H(z)$ is a holomorphic function on $\mathbb{C} \setminus \sigma(H)$ and $\|\mathcal{R}_H(z)\| \leq \operatorname{dist}(z, \sigma(H))^{-1} \leq (\Im[z])^{-1}$ for every $z \in \mathbb{C} \setminus \sigma(H)$. Furthermore, $\mathcal{R}_H(z)$ is a Herglotz function. Finally, if $H = U\Lambda U^T$ for some orthonormal matrix U and diagonal matrix Λ and Γ is a positively oriented simple closed curve with interior Γ°, then*

$$U \operatorname{diag}\{f(\lambda_j(H))\mathbb{I}_{\lambda_j(H) \in \Gamma^\circ}\} U^T = \frac{1}{2\pi i} \oint_{\Gamma} f(z) \mathcal{R}_H(z) \mathrm{d}z$$

for every complex function f analytic on a region containing Γ and its inside.

The proof of Lemma 2.3.1 mostly follows from the properties introduced above. The conclusion concerning the contour integral is a consequence of Cauchy's integral formula and is inspired by [CL22, Theorem 2.2 and the discussion that follows]. We refer the reader to this reference for a good introduction to the topics discussed here.

2.3.2 Deterministic Equivalent

When examining random matrices, we are generally not interested in individual matrix entries, as their contributions often prove to be negligible. Rather, we are focused on the collective impact of all entries. Consequently, we tend to analyze scalar observations. While exploring the distribution of eigenvalues is a common practice, it is also important to understand the eigenvector structure. For instance, in principal component analysis (PCA), we pay close attention to the largest eigenvectors of the sample covariance matrix. For this reason, we will be interested in deriving *deterministic equivalents*. Given a potential random matrix $A \in \mathbb{C}^{n \times n}$, we say that $B \in \mathbb{C}^{n \times n}$ is a deterministic equivalent of A if B is a deterministic matrix and $\operatorname{tr} U(A - B) \to 0$ almost surely as $n \to \infty$ for any sequence of unit matrices $U \in \mathbb{R}^{n \times n}$ with $\|U\|_* \leq 1$. For example, by finding a deterministic equivalent for the resolvent of a random matrix, one may use Lemma 2.3.1 to study the asymptotic properties of the spectrum of the random matrix and Lemma 2.2.9 to study the density of the limiting empirical spectral distribution.

2.3.3 Matrix Dyson Equation

Various approaches have been developed to derive deterministic equivalents for random matrices. Each method has its own advantages and disadvantages, and the choice of method often depends on the specific problem at hand. One popular method is the *leave-one-out* approach, which is a perturbation method that involves removing a relatively small set of entries from the matrix and studying the impact on the resolvent. This method is particularly useful when the matrix is a sum of independent random variables or when the matrix has independent rows or columns. The leave-one-out method is often simple to use and relies on the concentration of quadratic forms [BZ08; BS10]. This concentration can be established using a variety of techniques, for instance using concentration of measure arguments [Cho22; LLC18]. Indeed, when the entries of the random matrix of interest can be expressed as the Lipschitz image of Gaussian random variables, one may use a powerful Gaussian concentration inequality to establish the concentration of various functionals.

Proposition 2.3.1 (Gaussian concentration inequality for Lipschitz functions [Tao12, Theorem 2.1.12] and [Led01, Proposition 1.10])**.** *Suppose that $x \sim \mathcal{N}(0, I_\gamma)$ and let $f : \mathbb{R}^\gamma \mapsto \mathbb{R}$ be a λ-Lipschitz function. Then, for every $t \in \mathbb{R}$,*

$$\mathbb{P}\left(|f(x) - \mathbb{E}[f(x)]| \geq t\right) \leq c_1 e^{-\frac{c_2 t^2}{\lambda^2}}$$

for some absolute constants $c_1, c_2 \in \mathbb{R}_{>0}$.

Another approach is the moment method, which can be more combinatorial and differs from resolvent analysis [FM19; BP21].

The “Gaussian method”, utilizing Stein’s lemma and a Lindeberg’s interpolation trick, has also proven to be effective [PS11]. This method relies on Stein’s lemma to study the expected resolvent of random matrices with Gaussian entries.

Proposition 2.3.2 (Stein's lemma [Ste81, Lemma 1])**.** *Let $x \sim \mathcal{N}(0,1)$ be a real random variable and let $f : \mathbb{R} \mapsto \mathbb{R}$ be an indefinite integral of the Lebesgue measurable function f'. Suppose that $\mathbb{E}|f'(x)| < \infty$. Then, $\mathbb{E}[xf(x)] = \mathbb{E}[f'(x)]$.*

Then, one may use a Nash-Poincaré inequality, as shown in [Pas05, Proposition 2.4], to establish the concentration of the functional of the resolvents around their mean. It is worth noting that while the aforementioned approach is specific to Gaussian matrices, it can be extended to more general matrices using Lindeberg's interpolation trick [PS11].

The method that we focus on in this thesis is the *matrix Dyson equation (MDE)*. The matrix Dyson equation is a crucial method used in constructing a deterministic equivalent for the resolvent of a random matrix. Given a symmetric random matrix $H \in \mathbb{R}^{n\times n}$, the matrix Dyson equation takes the form

$$(\mathbb{E}H - \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)] - zI_n)M = I_n$$

for $z \in \mathbb{H}$. The equation contains the mean matrix $\mathbb{E}H$ and encodes the covariance structure in the so-called superoperator $M \mapsto \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)]$. In order to use the matrix Dyson equation framework, the first step is to demonstrate the existence of a unique solution within a suitable admissible set. As we aim for the solution to serve as a deterministic equivalent for the random resolvent $(H - zI_n)^{-1}$, it must share certain properties with the resolvent, such as analyticity and a positive definite imaginary part. To show the existence of this unique solution, we rely on the Earle-Hamilton fixed-point theorem [HFS07]. The unique solution becomes the candidate deterministic equivalent for the random resolvent. The next step is to confirm that the solution of the MDE closely matches the random resolvent. We accomplish this in two substeps. First, we demonstrate that the random resolvent nearly solves the MDE, up to an additive perturbation matrix. If we can show that this perturbation matrix vanishes in a suitable sense, then the random resolvent asymptotically solves the MDE. Lastly, we must ensure that the matrix Dyson equation is stable. This means that two functions nearly solving the MDE should be close to each other, which is a standard consideration in the fixed-point literature.

For an in-depth introduction to this subject, we recommend reading [Erd19]. The vector version of the matrix Dyson equation has proven invaluable in establishing local laws for Generalized Wigner matrices [AEK17; AEK19a; AKE17]. Extending its applicability, the matrix Dyson equation has been instrumental in investigating local laws for Hermitian matrices with correlations featuring fast decay, as well as those with slower correlation decay, particularly focusing on regular edges [Alt+20] and regions away from the support edges [EKS19]. A notable advantage of the matrix Dyson equation is that its solution is given by a Stieltjes representation. Leveraging this equation, detailed regularity properties of the self-consistent density of states have been explored [AEK18], and bounds on the spectrum of Kronecker random matrices have been established [Alt+19]. In this thesis, we will explore a broader version of the matrix Dyson equation. Thus, we reserve the detailed discussion of the matrix Dyson equation's various properties for our theoretical exploration.

For now, a constructive approach is to heuristically derive the matrix Dyson equation for a Gaussian Orthogonal Ensemble (GOE) matrix. A matrix $G \in \mathbb{R}^{n\times n}$ is considered a GOE matrix if its entries are independent and identically distributed (i.i.d.) up to symmetry, meaning that $G_{j,k} \sim \mathcal{N}(0,2)$ and $G_{j,k} \sim \mathcal{N}(0,1)$ for every $1 \leq j,k \leq n$ and $j \neq k$. Particularly, to generate a GOE matrix, we sample a matrix of i.i.d. Gaussian $Z \in \mathbb{R}^{n\times n}$ and set $G = \frac{Z+Z^T}{\sqrt{2}}$. Let $H = n^{-\frac{1}{2}}G$ be a normalized GOE matrix. We present the following lemma and provide a brief sketch of the proof because it is a good illustration of how we will eventually apply our extension of the matrix Dyson equation framework to study random matrices.

Before stating the result, we recall that the notation $o_n(f(n))$ refers to any function $g(n)$ such that $\lim_{n\to\infty} \frac{g(n)}{f(n)} = 0$.

Lemma 2.3.2. *Let $H \in \mathbb{R}^{n\times n}$ be a GOE matrix. Then, $-(\mathbb{E}[H\mathbb{E}[\mathcal{R}_H(z)]H]+zI_n)\mathbb{E}\mathcal{R}_H(z) = I_n + o_n(1)$.*

Sketch of proof. Rearranging the equation $\mathbb{E}[(H - zI_n)\mathcal{R}_H(z)] = I_n$, we get

$$-(\mathbb{E}[H\mathbb{E}[\mathcal{R}_H(z)]H] + zI_n)\mathbb{E}\mathcal{R}_H(z) = I_n - \mathbb{E}[H\mathcal{R}_H(z)] - \mathbb{E}[H\mathbb{E}[\mathcal{R}_H(z)]H]\mathbb{E}\mathcal{R}_H(z).$$

By Stein's lemma, we have

$$\mathbb{E}[G\mathcal{R}_H(z)]_{j,k} = \mathbb{E}[G_{j,a}\mathcal{R}_H(z)_{a,k}] = \sum_{a\neq j}\mathbb{E}\left[\frac{\partial \mathcal{R}_H(z)}{\partial G_{j,a}}\right]_{a,k} + \sqrt{2}\mathbb{E}\left[\frac{\partial \mathcal{R}_H(z)}{\partial Z_{j,j}}\right]_{j,k}.$$

Let $E_{j,k}$ is the matrix which has all zero entries except the (j,k) entry, which is 1. By Lemma 2.2.1,

$$\begin{aligned}\lim_{h\to 0}\frac{\mathcal{R}_{H+\sqrt{\frac{2}{n}}hE_{j,j}}(z) - \mathcal{R}_H(z)}{h} &= \lim_{h\to 0} -\sqrt{\frac{2}{n}}\mathcal{R}_{H+\sqrt{\frac{2}{n}}hE_{j,j}}(z)E_{j,j}\mathcal{R}_H(z)\\ &= -\sqrt{2}n^{-\frac{1}{2}}\mathcal{R}_H(z)E_{j,j}\mathcal{R}_H(z)\end{aligned}$$

for every j. Similarly, for $j \neq a$,

$$\begin{aligned}\lim_{h\to 0}\frac{\mathcal{R}_{H+\frac{h}{\sqrt{n}}(E_{j,a}+E_{a,j})}(z) - \mathcal{R}_H(z)}{h} &= \lim_{h\to 0} -n^{-\frac{1}{2}}\mathcal{R}_{H+\frac{h}{\sqrt{n}}(E_{j,a}+E_{a,j})}(z)(E_{j,a}+E_{a,j})\mathcal{R}_H(z)\\ &= -n^{-\frac{1}{2}}\mathcal{R}_H(z)(E_{j,a}+E_{a,j})\mathcal{R}_H(z)\end{aligned}$$

Therefore, $-\mathbb{E}[H\mathcal{R}_H(z)] = n^{-1}\mathbb{E}[(\mathcal{R}_H(z))^2] + n^{-1}\mathbb{E}[\operatorname{tr}(\mathcal{R}_H(z))(H - zI_n)^{-1}] + o_n(1)$. Since the resolvent $\mathcal{R}_H(z)$ is bounded in spectral norm, it follows from Jensen's inequality that the term $n^{-1}\mathbb{E}[(\mathcal{R}_H(z))^2]$ is of order n^{-1} in spectral norm. For any $H_1, H_2 \in \mathbb{R}^{n\times n}$, it follows from Lemma 2.2.1 and the Cauchy-Schwarz inequality for the Frobenius norm that $n^{-1}|\operatorname{tr}(\mathcal{R}_{H_1}(z) - \mathcal{R}_{H_2}(z))| \leq n^{-\frac{1}{2}}\|\mathcal{R}_{H_1}(z)(H_1 - H_2)\mathcal{R}_{H_2}(z)\|_F$. Using Lemma 2.2.7 and the fact that resolvent is bounded in spectral norm, we have $n^{-1/2}\|\mathcal{R}_{H_1}(z)(H_1-H_2)\mathcal{R}_{H_2}(z)\|_F \leq n^{-1/2}(\Im[z])^{-2}\|(H_1 - H_2)\|_F$. By Proposition 2.3.1, the probability that $n^{-1}\operatorname{tr}(\mathcal{R}_H(z))$ deviates from its mean by more than t is bounded above by $c_1\exp(-c_2\Im[z]^2n^2t^2)$ for some absolute constants $c_1, c_2 \in \mathbb{R}_{>0}$. This implies that $n^{-1}\operatorname{tr}(\mathcal{R}_H(z))$ concentrates around its mean. Hence, $n^{-1}\mathbb{E}[\operatorname{tr}(\mathcal{R}_H(z))\mathcal{R}_H(z)] \approx n^{-1}\operatorname{tr}(\mathbb{E}\mathcal{R}_H(z))\mathbb{E}\mathcal{R}_H(z)$. On the other hand, straightforward calculations gives $\mathbb{E}[H\mathbb{E}[\mathcal{R}_H(z)]H]\mathbb{E}\mathcal{R}_H(z) = n^{-1}(\mathbb{E}\mathcal{R}_H(z))^2 + n^{-1}\operatorname{tr}(\mathbb{E}\mathcal{R}_H(z))\mathbb{E}\mathcal{R}_H(z)$. This gives the claim. □

If we have stability of the matrix Dyson equation, then we can conclude that the solution of the matrix Dyson equation is an asymptotic deterministic equivalent for the expected random resolvent $\mathbb{E}(H - zI_n)^{-1}$. Using the Gaussian concentration inequality for Lipschitz functions, we can show that the expected random resolvent is a deterministic equivalent for the random resolvent itself. Hence, we can use the solution of the matrix Dyson equation as a deterministic equivalent for the random resolvent. This is a broad and incomplete overview of how the matrix Dyson equation is derived. We will provide a more detailed discussion regarding using the matrix Dyson equation to derive deterministic equivalents for random matrices later in this thesis.

2.3.4 Linearization Trick

The matrix Dyson equation is a powerful tool for deriving deterministic equivalents, but it is limited to certain classes of matrices. To illustrate this point, consider two examples where we compute the superoperator for Wigner and Wishart matrices.

Example 2.3.1 (Superoperator for Wigner matrix). Let $Z \in \mathbb{R}^{n\times n}$ be a matrix with i.i.d. standard Gaussian entries and define the Wigner matrix $W = (2n)^{-\frac{1}{2}}(Z + Z^T) \in \mathbb{R}^{n\times n}$. Let $M \in \mathbb{C}^{d\times d}$ be any deterministic matrix. Then, for $1 \leq j, k \leq d$, we have $\mathbb{E}[WMW]_{j,k} = \sum_{a,b} \mathbb{E}W_{j,a}M_{a,b}W_{b,k}$. Indeed, $\mathbb{E}W_{j,a}M_{a,b}W_{b,k} \neq 0$ if and only if either $(j, a) = (b, k)$ or $(j, a) = (k, b)$. Thus, $\mathbb{E}[WMW]_{j,k} = \mathbb{I}_{j=k}\frac{n-1}{n}M_{j,j} + \mathbb{I}_{j=k}2n^{-1}M_{j,j} + \mathbb{I}_{j\neq k}n^{-1}M_{k,j}$. Hence, if M is bounded in norm as $n \to \infty$, we have $\mathbb{E}[WMW] = n^{-1}\operatorname{tr}(M)I_n + o_n(1)$.

Example 2.3.2 (Superoperator for Wishart matrix). Let $Z \in \mathbb{R}^{d\times n}$ be a matrix with i.i.d. standard Gaussian entries and define the Wishart matrix $W = n^{-1}ZZ^T \in \mathbb{R}^{d\times d}$. Let $M \in \mathbb{C}^{d\times d}$ be any deterministic matrix. Indeed, $\mathbb{E}W = I_d$ and $\mathbb{E}[(W - \mathbb{E}W)M(W - \mathbb{E}W)] = \mathbb{E}[WMW] - M$. For $1 \leq j, k \leq d$, we have $\mathbb{E}[WMW]_{j,k} = n^{-2}\sum_{a,b,p,q} \mathbb{E}Z_{j,a}Z_{b,a}M_{b,p}Z_{p,q}Z_{k,q}$. Since the entries of Z are i.i.d. centered Gaussian, $\mathbb{E}Z_{j,a}Z_{b,a}M_{b,p}Z_{p,q}Z_{k,q}$ is not zero if and only we can group the indices into even groups. This gives $\mathbb{E}[WMW]_{j,k} = \mathbb{I}_{j=k}n^{-1}\mathbb{E}[Z_{1,1}^4]M_{j,j} + \mathbb{I}_{j=k}\frac{n(n-1)}{n^2}M_{j,j} + \mathbb{I}_{j\neq k}M_{j,k} + \mathbb{I}_{j\neq k}n^{-1}M_{k,j} + \mathbb{I}_{j=k}n^{-1}\sum_{a\neq j} M_{a,a}$. Hence, if M is bounded in norm as $n \to \infty$, we have $\mathbb{E}[(W - \mathbb{E}W)M(W - \mathbb{E}W)] = n^{-1}\operatorname{tr}(M)I_d + o_n(1)$.

Examples 2.3.1 and 2.3.2 illustrate that the superoperator for Wigner and Wishart matrices are asymptotically equal. Substituting the superoperator $\mathbb{E}[WMW] = n^{-1}\operatorname{tr}(M)I_n$ for the Wigner matrix into the matrix Dyson equation, we obtain $m = -(m + z)^{-1}$ for $m = n^{-1}\operatorname{tr}(M)$. Solving for m using the fact that as $\Im[z] \to \infty$ we have $m(z) \to 0$ to choose the right root, we get $m = \frac{\sqrt{z^2-4}-z}{2}$. This is precisely the Stieltjes transform of the semicircular distribution in Figure 2.1. On the other hand, substituting the same superoperator for the matrix Dyson equation associated with the Wishart matrix, we get $m = \frac{d}{n}(1-m-z)^{-1}$. In particular, m satisfies $\frac{d}{n}m^2 - (1-z)m + 1 = 0$. The equation for the Stieltjes transform of the Marchenko-Pastur law, as shown in Figure 2.1, is $\frac{d}{n}m^2 - (1 - \frac{d}{n} - z)m + 1 = 0$.

To summarize, while the matrix Dyson equation correctly recovers the limiting behavior of the Stieltjes transform for the Wigner matrix, it fails to capture the correct limiting behavior for the Wishart matrix. This is because the matrix Dyson equation is only applicable to a certain class of matrices, essentially those that are “Wigner-like”. This is a significant limitation of the matrix Dyson equation. In order to extend its applicability to a broader class of matrices, we employ a *linearization trick*. The idea behind linearizations, which are also referred to as linear pencils or realizations, is to represent rational functions of random matrices as blocks of inverse of larger random matrices which depend linearly on their blocks. These linearizations possess simpler correlation structures, rendering them more amenable to certain types of analysis.

The concept of linearization became particularly important following the influential work of Haagerup and Thorbjørnsen. This work essentially demonstrated that to analyze a polynomial expression in matrices, it is sufficient to consider a linear polynomial with matrix coefficients [HT05]. However, a limitation of [HT05] is that their linearization approach does not inherently preserve the self-adjointness of the original polynomial expression. Anderson addressed this problem, proving that a self-adjoint polynomial expression allows the linearization's coefficients to be chosen in a way that retains self-adjointness [And13]. This finding was later generalized to include rational expressions [HMS18]. The fact that linearizations are not unique has inspired other linearization methods. For example, [EKN20] developed the concept of a minimal linearization. A key strength of these linearization techniques is that their supporting arguments are often constructive, giving explicit instructions for creating appropriate linearizations.

Examples 2.3.3 to 2.3.5 illustrate how the linearization trick can encode the resolvent of common random matrices as the inverse of a larger random matrix. We use the notation $M^{-1}_{1,1} \equiv [M^{-1}]_{1,1}$ to refer to the $(1,1)$ sub-block of the inverse of a block matrix M.

Example 2.3.3 (Linearization for Wishart matrix)*.* Let $X \in \mathbb{C}^{n\times d}$ and $z \in \mathbb{H}$. Then,

$$\begin{bmatrix} -zI_n & X \\ X^* & -I_d \end{bmatrix}^{-1}_{1,1} = (XX^* - zI_n)^{-1}.$$

Example 2.3.4 (Linearization for sample covariance matrix)*.* Let $X \in \mathbb{C}^{n\times d}$, $Y \in \mathbb{C}^{d\times m}$ and $z \in \mathbb{H}$. Then,

$$\begin{bmatrix} -zI_n & 0 & 0 & X \\ 0 & 0 & Y & -I_d \\ 0 & Y^* & -I_m & 0 \\ X^* & -I_d & 0 & 0 \end{bmatrix}^{-1}_{1,1} = (XYY^*X^* - zI_n)^{-1}.$$

Example 2.3.5 (Linearization for the anticommutator)*.* Let $X, Y \in \mathbb{C}^{n\times d}$ and $z \in \mathbb{H}$. Then,

$$\begin{bmatrix} -zI_n & X & Y \\ X^* & 0 & -I_d \\ Y^* & -I_d & 0 \end{bmatrix}^{-1}_{1,1} = (XY^* + YX^* - zI_n)^{-1}.$$

The linearization trick is a powerful tool within free probability, enabling the study of random matrix polynomials on the global scale [And13; BMS17; HT05; HMS18; HMV06] and the local scale [EKN20; FKN23; And15]. Combined with operator-valued free probability, the linearization trick—referred to as the pencil method in this context—has found successful applications in the study of simple neural networks [MP22; ALP22; AP20a; AP20b; TAP21]. We will revisit the linearization technique and its integration with our theoretical framework when studying the asymptotic empirical test error of random features ridge regression.

The linearization trick, illustrated in Examples 2.3.3 to 2.3.5, naturally leads to the study of pseudo-resolvents, or generalized resolvents. Given a linearization $L \in \mathbb{C}^{\ell\times\ell}$ for $\ell = n + d$, we aim to understand the asymptotic properties of the pseudo-resolvent $(L - z\Lambda_\ell)^{-1}$, where $\Lambda = \mathrm{diag}\{I_n, 0_{d\times d}\}$. As a generalization of resolvents, pseudo-resolvents are inherently more challenging to analyze than resolvents due to the absence of a spectral parameter spanning the entire diagonal. For example, unlike Hermitian matrix resolvents which are bounded by the inverse of the imaginary part of the spectral parameter as stated in Lemma 2.3.1, pseudo-resolvents lack such an a priori bound. Despite these challenges, the matrix Dyson equation has been extended for pseudo-resolvent analysis. Anderson derived global laws for linearizations of polynomials in independent Wigner matrices, terming their matrix Dyson equation the *Schwinger-Dyson equation* [And13]. This work was expanded to study the anti-commutator (see Example 2.3.5) on a local scale [And15] and more generally to polynomials of matrices with independent centered entries and suitable normalization [EKN20; FKN23] under the name *Dyson equation for linearization (DEL).*

In this work, we develop a framework based on an extension of the matrix Dyson equation to study asymptotic properties of linearizations with a general correlation structure. This provides an alternative to the use of operator-valued free probability, and we believe that our framework could find multiple applications in machine learning. Our approach differs significantly from previous extension of the matrix Dyson equation for linearizations. While previous research has focused on linearizations with blocks of independent generalized Wigner matrices, we consider linearizations with arbitrary correlation structures. We are interested in studying pseudo-resolvents on a global scale, which, although less precise than the local scale, allows us to relax the assumptions of previous work. Additionally, global laws are sufficient to make assertions about machine learning models in many cases. Our approach also provides a novel perspective on studying the matrix Dyson equation. Notably, we analyze the Carathéodory-Riffen-Finsler pseudometric to demonstrate that the matrix Dyson equation for linearizations is asymptotically stable under general assumptions. We will rigorously state our settings in the next section and relate our approach to the literature while developing our theoretical framework.

3

Matrix Dyson Equation for Correlated Linearizations

In this chapter, we extend the matrix Dyson equation to correlated linearizations, demonstrating that key properties such as existence of a unique solution, asymptotic stability, and other desirable characteristics persist under these general conditions. Our approach faces two key challenges. First, the general correlation structures in our study preclude the direct use of free probability tools commonly leveraged in this field [And13; EKN20]. The second challenge emerges from the inherent instability of the matrix Dyson equation for linearization and the pseudo-resolvent when compared to their counterparts involving a spectral parameter spanning a full-rank identity matrix. To overcome these challenges, drawing inspiration from [EKN20], we introduce a regularized version of the problem in which properties such as existence and stability can be more easily established. Subsequently, a significant portion of our efforts is focused on demonstrating that we can gradually eliminate the regularization while simultaneously increasing the dimension of the problem, thereby preserving these key properties. This endeavor is challenging due to the fact that, initially, numerous bounds and attributes related to the regularized problem deteriorate as the regularization term approaches zero. An insightful perspective on this is to consider that, with the results obtained for the full resolvent case, the removal of the regularization essentially corresponds to driving the spectral parameter towards zero.

Unlike much of the existing MDE literature, we focus on deriving *global laws,* studying eigenvalue behavior on a macroscopic scale. While potentially less precise, this allows us to work under broader assumptions. We believe that generalizing the MDE to obtain global laws for pseudo-resolvents of correlated linearizations has significant potential, particularly in machine learning where global laws often provide sufficient insight.

We begin by introducing the settings for the chapter, then outline the main properties of the MDE for correlated linearizations. Using these, we establish the existence and uniqueness of the MDE solution within a carefully chosen set. As far as we are aware, our method for establishing existence represents one of the most comprehensive approach within the existing literature. Next, we demonstrate asymptotic stability of the solution under appropriate conditions, offering a novel perspective on this proof. Finally, we show that the unique solution to the MDE is an asymptotic deterministic equivalent for the pseudo-resolvent of random linearizations under suitable assumptions.

The results outlined in this chapter are a slight extension of the findings presented by the current author in [LP23], but we add a more detailed discussion of the results and their implications.

3.1 Settings

Using the linearization trick introduced in Section 2.3.4, we study a class of real self-adjoint linearizations of the form

$$L = \begin{bmatrix} A & B^T \\ B & Q \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}. \tag{3.1}$$

Here, $A \in \mathbb{R}^{n\times n}$ is self-adjoint, $Q \in \mathbb{R}^{d\times d}$ is self-adjoint with both Q and $\mathbb{E}Q$ non-singular, and $B \in \mathbb{R}^{d\times n}$ is arbitrary. Our primary focus is analyzing the high-dimensional behavior of the pseudo-resolvent $(L - z\Lambda)^{-1}$, where $\Lambda := \operatorname{diag}\{I_{n\times n}, 0_{d\times d}\}$ and $z \in \mathbb{H} := \{z \in \mathbb{C} : \Im[z] > 0\}$ represents the upper half of the complex plane. Our framework relies on the *linearized* *matrix Dyson equation (MDE)*

$$(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)M = I_\ell, \tag{3.2}$$

where the spectral parameter $z \in \mathbb{H}$ and the *superoperator*

$$\mathcal{S} : M \in \mathbb{C}^{\ell\times\ell} \mapsto \begin{bmatrix} \mathcal{S}_{1,1}(M) & \mathbb{E}[(A-\mathbb{E}A)M_{1,1}(B-\mathbb{E}B)^T] \\ \mathbb{E}[(B-\mathbb{E}B)M_{1,1}(A-\mathbb{E}A)] & \mathbb{E}[(B-\mathbb{E}B)M_{1,1}(B-\mathbb{E}B)^T] \end{bmatrix} \in \mathbb{C}^{\ell\times\ell} \tag{3.3}$$

is a symmetric positivity-preserving linear map encoding the correlation structure of the linearization. Here, by positivity-preserving, we mean that $\mathcal{S}$ maps positive semidefinite matrices to positive semidefinite matrices. On the other hand, by symmetric, we mean that $(\mathcal{S}(M))^* = \mathcal{S}(M^*)$. The sub-superoperator $\mathcal{S}_{1,1}$ also preserves positivity. Importantly, the expectation in the superoperator is taken only over the linearization, and not over the superoperator's argument. Throughout the rest of this thesis, we will refer to (3.2) as the matrix Dyson equation, or MDE for short. For conciseness, let $s \in \mathbb{R}_{>0}$ be such that $\|\mathcal{S}(M)\| \leq s\|M\|$ and $\|\mathcal{S}_{i,j}(M)\| \leq s\|M_{1,1}\|$ for $(i,j) \in \{(1,2),(2,1),(2,2)\}$ and all $M \in \mathbb{C}^{\ell\times\ell}$. This condition resembles the upper bound in the *flatness* assumption common in MDE literature [Erd19; Alt18; Alt+19]. Consequently, we will adopt the term flatness to describe this property.

Remark 3.1.1. Under the setting presented above, Q is non-singular and the Schur complement of the lower-right $d \times d$ block of $L - z\Lambda$ is given by $A - B^TQ^{-1}B - zI_n$. By Lemma 2.2.4, the Schur complement is non-singular. Hence, we may apply Lemma 2.2.11 to conclude that the matrix $L - z\Lambda$ is non-singular, and the pseudo-resolvent $(L - z\Lambda)^{-1}$ is well-defined for every $z \in \mathbb{H}$.

We consider $\mathcal{S}$ as a perturbation of the full superoperator $M \in \mathbb{C}^{\ell\times\ell} \mapsto \mathbb{E}[(L-\mathbb{E}L)M(L-\mathbb{E}L)] \in \mathbb{C}^{\ell\times\ell}$. While the full superoperator is appropriate for resolvent approximation, as evidenced by the derivation in Example 2.3.1, the perturbed superoperator proves more convenient for analyzing the MDE. This is because we can first establish good control over the upper-left block of the matrix Dyson equation and demonstrate the existence of a solution there. We then leverage the form of the superoperator to extend the solution to the full matrix. However, this approach requires a trade-off. Define

$$\tilde{\mathcal{S}} : M \in \mathbb{C}^{\ell\times\ell} \mapsto \mathbb{E}\left[(L - \mathbb{E}L)M(L - \mathbb{E}L)\right] - \mathcal{S}(M) \in \mathbb{C}^{\ell\times\ell}. \tag{3.4}$$

Then $\tilde{\mathcal{S}}$ must become asymptotically negligible for our approach to be valid.

In order to ensure the existence of a unique solution to the matrix Dyson equation, we need to restrict (3.2) to a suitable set. Consequently, based on properties of the pseudo-resolvent $(L - z\Lambda)^{-1}$, we introduce the *admissible set*

$$\mathscr{M} := \mathrm{Hol}(\mathbb{H}, \mathscr{A}), \quad \mathscr{A} := \{W \in \mathbb{C}^{\ell\times\ell} : \Im[W] \succeq 0 \text{ and } \Im[W_{1,1}] \succ 0\}. \tag{3.5}$$

Our primary strategy for analyzing (3.2) involves initially establishing analogous results for a regularized version of the equation. This regularization typically simplifies the problem, enabling us to leverage existing knowledge. Subsequently, we demonstrate the feasibility of setting the regularization parameter to zero, effectively reverting to the original equation. Importantly, we ensure that the statements derived for the regularized variant remain valid in this limit, thereby providing valuable insights into the properties of (3.2). For this reason, we introduce the *regularized matrix Dyson equation (RMDE)*

$$(\mathbb{E}L - \mathcal{S}(M^{(\tau)}) - z\Lambda - i\tau I_\ell)M^{(\tau)} = I_\ell \tag{3.6}$$

for every $\tau > 0$. The corresponding admissible set is given by

$$\mathscr{M}_+ := \mathrm{Hol}(\mathbb{H}, \mathscr{A}_+), \quad \mathscr{A}_+ := \{W \in \mathbb{C}^{\ell\times\ell} : \Im[W] \succ 0\} \cap \mathscr{A}. \tag{3.7}$$

It will be convenient to view (3.2) as a fixed point equation, so we introduce the *MDE map*

$$\mathcal{F} : f \in \mathscr{M} \mapsto (\mathbb{E}L - \mathcal{S}(f(\cdot)) - (\cdot)\Lambda)^{-1} \in \mathscr{M}, \tag{3.8}$$

assuming its well-definedness, which we establish in lemmas 3.2.1 to 3.2.3. With this definition, we can reexpress the MDE (3.2) as $M = \mathcal{F}(M)$. Whenever convenient, we will fix a spectral parameter $z \in \mathbb{H}$ and operate with $\mathcal{F}$ over $\mathscr{A}$. Similarly, the formulation of (3.6) becomes $M^{(\tau)} = \mathcal{F}^{(\tau)}(M^{(\tau)})$, where

$$\mathcal{F}^{(\tau)} : f \in \mathscr{M}_+ \mapsto \left(\mathbb{E}L - \mathcal{S}(f(\cdot)) - (\cdot)\Lambda - i\tau I_\ell\right)^{-1} \in \mathscr{M}_+ \tag{3.9}$$

is the *RMDE map.*

3.1.1 Notation

To maintain consistency between the regularized matrix Dyson equation and the matrix Dyson equation, we will denote solutions with a regularization parameter of zero (i.e., $\tau = 0$) as $M^{(0)} = M$, $\mathcal{F}^{(0)} = \mathcal{F}$, etc. Throughout this section, the spectral parameter $z \in \mathbb{H}$ is fixed, unless otherwise specified. For notational brevity, we will represent the solution of the MDE as $M \equiv M(z)$, omitting the explicit dependence on z. Matrices $M \in \mathbb{C}^{\ell \times \ell}$ and operators on this space will be represented in block form

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix}$$

where $M_{1,1} \in \mathbb{C}^{n \times n}$, $M_{1,2} \in \mathbb{C}^{n \times d}$, $M_{2,1} \in \mathbb{C}^{d \times n}$, and $M_{2,2} \in \mathbb{C}^{d \times d}$. Sub-block operations follow the conventions: $M^*_{j,k} = (M_{j,k})^*$ denotes the conjugate transpose of the (j,k) sub-block, and $M^{-1}_{j,k} = (M^{-1})_{j,k}$ denotes the (j,k) block of the inverse of M. We will use $\|\cdot\|$ to denote the Euclidean norm for vectors and the operator norm for matrices and complex-valued matrix functions. Additionally, $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|_*$ denotes the nuclear norm.

3.2 Main Properties

In this section, we establish the main properties of the MDE and RMDE. We begin by demonstrating general properties of the MDE map $\mathcal{F}$ and the RMDE map $\mathcal{F}^{(\tau)}$, such as their well-definedness. Subsequently, we consider the behavior of the fixed point equations for large spectral parameters, demonstrating that they are well-behaved in this limit. Then, we establish a Stieltjes transform representation, as well as a power series representation for the solution of the matrix Dyson equation. These properties will be crucial to establish the existence of a unique solution to (3.2) in Section 3.3.

3.2.1 General Properties

The primary challenge in analyzing (3.2) lies in the fact that the spectral parameter does not span the entire diagonal. Consequently, obtaining certain desirable properties that are typically straightforward to establish for the full resolvent case requires additional effort. As an initial example, it is not immediately evident whether the matrix $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$ is non-singular. The following lemma confirms that this is indeed the case.

Lemma 3.2.1. *Let $\tau \in \mathbb{R}_{\geq 0}$ and $M \in \mathscr{A}$. Then, $\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ is invertible.*

Proof. Let $M \in \mathscr{A}$ be arbitrary. By definition of $\mathscr{A}$, $\Im[M] \succeq 0$. Since $\mathcal{S}$ is symmetric, linear and positivity-preserving, $\Im[\mathcal{S}(M)] = \mathcal{S}(\Im[M]) \succeq 0$. If $\tau > 0$, it follows directly that $\Im[\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell] \preceq -\tau$, which implies that $\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ is non-singular by Lemma 2.2.5.

Assume that $\tau = 0$ and let $v^* = (v_1^*, v_2^*)$ with $v_1 \in \mathbb{C}^n$ and $v_2 \in \mathbb{C}^d$ be a unitary vector in the kernel of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$. We will show that $v = 0$ and conclude that the kernel of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$ is trivial. Decomposing $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$ into its real and imaginary parts, we have

$$0 = v^*(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)v = v^*\Re[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]v + iv^*\Im[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]v.$$

Since both $\Re[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]$ and $\Im[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]$ are Hermitian, the quadratic forms are real and $v^*\Re[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]v = v^*\Im[\mathbb{E}L - \mathcal{S}(M) - z\Lambda]v = 0$. Since $\Im[\mathcal{S}(M)] \succeq 0$, the imaginary part of the upper-left $n \times n$ block of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$ is negative definite and the imaginary part of the whole matrix is negative semidefinite. Consequently, it must be the case that $v_1 = 0$.

Returning to the equation $(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)v = 0$, we have in particular that $(\mathbb{E}Q - \mathcal{S}_{2,2}(M))v_2 = 0$. Left-multiplying by v_2^* and decomposing the matrix $\mathbb{E}Q - \mathcal{S}_{2,2}(M)$ into its real and imaginary parts,

$$0 = v_2^*\Re[\mathbb{E}Q - \mathcal{S}_{2,2}(M)]v_2 + iv_2^*\Im[\mathbb{E}Q - \mathcal{S}_{2,2}(M)]v_2.$$

Again, since the real and imaginary parts of a matrix are Hermitian, the quadratic forms are real and $v_2^*\Im[\mathbb{E}Q - \mathcal{S}_{2,2}(M)]v_2 = v_2^*\Re[\mathbb{E}Q - \mathcal{S}_{2,2}(M)]v_2 = 0$. In particular, since $M \in \mathscr{A}$,

$\Im[M_{1,1}] \succ 0$ and $0 = v_2^* \Im[\mathbb{E}Q - \mathcal{S}_{2,2}(M)]v_2 \leq -v_2^* \mathcal{S}_{2,2}(\Im[M])v_2 \leq 0$. By definition of $\mathcal{S}_{2,2}$, $v_2^* \mathcal{S}_{2,2}(\Im[M])v_2 = \mathbb{E}v_2^*(B - \mathbb{E}B)\Im[M_{1,1}](B - \mathbb{E}B)^T v_2$. Consequently, as $\Im[M_{1,1}]$ is positive definite, it must be the case that $(B - \mathbb{E}B)^T v_2 = 0$ almost surely. Going back to the equation $(\mathbb{E}Q - \mathcal{S}_{2,2}(M))v_2 = 0$, we obtain $\mathbb{E}Qv_2 = 0$. However, $\mathbb{E}Q$ is non-singular, which implies that $v_2 = 0$. This is a contradiction, as v is a unitary vector. Consequently, the kernel of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda$ is trivial, and the matrix is non-singular. □

For every $M \in \mathscr{A}$, $\tau \in \mathbb{R}_{\geq 0}$ and $z \in \mathbb{H}$, the upper-left $n \times n$ block of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ has negative definite imaginary part. Hence, by Lemma 2.2.5, the upper-left $n \times n$ block is non-singular. By Lemma 2.2.11, this implies that the full matrix $\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ is non-singular if and only if the associated Schur complement is non-singular. Since we established non-singularity of the full matrix, this implies that the lower-right block of $(\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell)^{-1}$ is non-singular. Similarly, the previous lemma establishes that the lower-right $d \times d$ block of $\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ is non-singular. Additionally, the associated Schur complement has a negative definite imaginary part, which, by Lemma 2.2.5, implies that the Schur complement is non-singular. Consequently, we obtain the following corollary.

Corollary 3.2.1. *Let $\tau \in \mathbb{R}_{\geq 0}$ and $M \in \mathscr{A}$. Then, the diagonal blocks of $(\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell)^{-1}$ are invertible.*

Lemma 3.2.1 is a first step towards considering the MDE (3.2) as a fixed point equation $\mathcal{F}(M) = M$, along with its regularized counterpart. This perspective allows us to explore the existence and uniqueness of solutions by leveraging the extensive theory on fixed points. A second step in this direction is showing that $\mathcal{F}$ and $\mathcal{F}^{(\tau)}$ both map their respective domains to themselves. We adapt the argument from [HFS07].

Lemma 3.2.2. *Let $\tau \in \mathbb{R}_{\geq 0}$, $z \in \mathbb{H}$ and $M \in \mathscr{A}$. Then,*

$$\Im[\mathcal{F}^{(\tau)}(M)] \succeq \tau \mathcal{F}^{(\tau)}(M)(\mathcal{F}^{(\tau)}(M))^*, \quad \Im[\mathcal{F}_{1,1}^{(\tau)}(M)] \succeq \Im[z]\mathcal{F}_{1,1}^{(\tau)}(M)(\mathcal{F}_{1,1}^{(\tau)}(M))^*$$

and $\|\mathcal{F}_{1,1}^{(\tau)}(M)\| \leq (\Im[z])^{-1}$. Furthermore, if $\tau > 0$, then $\|\mathcal{F}^{(\tau)}(M)\| \leq \tau^{-1}$.

Proof. By Lemma 2.2.6, $\Im[\mathcal{F}^{(\tau)}(M)] = -\mathcal{F}^{(\tau)}(M)\Im[\mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell](\mathcal{F}^{(\tau)}(M))^* \succeq \mathcal{F}^{(\tau)}(M)(\Im[z]\Lambda + \tau I_\ell)(\mathcal{F}^{(\tau)}(M))^* \succeq \tau \mathcal{F}^{(\tau)}(M)(\mathcal{F}^{(\tau)}(M))^* \succeq 0$. Leveraging the observation that the spectral norm of positive semidefinite matrices adheres to the Loewner partial ordering of positive semidefinite matrices, we obtain $\|\Im[\mathcal{F}^{(\tau)}(M)]\| \geq \tau\|\mathcal{F}^{(\tau)}(M)(\mathcal{F}^{(\tau)}(M))^*\| = \tau\|\mathcal{F}^{(\tau)}(M)\|^2$. By Lemma 2.2.4, $\|\mathcal{F}^{(\tau)}(M)\| \leq \tau^{-1}$ whenever $\tau > 0$. For the other inequalities, extract the upper-left $n\times n$ block of the equation $\Im[\mathcal{F}^{(\tau)}(M)] \succeq \Im[z]\mathcal{F}^{(\tau)}(M)\Lambda(\mathcal{F}^{(\tau)}(M))^*$ to obtain $\Im[\mathcal{F}_{1,1}^{(\tau)}(M)] \succeq \Im[z]\mathcal{F}_{1,1}^{(\tau)}(M)(\mathcal{F}_{1,1}^{(\tau)}(M))^*$ and $\|\mathcal{F}_{1,1}^{(\tau)}(M)\| \leq (\Im[z])^{-1}$. □

It is important to note that Lemma 3.2.2 reveals a weaker control of the MDE in comparison to the RMDE. Specifically, we only have an a priori norm bound for the upper-left $n \times n$ block of $\mathcal{F}$.

Combining lemmas 3.2.1 and 3.2.2, it only remains to show that the MDE map $\mathcal{F}$ and the RMDE map $\mathcal{F}^{(\tau)}$ preserve holomorphicity to establish that they map their respective domains to themselves.

Lemma 3.2.3. *Let $\tau \in \mathbb{R}_{\geq 0}$. Then, $\mathcal{F}$ and $\mathcal{F}^{(\tau)}$ are well-defined. In particular, they map their respective domains into themselves.*

Proof. As mentioned above, it suffices to prove that $\mathcal{F}$ and $\mathcal{F}^{(\tau)}$ preserve holomorphicity. We will only prove this for $\mathcal{F}$, as the proof for $\mathcal{F}^{(\tau)}$ is analogous. Let $M \in \mathscr{M}$ be arbitrary and $z, h \in \mathbb{H}$. Since M is holomorphic on $\mathbb{H}$, let $\mathrm{D}M(z) : h \in \mathbb{H} \mapsto \mathrm{D}M(z)h$ be the Fréchet derivative of M at z. Furthermore, let $\mathrm{D}[\mathcal{F}(M(z))(z)] : h \in \mathbb{H} \mapsto \mathcal{F}(M(z))(z)\mathcal{S}(\mathrm{D}M(z)h)\mathcal{F}(M(z))(z) + h[\mathcal{F}(M(z))(z)]^2$. Clearly, $\mathrm{D}[\mathcal{F}(M(z))(z)]$ is a bounded linear operator. Furthermore, $\mathcal{F}(M(z+h))(z+h) - \mathcal{F}(M(z))(z) - \mathrm{D}[\mathcal{F}(M(z))(z) = X_1 + X_2$ with

$$X_1 = -\mathcal{F}(M(z+h))(z+h)\mathcal{S}(M(z+h) - M(z) - \mathcal{S}(\mathrm{D}M(z)h))\mathcal{F}(M(z))(z+h)$$

and $X_2 = h([\mathcal{F}(M(z))(z)]^2 - \mathcal{F}(M(z))(z+h)\mathcal{F}(M(z))(z))$. On one hand, since M is holomorphic,

$$\lim_{h\to 0} \frac{\|X_1\|}{|h|} \leq s\|\mathcal{F}(M(z))(z)\|^2 \lim_{h\to 0} \frac{\|M(z+h) - M(z) - \mathcal{S}(\mathrm{D}M(z)h)\|}{|h|} = 0.$$

On the other hand, by continuity of the map $z \mapsto \mathcal{F}(M(z))(z)$ on $\mathbb{H}$, we have $\lim_{h\to 0}\|X_2\|/h = 0$. Consequently, $\mathcal{F}(M(z+h))(z+h) - \mathcal{F}(M(z))(z) - \mathrm{D}[\mathcal{F}(M(z))(z)]h = o(h)$, which implies that $z \mapsto \mathcal{F}(M(z))(z)$ is holomorphic on $\mathbb{H}$. This concludes the proof. □

The insights garnered from Lemma 3.2.3, which relies on lemmas 3.2.1 and 3.2.2, leads us to conceptualize the MDE and RMDE as fixed point equations. This characterization is instrumental, as we rely on it to assert the existence of a unique solution for (3.6) using the contractive property of the RMDE map $\mathcal{F}^{(\tau)}$ with respect to the CRF-pseudometric. Furthermore, this perspective will be crucial to establish the stability of the MDE in Section 3.4.2.

In what follows, we will take advantage of the block structure of the MDE and RMDE. Using Lemma 2.2.11, we decompose (3.2) as

$$M_{1,1} = \left(\mathcal{T}_{1,1}(M) - \mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M)\right)^{-1}, \tag{3.10a}$$

$$M_{2,2} = \left(\mathcal{T}_{2,2}(M) - \mathcal{T}_{2,1}(M)(\mathcal{T}_{1,1}(M))^{-1}\mathcal{T}_{1,2}(M)\right)^{-1}, \tag{3.10b}$$

$$M_{1,2} = -\,\mathcal{F}_{1,1}(M)\mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1} \text{ and} \tag{3.10c}$$

$$M_{2,1} = -\,(\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M)\mathcal{F}_{1,1}(M) \tag{3.10d}$$

where we write $\mathcal{T}(M) = \mathbb{E}L - \mathcal{S}(M) - z\Lambda$ for notational convenience. It may sometimes be practical to work with the equivalent form

$$M_{2,2} = (\mathcal{T}_{2,2}(M))^{-1} + (\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M)\mathcal{F}_{1,1}(M)\mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1}. \tag{3.10e}$$

We may decompose (3.6) similarly. In this case, we will write $\mathcal{T}^{(\tau)}(M) = \mathbb{E}L - \mathcal{S}(M^{(\tau)}) - z\Lambda - i\tau I_d$.

3.2.2 Large Spectral Parameter

For any solution M to (3.2), Lemma 3.2.2 implies that $\|M_{1,1}(z)\| \leq (\Im[z])^{-1}$. This bound is particularly useful when $\Im[z]$ is large, as it allows us to ensure that the norm of the upper-left $n \times n$ block of M is arbitrarily small. In fact, as this suggests, it will be beneficial to analyze the limit of the MDE map as $\Im[z]$ grows large. For every $\tau \in \mathbb{R}_{\geq 0}$, we define

$$M_{\star}^{(\tau)} = (\mathbb{E}Q - i\tau I_d)^{-1} \quad \text{and} \quad M_{\infty}^{(\tau)} = \begin{bmatrix} 0_{n\times n} & 0_{n\times d} \\ 0_{d\times n} & M_{\star}^{(\tau)} \end{bmatrix} \tag{3.11}$$

and denote $M_\star = M_\star^{(0)}$, $M_\infty = M_\infty^{(0)}$. Indeed, by Lemma 2.2.6, $\Im[M_*^{(\tau)}] \succeq 0$. We demonstrate in Lemma 3.2.4 that $M_\star^{(\tau)}$ corresponds precisely to the limit of $M^{(\tau)}(z)$ as $\Im[z]$ diverges to infinity.

Lemma 3.2.4. *Fix $\tau \in \mathbb{R}_{\geq 0}$ and assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the RMDE* (3.6)*. Then, $\|M(z) - M_\infty\| \to 0$ as $\Im[z] \to \infty$.*

Proof. We proceed block-wise. By Lemma 3.2.2, $\|M_{1,1}(z)\| \leq (\Im[z])^{-1}$. Consequently, $\|M_{1,1}(z)\| \to 0$ as $\Im[z] \to \infty$. Furthermore, it follows from Lemma 2.2.6 and the assumed properties of the superoperator that $\Im[(\mathcal{T}_{1,1}^{(\tau)}(M))^{-1}] \succeq \Im[z](\mathcal{T}_{1,1}^{(\tau)}(M))^{-1}(\mathcal{T}_{1,1}^{(\tau)}(M))^{-*}$. Here, we use the notation $\mathcal{T}^{(\tau)}(M) = \mathbb{E}L - \mathcal{S}(M) - z\Lambda - i\tau I_\ell$ introduced in (3.10). Taking the norm of both sides and rearranging, we obtain $\|(\mathcal{T}_{1,1}^{(\tau)}(M))^{-1}\| \leq (\Im[z])^{-1}$. Hence, $\|(\mathcal{T}_{1,1}^{(\tau)}(M))^{-1}\| \to 0$ as $\Im[z] \to \infty$. Furthermore, using the flatness of the superoperator, we have $\|\mathcal{S}_{1,2}(M)\| \vee \|\mathcal{S}_{2,1}(M)\| \vee \|\mathcal{S}_{2,2}(M)\| \to 0$ as $\Im[z] \to \infty$. This implies that $\mathcal{T}_{1,2}(M) \to \mathbb{E}B^T$, $\mathcal{T}_{2,1}(M) \to \mathbb{E}B$ and $\mathcal{T}_{2,2}(M) \to \mathbb{E}Q - i\tau I_d$ as $\Im[z] \to \infty$. Since $\mathbb{E}Q - i\tau I_d$ is non-singular and the taking a matrix inverse is a continuous operation, $(\mathcal{T}_{2,2}(M))^{-1} \to M_*^{(\tau)}$ as $\Im[z] \to \infty$. Finally, using (3.10), we conclude that $M(z) \to M_\infty$ as $\Im[z] \to \infty$. □

Remark 3.2.1. Similarly to Lemma 3.2.4, it follows from lemmas 2.2.5 and 2.2.11 that $\|(L - z\Lambda - i\tau I_\ell)^{-1} - \operatorname{diag}\{0_{n\times n}, (Q - i\tau I_d)^{-1}\}\|$ as $\Im[z] \to \infty$.

Although the pseudo-resolvent and the solution to the (R)MDE display favorable properties when the spectral parameter moves far from the real axis, it is crucial to understand their behavior near the real axis, as this region contains the spectral information. Hence, we need to bring the spectral parameter closer to the real axis. The next lemma constructs a loose bound on the norm of any solution to (3.6). This bound holds for every spectral parameter large enough in norm, regardless of the magnitude of its imaginary part. As a result, we can explore the behavior of the solution of the (R)MDE for large spectral values that are close to the real line.

Lemma 3.2.5. *Fix $\tau \in \mathbb{R}_{\geq 0}$ and assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the RMDE* (3.6)*. Then, there exists some constant $c \in \mathbb{R}_{>0}$ such that $\|M(z) - M_\infty^{(\tau)}\| \leq c(|z| - \kappa)^{-1}$ for all $z \in \{z \in \mathbb{H} : |z| > \kappa + c\kappa^{-1}\}$ with $\kappa := 2\|(\mathbb{E}Q - i\tau I_d)^{-1}\|(\|\mathbb{E}B\| + (2\|(\mathbb{E}Q - i\tau I_d)^{-1}\|)^{-1})^2 + \|\mathbb{E}A\| + (2\|(\mathbb{E}Q - i\tau I_d)^{-1}\|)^{-1} + s\|(\mathbb{E}Q - i\tau I_d)^{-1}\|$.*

Proof. Fix $z \in \mathbb{H}$ with $|z| > \kappa$ and let $M \equiv M(z)$. For notational convenience, we denote $a = \|\mathbb{E}A\|$, $b = \|\mathbb{E}B\|$ and $m_\star = \|M_\star^{(\tau)}\|$. We will show that there exists $c \in \mathbb{R}_{>0}$ such that $\|M(z) - M_\infty^{(\tau)}\| \notin (c(|z| - \kappa)^{-1}, \kappa]$ for every $z \in \mathbb{H}$ with $|z| > \kappa + c\kappa^{-1}$. By Lemma 3.2.4, $\|M(z) - M_\infty^{(\tau)}\|$ is in a neighborhood of the origin for every $z \in \mathbb{H}$ with $\Im[z]$ large enough. Since $z \mapsto \|M(z) - M_\infty^{(\tau)}\|$ is a continuous function on $\{z \in \mathbb{H} : |z| > \kappa + c\kappa^{-1}\}$, this will imply that $\|M(z) - M_\infty^{(\tau)}\| \in [0, c(|z| - \kappa)^{-1}]$ for every $z \in \{z \in \mathbb{H} : |z| > \kappa + c\kappa^{-1}\}$.

Suppose that $\|M - M_\infty^{(\tau)}\| \leq (2sm_\star)^{-1}$ such that $\|M\| \leq \|M - M_\infty^{(\tau)}\| + \|M_\infty^{(\tau)}\| \leq (2sm_\star)^{-1} + m_\star$. We consider the blocks separately using (3.10). By definition of $\mathcal{S}_{2,2}$, $\|\mathcal{S}_{2,2}(M)\| \leq s\|M_{1,1}\| \leq (2m_\star)^{-1}$. It follows from Lemma 2.2.2 that

$$\|(\mathcal{T}_{2,2}(M))^{-1}\| = \|(\mathbb{E}Q - i\tau I_d)^{-1} \left(\mathcal{S}_{2,2}(M)(\mathbb{E}Q - i\tau I_d)^{-1} - I_d\right)^{-1}\| \leq 2m_\star.$$

By subadditivity of the spectral norm,

$$\|\mathbb{E}A - \mathcal{S}_{1,1}(M) - \mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M)\| \leq a + (2m_\star)^{-1} + 2m_\star(b + (2m_\star)^{-1})^2.$$

For $z \in \mathbb{H}$ with $|z| > \kappa \geq a + (2m_\star)^{-1} + 2m_\star(b + (2m_\star)^{-1})^2$, it follows from Lemma 2.2.2 and (3.10a) that $\|M_{1,1}\| \leq (|z| - \kappa)^{-1}$. We now turn our attention to $M_{2,2}$. Using Lemma 2.2.1 and (3.10b),

$$M_{2,2} - M_\star^{(\tau)} = M_{2,2} \left(\mathcal{S}_{2,2}(M) + \mathcal{T}_{2,1}(M)(\mathcal{T}_{1,1}(M))^{-1}\mathcal{T}_{1,2}(M)\right)^{-1} M_*^{(\tau)}.$$

By Lemma 2.2.2, $\|(\mathcal{T}_{1,1}(M))^{-1}\| \leq (|z| - a - (2m_\star)^{-1} - sm_\star)^{-1}$. Hence,

$$\|M_{2,2} - M_\star^{(\tau)}\| \leq sm_\star((2sm_\star)^{-1} + m_\star)\|M_{1,1}\| + m_\star((2sm_\star)^{-1} + m_\star)(b + (2m_\star)^{-1})^2 \left(|z| - \kappa\right)^{-1}.$$

Plugging the bound for $\|M_{1,1}\|$ derived above and simplifying,

$$\|M_{2,2} - M_\star^{(\tau)}\| \leq m_\star(s + (b + (2m_\star)^{-1})^2)((2sm_\star)^{-1} + m_\star)(|z| - \kappa)^{-1}.$$

It only remains to treat $\|M_{1,2}\|$ and $\|M_{2,1}\|$, which we directly bound by

$$\max\{\|M_{1,2}\|, \|M_{2,1}\|\} \leq 2m_\star(b + (2m_\star)^{-1})\|M_{1,1}\| \leq 2m_\star(b + (2m_\star)^{-1})(|z| - \kappa)^{-1}$$

using (3.10c) and (3.10d). To summarize, we showed that for every $z \in \mathbb{H}$ with $|z| > a + (2m_\star)^{-1} + 2m_\star(b + (2m_\star)^{-1})^2 + sm_\star$, $\|M(z) - M_\infty^{(\tau)}\| \leq (2sm_\star)^{-1}$ implies that

$$\|M(z) - M_\infty^{(\tau)}\| \leq \|M_{1,1}(z)\| + \|M_{1,2}\| + \|M_{2,1}\| + \|M_{2,2} - M_\star^{(\tau)}\| \leq c(|z| - \kappa)^{-1}$$

with $c := 1 + m_\star(s + (b + (2m_\star)^{-1})^2)((2sm_\star)^{-1} + m_\star) + 4m_\star(b + (2m_\star)^{-1})$. Choosing $|z| > \kappa + c\kappa^{-1}$ completes the proof. □

The previous lemma is a key step in controlling the norm of the solution to the RMDE for large spectral parameters. We will now proceed with our analysis of the RMDE by establishing an upper bound on the imaginary part of any solution when the spectral parameter large.

Lemma 3.2.6. *Fix $\tau \in \mathbb{R}_{\geq 0}$ and assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the RMDE* (3.6)*. Let κ and c be defined as in Lemma 3.2.5. Then, there exists $\kappa_+ \geq \kappa + c\kappa^{-1}$ and constant $c_+ \in \mathbb{R}_{>0}$ such that $\|\Im[M_{1,1}(z)]\| \leq c_+(|z| - \kappa)^{-2}(\tau + \Im[z])$ for every $z \in \{z \in \mathbb{H} : |z| \geq \kappa_+\}$. In particular, if $\tau = 0$, then $\|\Im[M(z)]\|$ converges uniformly to* 0 *as $\Im[z] \downarrow 0$ on $\{z \in \mathbb{H} : \Re[z] \geq \kappa_+\}$.*

Proof. Let $m \equiv m(z) = c(|z| - \kappa)^{-1}$ be the bound in Lemma 3.2.5 and choose $\kappa_+ \geq \kappa + c\kappa^{-1}$ such that $3sm^2 \leq 2^{-1}$ and $4sm^2(1 + 2(m + m_\star)^2) \leq 2^{-1}$ for all $z \in \mathbb{H}$ with $|z| \geq \kappa_+$. Fix $z \in \mathbb{H}$ with $|z| \geq \kappa_+$ and denote $M = M(z)$.

By Lemma 2.2.6, we may write $\Im[M] = M(\Im[z]\Lambda + i\tau I_\ell + \mathcal{S}(\Im[M]))M^*$. By Lemma 3.2.5, $\|M\| \leq \|M - M_\infty\| + \|M_\infty\| \leq m + m_\star$ where we denote $m_\star = \|M_*^{(\tau)}\| = \|(\mathbb{E}Q - i\tau I_d)^{-1}\|$. Furthermore, by flatness, $\|\mathcal{S}_{1,2}(\Im[M])\| \vee \|\mathcal{S}_{2,1}(\Im[M])\| \vee \|\mathcal{S}_{2,2}(\Im[M])\| \leq s\|\Im[M_{1,1}]\|$ and $\|\mathcal{S}_{1,1}(\Im[M])\| \leq s\|\Im[M]\|$. Let $N \in \mathbb{R}^{2\times 2}$ such that $N_{j,k} = \|\Im[M_{j,k}]\|$. Then,

$$N \leq \begin{bmatrix} m & m \\ m & m + m_\star \end{bmatrix} \begin{bmatrix} \Im[z] + \tau + s\|\Im[M]\| & s\|\Im[M_{1,1}]\| \\ s\|\Im[M_{1,1}]\| & \tau + s\|\Im[M_{1,1}]\| \end{bmatrix} \begin{bmatrix} m & m \\ m & m + m_\star \end{bmatrix},$$

where the inequality is entry-wise. Expanding the product, we get that

$$\|\Im[M_{j,k}]\| \leq m^2\Im[z] + 2(m + m_\star)^2\tau + 3s(m + m_\star)^2\|\Im[M_{1,1}]\| + sm^2\|\Im[M]\|$$

for every $(j,k) \in \{(1,2),(2,1),(2,2)\}$. In particular, since $\|\Im[M]\| \le \sum_{j,k=1}^{2} \|\Im[M_{j,k}]\|$,

$$x \le m^2\Im[z] + 2(m+m_\star)^2\tau + 4s(m+m_\star)^2\|\Im[M_{1,1}]\| + 3sm^2x$$

where $x = \|\Im[M_{1,2}]\| \vee \|\Im[M_{2,1}]\| \vee \|\Im[2,2]\|$. Given our choice of κ_+, we can rearrange to obtain $x \le 2m^2\Im[z] + 4(m+m_\star)^2\tau + 8s(m+m_\star)^2\|\Im[M_{1,1}]\|$. Using the bound for N above,

$$\begin{aligned}\|\Im[M_{1,1}]\| &\le m^2\Im[z] + 2m^2\tau + 4sm^2\|\Im[M_{1,1}]\| + m^2x \\ &\le m^2(1+2m^2)\Im[z] + 2m^2(1+2(m+m_\star)^2)\tau + 4sm^2(1+2(m+m_\star)^2)\|\Im[M_{1,1}]\|.\end{aligned}$$

Rearranging, we obtain that $\|\Im[M_{1,1}]\| \le 2m^2(1+2m^2)\Im[z] + 4m^2(1+2(m+m_\star)^2)\tau$. This proves the first part of the lemma. The second part follows from the first part and the derived bound on x. □

3.2.3 Stieltjes Transform Representation

Given our chosen admissible set, it is evident that any solution to (3.2) or (3.6) is a matrix-valued Herglotz function. Specifically, every solution possesses a Stieltjes transform representation, as guaranteed by Theorem 2.2.1. This representation will prove particularly advantageous, as the positive semidefinite measure in the Stieltjes transform representation of the solution to the MDE is compactly supported.

Lemma 3.2.7 (Stieltjes transform representation). *Assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the MDE* (3.2)*. Then,*

$$M(z) = M_\infty + \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{\lambda - z}$$

for all $z \in \mathbb{H}$, where Ω is a matrix-valued measure on bounded Borel subsets of $\mathbb{R}$. Additionally, Ω is compactly supported and satisfies

$$\int_{\mathbb{R}} \Omega(\mathrm{d}\lambda) = \begin{bmatrix} I_n & -\mathbb{E}B^T(\mathbb{E}Q)^{-1} \\ -(\mathbb{E}Q)^{-1}\mathbb{E}B & (\mathbb{E}Q)^{-1}\mathbb{E}[BB^T](\mathbb{E}Q)^{-1}. \end{bmatrix}.$$

Proof. By Lemma 3.2.4, $\|M - M_\infty\| \to 0$ as $\Im[z] \to \infty$. Using (3.10a), we have

$$zM_{1,1} = -I_n + (\mathbb{E}A - \mathcal{S}_{1,1}(M) - \mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M))M_{1,1}.$$

Hence, $zM_{1,1}(z) \to -I_n$ as $\Im[z] \to \infty$. In addition, by (3.10c), (3.10d) and Lemma 3.2.4, $zM_{1,2} \to \mathbb{E}B^T(\mathbb{E}Q)^{-1}$ and $zM_{2,1} \to (\mathbb{E}Q)^{-1}\mathbb{E}B$ as $\Im[z] \to \infty$. Finally, by (3.10e) and Lemma 2.2.1,

$$z(M_{2,2} - M_\star) = (\mathcal{T}_{2,2}(M))^{-1}\mathcal{S}_{2,2}(zM)M_\star + (\mathcal{T}_{2,2}(M))^{-1}\mathcal{T}_{2,1}(M)zM_{1,1}\mathcal{T}_{1,2}(M)(\mathcal{T}_{2,2}(M))^{-1}.$$

By definition of the superoperator,

$$\lim_{\Im[z]\to\infty} \mathcal{S}_{2,2}(zM) = \mathbb{E}[(B - \mathbb{E}B) \lim_{\Im[z]\to\infty} zM_{1,1}(B - \mathbb{E}B)^T] = -\mathbb{E}[(B - \mathbb{E}B)(B - \mathbb{E}B)^T].$$

Since $\mathcal{S}_{2,2}(M)$ approach 0 as $\Im[z]$ approaches infinity, $(\mathcal{T}_{2,2}(M))^{-1} \to (\mathbb{E}Q)^{-1}$ as $\Im[z] \to \infty$. Also, $\mathcal{T}_{1,2}(M) \to \mathbb{E}B^T$ and $\mathcal{T}_{2,1}(M) \to \mathbb{E}B$ as $\Im[z] \to \infty$. We get $z(M_{2,2} - M_\star) \to (\mathbb{E}Q)^{-1}\mathbb{E}[BB^T](\mathbb{E}Q)^{-1}$ as $\Im[z] \to \infty$. Since M_∞ is real, it is clear that $M - M_\infty$ is a matrix-valued Herglotz function. Hence, by Theorem 2.2.1,

$$M(z) = M_\infty + \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{\lambda - z}$$

for all $z \in \mathbb{H}$, where Ω is a matrix-valued measure on bounded Borel subsets of $\mathbb{R}$ satisfying

$$\int_{\mathbb{R}} \Omega(\mathrm{d}\lambda) = -\lim_{\eta\to\infty} i\eta M(i\eta) = \begin{bmatrix} I_n & -\mathbb{E}B^T(\mathbb{E}Q)^{-1} \\ -(\mathbb{E}Q)^{-1}\mathbb{E}B & (\mathbb{E}Q)^{-1}\mathbb{E}[BB^T](\mathbb{E}Q)^{-1}. \end{bmatrix}$$

It only remains to show that Ω is compactly supported. By Lemma 3.2.6, the imaginary part $\Im[M(z)]$ converges uniformly to 0 as $\Im[z] \to 0$ on $\{z \in \mathbb{H} : \Re[z] \geq \kappa_+\}$. Hence, by Lemma 2.2.9, Ω is compactly supported on $[-\kappa_+, \kappa_+]$, where κ_+ is the constant defined in Lemma 3.2.6. □

We can interpret the integral representation in Lemma 3.2.7 as a matrix-valued Stieltjes transform. Hence, we will refer to it using this terminology. Furthermore, given the normalization of Ω in Lemma 3.2.7, we say that $\Omega_{1,1}$ is a matrix-valued probability measure in the sense that $v^*\Omega_{1,1}v$ is a real Borel measure satisfying $\int_{\mathbb{R}} v^*\Omega_{1,1}(\mathrm{d}\lambda)v = 1$ for every $v \in \mathbb{C}^n$ with $\|v\| = 1$.

Lemma 3.2.7 provides an explicit bound on the solution to (3.2), which we state in the following corollary.

Corollary 3.2.2. *Assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the MDE* (3.2)*. Then, for every $z \in \mathbb{H}$,*

$$\|M(z)\| \leq \|M_\infty\| + \operatorname{dist}(z, \operatorname{supp}(\Omega))^{-1} \left\| \int_{\mathbb{R}} \Omega(\mathrm{d}\lambda) \right\|. \tag{3.12}$$

It is tempting to try to directly generalize Lemma 3.2.7 to the solution of the regularized matrix equation (3.6). However, we encounter a challenge in applying the same procedure to obtain a bound on the solution to the regularized version. The issue arises from the fact that, when the regularization parameter τ is strictly positive, $M_\infty^{(\tau)}$ has a positive semidefinite imaginary part, which implies that the function $z \mapsto M^{(\tau)}(z) - M_\infty$ may not be Herglotz. One potential alternative approach is to utilize a multivariate Herglotz representation, as discussed in [LN17]. This representation provides an integral representation for the function $(z, i\tau) \mapsto M^{(\tau)}(z)$ involving a multivariate measure. However, it should be noted that in such representations, the measure cannot be finite unless it is trivial. Nonetheless, an analogue of Lemma 3.2.7 holds for the upper-left $n \times n$ block of the solution to the RMDE. The result is obtained via a similar argument, so we omit the proof.

Lemma 3.2.8. *Fix $\tau \in \mathbb{R}_{\geq 0}$ and assume that $M \in \mathscr{M}$ such that, for all $z \in \mathbb{H}$, $M(z)$ solves the RMDE* (3.6)*. Then, $M_{1,1}(z) = \int_{\mathbb{R}} \frac{\Omega_{1,1}(\mathrm{d}\lambda)}{\lambda - z}$ for all $z \in \mathbb{H}$, where $\Omega_{1,1}$ is a $n \times n$ matrix-valued measure on Borel subsets of $\mathbb{R}$ satisfying $\int_{\mathbb{R}} \Omega_{1,1}(\mathrm{d}\lambda) = I_n$.*

Aside from their inherent value as results, lemmas 3.2.7 and 3.2.8 are particularly significant because they allow us to treat the solution of the MDE as the limit of solutions to the RMDE as τ approaches zero. The tightness of the family of measures induced by the Stieltjes representation of RMDE solutions plays a key role in this step.

Corollary 3.2.3. *For every $\tau \in \mathbb{R}_{>0}$, let $M^{(\tau)} \in \mathscr{M}_+$ such that, for all $z \in \mathbb{H}$, $M^{(\tau)}(z)$ solves the RMDE* (3.6)*. For all $\tau \in \mathbb{R}_{>0}$, denote by $\Omega_{1,1}^{(\tau)}$ the positive semidefinite measure* *in the Stieltjes transform representation of* $M_{1,1}^{(\tau)}$. *Then, for every* $v \in \mathbb{C}^n$ *and* $\tau_+ \in \mathbb{R}_{>0}$, *the family of measures* $\{v^*\Omega_{1,1}^{(\tau)}v \,:\, \tau \in [0,\tau_+]\}$ *is tight.*

Proof. Let $\tau \in \mathbb{R}_{>0}$. By lemmas 3.2.5 and 3.2.6, there exists $\kappa, c, \kappa_+ \in \mathbb{R}_{>0}$ such that $\|\Im[M_{1,1}(z)]\| \leq c(|z|-\kappa)^{-2}(\tau + \Im[z])$ for every $z \in \mathbb{H}$ with $|z| \geq \kappa_+$. Then,

$$\|\Im[M_{1,1}(\lambda + i\epsilon)]\| \leq c(\sqrt{\lambda^2+\epsilon^2}-\kappa)^{-2}(\tau+\epsilon) \leq c(\lambda-\kappa)^{-2}(\tau+\epsilon)$$

for every $\lambda > \kappa_+$ and $\epsilon \in [0,1]$. Here, c is some constant independent of λ and τ. Hence, for every $\lambda_+ > \kappa_+$, according to the Stieltjes inversion formula for $\Omega^{(\tau)}$ as stated in Lemma 2.2.9,

$$\begin{aligned}\Omega_{1,1}^{(\tau)}\left((\lambda_+,\infty)\right) &\preceq \pi^{-1}\lim_{\epsilon\downarrow 0}\int_{\lambda_+}^{\infty}\Im[M_{1,1}^{(\tau)}(\lambda+i\epsilon)]\mathrm{d}\lambda \\ &\preceq \pi^{-1}\lim_{\epsilon\downarrow 0}\int_{\lambda_+}^{\infty}\|\Im[M_{1,1}^{(\tau)}(\lambda+i\epsilon)]\|\mathrm{d}\lambda \\ &\preceq c\pi^{-1}\tau\int_{\lambda_+}^{\infty}(\lambda-\kappa)^{-2}\mathrm{d}\lambda.\end{aligned}$$

Therefore, if τ is bounded, we may pick $\lambda_+ > \kappa_+$ arbitrarily large to ensure that $\int_{\lambda_+}^{\infty}(\lambda-\kappa)^{-2}\mathrm{d}\lambda$ is arbitrarily small. □

3.2.4 Power Series Representation

As the set of admissible solutions $\mathscr{M}$ comprises analytic matrix-valued functions, any solution to equation (3.2) can be locally expressed as a power series. Utilizing the Stieltjes transform representation provided in Lemma 3.2.7, we can derive a solvable recurrence relation that determines the coefficients in such an expansion. This recurrence relation facilitates the systematic computation of the coefficients in the power series representation of the solution.

Lemma 3.2.9. *Let* $M \in \mathscr{M}$ *such that, for all* $z \in \mathbb{H}$, $M(z)$ *solves the MDE* (3.2) *and let* Ω *be the positive semidefinite measure in Lemma 3.2.7. Then, there exists* $\lambda_+ > \sup\{|\lambda| : \lambda \in \operatorname{supp}(\Omega)\}$ *such that*

$$M(z) = \sum_{j=0}^{\infty} z^{-j}M_j = (\mathbb{E}L - z\Lambda)^{-1}\sum_{j=0}^{\infty}\left(\sum_{k=0}^{\infty} z^{-k}\mathcal{S}(M_k)(\mathbb{E}L - z\Lambda)^{-1}\right)^j$$

for every $z \in \mathbb{H}$ with $|z| \geq \lambda_+$. Here, $M_0 = M_\infty$ and $M_j = -\int_{\mathbb{R}} \lambda^{j-1}\Omega(\mathrm{d}\lambda)$ for every $j \in \mathbb{N}$.

Proof. Since Ω is compactly supported by Lemma 3.2.7, $\sup\{|\lambda| : \lambda \in \operatorname{supp}(\Omega)\}$ is finite. Let $z \in \mathbb{H}$ with $|z| > \sup\{|\lambda| : \lambda \in \operatorname{supp}(\Omega)\}$ and write

$$M(z) = M_\infty + \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{\lambda - z} = M_\infty - z^{-1} \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{1 - \lambda/z}.$$

We recognize $(1-\lambda/z)^{-1}$ as a geometric series and write $(1-\lambda/z)^{-1} = \sum_{j=0}^{\infty} \frac{\lambda^j}{z^j}$. By Fubini's theorem,

$$\int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{1 - \lambda/z} = \sum_{j=0}^{\infty} z^{-j} \int_{\mathbb{R}} \lambda^j \Omega(\mathrm{d}\lambda)$$

which implies that

$$M(z) = M_\infty - \sum_{j=0}^{\infty} z^{-j-1} \int_{\mathbb{R}} \lambda^j \Omega(\mathrm{d}\lambda).$$

On the other hand, by definition, $M(z)$ solves (3.2), and we may write $M(z) = \mathcal{F}(M(z)) = (\mathbb{E}L - \mathcal{S}(M(z)) - z\Lambda)^{-1}$. Using Lemma 2.2.11,

$$(\mathbb{E}L - z\Lambda)^{-1} = \begin{bmatrix} R & -R\mathbb{E}[B^T]Q^{-1} \\ -Q^{-1}\mathbb{E}[B]R & Q^{-1} + Q^{-1}\mathbb{E}[B]R\mathbb{E}[B^T]Q^{-1} \end{bmatrix} \tag{3.13}$$

with $R = \mathcal{R}_{\mathbb{E}[B^T](\mathbb{E}Q)^{-1}\mathbb{E}[B]-\mathbb{E}[A]}(z)$. Since $\|R\| \leq \operatorname{dist}(z, \sigma\left(\mathbb{E}[A] - \mathbb{E}[B^T]Q^{-1}\mathbb{E}[B]\right))^{-1}$ by Lemma 2.3.1, we obtain $(\mathbb{E}L - z\Lambda)^{-1} \to M_\infty$ as $|z| \to \infty$. Because M_∞ is non-zero only in its lower-right $d \times d$ block, it follows from Lemma 3.2.5 and the flatness of the superoperator that $\|\mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}\| \to 0$ as $|z| \to \infty$. Let $\lambda_+ > \max\{|\lambda| : \lambda \in \operatorname{supp}(\Omega)\}$ such that $\|\mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}\| < 1$ for all $z \in \mathbb{H}$ with $|z| \geq \lambda_+$. Then, $I_\ell - \mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}$ is non-singular with Neumann series

$$\left(I_\ell - \mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}\right)^{-1} = \sum_{j=0}^{\infty} \left(\mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}\right)^j.$$

In particular,

$$(\mathbb{E}L - \mathcal{S}(M(z)) - z\Lambda)^{-1} = (\mathbb{E}L - z\Lambda)^{-1} \sum_{j=0}^{\infty} \left(\mathcal{S}(M(z))(\mathbb{E}L - z\Lambda)^{-1}\right)^{j}.$$

We obtain the result by plugging the series expansion for $M(z)$ and using linearity of the superoperator. □

3.3 Existence and Uniqueness

In this section, we establish the existence and uniqueness of a solution to the MDE (3.2). As discussed earlier, we begin by proving the existence of a solution to the RMDE (3.6) for every $\tau \in \mathbb{R}_{>0}$, followed by employing a continuity argument to consider the vanishing regularization.

3.3.1 Solution to the Regularized Matrix Dyson Equation

For every $\tau \in \mathbb{R}_{>0}$, the existence of a unique $M \in \mathscr{M}_+$ satisfying (3.6) for every $z \in \mathbb{H}$ follows directly from [HFS07]. At a high level, the proof of the existence of a solution to the RMDE (3.6) proceeds by demonstrating that the map $M \mapsto \mathcal{F}^{(\tau)}(M)$ is strictly holomorphic on bounded domains of $\mathscr{M}_+$. By the Earle-Hamilton fixed-point theorem, strict holomorphicity implies that the mapping is contractive with respect to the CRF-pseudometric. For further details on the CRF-pseudometric and a proof of the Earle-Hamilton fixed-point theorem, we refer the reader to sections 2.1.2 and 2.1.3.

Define

$$\mathscr{M}_b := \mathrm{Hol}(\mathbb{H} \cap b\mathbb{B}, \mathscr{A}_b), \qquad \text{and} \qquad \mathscr{A}_b := \mathscr{A}_+ \cap \mathscr{B}_b(0) \tag{3.14}$$

for every $b > 0$. Indeed, for every $b \in \mathbb{R}_{>0}$, $\mathscr{M}_b$ is a domain in the Banach space of matrix-valued bounded holomorphic functions on $\mathbb{H}$ with the canonical supremum norm. Also, $\mathscr{A}_b$ is a domain in the Banach space of complex symmetric $\ell \times \ell$ matrices with the operator norm. Using the work we did above, we can easily show that $\mathcal{F}^{(\tau)}$ is indeed a strict holomorphic function on $\mathscr{A}_b$ for every $\tau \in \mathbb{R}_{>0}$. The following lemma is a direct adaptation of [HFS07, Proposition 3.2].

Lemma 3.3.1. *Let $z \in \mathbb{H}$, $\tau, b \in \mathbb{R}_{>0}$ and define $m_b = \|\mathbb{E}L\| + sb + |z| + \tau$. Then, for every $W \in \mathscr{A}_b$, $\|\mathcal{F}^{(\tau)}(W)\| \leq \tau^{-1}$ and $\Im[\mathcal{F}^{(\tau)}(W)] \succeq \tau m_b^{-2} I_\ell \succ 0$. In particular, if $b > \tau^{-1}$, then $\mathcal{F}^{(\tau)}$ maps $\mathscr{M}_b$ strictly into itself.*

Proof. Let $W \in \mathscr{A}_b$. By Lemma 3.2.2, $\|\mathcal{F}^{(\tau)}(W)\| \leq \tau^{-1}$ and $\Im[\mathcal{F}^{(\tau)}(W)] \succeq \tau \mathcal{F}^{(\tau)}(W)[\mathcal{F}^{(\tau)}(W)]^*$. Let $v \in \mathbb{C}^\ell$ such that $\|v\| = 1$. By Cauchy-Schwarz inequality,

$$1 = v^*(\mathcal{F}^{(\tau)}(W))^{-1}\mathcal{F}^{(\tau)}(W)v \leq \|\mathcal{F}^{(\tau)}(W)v\| \|(\mathcal{F}^{(\tau)}(W))^{-*}v\|$$

which implies that $\|(\mathcal{F}^{(\tau)}(W))^{-1}\|^{-2} \leq \|(\mathcal{F}^{(\tau)}(W))^{-*}v\|^{-2} \leq \|\mathcal{F}^{(\tau)}(W)v\|^2$. Additionally,

$$\|(\mathcal{F}^{(\tau)}(W))^{-1}\| = \|\mathbb{E}L - \mathcal{S}(W) - z\Lambda - i\tau I_\ell\| \leq m_b.$$

Thus, $\Im[\mathcal{F}^{(\tau)}(W)] \succeq \tau m_b^{-2} I_\ell$. □

The existence of a unique solution to (3.6) then follows directly from an application of the Earle-Hamilton fixed-point theorem stated in Theorem 2.1.1. Indeed, for every $b \in \mathbb{R}_{>0}$ large enough, $\mathcal{F}^{(\tau)}$ has exactly one fixed point on $\mathscr{M}_b$. Since $\mathscr{M}_+ = \bigcup_{b \in \mathbb{R}_{>0}} \mathscr{M}_b$, we obtain the following result.

Lemma 3.3.2 ([HFS07, Theorem 2.1])**.** *There exists a unique solution $M \in \mathscr{M}_+$ such that $M^{(\tau)}(z)$ solves (3.6) for every $\tau \in \mathbb{R}_{>0}$ and $z \in \mathbb{H}$. Furthermore, for every $W_0 \in \mathscr{M}_+$, the iterates $W_{k+1} = \mathcal{F}^{(\tau)}(W_k)$ converge in norm to $M^{(\tau)}$.*

In what follows, we will denote the unique solution of the RMDE with $\tau \in \mathbb{R}_{>0}$ by $M^{(\tau)}$. While not explicitly stated, the analyticity of $M^{(\tau)}(z)$ in τ can be inferred using an implicit function theorem, as demonstrated in [EKN20, Theorem 2.14]. We state this result in the following lemma but omit the proof, directing the reader to the aforementioned reference for further details.

Lemma 3.3.3. *For every $z \in \mathbb{H}$, the map $\tau \in \mathbb{H} \mapsto M^{(\tau)}(z)$ is analytic.*

3.3.2 Solution to the Matrix Dyson Equation

We now establish the existence and uniqueness of a solution to the MDE (3.2).

Theorem 3.3.1 (Existence and Uniqueness). *There exists a unique matrix-valued function $M \in \mathscr{M}$ such that $M(z)$ solves the MDE* (3.2) *for every $z \in \mathbb{H}$.*

For clarity, we will separate the proof into two distinct sub-proofs: proof of existence and proof of uniqueness. Together, the following two proofs prove Theorem 3.3.1.

Proof of existence in Theorem 3.3.1. For every $k \in \mathbb{N}$, let $M^{(k^{-1})}$ be the unique solution to the RMDE and write

$$M_{1,1}^{(k^{-1})}(z) = \int_{\mathbb{R}} \frac{\Omega_{1,1}^{(k^{-1})}(\mathrm{d}\lambda)}{\lambda - z}$$

the Stieltjes transform representation guaranteed by Lemma 3.2.8. Additionally, let $\{v_j : j \in \mathbb{N}\} \subseteq \mathbb{C}^n$ be a countable dense subset of the ball of n-dimensional complex unit vectors.

By Corollary 3.2.3, the family of measures $\{v_1^* \Omega_{1,1}^{(k^{-1})} v_1 : k \in \mathbb{N}\}$ is tight. Consequently, by Prokhorov's theorem, there exists a probability measure ω_1 and a subsequence $\{\tau_{1,k} : k \in \mathbb{N}\} \subseteq \{k^{-1} : k \in \mathbb{N}\}$ such that $v_1^* \Omega_{1,1}^{(\tau_{1,k})} v_1$ converges weakly to ω_1 as k approaches infinity.

We now proceed inductively. Assume that there exists $m \in \mathbb{N}$ and a collection of compactly supported measures $\{\omega_j : 1 \leq j \leq m\}$ such that $v_j^* \Omega_{1,1}^{(\tau_{m,k})} v_j$ converges weakly to ω_j for all $1 \leq j \leq m$ as k approaches infinity. By Corollary 3.2.3 and Prokhorov's theorem, there exists a probability measure ω_{m+1} and a subsequence $\{\tau_{m+1,k} : k \in \mathbb{N}\} \subseteq \{\tau_{m,k} : k \in \mathbb{N}\}$ such that $v_{m+1}^* \Omega_{1,1}^{(\tau_{m+1,k})} v_{m+1}$ converges weakly to ω_{m+1} as k approaches infinity. Also, by construction of the subsequence, $v_j^* \Omega_{1,1}^{(\tau_{m+1,k})} v_j$ converges weakly to ω_j for all $1 \leq j \leq m+1$ as k approaches infinity.

Let $\tau_k = \tau_{k,k}$ for all $k \in \mathbb{N}$. By construction, $v_j^* \Omega^{(\tau_k)} v_j$ converges weakly to a probability measure ω_j for every $j \in \mathbb{N}$ as $k \to \infty$. Furthermore, by Lemma 3.2.2, $\{M_{1,1}^{(\tau_k)} : k \in \mathbb{N}\}$ is a locally uniformly bounded sequence of analytic functions. Hence, Montel's theorem guarantees the existence of a subsequence of $\{\tau_k : k \in \mathbb{N}\}$, which we will assume WLOG to be $\{\tau_k : k \in \mathbb{N}\}$ for notational convenience, such that $M_{1,1}^{(\tau_k)}$ converges to an analytic function $M_{1,1}$. By the proof of Corollary 3.2.3, there exists $\kappa_+ \in \mathbb{R}_{>0}$ and a constant $c \in \mathbb{R}_{>0}$ such that

$$\int_{\lambda_+}^{\infty} \omega_j(\mathrm{d}\lambda) = \lim_{k\to\infty} \int_{\lambda_+}^{\infty} v_j^* \Omega_{1,1}^{(\tau_k)}(\mathrm{d}\lambda) v_j \leq c \lim_{k\to\infty} \tau_k \int_{\lambda_+}^{\infty} (\lambda - \kappa)^{-2} \mathrm{d}\lambda = 0$$

for every $\lambda_+ \geq \kappa_+$ and $j \in \mathbb{N}$. By Lemma 3.2.8,

$$v_j^* \Im[M_{1,1}] v_j = \lim_{k \to \infty} v_j^* \Im[M_{1,1}^{(\tau_k)}] v_j = \Im[z] \int_{\mathbb{R}} \frac{\omega_j(\mathrm{d}\lambda)}{|\lambda - z|^2}.$$

Since ω_j is a probability measure,

$$\int_{\mathbb{R}} \frac{\omega_j(\mathrm{d}\lambda)}{|\lambda - z|^2} = \int_{[-\kappa_+, \kappa_+]} \frac{\omega_j(\mathrm{d}\lambda)}{|\lambda - z|^2} \geq \left(\max_{\lambda \in [-\kappa_+, \kappa_+]} |\lambda - z| \right)^{-2}$$

which implies that $v_j^* \Im[M_{1,1}] v_j \geq \Im[z] \left(\max_{\lambda \in [-\kappa_+, \kappa_+]} |\lambda - z| \right)^{-2}$ for every $j \in \mathbb{N}$. Fix $z \in \mathbb{H}$, $\epsilon = 3^{-1} (\Im[z])^2 \left(\max_{\lambda \in [-\kappa_+, \kappa_+]} |\lambda - z| \right)^{-2} \in \mathbb{R}_{>0}$. Let $v \in \mathbb{C}^n$ be any unit vector and let $j \in \mathbb{N}$ such that $\|v - v_j\| \leq \epsilon$. Then,

$$v^* \Im[M_{1,1}] v \geq v^* \Im[M_{1,1}] v - 2 \|v_j - v\| \|M_{1,1}\| \|v\| \geq \frac{\epsilon}{3 \Im[z]} > 0.$$

In particular, $\Im[M_{1,1}(z)] \succ 0$ for all $z \in \mathbb{H}$.

Define $M_{1,2}, M_{2,1}$ and $M_{2,2}$ as functions of $M_{1,1}$ using (3.10c), (3.10d) and (3.10e) respectively and let M be the block matrix with j, k block given by $M_{j,k}$ for all $(j,k) \in \{1,2\}^2$. It follows from Lemma 3.2.1, that $\mathbb{E}Q - \mathcal{S}_{2,2}(M)$ is non-singular and that M is well-defined. By construction, it is clear that $M \in \mathscr{M}$ and that $M(z)$ solves (3.2) for all $z \in \mathbb{H}$. □

Proof of uniqueness in Theorem 3.3.1. Uniqueness of the solution follows from analycity and the power series representation in Lemma 3.2.9. Let $\lambda_+ \in \mathbb{R}_{>0}$ such that

$$M(z) = \sum_{j=0}^{\infty} z^{-j} M_j = (\mathbb{E}L - z\Lambda)^{-1} \sum_{j=0}^{\infty} \left(\sum_{k=0}^{\infty} z^{-k} \mathcal{S}(M_k) (\mathbb{E}L - z\Lambda)^{-1} \right)^j$$

for every $z \in \mathbb{H}$ with $|z| \geq \lambda_+$.

Since resolvent of Hermitian matrices are analytic when the spectral parameter is away from the support, is follows from the decomposition in (3.13) that $(\mathbb{E}L - z\Lambda)^{-1}$ is analytic. Write $(\mathbb{E}L - z\Lambda)^{-1} = \sum_{j=0}^{\infty} z^{-j} C_j$ for some complex matrices $\{C_j : j \in \mathbb{N}_{\geq 0}\} \subseteq \mathbb{C}^{\ell \times \ell}$. Plugging this in the power series expansion of M and gathering coefficients of z^{-1}, we get that $M_1 = C_1 + C_0 \mathcal{S}(M_0) C_1 + C_0 \mathcal{S}(M_1) C_0$. We computed in Lemma 3.2.9 that $(\mathbb{E}L - z\Lambda)^{-1} \to M_\infty$

as $|z| \to \infty$ and similarly for $M(z)$. In other words, $C_0 = M_0 = M_\infty$. Looking at the structure of the superoperator, $\mathcal{S}_{2,2}(M_0) = 0$, which gives us $C_0\mathcal{S}(M_0) = 0$. In particular, $M_1 = C_1 + C_0\mathcal{S}(M_0)C_1$ is expressible solely in terms of C_0 and C_1.

This proves the base case. Let $k \in \mathbb{N}$ and assume that $\{M_j : j \in \{0, 1, \dots, k\}\}$ are fully determined by $\{C_j : j \in \mathbb{N}_{\geq 0}\}$. Gathering the coefficients for $z^{-(k+1)}$ in the power series expansion, we get that

$$M_{k+1} = f(M_0, M_1, \dots, M_k) + C_0\mathcal{S}(M_{k+1})C_0$$

for some analytic function f. By induction hypothesis, $f(M_0, M_1, \dots, M_k)$ may be expressed as an analytic function of $\{C_j : j \in \mathbb{N}_{\geq 0}\}$. Furthermore, since $C_0 = M_\infty$ is 0 everywhere outside its lower $d \times d$ block,

$$C_0\mathcal{S}(M_{k+1})C_0 = \begin{bmatrix} 0_{n\times n} & 0_{n\times d} \\ 0_{d\times n} & M_\star\mathcal{S}_{2,2}(M_{k+1})M_\star. \end{bmatrix}$$

Therefore, extracting the upper-left $n \times n$ block, we obtain that the upper-left $n \times n$ block along with both off-diagonal blocks of M_{k+1} are determined by the coefficient matrices $\{C_j : j \in \mathbb{N}_{\geq 0}\}$. Since $\mathcal{S}_{2,2}(M)$ does not depend on the lower-right block of M, we may also determine the lower-right block of M_{k+1}. Inducting, we get that any two solution to (3.2) must be equal for all $z \in \mathbb{H}$ with $|z| > \lambda_+$ for some $\lambda_+ \in \mathbb{R}_{>0}$. By analytic continuation, it follows that any two solution must be equal for all $z \in \mathbb{H}$. □

For the rest of this document, we will denote the unique solution of the MDE with by M, and we will omit the explicit mention of z when the context confines it to a fixed $z \in \mathbb{H}$.

Remark 3.3.1. The rationale behind our choices regarding the settings is now quite evident. By excluding $\tilde{\mathcal{S}}$ from the superoperator, we gain the advantage that each block in the block decomposition of the MDE can be determined by the upper-left $n \times n$ block. This upper-left block exhibits favorable properties, including an a priori norm bound due to the position of the spectral parameter. By leveraging these properties, we establish the existence of a solution for the upper-left block of (3.2). Subsequently, we utilize the existence of this sub-solution to construct the remaining part of the solution.

However, our selection of superoperator also imposes implicit restrictions on the distribution of the entries of L. To treat the term $\tilde{\mathcal{S}}$ as a perturbation, we need to ensure that the perturbation is small. It would be intriguing to explore the possibility of extending our results to a more general superoperator by adapting the proof, thus removing this restriction. For instance, by redefining the lower-right block of the superoperator as $\mathcal{S}_{2,2}(M) = \mathbb{E}[(B-\mathbb{E}B)M_{1,1}(B-\mathbb{E}B)] + \mathbb{E}[(Q-\mathbb{E}Q)M_{2,2}(Q-\mathbb{E}Q)]$, we believe that a promising approach would involve applying similar techniques as in the proof of Lemma 3.3.2 to a similar MDE but with a slightly different admissible set. Specifically, we propose considering admissible solutions that admit the limit $\lim_{\Im[z]\to\infty} M(z) = \operatorname{diag}\{0_{n\times n}, M_\star\}$, where $M_\star$ solves the fixed-point equation $M_\star = (\mathbb{E}Q - \mathbb{E}[(Q-\mathbb{E}Q)M_\star(Q-\mathbb{E}Q)])^{-1}$.

Upon initial inspection, the general properties established in Section 3.2.1 seem applicable, with the main adaptation being that Lemma 3.2.4 would require adjustments to demonstrate that the MDE map preserves the admissible set. We believe that pursuing this avenue holds promise, and it may be possible to extend the existence of a solution from the upper-left block to the entire solution by carefully analyzing the stability of the limiting equation which defines $M_\star$.

During the proof of the uniqueness of the solution to the linearized matrix Dyson equation, we encounter the stability operator evaluated at $z = \infty$. The stability operator is a fundamental concept in the literature on matrix Dyson equations, and we will delve into it further in Section 3.4.2. Its significance lies in the fact that our proof of uniqueness is equivalent to establishing the invertibility of the stability operator at infinity. This result effectively enables us to recursively determine the terms within the power series expansion of M.

In contrast to the assurance provided by Lemma 3.3.2, it is crucial to acknowledge that Theorem 3.3.1 does not guarantee pointwise convergence for the fixed-point iteration $M_{k+1} = \mathcal{F}(M_k)$ with an initial condition $M_0 \in \mathscr{M}$ to the solution of the matrix Dyson equation. This underscores one of the primary reasons we rely on the solution to the regularized MDE as a means to establish stability, effectively treating it as a surrogate for the solution to the MDE. Nonetheless, in our example application, will show in Lemma 4.3.11 that a fixed-point iteration converges to the solution of the MDE.

3.4 Asymptotic Equivalence

Now that we have existence of a unique solution M to (3.2), we want to show that $M(z)$ serves as a favorable asymptotic approximation for the pseudo-resolvent $(L - z\Lambda)^{-1}$. We will compare the pseudo-resolvent with the solution of the MDE using the following pairwise comparisons:

$$\begin{aligned}(L - z\Lambda)^{-1} - M(z) = {} & (L - z\Lambda)^{-1} - \mathbb{E}(L - z\Lambda)^{-1} && \text{(3.15a)}\\ & + \mathbb{E}(L - z\Lambda)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} && \text{(3.15b)}\\ & + \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} - M^{(\tau)}(z) && \text{(3.15c)}\\ & + M^{(\tau)}(z) - M(z). && \text{(3.15d)}\end{aligned}$$

The first comparison in (3.15a) corresponds to the concentration step of our argument. Although this difference may not generally be controlled in norm, we have the capability to demonstrate concentration, either in probability or almost surely, of generalized trace entries of the regularized pseudo-resolvent around its mean. By separating the concentration step from the rest of the method, we adopt a strategy that enables us to primarily work with *deterministic* objects throughout the analysis. This approach offers significant simplifications in various steps and allows us to work with norm bounds.

The second comparison in (3.15b) assesses the proximity of the pseudo-resolvent to its regularized counterpart, measured in norm. We show in Lemma 3.4.1 that this difference can be easily controlled by the parameter τ and $\mathbb{E}\|(L-z\Lambda)^{-1}\|^2$. Consequently, if $\mathbb{E}\|(L-z\Lambda)^{-1}\|^2$ is bounded, we can employ the regularized pseudo-resolvent with small $\tau \in \mathbb{R}_{>0}$ as an approximation for the pseudo-resolvent.

The third comparison, (3.15c), is directly linked with the stability properties of the RMDE. We will use the CRF-pseudometric to show that (3.6) is stable under small additive perturbation. Then, we will show that the expected regularized pseudo-resolvent almost satifies the RMDE, up to a small additive perturbation which vanishes as the dimension of the problem increases. The convergence of the expected regularized pseudo-resolvent to the solution of the regularized matrix Dyson equation depends intricately on the rate at which τ approaches zero while ℓ increases to infinity.

The fourth and final comparison, (3.15d), simply states that the solution to (3.6) should be a good approximation for (3.2) for small τ. For a fixed $\tau \in \mathbb{R}_{>0}$, it follows from the construction of M that $\|M^{(\tau)}(z) - M(z)\| \to 0$ as $\tau \to 0$. However, because we are taking $\ell \to \infty$ and $\tau \to 0$, we rely on Assumption 2 to control this term.

In the upcoming sections, we will demonstrate the convergence of the pseudo-resolvent to the solution of the matrix Dyson equation. Initially, in Section 3.4.1, we will focus on proving the convergence of the expected regularized pseudo-resolvent to the expected pseudo-resolvent, alongside establishing the convergence of the unique solution to the regularized matrix Dyson equation to the solution of the matrix Dyson equation. This involves confirming the convergence of equations (3.15b) and (3.15d) to zero. Following this, we will delve into section 3.4.2 to establish the stability of the regularized matrix Dyson equation with respect to small additive perturbations. Moving forward to section 3.4.3, we will introduce general distributional assumptions on the entries of L to derive simple conditions to establish that the perturbation vanishes as the problem dimension increases. Leveraging a similar approach, we will proceed to section 3.4.4 to illustrate the concentration of the pseudo-resolvent around its mean. Finally, in section 3.4.5, we will integrate these findings to establish the convergence of the pseudo-resolvent to the solution of the matrix Dyson equation. Throughout all of this, we will accumulate various assumptions critical to our analysis.

3.4.1 Regularization

In the proof of Theorem 3.3.1, we define $M_{1,1}$ as the limit point of the normal family $\{M_{1,1}^{(\tau)} : \tau > 0\}$ as $\tau \to 0$. Decomposing (3.2) block-wise, we then observe that $M^{(\tau)}(z)$ converges to $M(z)$ in spectral norm for any fixed $z \in \mathbb{H}$ as τ approaches the origin from above. However, this statement is derived for a fixed dimension $\ell \in \mathbb{N}$. To derive asymptotic global laws, we need to consider the solution of the matrix Dyson equation in the limit as $\ell \to \infty$. Therefore, it would be considerably beneficial to quantify the extent to which $\|M(z) - M^{(\tau)}(z)\|$ varies with respect to $\tau \in \mathbb{R}_{>0}$ and $\ell \in \mathbb{N}$. Ideally, we would like to establish that $\|M(z) - M^{(\tau)}(z)\|$ converges to zero as τ approaches zero from above, and that this convergence is uniform with respect to $\ell \in \mathbb{N}$. This idea is not far-fetched, as the associated pseudo-resolvent $(L - z\Lambda)^{-1}$ satisfies it under reasonable assumptions.

Lemma 3.4.1. *For every $\tau \in \mathbb{R}_{\geq 0}$ and $z \in \mathbb{H}$, $\|(L - z\Lambda - i\tau I_\ell)^{-1} - (L - z\Lambda)^{-1}\| \leq \tau\|(L - z\Lambda)^{-1}\|^2$.*

Proof. By Lemma 2.2.1, $\|(L-z\Lambda-i\tau I_\ell)^{-1}-(L-z\Lambda)^{-1}\| \leq \tau\|(L-z\Lambda-i\tau I_\ell)^{-1}\|\|(L-z\Lambda)^{-1}\|$. Let $v \in \mathbb{C}^\ell$ be arbitrary and decompose $L - z\Lambda - i\tau I_\ell = X + iY - i\tau I_\ell$ with $X = \Re[L - z\Lambda]$ and $Y = \Im[L-z\Lambda]$. Then, using the fact that $(L-z\Lambda-i\tau I_\ell)^* = X-iY+i\tau I_\ell$ and $\Im[Y] \preceq 0$,

$$\begin{aligned} v^* \left(X + iY - i\tau I_\ell\right)^* \left(X + iY - i\tau I_\ell\right) v &\geq v^*(X + iY)^*(X + iY)v + \tau^2 v^* v - 2\tau v^* Y v \\ &\geq v^*(X + iY)^*(X + iY)v. \end{aligned}$$

Because taking the inverse reverses the Loewner partial ordering, it follows that $\|(L - z\Lambda - i\tau I_\ell)^{-1}\| \leq \|(L - z\Lambda)^{-1}\|$. □

By Lemma 3.4.1 and Jensen's inequality, the expected pseudo-resolvent $\mathbb{E}(L - z\Lambda)^{-1}$ is well-approximated by its regularized version for small τ as long as $\mathbb{E}\|(L-z\Lambda)^{-1}\|^2$ is bounded. In fact, if the norm squared of the expected pseudo-resolvent is bounded in expectation, then the regularized expected pseudo-resolvent converges to the expected pseudo-resolvent in operator norm as τ approaches zero from above, uniformly in the dimension. Since our primary objective is to investigate the behavior in the high-dimensional limit, it is essential for the superoperator $\mathcal{S}$, among other objects, to remain bounded as the problem dimension increases.

Assumption 1. Suppose there exists $s \in \mathbb{R}_{>0}$ such that $\|\mathcal{S}(W)\| \leq s\|W\|$ for every $W \in \mathbb{C}^{\ell\times\ell}$ and $\limsup_{\ell\to\infty} s < \infty$. Furthermore, assume that $\limsup_{\ell\to\infty} \|\mathbb{E}L\| < \infty$ and $\limsup_{\ell\to\infty} \mathbb{E}\|(L - z\Lambda)^{-1}\|^2 < \infty$.

As we expect the solution of the matrix Dyson equation to be a deterministic equivalent for the (expected) pseudo-resolvent, we should anticipate a similar behavior from it. We formalize this expectation as follows.

Assumption 2. For every $z \in \mathbb{H}$, there exists a function f and a subsequence $\{\tau_k\} \subseteq \mathbb{R}_{>0}$ such that $\tau_k \to 0$, $f(\tau_k) \to 0$, and $\|M^{(\tau_k)}(z) - M(z)\| \leq f(\tau_k) + o_\ell(1)$ for all $k \in \mathbb{N}$ and every $\ell \in \mathbb{N}$ large enough.

It is noteworthy that Assumption 2 is fulfilled within the frameworks based on the matrix Dyson equation for linearization as detailed in [EKN20; And13; FKN23]. This is explicitly indicated by [EKN20, Equation 4.11], [And13, Estimates 6.3.3.], and [FKN23, Equation A.25]. In general, the validity of Assumption 2 in these cases stems from the ability to construct a dimension-independent representation of the solution to the (R)MDE using tools from free probability. For instance, as asserted by [HT05, Lemma 5.4], such a representation exists whenever L takes the form $L = A_0 \otimes I_n + \sum_{j=1}^{k} A_i \otimes X_j$, where $\{A_j\}_{j=0}^{k}$ forms a collection of complex $d \times d$ self-adjoint matrices, and $\{X_j\}_{j=1}^{k}$ forms a collection of independent random matrices with $\{(X_j)_{a,a}\}_{a=1}^{n} \cup \{(\sqrt{2}\Re X_j)_{a,b}\}_{a<b} \cup \{(\sqrt{2}\Im X_j)_{a,b}\}_{a<b}$ being a collection of n^2 i.i.d. centered Gaussian random variables for every $j \in \{1, 2, \ldots, k\}$.

3.4.2 Stability

In this section, we establish the asymptotic stability of the RMDE (3.6) with respect to small additive perturbations. The stability of the matrix Dyson equation is a fundamental concept in the matrix Dyson equation literature, typically analyzed through the use of the *stability operator*. Following the notation in [Alt+19], the stability operator is defined as $\mathcal{L} : X \in \mathbb{C}^{\ell \times \ell} \mapsto X - M\mathcal{S}(X)M$, where $M \equiv M(z)$ denotes the unique solution to a matrix Dyson equation. The concept of the stability operator is intrinsically linked to the analysis of the matrix Dyson equation [Alt+19; Erd19; AEK19b; FKN23]. The term "stability operator" is aptly chosen because, when it is both invertible and its inverse is bounded, it provides a means to establish the stability of the matrix Dyson equation through techniques like the implicit function theorem, as demonstrated in the work of [AEK19b, Lemma 4.10], and by [Erd19; EKN20]. The stability operator naturally emerges in the uniqueness part of the proof for Section 3.3, where its invertibility at infinity enables us to uniquely and recursively determine the coefficients in the power series expansion of the solution.

The connection between the stability operator and Assumption 2 becomes apparent when we consider the derivative of $M^{(\tau)}(z)$ with respect to $i\tau$, which yields $\mathcal{L}(\partial_{i\tau} M(z)) = (M(z))^2$. Then, under reasonable assumptions, because $M(z)$ is bounded in operator norm by Corollary 3.2.2, we can conclude that Assumption 2 is implied by the requirement of having an invertible stability operator with a bounded inverse. While Assumption 2 may be considered weaker than the requirement of having an invertible stability operator with a bounded inverse, we will eventually delve into the study of the invertibility of a small stability operator when applying our framework to analyze the empirical test error of random features ridge regression.

In order to maintain a certain level of generality, we will fix $z \in \mathbb{H}$ and consider matrices in $\mathscr{A}_b$ for some $b \in \mathbb{R}_{>0}$, as defined in (3.14). When considering the class of matrices $\mathscr{A}_b$, it is helpful to keep in mind the (expected) regularized pseudo-resolvent.

Lemma 3.4.2. *Fix $z \in \mathbb{H}$ and let $\tau, b \in \mathbb{R}_{>0}$ with $\tau^{-1} < b$. Then, $(L - z\Lambda - i\tau I_\ell)^{-1}, \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} \in \mathscr{A}_b$.*

Proof. By Lemma 2.2.5 and Jensen's inequality, $\|(L - z\Lambda - i\tau I_\ell)^{-1}\| \vee \|\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\| \leq \tau^{-1}$. Furthermore, by Lemma 2.2.6, we can follow a similar argument as in the proof of Lemma 3.3.1 to obtain $\Im[(L - z\Lambda - i\tau I_\ell)^{-1}] \succeq \tau(L - z\Lambda - i\tau I_\ell)^{-1}(L - z\Lambda - i\tau I_\ell)^{-*} \succeq \tau\|L - z\Lambda - i\tau I_\ell\|^{-2} \succeq \tau(\|L\| + |z| + \tau)^{-2} \succ 0$. By monotonicity of the expectation, $\Im[\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}] \succeq \tau\mathbb{E}(\|L\| + |z| + \tau)^{-2}$. Since the function $x \mapsto x^{-2}$ is convex for $x \in \mathbb{R}_{>0}$, we may apply Jensen's inequality to obtain $\Im[\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}] \succeq \tau(\mathbb{E}\|L\| + |z| + \tau)^{-2} \succ 0$. □

Fix $z \in \mathbb{H}$, let $b, \tau \in \mathbb{R}_{>0}$ and assume that $F^{(\tau)} \in \mathscr{A}_b$ satisfies

$$(\mathbb{E}L - \mathcal{S}(F^{(\tau)}) - z\Lambda - i\tau I_\ell)F^{(\tau)}(z) = I_\ell + D^{(\tau)}, \tag{3.16}$$

where $D^{(\tau)}$ is a perturbation term. In particular, $F^{(\tau)}$ almost solves (3.6) up to an additive perturbation term $D^{(\tau)}$. For a fixed $z \in \mathbb{H}$, let $E_\tau = \mathcal{F}^{(\tau)}(F^{(\tau)})D^{(\tau)}$ for every $F \in \mathscr{A}_b$ and $\tau \in \mathbb{R}_{>0}$, defining the *error matrix*, and $\epsilon_\tau = \|E_\tau\|$ representing the *magnitude of the error* at τ. The objective for the rest of this section is to show that if ϵ_τ is small, then $F^{(\tau)}$ is close to $M^{(\tau)}$. We will establish this result using properties of the CRF-pseudometric introduced in Section 2.1.2.

Before stating the first lemma, we recall from (3.14) and (3.3.1) that $\mathscr{A}_b := \mathscr{A}_+ \cap \mathscr{B}_b(0)$ is a domain in the Banach space of $\ell \times \ell$ complex matrices for every $b \in \mathbb{R}_{>0}$. The CRF-pseudometric is a crucial tool because $\mathcal{F}^{(\tau)}$ is a strict contraction on $\mathscr{A}_b$ with respect to the CRF-pseudometric. This can be observed and quantified by combining Lemma 3.3.1 with the proof of Theorem 2.1.1.

Lemma 3.4.3. *Fix $z \in \mathbb{H}$ and $\tau \in \mathbb{R}_{>0}$. For every $b \in \mathbb{R}_{>0}$, let $m_b = \|\mathbb{E}L\| + sb + |z| + \tau + 1$ and $\delta = (m_b^2\tau^{-2} - 1)^{-1}$. Suppose that $\tau^{-1}(1+2\delta) < b$ and let ρ denotes the CRF-pseudometric on $\mathscr{A}_b$. Then, for every $X, Y \in \mathscr{A}_b$, $\rho(\mathcal{F}^{(\tau)}(X), \mathcal{F}^{(\tau)}(Y)) \leq (1+\delta)^{-1}\rho(X, Y)$.*

Proof. Let $X \in \mathscr{A}_b$ and define $\mathcal{G} : Y \in \mathscr{A}_b \mapsto \mathcal{F}^{(\tau)}(Y) + \delta(\mathcal{F}^{(\tau)}(Y) - \mathcal{F}^{(\tau)}(X))$. By Lemma 3.3.1, $\|\mathcal{F}^{(\tau)}(Y)\| \leq \tau^{-1}$ and $\Im[\mathcal{F}^{(\tau)}(Y)] \succ \tau m_b^{-2}$ for every $Y \in \mathscr{A}_b$. Hence, $\|\mathcal{G}(Y)\| \leq \tau^{-1} + 2\delta\tau^{-1}$ and $\Im[\mathcal{G}(Y)] \succeq (1+\delta)\Im[\mathcal{F}^{(\tau)}(Y)] - \delta\|\mathcal{F}^{(\tau)}(X)\| \succ (1+\delta)\tau m_b^{-2} - \delta\tau^{-1}$. By our choice of δ and b, $\|\mathcal{G}(Y)\| < b$ and $\Im[G(Y)] \succ 0$ for every $Y \in \mathscr{A}_b$. In fact, $\mathcal{G}$ is a strict holomorphic function on $\mathscr{A}_b$. By the proof of Theorem 2.1.1, $\rho(\mathcal{F}^{(\tau)}(X), \mathcal{F}^{(\tau)}(Y)) \leq (1+\delta)^{-1}\rho(X, Y)$. □

While the MDE map exhibits favorable properties with respect to the CRF-pseudometric, we would like to know whether the CRF-pseudometric captures the convergence relevant to the problem at hand. Specifically, we aim to understand the behavior of the operator norm with respect to the CRF-pseudometric. The subsequent lemma notably demonstrates that the CRF-pseudometric can be employed to establish convergence in terms of generalized trace entries.

Lemma 3.4.4. *Fix $z \in \mathbb{H}$ and $\tau \in \mathbb{R}_{>0}$. Let $b \in \mathbb{R}_{>0}$ with $b > \tau^{-1}$, $F \in \mathscr{A}_b$ and ρ be the CRF-pseudometric on $\mathscr{A}_b$. Then, $\operatorname{tr}(U(M^{(\tau)} - F)) \leq (b + \tau^{-1})\tanh(\rho(M^{(\tau)}, F))$ for every $U \in \mathbb{C}^{\ell\times\ell}$ with $\|U\|_* \leq 1$.*

Proof. Let m_b be defined as in the proof of Lemma 3.4.3 and recall that $\|M^{(\tau)}\| \leq \tau^{-1}$ as well as $\Im[M^{(\tau)}] \succ \tau m_b^{-2}$. In particular, $M^{(\tau)}, F \in \mathscr{A}_b$. Define the holomorphic function $f : X \in \mathscr{A}_b \mapsto \operatorname{tr}(U(X - M^{(\tau)}))(b + \tau^{-1})^{-1} \in \mathbb{B}$. By Lemma 2.2.8, $|\operatorname{tr}(U(X - M^{(\tau)}))| \leq \|X - M^{(\tau)}\| < b + \tau^{-1}$, which ensures that f is well-defined. By Proposition 2.1.1 and (2.1),

$$\operatorname{arctanh}\left|\frac{\operatorname{tr}(U(X - M^{(\tau)}))}{(b + \tau^{-1})}\right| = \rho_\Delta\left(f(M^{(\tau)}), f(X)\right) \leq \rho\left(M^{(\tau)}, X\right).$$

Using the fact that the hyperbolic tangent is increasing, we obtain the result. □

Since the dual norm of the operator norm is the nuclear norm, Lemma 3.4.4 implies that the CRF-pseudometric captures the convergence of the operator norm.

Corollary 3.4.1. *Under the settings of Lemma 3.4.4,* $\|M^{(\tau)} - F\| \leq (b + \tau^{-1}) \tanh(\rho(M^{(\tau)}, F))$.

In Lemma 3.3.2, we established that the solution to the regularized matrix Dyson equation can be obtained using a fixed-point iteration scheme. Using this idea, we will recursively define a sequence of matrices and use the contraction property in Lemma 3.4.3 to control the distance between $M^{(\tau)}(z)$ and $F \in \mathscr{A}_b$ in the CRF-pseudometric. Since the CRF-pseudometric dominates the operator norm, we will obtain convergence in norm. The only remaining ingredient is a loose control of the CRF-pseudometric. In fact, while the norm $\|M^{(\tau)}(z)\|$ may be easily bounded uniformly in ℓ, transferring this bound to the CRF-pseudometric poses additional difficulties which we address in the following lemma.

Lemma 3.4.5. *Let* $z \in \mathbb{H}$ *and* $\tau, b \in \mathbb{R}_{>0}$ *such that* $b > \tau^{-1} + \tau m_b^{-2}$. *Additionally, let* $F^{(\tau)} \in \mathscr{A}_b$ *satisfying* (3.16) *with* $\epsilon_\tau < \tau m_b^{-2}$, ρ *be the CRF-pseudometric on* $\mathscr{A}_b$ *and* $m_b = \|\mathbb{E}L\| + sb + |z| + \tau + 1$ *be defined as in Lemma 3.4.3. Then,* $\rho(\mathcal{F}^{(\tau)}(F^{(\tau)}), F^{(\tau)}) \leq \operatorname{arctanh}(\epsilon_\tau m_b^2/\tau)$.

Proof. We may assume WLOG that $\mathcal{F}^{(\tau)}(F^{(\tau)}) \neq F^{(\tau)}$, as otherwise the claim is trivial. By Lemma 3.3.1, $\|\mathcal{F}^{(\tau)}(F^{(\tau)})\| \leq \tau^{-1}$ and $\Im[\mathcal{F}^{(\tau)}(F^{(\tau)})] \succ \tau m_b^{-2}$. Define the holomorphic function

$$g : w \in \mathbb{B} \mapsto \mathcal{F}^{(\tau)}(F^{(\tau)}) + \frac{w \tau m_b^{-2}}{\|\mathcal{F}^{(\tau)}(F^{(\tau)}) - F^{(\tau)}\|}(F^{(\tau)} - \mathcal{F}^{(\tau)}(F^{(\tau)})) \in \mathscr{A}_b.$$

Then, it is straightforward to check that $\|g(w)\| \leq \tau^{-1} + \tau m_b^{-2} < b$ and $\Im[g(w)] \succ 0$ for every $w \in \mathbb{B}$.

Let $\rho_{\mathbb{B}}$ denote the CRF-pseudometric on $\mathbb{B}$. Using (3.16), we may write $\mathcal{F}^{(\tau)}(F) - F^{(\tau)} = -E_\tau = -\mathcal{F}^{(\tau)}(F^{(\tau)}) D^{(\tau)}$ which implies that $\|\mathcal{F}^{(\tau)}(F) - F^{(\tau)}\| \leq \epsilon_\tau$. Hence, by Proposition 2.1.1,

$$\rho(\mathcal{F}^{(\tau)}(F^{(\tau)}), F^{(\tau)}) = \rho\left(g(0), g\left(\frac{\|\mathcal{F}^{(\tau)}(F) - F^{(\tau)}\|}{\tau m_b^{-2}}\right)\right) \leq \rho_{\mathbb{B}}\left(0, \frac{\|\mathcal{F}^{(\tau)}(F) - F^{(\tau)}\|}{\tau m_b^{-2}}\right).$$

By (2.1), $\rho_{\mathbb{B}}(0, \|\mathcal{F}^{(\tau)}(F) - F^{(\tau)}\| m_b^2/\tau) \leq \operatorname{arctanh}(\epsilon_\tau m_b^2/\tau)$. □

Lemma 3.4.5 controls the discrepancy between the matrix $F^{(\tau)} \in \mathscr{A}_b$, which approximately solves the RMDE up to an additive perturbation term $D^{(\tau)}$, before and after applying the RMDE map once, in terms of the magnitude of the error ϵ_τ.

Combining lemmas 3.4.3 to 3.4.5, we obtain the main stability result.

Theorem 3.4.1. *Fix $z \in \mathbb{H}$, $\tau \in \mathbb{R}_{>0}$ and let $b = \tau^{-1} + 2\tau$ and $m_b = \|\mathbb{E}L\| + sb + |z| + \tau + 1$. Let $F^{(\tau)} \in \mathscr{A}_b$ satisfying* (3.16) *with $\epsilon_\tau < \tau m_b^{-2}$. Then, $\|M^{(\tau)} - F^{(\tau)}\| \leq 2(\tau + \tau^{-1}) \tanh((1 + \tau^2)(m_b^2 \tau^{-2} - 1)\mathrm{arctanh}(\epsilon_\tau m_b^2/\tau))$.*

Proof. Recall that $\delta = (m_b^2 \tau^{-2} - 1)^{-1}$. We begin by verifying that this choice of b satisfies the assumptions in lemmas 3.4.3 and 3.4.5. That is, we show that $\tau^{-1}(1 + 2\delta) < b$ and $b > \tau^{-1} + \tau m_b^{-2}$. To this end, notice that $m_b^2 \tau^{-2} > \tau^{-2} + 1$. Therefore, $\delta < \tau^2$ and $\tau^{-1}(1 + 2\delta) < b$. Furthermore, $m_b^{-2} < 1$ so $\tau^{-1} + \tau m_b^{-2} < b$.

By Corollary 3.4.1, $\|M^{(\tau)} - F^{(\tau)}\| \leq (b + \tau^{-1}) \tanh(\rho(M^{(\tau)}, F^{(\tau)}))$. Recursively define a sequence $\{M_k : k \in \mathbb{N}_0\} \subseteq \mathscr{A}_b$ such that $M_0 = F^{(\tau)}$ and $M_k = \mathcal{F}^{(\tau)}(M_{k-1})$ for every $k \in \mathbb{N}$. By lemmas 3.4.3 and 3.4.5,

$$\begin{aligned}
\rho(M^{(\tau)}, F^{(\tau)}) &\leq \rho(M^{(\tau)}, M_k) + \sum_{j=1}^{k} \rho(M_j, M_{j-1}) \\
&\leq (1 + \delta)^{-k} \rho(M^{(\tau)}, F^{(\tau)}) + \rho(\mathcal{F}^{(\tau)}(F^{(\tau)}), F^{(\tau)}) \sum_{j=1}^{k} (1 + \delta)^{-j} \\
&\leq (1 + \delta)^{-k} \rho(M^{(\tau)}, F^{(\tau)}) + \frac{\mathrm{arctanh}(\epsilon_\tau m_b^2/\tau)}{1 - (1 + \delta)^{-1}}.
\end{aligned}$$

Since the above inequality hold for every $k \in \mathbb{N}$, we may take the limit as $k \to \infty$ to obtain $\rho(M^{(\tau)}, F^{(\tau)}) \leq (1+\delta)\mathrm{arctanh}(\epsilon_\tau m_b^2/\tau)/\delta$. Combining everything, using the fact that $\delta < \tau^2$, we obtain the result. □

To summarize, if $F^{(\tau)} \in \mathscr{A}_{\tau^{-1}+2\tau}$ approximately solves (3.6) up to an additive perturbation term $D^{(\tau)}$ in the sense of (3.16), and $\|\mathcal{F}^{(\tau)}(F^{(\tau)}) D^{(\tau)}\| \leq \tau^{-1} \|D^{(\tau)}\|$ vanishes as $\ell \to \infty$, then $\|M^{(\tau)} - F^{(\tau)}\|$ converges to 0 as $\ell \to \infty$. Here, $M^{(\tau)}$ refers to the unique solution to the regularized matrix Dyson equation. Notably, if we take $F^{(\tau)} = \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}$, then we obtain a way to establish pointwise convergence in z and τ. Utilizing assumptions 1 and 2 and lemmas 3.4.1 and 3.4.2, along with a diagonalization argument, we can establish the following result.

Corollary 3.4.2. *Let $z \in \mathbb{H}$, $M \in \mathscr{M}$ be the unique solution to* (3.2) *and assume that assumptions 1 and 2 hold. For every $\tau \in \mathbb{R}_{>0}$, let $D^{(\tau)}$ be defined by* (3.16) *with $F^{(\tau)} = \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}$. If $\|D^{(\tau)}\| \to 0$ as $\ell \to \infty$ for every $\tau \in \mathbb{R}_{>0}$, then $\|\mathbb{E}(L - z\Lambda)^{-1} - M(z)\| \to 0$ as $\ell \to \infty$.*

Proof. The result follows from a combination of Theorem 3.4.1 along with assumptions 1 and 2 and lemmas 3.4.1 and 3.4.2. We write

$$\begin{aligned}\|\mathbb{E}(L - z\Lambda)^{-1} - M(z)\| \leq \, & \|\mathbb{E}(L - z\Lambda)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\| \\ & + \|\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} - M^{(\tau)}(z)\| \\ & + \|M^{(\tau)}(z) - M(z)\|.\end{aligned}$$

For the first term, we use Lemma 3.4.1 and Assumption 1 to obtain $\|\mathbb{E}(L - z\Lambda)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\| \lesssim \tau$. For the second term, it follows from Theorem 3.4.1 and Assumption 2 that $\|\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} - M^{(\tau)}(z)\| \lesssim \tau^{-1} \tanh(\tau^{-4} \text{arctanh}(\tau^{-4}\|D^{(\tau)}\|))$. For the third term, it follows directly from Assumption 2 that there exists a sequence $\{\tau_k\}_{k\in\mathbb{N}}$ and a function $f : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ such that $\tau_k \to 0$ as $k \to \infty$, $f(\tau_k) \to 0$ as $k \to \infty$ and $\|M^{(\tau)}(z) - M(z)\| \leq f(\tau_k) + o_\ell(1)$. We obtain the result by letting $\tau \to 0$ and $\ell \to \infty$ simultaneously such that the rate of convergence of τ is chosen in terms of the rate of convergence of $D^{(\tau)}$. □

The proof of Corollary 3.4.2 highlights a contrast between τ and ℓ regarding their impact on convergence behavior. On one hand, as τ approaches 0, the solution to the regularized matrix Dyson equation converges to the solution of the matrix Dyson equation. However, the solution to the matrix Dyson equation possesses less desirable properties compared to the regularized solution, notably because the imaginary part is not guaranteed to be positive definite. On the other hand, as ℓ approaches ∞, we establish stability for fixed τ by controlling the magnitude of the error ϵ_τ. We leverage this stability to establish convergence between the expected regularized pseudo-resolvent and the solution to the MDE. Therefore, we need to allow τ to approach 0 slowly enough as $\ell \to \infty$ to preserve the stability property.

3.4.3 Perturbation

In view of Theorem 3.4.1 and Corollary 3.4.2, the focus shifts to proving that the perturbation matrix vanishes in norm as the problem dimension grows for every regularization parameter. For each $\tau \in \mathbb{R}_{>0}$, we consider the expected regularized pseudo-resolvent $F^{(\tau)} \equiv F^{(\tau)}(z) = \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1} \in \mathscr{A}_+$ which satisfies $(\mathbb{E}L - \mathcal{S}(F^{(\tau)}(z)) - z\Lambda - i\tau I_\ell)F^{(\tau)}(z) = I_\ell + D^{(\tau)}$ where $D^{(\tau)}$ is a regularized perturbation term explicitly given by

$$D^{(\tau)} = \mathbb{E}\left[\left(\mathbb{E}L - L - \mathcal{S}(\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1})\right)(L - z\Lambda - i\tau I_\ell)^{-1}\right]. \tag{3.17}$$

There are various methods to establish that the perturbation term is vanishing, depending on the assumptions about the linearization L. To apply our framework and study random features ridge regression, we naturally choose a route based on Gaussian concentration inequalities inspired by works such as [LLC18; Cho22]. This choice confines our theoretical considerations to linearizations characterized by Gaussian-concentrated entries.

Assumption 3. Suppose that $\gamma \in \mathbb{N}$, $g \sim \mathcal{N}(0, I_\gamma)$ and that there exists a map $\mathcal{C} : \mathbb{R}^\gamma \mapsto \mathbb{R}^{\ell\times\ell}$ such that $L \equiv L(g) = \mathcal{C}(g) + \mathbb{E}L$. Furthermore, assume that $\mathcal{C}$ is symmetric in the sense that $\mathcal{C}(x) = (\mathcal{C}(x))^T$ for every $x \in \mathbb{R}^\gamma$.

This allows us to derive straightforward conditions on the function $\mathcal{C}$, ensuring $D^{(\tau)} \to 0$ as $\ell \to \infty$ for all $\tau \in \mathbb{R}_{>0}$. Additionally, we employ a Gaussian concentration inequality as in Proposition 2.3.1 to show that Lipschitz functionals of the regularized pseudo-resolvent $(L - z\Lambda - i\tau I_\ell)^{-1}$ concentrate around their mean. Alternatively, we could utilize the Nash-Poincaré inequality [Pas05, Proposition 2.4], a consequence of Stein's lemma, to establish concentration.

Under Assumption 3, we aim to decompose the perturbation matrix $D^{(\tau)}$ into terms that are amenable to analysis. To achieve this, define

$$\begin{aligned}\Delta(L, \tau; z) = {}& \mathbb{E}[(L - \mathbb{E}L)(L - z\Lambda - i\tau I_\ell)^{-1}] \\ & + \mathbb{E}[(\tilde{L} - \mathbb{E}L)(L - z\Lambda - i\tau I_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - z\Lambda - i\tau I_\ell)^{-1}]\end{aligned} \tag{3.18}$$

where $\tilde{L}$ is an i.i.d. copy of L and consider the decomposition

$$\begin{aligned} D^{(\tau)} &= \mathbb{E}\left[\mathcal{S}((L - z\Lambda - i\tau I_\ell)^{-1})(L - z\Lambda - i\tau I_\ell)^{-1}\right] - \mathcal{S}(F^{(\tau)})F^{(\tau)} && \text{(3.19a)} \\ &+ \mathbb{E}\left[\tilde{\mathcal{S}}((L - z\Lambda - i\tau I_\ell)^{-1})(L - z\Lambda - i\tau I_\ell)^{-1}\right] && \text{(3.19b)} \\ &- \Delta(L, \tau). && \text{(3.19c)} \end{aligned}$$

The first perturbation term in (3.19a) arises from the use of the expected pseudo-resolvent in Theorem 3.4.1. To ensure that this perturbation term is asymptotically small, we require the superoperator $\mathcal{S}$ to be *averaging.* This implies that $\mathcal{S}((L - z\Lambda - i\tau I_\ell)^{-1})$ should exhibit a "law of large numbers" behavior and converge around a deterministic limit. While working directly with the pseudo-resolvent would eliminate this specific perturbation term from the expectation of $D^{(\tau)}$, such an approach would have its disadvantages. Indeed, utilizing the expected pseudo-resolvent allows us to work with deterministic objects and leverage norm bounds. We derive a condition for $\mathcal{S}((L - z\Lambda - i\tau I_\ell)^{-1})$ to concentrate around its mean based on Gaussian concentration.

The second perturbation term, as expressed in (3.19b), arises from our specific definition of the superoperator and would not be present if we defined the superoperator as $\mathbb{E}[(L - \mathbb{E}L)M(L - \mathbb{E}L)]$. However, our chosen definition of the superoperator ensures that the MDE can be expressed solely in terms of its upper-left $n \times n$ block. This distinction allows us to establish the existence of a solution to (3.2). Consequently, we view $\tilde{\mathcal{S}}$ as a correction term that should be vanishing in ℓ.

Finally, (3.19c) posits that the matrix L should approximate a Gaussian distribution in the sense that it should asymptotically satisfy a matrix Stein lemma with a vanishing error. The quantity $\|\Delta(L, \tau)\|$ serves informally as a metric characterizing the distance between L and a matrix with Gaussian entries, since $\Delta(L, \tau) = 0$ holds whenever L has Gaussian entries.

Lemma 3.4.6. *If $\tau \in \mathbb{R}_{>0}$, $z \in \mathbb{H}$ and Assumption 3 holds with a linear map $\mathcal{C}$, then $\Delta(L, \tau; z) = 0$.*

Proof. Let $j, k \in \{1, 2, \dots, \ell\}$ be arbitrary. Consider $\mathcal{C}$ as a $\ell \times \ell \times \gamma$ tensor such that $[\mathcal{C}(g)]_{j,k} = \mathcal{C}_{j,k,\alpha} g_\alpha$. Here, we use Einstein's notation which means that we sum over every subscript appearing at least two times in a given expression. By Stein's lemma, which we stated in Proposition 2.3.2,

$$\begin{aligned}\mathbb{E}\left[(L-\mathbb{E}L)(L-z\Lambda-i\tau I_\ell)^{-1}\right]_{j,k} &= \mathbb{E}\left[\mathcal{C}_{j,m,\alpha}g_\alpha(L-z\Lambda-i\tau I_\ell)^{-1}_{m,k}\right]\\ &= \mathbb{E}\left[\mathcal{C}_{j,m,\alpha}\frac{\partial(L-z\Lambda-i\tau I_\ell)^{-1}_{m,k}}{\partial g_\alpha}\right]\end{aligned}$$

Let $e_\alpha \in \mathbb{R}^\gamma$ be the α-th canonical basis vector, $\delta \in \mathbb{R}_{>0}$ and $L_\delta = \mathcal{C}(g+\delta e_\alpha)+\mathbb{E}L$. Then,

$$\begin{aligned}(L_\delta-z\Lambda-i\tau I_\ell)^{-1}_{m,k}-(L-z\Lambda-i\tau I_\ell)^{-1}_{m,k} &= \left[(L_\delta-z\Lambda-i\tau I_\ell)^{-1}(L-L_\delta)(L-z\Lambda-i\tau I_\ell)^{-1}\right]_{m,k}\\ &= -\delta\left[(L_\delta-z\Lambda-i\tau I_\ell)^{-1}\mathcal{C}(e_\alpha)(L-z\Lambda-i\tau I_\ell)^{-1}\right]_{m,k}.\end{aligned}$$

Taking the limit of the quotient of this difference with δ as δ approaches 0, we get that

$$\frac{\partial(L-z\Lambda-i\tau I_\ell)^{-1}_{m,k}}{\partial g_\alpha} = -\left[(L-z\Lambda-i\tau I_\ell)^{-1}\mathcal{C}(e_\alpha)(L-z\Lambda-i\tau I_\ell)^{-1}\right]_{m,k}$$

and, consequently, $\mathbb{E}[(L-\mathbb{E}L)(L-z\Lambda-i\tau I_\ell)^{-1}]_{j,k} = -\mathbb{E}[\mathcal{C}_{j,m,\alpha}(L-z\Lambda-i\tau I_\ell)^{-1}_{m,a}\mathcal{C}_{a,b,\alpha}(L-z\Lambda-i\tau I_\ell)^{-1}_{b,k}]$. Note that $\mathbb{E}[(L-\mathbb{E}L)^T W(L-\mathbb{E}L)]_{j,k} = \mathbb{E}[\mathcal{C}_{j,a,\alpha}g_\alpha W_{a,b}\mathcal{C}_{b,k,\beta}g_\beta] = \mathbb{E}[\mathcal{C}_{j,a,\alpha}W_{a,b}\mathcal{C}_{b,k,\alpha}]$ for every $W \in \mathbb{R}^{\ell\times\ell}$ independent of L. The result follows. □

Thus, it is trivial to control $\|\Delta(L,\tau)\|$ when L has Gaussian entries. Alternatively, an interpolation approach based on cumulant bounds in the spirit of [LP09, Proposition 3.1] appears to be a suitable avenue to extend the result to other distributions. In Section 4.3.1, we employ a leave-one-out strategy to demonstrate that $\|\Delta(L,\tau)\|$ is vanishing in ℓ for every $\tau \in \mathbb{R}_{>0}$.

In order to maintain an adequate level of abstraction, we will directly assume that the mapping $g \mapsto \mathcal{S}(L(g)-z\Lambda-i\tau I_\ell)^{-1}$ is λ-Lipschitz with respect to the operator norm and employ an ϵ-net argument to obtain bounds $\mathbb{E}_{\tilde{L}}[\|(\tilde{L}-\mathbb{E}L)((L-z\Lambda-i\tau I_\ell)^{-1}-\mathbb{E}(L-z\Lambda-i\tau I_\ell)^{-1})(\tilde{L}-\mathbb{E}L)\|]$ for $k \in \mathbb{N}$.

Lemma 3.4.7. *Fix $z \in \mathbb{H}$ and $\tau \in \mathbb{R}_{>0}$. Assume that the mapping $g \in (\mathbb{R}^\gamma, \|\cdot\|_2) \mapsto$*

$\mathcal{S}((L(g) - z\Lambda - i\tau I_\ell)^{-1}) \in (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_2)$ *is* λ*-Lipschitz. Then, for every* $k \in \mathbb{N}$*, there exists an absolute constant* $c \in \mathbb{R}_{>0}$ *such that*

$$\mathbb{E}\left[\|\mathcal{S}\left((L - z\Lambda - i\tau I_\ell)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\right)\|^k\right] \leq c\ell^{k/2}\lambda^k.$$

Proof. By Proposition 2.3.1, there exists some absolute constant $c_1, c_2 \in \mathbb{R}_{>0}$ such that

$$\mathbb{P}\left(\lambda^{-1}\left|u^*\mathcal{S}\left((L - z\Lambda - i\tau I_\ell)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\right)v\right| \geq x\right) \leq c_1 e^{-c_2 x^2}$$

for all unit vectors $u, v \in \mathbb{C}^\ell$. Suppose that $\epsilon \in (0, 2^{-3/2})$ and let $\mathscr{N}$ be an ϵ-net for the unit ball of ℓ-dimensional real vectors. Then, given $u \in \mathbb{C}^\ell$, we may find $v_1, v_2 \in \mathscr{N}$ such that $\|u - v_1 - iv_2\|^2 = \|\Re[u] - v_1\|^2 + \|\Im[u] - v_2\|^2 \leq 2\epsilon^2$. In particular, $\mathscr{N} + i\mathscr{N} := \{v_1 + iv_2 : v_1, v_2 \in \mathscr{N}\}$ forms a $\sqrt{2}\epsilon$-net for the unit sphere of ℓ-dimensional complex unitary vectors. By [Ver18, Corollary 4.2.13], $|\mathscr{N} + i\mathscr{N}| \leq (2\epsilon^{-1} + 1)^{2\ell}$.

Let $u, v \in \mathbb{C}^\ell$ be unitary and let $u_0, v_0 \in \mathscr{N} + i\mathscr{N}$ such that $\|u - u_0\| \leq \sqrt{2}\epsilon$ and $\|v - v_0\| \leq \sqrt{2}\epsilon$. Let $X \in \mathbb{C}^{\ell\times\ell}$ be any matrix. Using the identity $u^*Xv = u_0^*Xv_0 + (u^* - u_0^*)Xv + u_0^*X(v - v_0)$, we obtain $|u^*Xv| \leq \sup_{u_0, v_0 \in \mathscr{N} + i\mathscr{N}} |u_0^*Xv_0| + 2^{3/2}\epsilon\|X\|$. Taking the supremum over unitary complex vectors u and v, we get that $\|X\| \leq (1 - 2^{3/2}\epsilon)^{-1}\sup_{u_0, v_0 \in \mathscr{N} + i\mathscr{N}} |u_0^*Xv_0|$. In particular, for $X = \mathcal{S}\left((L - z\Lambda - i\tau I_\ell)^{-1} - \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}\right)$, we apply a union bound to obtain $\mathbb{P}\left(\lambda^{-1}\|X\| \geq y\right) \leq c_1(2\epsilon^{-1} + 1)^{4\ell}e^{-c_3 y^2}$ for every $y \in \mathbb{R}$, where $c_3 = c_2(1 - 2^{3/2}\epsilon)$. Let $c_4 \in \mathbb{R}_{>0}$ and $y = c_4(2\sqrt{\ell} + x)$ for all $x \in \mathbb{R}_{\geq 0}$ such that $y^2 \geq c_4^2(4\ell + x^2)$. Choosing c_4 large enough such that $c_3^2 y^2 \geq \ln(2\epsilon^{-1} + 1)4\ell + x^2$ for every $x \in \mathbb{R}_{\geq 0}$, we have $c_1(2\epsilon^{-1} + 1)^{4\ell}e^{-c_3 y^2} \leq c_1(2\epsilon^{-1} + 1)^{4\ell}e^{-\ln(2\epsilon^{-1}+1)4\ell - x^2} = c_1 e^{-x^2}$. Let $k \in \mathbb{N}$ be arbitrary. Then,

$$\begin{aligned}\mathbb{E}[\lambda^{-k}\|X\|^k] &= k\int_0^\infty \mathbb{P}\left(\lambda^{-1}\|X\| \geq y\right)y^{k-1}\mathrm{d}y \\ &= k\int_0^{2c_4\sqrt{\ell}} \mathbb{P}\left(\lambda^{-1}\|X\| \geq y\right)y^{k-1}\mathrm{d}y + k\int_{2c_4\sqrt{\ell}}^\infty \mathbb{P}\left(\lambda^{-1}\|X\| \geq y\right)y^{k-1}\mathrm{d}y.\end{aligned}$$

On one hand, it is straightforward to bound $k\int_0^{2c_4\sqrt{\ell}} \mathbb{P}\left(\lambda^{-1}\|X\| \geq y\right)y^{k-1}\mathrm{d}y \leq 2^k c_4^k \ell^{k/2}$. On the other hand, we have

$$\begin{aligned}
k\int_{2c_4\sqrt{\ell}}^{\infty}\mathbb{P}\left(\lambda^{-1}\|X\|\geq y\right)y^{k-1}\mathrm{d}y &= c_4^k k\int_{2c_4\sqrt{\ell}}^{\infty}\mathbb{P}\left(\lambda^{-1}\|X\|\geq c_4(2\sqrt{\ell}+x)\right)(2\sqrt{\ell}+x)^{k-1}\mathrm{d}x\\
&\leq c_1c_4^k k\int_{2c_4\sqrt{\ell}}^{\infty}e^{-x^2}(2\sqrt{\ell}+x)^{k-1}\mathrm{d}x\\
&= c_1c_4^k k\int_{0}^{\infty}e^{-(x+2c_4\sqrt{\ell})^2}(2(1+c_4)\sqrt{\ell}+x)^{k-1}\mathrm{d}x.
\end{aligned}$$

In particular, writing $\int_0^\infty e^{-(x+2c_4\sqrt{\ell})^2}(2(1+c_4)\sqrt{\ell}+x)^{k-1}\mathrm{d}x \leq e^{-2c^4\ell}\int_0^\infty e^{-x^2}(2(1+c_4)\sqrt{\ell}+x)^{k-1}\mathrm{d}x$ and noting that $\int_0^\infty e^{-x^2}(2(1+c_4)\sqrt{\ell}+x)^{k-1}\mathrm{d}x$ is polynomial in ℓ, we get that $\int_{2c_4\sqrt{\ell}}^\infty \mathbb{P}\left(\lambda^{-1}\|X\|\geq y\right)y^{k-1}\mathrm{d}y = o_\ell(1)$. The result follows. □

The practicality of Lemma 3.4.7 relies on the Lipschitz constant λ satisfying $\lim_{\ell\to\infty}\lambda\sqrt{\ell} = 0$. Under this condition, we may show that the perturbation term $D^{(\tau)}$ vanishes in norm as $\ell\to\infty$ for every $\tau\in\mathbb{R}_{>0}$.

Lemma 3.4.8. *Let $\tau\in\mathbb{R}_{>0}$, $z\in\mathbb{H}$ and $D^{(\tau)}$ be the perturbation matrix in* (3.17). *Under Assumption 3, assume that the mapping $g\in(\mathbb{R}^\gamma,\|\cdot\|_2)\mapsto\mathcal{S}((L(g)-z\Lambda-i\tau I_\ell)^{-1})\in(\mathbb{C}^{\ell\times\ell},\|\cdot\|_2)$ is λ-Lipschitz. Then, there exists an absolute constant $c\in\mathbb{R}_{>0}$ such that*

$$\|D^{(\tau)}\|\leq c\tau^{-1}\sqrt{\ell}\lambda+\tau^{-1}\mathbb{E}\|\tilde{\mathcal{S}}((L-z\Lambda-i\tau I_\ell)^{-1})\|+\|\Delta(L,\tau)\|.$$

Proof. By (3.19), we have

$$\begin{aligned}
\|D^{(\tau)}\| &\leq \|\mathbb{E}[\mathcal{S}((L-z\Lambda-i\tau I_\ell)^{-1}-\mathbb{E}(L-z\Lambda-i\tau I_\ell)^{-1})(L-z\Lambda-i\tau I_\ell)^{-1}]\|\\
&\quad+\|\mathbb{E}[\tilde{\mathcal{S}}((L-z\Lambda-i\tau I_\ell)^{-1})(L-z\Lambda-i\tau I_\ell)^{-1}]\|+\|\Delta(L,\tau)\|.
\end{aligned}$$

By Jensen's inequality, submultiplicativity of the operator norm, Lemma 3.2.2 and Lemma 3.4.7, there exists an absolute constant $c\in\mathbb{R}_{>0}$ such that $\|\mathbb{E}[\mathcal{S}((L-z\Lambda-i\tau I_\ell)^{-1}-\mathbb{E}(L-z\Lambda-i\tau I_\ell)^{-1})(L-z\Lambda-i\tau I_\ell)^{-1}]\|\leq c\tau^{-1}\sqrt{\ell}\lambda$. Similarly, $\|\mathbb{E}[\tilde{\mathcal{S}}((L-z\Lambda-i\tau I_\ell)^{-1})(L-z\Lambda-i\tau I_\ell)^{-1}]\|\leq\tau^{-1}\mathbb{E}\|\tilde{\mathcal{S}}((L-z\Lambda-i\tau I_\ell)^{-1})\|$. The result follows. □

As a direct outcome of lemmas 3.2.2 and 3.4.8, it follows that as the dimension ℓ tends towards infinity, $\|D^{(\tau)}\|$ diminishes, provided that $\lim_{\ell\to\infty}\sqrt{\ell}\lambda = 0$, $\lim_{\ell\to\infty}\|\mathcal{S}\| = 0$, and

$\lim_{\ell\to\infty} \|\Delta(L,\tau)\| = 0$ for every $\tau \in \mathbb{R}_{>0}$ sufficiently small. Before we proceed, we highlight two significant observations.

Remark 3.4.1. Despite the possibility to simplify the upper bound $\tau^{-1}\mathbb{E}\|\tilde{\mathcal{S}}((L - z\Lambda - i\tau I_\ell)^{-1})\|$ in Lemma 3.4.8 to $\tau^{-2}\|\tilde{\mathcal{S}}\|$ using Lemma 3.2.2, we choose to retain its current form since we consider it as more insightful and convenient in certain scenarios. For example, if $B = 0$ in the linearization, the regularized pseudo-resolvent becomes block-diagonal. Hence, the chosen bound allows us to focus solely on demonstrating the vanishing operator norm of $\tilde{\mathcal{S}}$ within the set of block diagonal matrices as ℓ grows.

Remark 3.4.2. In the context of our application concerning the test error of random features ridge regression, we upper-bound the Lipschitz constant λ in Lemma 3.4.8 by $\lambda \leq \tau^{-2}\|\mathcal{S}\|_{F\to 2}\lambda_{\mathcal{C}}$. Here, $\|\mathcal{S}\|_{F\to 2}$ denotes the operator norm of the map $\mathcal{S} : (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_F) \mapsto (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_2)$, while $\lambda_{\mathcal{C}}$ represents the Lipschitz constant associated with the map $\mathcal{C} : (\mathbb{R}^\gamma, \|\cdot\|_2) \mapsto (\mathbb{R}^{\ell\times\ell}, \|\cdot\|_F)$. Consequently, $\lim_{\ell\to\infty} \sqrt{\ell}\lambda = 0$ ensues from $\|\mathcal{S}\|_{F\to 2} \lesssim \ell^{-\frac{1}{2}}$ and $\lambda_{\mathcal{C}} \lesssim \ell^{-\frac{1}{2}}$.

The convergence results from Corollary 3.4.2 and Lemma 3.4.8, coupled with the outlined assumptions, establish that $M(z)$ serves as a deterministic equivalent for the expected pseudo-resolvent $\mathbb{E}(L - z\Lambda)^{-1}$ across the entire upper-half complex plane $\mathbb{H}$.

Corollary 3.4.3. *Let $z \in \mathbb{H}$ and λ be defined as in Lemma 3.4.8. Under assumptions 1 to 3, suppose that $\lim_{\ell\to\infty} \sqrt{\ell}\lambda = \lim_{\ell\to\infty} \|\tilde{\mathcal{S}}\| = \lim_{\ell\to\infty} \|\Delta(L,\tau)\| = 0$ for every $\tau \in \mathbb{R}_{>0}$ small enough. Then, $\|\mathbb{E}(L - z\Lambda)^{-1} - M(z)\| \to 0$ as $\ell \to \infty$.*

3.4.4 Concentration

The only remaining task is to establish that the expected pseudo-resolvent is itself a deterministic equivalent for the pseudo-resolvent. We present one possible approach, which is based on Assumption 3.

Lemma 3.4.9. *Under Assumption 3, let $U \in \mathbb{C}^{\ell\times\ell}$ with $\|U\|_F \leq 1$ and assume that the map $g \in (\mathbb{R}^\gamma, \|\cdot\|_2) \mapsto (L(g) - z\Lambda)^{-1} \in (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_F)$ is λ-Lipschitz with $\lambda \asymp \ell^{-r}$ for some $r > 0$. Then, $\operatorname{tr}(U((L - z\Lambda)^{-1} - \mathbb{E}(L - z\Lambda)^{-1})) \to 0$ almost surely as $\ell \to \infty$.*

Proof. Let $g_1, g_2 \in \mathbb{R}^\gamma$. Then, by Cauchy-Schwarz's inequality $|\operatorname{tr}(U((L(g_1)-z\Lambda)^{-1}-(L(g_2)-z\Lambda)^{-1}))| \leq \|U\|_F \|(L(g_1)-z\Lambda)^{-1}-(L(g_2)-z\Lambda)^{-1}\|_F \leq \lambda\|g_1-g_2\|$. By Proposition 2.3.1, there exists absolute constants $c_1, c_2 \in \mathbb{R}_{>0}$ such that $\mathbb{P}(|\operatorname{tr}(U((L-z\Lambda)^{-1}-\mathbb{E}(L-z\Lambda)^{-1}))| \geq x) \leq c_1 e^{-c_2 x^2/\lambda^2} \leq c_1 e^{-c_3 x^2 \ell^{2r}}$ for every $x \in \mathbb{R}_{>0}$ and $\ell \in \mathbb{N}$. Here, $c_3 \in \mathbb{R}_{>0}$ is some constant satisfying $0 < c_3 \leq c_2/(\lambda \ell^r)^2$ for every $\ell \in \mathbb{N}$ large enough. In particular, for any $\epsilon \in \mathbb{R}_{>0}$, $\sum_{\ell=1}^\infty \mathbb{P}(|\operatorname{tr}(U((L-z\Lambda)^{-1}-\mathbb{E}(L-z\Lambda)^{-1}))| \geq \epsilon) \leq c_1 \sum_{\ell=1}^\infty e^{-c_3\epsilon^2\ell^{2r}}$. By the integral test, the series $\sum_{\ell=1}^\infty e^{-c_3\epsilon^2\ell^{2r}}$ converges, and the result follows from the Borel-Cantelli lemma. □

Lemma 3.4.9 ensures the concentration of the generalized trace entries of the pseudo-resolvent around its mean, assuming a Gaussian design. However, it is worth noting that this concentration phenomenon may not always hinge on Assumption 3. An alternative approach could involve invoking a universality result, such as [BH23, Lemma 6.11], which asserts that certain functionals of the resolvent of a class of random matrices remain unaffected by the distribution of the input. This notion is known as *universality.* This class of random matrices is typically characterized by low-order moments of the entry distribution. For instance, it could encompass all linearizations sharing a fixed mean and covariance structure. If we can identify a linearization within this class featuring Gaussian entries, then according to Corollary 3.4.3, we can conclude that the solution to the MDE serves as a deterministic equivalent for the pseudo-resolvent of linearizations within this class.

3.4.5 Convergence

Finally, we combine the results from Corollary 3.4.3 and Lemma 3.4.9 to establish the convergence of the pseudo-resolvent to the unique solution to (3.2).

Theorem 3.4.2. *Let $z \in \mathbb{H}$, $M \in \mathscr{M}$ be the unique solution to* (3.2)*, $\tilde{\mathcal{S}}$ be defined in* (3.4)*, $\Delta(L,\tau)$ be defined in* (3.18) *and assume that assumptions 1 to 3 hold. Suppose that the mapping $g \in (\mathbb{R}^\gamma, \|\cdot\|_2) \mapsto \mathcal{S}((L(g)-z\Lambda-i\tau I_\ell)^{-1}) \in (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_2)$ is λ-Lipschitz and $\lim_{\ell\to\infty}\sqrt{\ell}\lambda = \lim_{\ell\to\infty}\|\tilde{\mathcal{S}}\| = \lim_{\ell\to\infty}\|\Delta(L,\tau)\| = 0$ for every $\tau \in \mathbb{R}_{>0}$ small enough. Furthermore, suppose that the mapping $g \in (\mathbb{R}^\gamma, \|\cdot\|_2) \mapsto (L(g)-z\Lambda)^{-1} \in (\mathbb{C}^{\ell\times\ell}, \|\cdot\|_F)$ is $c\ell^{-r}$-Lipschitz for some $c, r \in \mathbb{R}_{>0}$. Then, $\operatorname{tr}(U(L-z\Lambda)^{-1}-M(z)) \to 0$ almost surely as $\ell\to\infty$ for every sequence of matrices $U \in \mathbb{C}^{\ell\times\ell}$ with $\|U\|_* \leq 1$.*

Proof. The result follows immediately from Corollary 3.4.3, as well as lemmas 2.2.8 and 3.4.9. □

By decomposing the analysis as shown in (3.15), we were able to segregate the concentration step from the stability analysis. This allows us to work with deterministic objects throughout sections 3.4.1 to 3.4.3. This approach conferred a notable advantage, enabling us to exploit norm bounds and streamline our analysis. This methodology proved adequate for deriving several results, including global anisotropic laws. However, for a more fine-grained examination, such as studying the behavior of eigenvalues at the edge of the spectrum, we would need to extend and adapt the stability argument to directly involve the pseudo-resolvent. We posit that the arguments presented in the stability section could be modified by considering convergence in generalized trace entries instead of operator norm. A key distinction would arise from the existence of correlations between the perturbation matrix $D^{(\tau)}$ and $\mathcal{F}^{(\tau)}((L - z\Lambda - i\tau I_\ell)^{-1})$, necessitating a refined approach to control the error term. This remains a topic for future research.

Example 3.4.1. In Example 2.3.1, we derive an expression for the superoperator $\mathcal{S}$ for GOE matrices. It is straightforward to verify that both assumption 1 and assumption 3 hold in this context. Furthermore, considering two Wigner matrices $W_1 = (2n)^{-\frac{1}{2}}(Z_1 + Z_1^T)$ and $W_2 = (2n)^{-\frac{1}{2}}(Z_2 + Z_2^T)$, as described in Example 2.3.1, we observe that $\|\mathcal{S}((W_1 - zI_n)^{-1} - (W_2 - zI_n)^{-1})\| \leq n^{-1/2}(\Im[z])^{-2}\|W_1 - W_2\|_F \leq \sqrt{2}n^{-1}(\Im[z])^{-2}\|Z_1 - Z_2\|_F$. Hence, by Corollary 3.4.3, the perturbation matrix vanishes in norm as $n \to \infty$. Leveraging the stability result from Theorem 3.4.1 alongside the concentration result from Lemma 3.4.9, we can conclude that the resolvent of Wigner matrices converges to the solution of the MDE. In doing so, we recover *Wigner's semicircle law* in the particular case of GOE matrices.

4

Random Features Ridge Regression

In Chapter 3, we developed a rigorous theoretical framework for the matrix Dyson equation involving correlated linearizations, designed for application to machine learning problems. In this chapter, we will leverage this theory to analyze the empirical test error of random features ridge regression. Specifically, we will demonstrate that the empirical test error of random features ridge regression concentrates around a deterministic quantity in the proportional regime, and we will characterize this deterministic quantity using the matrix Dyson equation. Furthermore, we will utilize the deterministic equivalent to derive insightful conclusions about the performance of random features ridge regression. Our analysis will be complemented by numerical experiments, a discussion of related work, and implications of our results. Importantly, we will substantiate our findings using the theory developed in the previous chapter, showcasing its applicability in addressing machine learning problems. This approach will provide valuable insights into how theoretical frameworks can be effectively applied to analyze real-world machine learning scenarios.

Consider a supervised training problem with a labeled dataset $\mathscr{D} = \{(x_j, y_j)\}_{j=1}^{n_{\text{train}}}$ with

$x_j \in \mathbb{R}^{n_0}$ and $y_j \in \mathbb{R}$ for every $j \in \{1, 2, \ldots, n_{\text{train}}\}$. For conciseness, let $X \in \mathbb{R}^{n_{n_{\text{train}}} \times n_0}$ be the matrix with jth rows corresponding to x_j^T and y be the vectors of labels. We wish to learn a relation between the inputs x_j and the outputs y_j by fitting the parametric function $x \mapsto n^{-\frac{1}{2}}\sigma(x^T W)w$ for some random matrix $W \in \mathbb{R}^{n_0 \times d}$, a λ_σ-Lipschitz activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ applied entrywise and some weights $w \in \mathbb{R}^d$. Following the setup of Louart, Liao, and Couillet, we will assume that $W = \varphi(Z)$ for some $Z \in \mathbb{R}^{n_0 \times d}$ with independent standard normal entries and $\varphi : \mathbb{R} \mapsto \mathbb{R}$ a λ_φ-Lipschitz function applied entrywise [LLC18]. The Lipschitz constants λ_σ and λ_φ should be independent of the dimension of the problem in the sense that, as $n \to \infty$ with $n_{\text{train}} \propto n_0 \propto d \propto n$, $\limsup_{n\to\infty}(\lambda_\varphi \vee \lambda_\sigma) < \infty$. As mentioned in the introduction, this is corresponds to the random features model of [RR07]. This model can be viewed as a two-layer neural network, where the first layer is frozen at random initialization, and only the second layer is trained.

In order to find suitable weights w, we minimize the ℓ_2-regularized norm squared loss

$$\min_{w\in\mathbb{R}^d} \|y - Aw\|^2 + \delta\|w\|^2 \tag{4.1}$$

where $A = n^{-\frac{1}{2}}\sigma(XW) \in \mathbb{R}^{n_{\text{train}} \times d}$ denotes the *random features matrix* and $\delta \in \mathbb{R}_{>0}$ is the *ridge parameter*. In other words, we are fitting a random features model using ridge regression. The minimization problem in (4.1) is strongly convex and admits the closed form solution $w_{\text{ridge}} = A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}y$ which is called the *ridge estimator*.

4.1 Empirical Test Error

Suppose that we computed the ridge estimator w_{ridge} which solves (4.1). To evaluate its performance, we can compute the empirical test error, or out-of-sample error, on a separate labeled dataset $\hat{\mathscr{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{n_{\text{test}}}$ using the squared norm of the residuals

$$E_{\text{test}} := \|\hat{y} - \hat{A}w_{\text{ridge}}\|^2 = \|\hat{y} - \hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}y\|^2 \tag{4.2}$$

with $\hat{A} = n^{-\frac{1}{2}}\sigma(\hat{X}W) \in \mathbb{R}^{n_{\text{test}} \times d}$. This measures the performance of the model $x \mapsto n^{-\frac{1}{2}}\sigma(x^T W)w_{\text{ridge}}$ on $\hat{\mathscr{D}}$. If $\hat{\mathscr{D}} = \mathscr{D}$, then (4.2) corresponds to the training error.

Since E_{test} is a scalar observation, we expect that it will concentrate around a deterministic quantity depending on the first and second moments of A and $\hat{A}$. Consequently, denote

$$\mathbb{E}[(a_1^T, \hat{a}_1^T)^T(a_1^T, \hat{a}_1^T)] = \begin{bmatrix} K_{AA^T} & K_{A\hat{A}^T} \\ K_{\hat{A}A^T} & K_{\hat{A}\hat{A}^T} \end{bmatrix}$$

where $\{(a_j^T, \hat{a}_j^T)^T\}_{j=1}^d$ represent the i.i.d. columns of A and $\hat{A}$. Indeed, K_{AA^T}, $K_{A\hat{A}^T}$, $K_{\hat{A}A^T}$, and $K_{\hat{A}\hat{A}^T}$ encode the covariance between the entries of A and $\hat{A}$. Our main result verifies [LLC18, Conjecture 1] under an additional boundedness assumption.

Theorem 4.1.1. *Assume that $\mathbb{E}A = \mathbb{E}\hat{A} = 0$. Furthermore, suppose that $n_{\text{train}}, d, n_{\text{test}}, n_0 \propto n$ such that λ_σ, λ_φ $\|X\|$, $\|\hat{X}\|$, $\|y\|$, $\|\hat{y}\|$, $\mathbb{E}[\|A\|^4]$ and $\mathbb{E}[\|\hat{A}\|^4]$ remain bounded as $n \to \infty$. Let α be the unique non-positive real number satisfying*

$$\alpha = -(1 + \operatorname{tr}(K_{AA^T}(\delta I_{n_{\text{train}}} - d\alpha K_{AA^T})^{-1}))^{-1} \in \mathbb{R}_{\leq 0}$$

and denote $M = (\delta I_{n_{\text{train}}} - d\alpha K_{AA^T})^{-1}$ as well as

$$\beta = \frac{\alpha^2 \operatorname{tr}(K_{\hat{A}\hat{A}^T} + d\alpha K_{\hat{A}A^T} M(I_{n_{\text{train}}} + \delta M)K_{A\hat{A}^T})}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M K_{AA^T}^{\frac{1}{2}}\|_F^2} \in \mathbb{R}_{\geq 0}.$$

Then, $d\beta\|K_{AA^T}^{\frac{1}{2}} My\|^2 + \|d\alpha K_{\hat{A}A^T} My + \hat{y}\|^2 - E_{\text{test}} \to 0$ almost surely as $n \to \infty$.

In the rest of this section, we will discuss some aspects pertaining to the assumptions and implications of Theorem 4.1.1. This will be followed by a discussion of related work in Section 4.2, and a proof of Theorem 4.1.1 in Section 4.3.

4.1.1 Discussion

Let us briefly discuss some aspects regarding Theorem 4.1.1.

Boundedness Assumptions and Extension to Deep Random Features

The conditions $\limsup_{n\to\infty} \mathbb{E}[\|A\|^4] < \infty$ and $\limsup_{n\to\infty} \mathbb{E}[\|\hat{A}\|^4] < \infty$ are satisfied when the data matrices exhibit approximate orthogonality, as discussed in [FW20; WWF24]. Theorem 4.1.1 extends naturally to deep random features models, which consider compositions of random feature layers

$$x \in \mathbb{R}^{n_0} \mapsto n_k^{-\frac{1}{2}}\sigma\left(n_{k-1}^{-\frac{1}{2}}\sigma\left(\cdots n^{-\frac{1}{2}}\sigma\left(n_1^{-\frac{1}{2}}\sigma(x^TW_1)W_2\right)\right)W_k\right) w \in \mathbb{R}^{n_k}.$$

Here, $k \in \mathbb{N}$ corresponds to the number of layers and $\{W_j \in \mathbb{R}^{n_{j-1}\times n_k}\}_{j=1}^k$ is a collection of random matrices with independent standard normal entries. Under the assumptions of [FW20], it follows from [FW20, Lemma D.4] and the equivalent characterizations of sub-exponential random variables [Ver18, Proposition 2.7.1] that the norm of the conjugate kernel matrices AA^T and $\hat{A}\hat{A}^T$ have bounded fourth moments. This allows for a direct extension of Theorem 4.1.1 to deep random features.

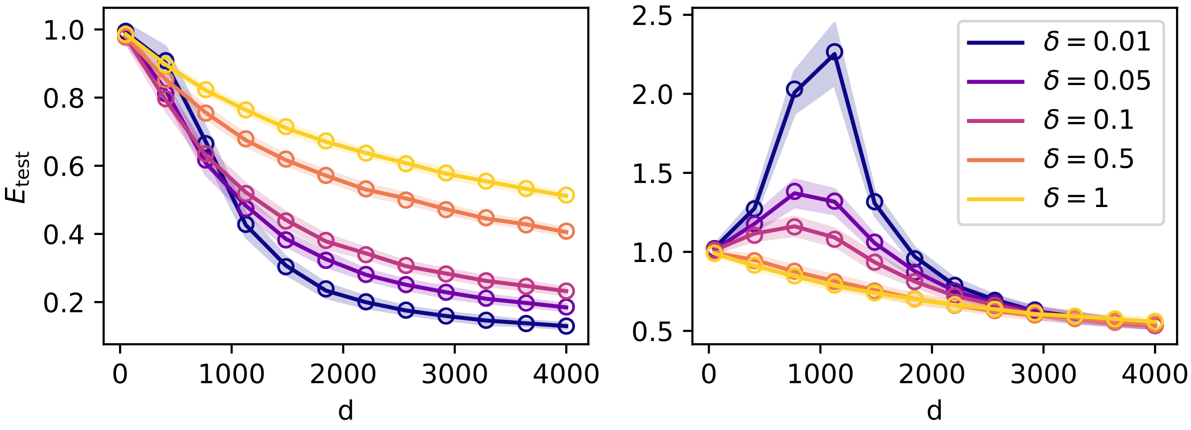

Figure 4.1: E_{test} vs the deterministic approximation given in Theorem 4.1.1 for various odd activation functions with different sizes of hidden layers d and ridge parameter δ. The data matrices, as well as the response variables, are sampled from a synthetic regression dataset, $n_{\text{train}} = n_{\text{test}} = n_0 = 1000$. Left: Error function activation ($\sigma(x) = \text{erf}(x)$); Right: Sign activation ($\sigma(x) = \text{sign}(x)$).

The boundedness conditions are also met with concentrated random vectors, as outlined in [LCM21, Assumption 2]. Notably, these assumptions include the common case of i.i.d. standard normal entries in independent data matrices, a widely studied scenario.

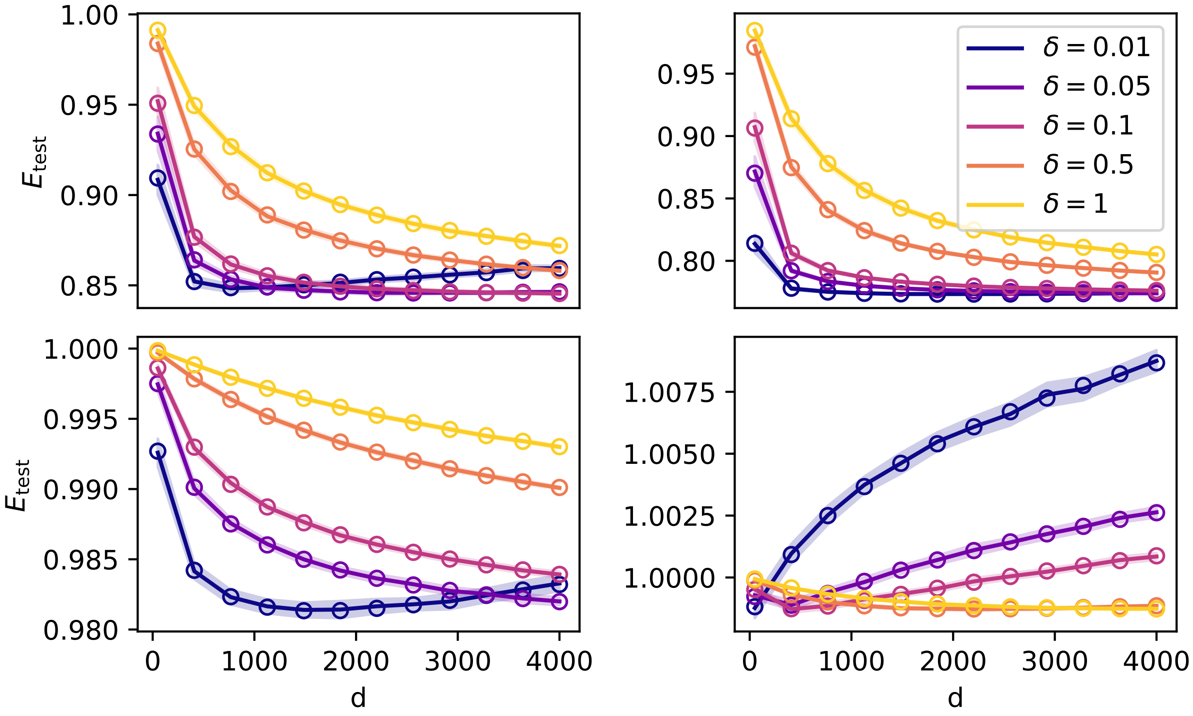

Figure 4.2: E_{test} vs the deterministic approximation given in Theorem 4.1.1 for various flattened image classification datasets with different sizes of hidden layers d and ridge parameter δ. Sine activation ($\sigma = \sin$), $n_{\text{train}} = 1500$, $n_{\text{test}} = 1000$. Upper left: MNIST [Den12]; Upper right: Fashion-MNIST [XRV17]; Lower left: CIFAR-10 [Kri09]; Lower right: CIFAR-100 [Kri09].

Bounded Denominator

While not obvious at first, we show in Lemma 4.3.9 that $1 - \|\sqrt{d}\alpha K_{AA^T}^{1/2} M K_{AA^T}^{1/2}\|_F^2$ is positive and bounded away from 0 as $n \to \infty$ in the setting of Theorem 4.1.1. This implies that β, and therefore E_{test}, is well-behaved in the proportional limit.

Data Assumptions and Real-World Relevance

Our assumptions concerning the norms of A and $\hat{A}$ implicitly impose constraints on the data matrices X and $\hat{X}$. However, conditioning on these matrices allows us to establish asymptotic equivalence without requiring restrictive distributional assumptions. Consequently, our results extend applicability to a broad spectrum of data matrices, offering a more accurate model for real-world datasets. For instance, as illustrated in Figure 4.2, we observe a striking alignment between the empirical test error E_{test} and the deterministic approximation provided by Theorem 4.1.1 across various dimensions and ridge parameters when the data is sourced from real-world flattened image classification datasets such as MNIST [Den12], Fashion-MNIST [XRV17], CIFAR-10 [Kri09], and CIFAR-100 [Kri09]. Notably, the agreement between empirical simulations and the theoretical prediction of Theorem 4.1.1 holds even for datasets with anisotropic features.

This adaptability also permits the analysis of random features ridge regression with test samples drawn from a distribution distinct from that of the training samples. Such flexibility opens up promising avenues for future research, particularly in the realm of privacy, where test samples may be deliberately chosen in an adversarial manner.

Numerical Considerations

Even though Theorem 4.1.1 is an asymptotic result, figures 4.1 and 4.2 demonstrates a close match between the empirical test error E_{test} and the deterministic approximation provided by Theorem 4.1.1 for various activation functions σ and datasets. Notably, the approximation remains accurate for realistic dimensions and even for the non-Lipschitz continuous sign function. Theorem 4.1.1 suggests that computing the asymptotic deterministic equivalent for E_{test} can be reduced to solving a scalar fixed-point equation. As shown in Lemma 4.3.11, the iterates $\{\alpha_k\}_{k\in\mathbb{N}_0}$ obtained by iterating $\alpha_{k+1} = -(1+\operatorname{tr}(K_{AA^T}(\delta I_{n_{\text{train}}} - \alpha_k d K_{AA^T})^{-1}))^{-1}$ for every $k \in \mathbb{N}$ with arbitrary $\alpha_0 \in \mathbb{R}_{\leq 0}$ converge to α as $k \to \infty$. Using the spectral decomposition $K_{AA^T} = U \operatorname{diag}\{\lambda_j(K_{AA^T})\}_{j=1}^{n_{\text{train}}} U^T$ for some orthonormal matrix U, we can rewrite the iteration as

$$\alpha_{k+1} = -\left(1 + \sum_{j=1}^{n_{\text{train}}} \frac{\lambda_j(K_{AA^T})}{\delta - d\alpha_k \lambda_j(K_{AA^T})}\right)^{-1}.$$

Hence, instead of performing a matrix inversion at each iteration, we compute one spectral decomposition and use the above formula to update α_k for every $k \in \mathbb{N}$. This allows us to efficiently compute the deterministic equivalent of E_{test} for various activation functions and dimensions. Moreover, when φ is the identity, the kernel matrices K_{AA^T}, $K_{A\hat{A}^T}$, and $K_{\hat{A}\hat{A}^T}$ can be efficiently computed using [LLC18, Table 1].

4.1.2 Implications

We now discuss some implications of Theorem 4.1.1.

Gaussian Equivalence

Theorem 4.1.1 establishes a Gaussian equivalence principle, indicating that every random features model trained with ridge regression, as described in the statement of Theorem 4.1.1, performs equivalently to a surrogate Gaussian model with a matching covariance structure. However, it is important to note, as mentioned in [LLC18], that the distribution of the input data can impact the performance of the random features model. This influence stems from the fact that, although there is Gaussian equivalence at the level of random feature matrices, the distribution of the input may influence the covariance matrices K_{AA^T}, $K_{A\hat{A}^T}$, $K_{\hat{A}A^T}$, and $K_{\hat{A}\hat{A}^T}$, which are directly linked to the performance of the random features model.

Implicit Regularization and Relation to Kernel Regression

Theorem 4.1.1 demonstrates the concentration of the empirical test error of random features ridge regression around the deterministic quantity $d\beta\|K_{AA^T}^{1/2}My\|^2 + \|d\alpha K_{\hat{A}A^T}My + \hat{y}\|^2$ as the dimensions increases in the proportional regime. The second term, $\|d\alpha K_{\hat{A}A^T}My + \hat{y}\|^2 = \|\hat{y} - dK_{\hat{A}A^T}(dK_{AA^T} + (-\delta/\alpha)I_{n_{\text{train}}})^{-1}y\|^2$, is equivalent to the squared norm of the empirical test error for kernel ridge regression with ridge parameter $-\delta/\alpha \in \mathbb{R}_{>0}$ and the conjugate kernel $K(x_1, x_2) = \frac{d}{n}\mathbb{E}_{z\sim\mathcal{N}(0,I_{n_0})}[\sigma(x_1^T\varphi(z))\sigma(x_2^T\varphi(z))]$. We will show in Section 4.3 that $-1 \leq \alpha < 0$, which implies that $\delta < -\delta/\alpha$. This reveals that the randomness in the random features matrix acts as a form of regularization, similar to increasing the ridge parameter in kernel ridge regression. This is reminiscent of [Jac+20] and the generalization in [Cho22], and is related to the implicit regularization of the random features model. In fact, the proof of Theorem 4.1.1 recovers both of those results.

However, the additional term $d\beta\|K_{AA^T}^{1/2}My\|^2$ represents the variance in the empirical test error due to the random weights. Consequently, despite its computational benefits for kernel approximation, Theorem 3.3.1 indicates that in the proportional regime, the random features model may still underperform kernel ridge regression if $d\beta\|K_{AA^T}^{1/2}My\|^2$ is significantly positive.

4.2 Related Work

This section connects our results to the existing body of research and discusses relevant prior work.

4.2.1 Conjugate Kernel

The *conjugate kernel*, defined as the Gram matrix of features produced by the final layer of a network [FW20], is central to the analysis of random features models. This connection stems from the fact that the ouput of a neural network is linear in those derived features. Thus, the conjugate kernel characterizes the training and test error of this linear model. In the case of shallow random features models, the conjugate kernel is equivalent to the Gram matrix of the random features themselves.

Numerous studies have employed random matrix theory to investigate the conjugate kernel in the proportional regime. Works such as [PW17; BP21] leverage the moment method to establish deterministic equivalents for random features models with isotropic data and weight matrices. Piccolo and Schröder extend these results to include an additive bias [PS21]. To address more realistic data distributions, Fan and Wang study the case of nearly orthogonal data. They introduce a notion of orthogonality that can propagate through network layers, providing control over key quantities related to the conjugate kernel [FW20]. Notably, their settings encompass the case of isotropic data, with the difference that the author conditioned on the data. The conjugate kernel is also studied in [Cho22; LLC18] utilizing concentration of measure and leave-one-out techniques.

Beyond bulk spectrum analysis, Benigni and Péché investigate outlier eigenvalues of the conjugate kernel using random rectangular matrices with i.i.d. centered entries for both data and weights [BP22]. They demonstrate that uninformative spikes can arise in the conjugate kernel when the activation function lacks odd symmetry. Building on this, [WWF24] proposes a spiked conjugate kernel model with low-rank informative structure in the data.

Analyzing deep networks under a near-orthogonality assumptions similar to that of [FW20], they study outlier propagation in the conjugate kernel. This work is motivated in part by results from [Ba+22; Cui+24], where a single gradient step in a neural network at random initialization is shown to be equivalent to a spiked random features model. The structure in random features model turns out to have important implication, notably in understanding the feature learning mechanisms.

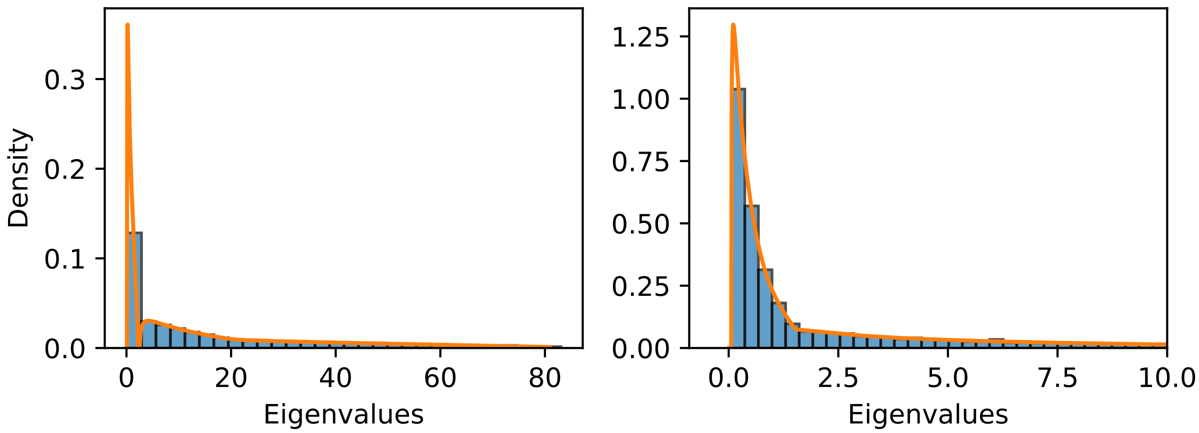

Figure 4.3: Spectrum of the conjugate kernel matrix AA^T. Empirical spectrum (blue) compared to the theoretical density obtained from Theorem 4.1.1 (orange). Left: Sample covariance matrix with $\sigma = \mathrm{Id}$, $n_{\text{train}} = 1000$, $n_0 = 1000$, and $d = 1500$. X is a diagonal matrix with entries drawn uniformly from $\{1, 3, 5\}$. Right: Conjugate kernel matrix with $\sigma = \mathrm{erf}$, $n_{\text{train}} = 1500$, $n_0 = 1000$, and $d = 3000$. The matrix $X = ZC$ for some matrix $Z \in \mathbb{R}^{n_{\text{train}} \times n_0}$ with i.i.d. standard Gaussian entries and $C \in \mathbb{R}^{n_0 \times n_0}$ a diagonal matrix with entries drawn uniformly from $\{1, 3\}$.

The proof of our main result recovers a deterministic equivalent for the conjugate kernel, consistent with the findings of [Cho22; LLC18], employing a significantly different approach. In particular, replacing the ridge parameter by a spectral parameter in Theorem 4.1.1 and using the Stieltjes inversion lemma, we can recover the density of the empirical spectral distribution of the conjugate kernel. This is illustrated in Figure 4.3. As a particular instance of the conjugate kernel, we recover the Marchenko-Pastur law for Wishart matrices, as discussed in examples 2.3.2 and 2.3.3 and illustrated in Figure 4.3.

4.2.2 Gaussian Equivalence

As a consequence of Theorem 4.1.1, we discussed in Section 4.1.2 that random features ridge regression follows a Gaussian equivalence principle. This means that in terms of training and empirical test error, it behaves equivalently to a surrogate linear Gaussian model with matching covariance. This phenomenon was initially established for random features ridge regression under the linear regime in the sense of test error [MM22]. Subsequent work proved its extension to broader loss functions and regularization, initially via non-rigorous replica methods [Ger+20] and later through rigorous analysis [Gol+22; HL23]. Gaussian equivalence has also been demonstrated for deep random features [Sch+23] and for random features ridge regression beyond the linear scaling regime [HLM24].

Our work extends this literature by establishing Gaussian equivalence in terms of the empirical test error of random features ridge regression. This contributes to the generalization of Gaussian equivalence under broader distributional assumptions, aligning with the direction explored by Schröder et al. [Sch+24].

4.2.3 Training and Test Error of Random Features

The random features model provides valuable insights into the behavior of more complex machine learning models and serves as a useful benchmark. Within the context of high-dimensional settings, random features ridge regression is a generalization of traditional ridge regression [DW18; Dic16; WX20; MG21]. With non-linear activation functions, previous studies have extensively examined the test error of random features ridge regression in the proportional regime both using non-rigorous replica methods [Ger+20] and rigorous analyses [MM22; ALP22; AP20b]. This line research characterizes the training and test error of random features ridge regression in order to offer insights into the impacts of various model choices, such as overparameterization.

Research has expanded to address anisotropic data [Has+22; MP22; MMM22; Sch+24] and covariate shift scenarios [TAP21]. A key motivation lies in overcoming the *kernel lower bound* of isotropic data in the proportional regime, where random features models only learn linear label components. By introducing anisotropy, the aim is to *feature learning*—the ability of machine learning models to extract meaningful representations from data. This enhances the expressiveness of random features models, enabling them to better capture the statistical properties of trained neural networks. Breaking the kernel lower bound can also be achieved using a random features model with a hidden layer size that scales polynomially with input size. Studies have investigated random features models beyond the linear scaling regime, revealing significant transitions in the degree of label learning as a function of the polynomial scaling exponent [HLM24; MMM22]. Additionally, some studies investigated training, test, and cross-validation errors in the highly overparameterized regime with nearly orthogonal data assumptions [WZ23].

It is worth noting that investigations into the test error of random features extend beyond ridge regression. Studies have explored generic convex losses [Gol+22] and alternative penalty terms [Lou+22; BPH23]. A recent trend focuses on *deep random features*, a multilayer generalization of random features. Empirical findings indicate that trained neural network outputs can be modeled by a deep random features model, with each layer's covariance corresponding to that of the neural network [Gut+23]. Research has delved into deep structured linear networks [ZP23] and deep non-linear networks [BPH23; Sch+23; Sch+24].

Our study diverges from these works by focusing on the empirical test error of random features ridge regression, without assuming specific data models or distributions beyond some boundedness conditions. Our main result regarding the test error of random features ridge regression is most similar to the work of Louart, Liao, and Couillet, who established an asymptotically exact expression for the training error of random features ridge regression [LLC18]. The authors conjectured that Theorem 4.1.1 holds without the additional conditions imposing bounded fourth moments for the norm of the random features matrices. [LCM21] resolves this conjecture in the special case of random Fourier features. Both [LCM21; LLC18] employ leave-one-out techniques and concentration of measure arguments in their approaches. Although we also utilize a leave-one-out argument to establish universality, our overall approach differs fundamentally, providing flexibility for addressing more complex scenarios where leave-one-out approaches may not be as straightforward to apply.

4.3 Proof of Theorem 4.1.1

In this section, we utilize the framework developed in Chapter 3 to prove Theorem 4.1.1. Before delving into the argument, note that we may expand the test error in (4.2) as

$$E_{\text{test}} = \|\hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}y\|^2 - 2\hat{y}^T\hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}y + \|\hat{y}\|^2.$$

Each term in the above equations is a bilinear form, aligning well with the framework of deterministic equivalence. In order to apply our framework, we have to find a linearization such that a matrix of interest is contained in one of the block of the inverse. To this end, let $\ell = n_{\text{train}} + d + 2n_{\text{test}}$ and consider the linearization

$$L = \begin{bmatrix} \delta I_{n_{\text{train}}} & A & 0_{n_{\text{train}}\times n_{\text{test}}} & 0_{n_{\text{train}}\times n_{\text{test}}} \\ A^T & -I_{d\times d} & 0_{d\times n_{\text{test}}} & \hat{A}^T \\ 0_{n_{\text{test}}\times n_{\text{train}}} & 0_{n_{\text{test}}\times d} & 0_{n_{\text{test}}\times n_{\text{test}}} & -I_{n_{\text{test}}} \\ 0_{n_{\text{test}}\times n_{\text{train}}} & \hat{A} & -I_{n_{\text{test}}} & 0_{n_{\text{test}}\times n_{\text{test}}} \end{bmatrix} \in \mathbb{R}^{\ell\times\ell}. \tag{4.3}$$

Taking $\Lambda := \operatorname{diag}\{I_{n_{\text{train}}+d}, 0_{2n_{\text{test}}\times 2n_{\text{test}}}\}$, we use Lemma 2.2.11 to express the pseudo-resolvent $(L - z\Lambda)^{-1}$ block-wise as

$$(L - z\Lambda)^{-1} = \begin{bmatrix} R & (1+z)^{-1}RA & (1+z)^{-1}RA\hat{A}^T & 0 \\ (1+z)^{-1}A^TR & \bar{R} & \bar{R}\hat{A}^T & 0 \\ (1+z)^{-1}\hat{A}A^TR & \hat{A}\bar{R} & \hat{A}\bar{R}\hat{A}^T & -I_{n_{\text{test}}} \\ 0 & 0 & -I_{n_{\text{test}}} & 0 \end{bmatrix}.$$

Here $R := ((1+z)^{-1}AA^T + (\delta - z)I_{n_{\text{train}}})^{-1}$ represents a resolvent and $\bar{R} := -((1+z)I_d + (\delta - z)^{-1}A^TA)^{-1}$ is a co-resolvent. Indeed, $\lim_{z\to 0}(L - z\Lambda)^{-1}_{3,1} = \hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}$ holds one of the relevant expression for which we want to find a deterministic equivalent. Therefore, it suffices to find a deterministic equivalent for the pseudo-resolvent $(L - z\Lambda)^{-1}$ and take the spectral parameter to zero in order to recover a deterministic equivalent for $\hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}$.

The linearization in (4.3) yields the superoperator

$$\mathcal{S} : M : \mathbb{C}^{\ell\times\ell} \mapsto \begin{bmatrix} \operatorname{tr}(M_{2,2})K_{AA^T} & 0 & 0 & \operatorname{tr}(M_{2,2})K_{A\hat{A}^T} \\ 0 & \rho(M)I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \operatorname{tr}(M_{2,2})K_{\hat{A}A^T} & 0 & 0 & \operatorname{tr}(M_{2,2})K_{\hat{A}\hat{A}^T} \end{bmatrix} \in \mathbb{C}^{\ell\times\ell}$$

where $\rho(M) := \operatorname{tr}(K_{AA^T}M_{1,1} + K_{A\hat{A}^T}M_{4,1} + K_{\hat{A}A^T}M_{1,4} + K_{\hat{A}\hat{A}^T}M_{4,4})$. Then, $\mathcal{S}(M) = \mathbb{E}[(L - \mathbb{E}L)M(L - \mathbb{E}L)] - \tilde{\mathcal{S}}(M)$ holds with

$$\tilde{\mathcal{S}}(M) := \mathbb{E}\begin{bmatrix} 0 & K_{AA^T}M_{2,1}^T + K_{A\hat{A}^T}M_{2,4}^T & 0 & 0 \\ M_{1,2}^T K_{AA^T} + M_{4,2}^T K_{\hat{A}A^T} & 0 & 0 & M_{1,2}^T K_{A\hat{A}^T} + M_{4,2}^T K_{\hat{A}\hat{A}^T} \\ 0 & 0 & 0 & 0 \\ 0 & K_{\hat{A}A^T}M_{2,1}^T + K_{\hat{A}\hat{A}^T}M_{2,4}^T & 0 & 0 \end{bmatrix}.$$

By Theorem 3.3.1, there exists a unique solution $M \in \mathscr{M}$ such that $M(z)$ solves (3.2) for every $z \in \mathbb{H}$. Plugging-in the expression for the superoperator above and using block inversion lemma, we find that

$$M(z) = \begin{bmatrix} ((\delta - z)I_{n_{\text{train}}} - \operatorname{tr}(M_{2,2})K_{AA^T})^{-1} & 0 & -\operatorname{tr}(M_{2,2})M_{1,1}K_{A\hat{A}^T} & 0 \\ 0 & d^{-1}\operatorname{tr}(M_{2,2})I_d & 0 & 0 \\ -\operatorname{tr}(M_{2,2})K_{\hat{A}A^T}M_{1,1} & 0 & \begin{matrix}(\operatorname{tr}(M_{2,2}))^2 K_{\hat{A}A^T}M_{1,1}K_{A\hat{A}^T} \\ + \operatorname{tr}(M_{2,2})K_{\hat{A}\hat{A}^T}\end{matrix} & -I_{n_{\text{test}}} \\ 0 & 0 & -I_{n_{\text{test}}} & 0 \end{bmatrix} \tag{4.4}$$

with $M_{2,2} = -(1 + z + \operatorname{tr}(K_{AA^T}M_{1,1}))^{-1}I_d$.

A key observation that greatly simplifies both the theoretical analysis of the MDE and enables us to derive an iterative procedure for computing its solution is the fact that we can treat the upper-left $n_{\text{train}} + d$ block of the MDE as a separate MDE. This insight allows us to effectively break down the problem and focus on a smaller sub-MDE. Let $L^{(\text{sub})}$ denote the upper-left $n_{\text{train}} + d$ block of L, define a new superoperator

$$\mathcal{S}^{(\text{sub})} : X \in \mathbb{C}^{(n_{\text{train}}+d)\times(n_{\text{train}}+d)} \mapsto \begin{bmatrix} \operatorname{tr}(X_{2,2})K_{AA^T} & 0 \\ 0 & \operatorname{tr}(K_{AA^T}X_{1,1}) \end{bmatrix} \in \mathbb{C}^{(n_{\text{train}}+d)\times(n_{\text{train}}+d)}$$

and a new sub-MDE mapping

$$\mathcal{F}^{(\text{sub})} : f \in \mathscr{M}_+^{(\text{sub})} \mapsto (\mathbb{E}L^{(\text{sub})} - \mathcal{S}^{(\text{sub})}(f(\cdot)) - (\cdot)I_{n_{\text{train}}+d})^{-1} \in \mathscr{M}_+^{(\text{sub})}.$$

Here, the set $\mathscr{M}_+^{(\text{sub})} = \text{Hol}(\mathbb{H}, \mathscr{A}^{(\text{sub})})$ and $\mathscr{A}^{(\text{sub})} = \{N \in \mathbb{C}^{(n_{\text{train}}+d)\times(n_{\text{train}}+d)} : \Im[N] \succ 0\}$. Given that the sub-MDE has a spectral parameter spanning its diagonal, the iteration scheme $N_{k+1} = \mathcal{F}^{(\text{sub})}(N_k)$ converges to the unique solution of the sub-MDE $M^{(\text{sub})} = \mathcal{F}^{(\text{sub})}(M^{(\text{sub})})$ for any $N_0 \in \mathscr{M}_+^{(\text{sub})}$, as per Lemma 3.3.2.

We will extensively use the MDE in (4.4) and the sub-MDE to establish Theorem 4.1.1. To apply our theoretical framework, it is necessary to demonstrate that $\|\Delta(L, \tau; z)\|$, as defined in (3.18), vanishes as $n \to \infty$ for every regularization parameter $\tau \in \mathbb{R}_{>0}$. To achieve this, we employ a leave-one-out method. While the ensuing argument involves detailed and intricate calculations, it is tedious and differs heavily from the rest of the argument. Therefore, we will establish that $\|\Delta(L, \tau; z)\|$ is vanishing in Section 4.3.1. Then, we will use the matrix Dyson equation for linearization framework to derive a deterministic equivalent for the matrix $\hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}$ in Section 4.3.2, and subsequently derive a deterministic equivalent for its square in Section 4.3.3.

4.3.1 Universality

Fix $z \in \mathbb{H}$, let $\{a_j\}_{j=1}^d$, $\{\hat{a}_j\}_{j=1}^d$ denote the columns of A and $\hat{A}$ respectively. Suppose that $l_j^T = (a_j^T, 0, 0, \hat{a}_j^T)$ and $L_j = l_j e_{n_{\text{train}}+j}^T + e_{n_{\text{train}}+j} l_j^T$ for every $j \in \{1, 2, \ldots, d\}$, where $\{e_j\}_{j=1}^{\ell}$ is the canonical basis of $\mathbb{R}^{\ell}$. In particular, we may write the linearization in (4.3) as $L = \mathbb{E}L + \sum_{j=1}^d L_j$. For every $j \in \{1, 2, \ldots, d\}$, let $P_j \in \mathbb{R}^{\ell\times\ell}$ be the orthogonal matrix permuting the first and $n_{\text{train}} + j$th entries exclusively and $C_j \in \mathbb{R}^{(\ell-1)\times(\ell-1)}$ be the matrix cycling from position $n_{\text{train}} + j - 1$ to 1. For instance, if $v = (v_k)_{k=1}^{\ell-1}$, then

$$v^T C_j^{-1} = (v_2, v_3, \ldots, v_{j-1}, v_1, v_j, v_{j+1}, \ldots, v_{\ell-1}).$$

We will rely heavily on a Schur complement decomposition of $(P_j L P_j - zI_\ell)^{-1}$. For every $j \in \{1, 2, \ldots, d\}$, let $l_{-j} \in \mathbb{R}^{\ell-1}$ be obtained by removing the $n_{\text{train}} + j$th entry of l_j and $L_{-j} \in \mathbb{R}^{(\ell-1)\times(\ell-1)}$ be obtained by removing the $n_{\text{train}} + j$th columns and $n_{\text{train}} + j$th row from L. Define the scalar $\xi_j := (1 + z + l_{-j}^T(L_{-j} - zI_{\ell-1})^{-1}l_{-j})^{-1}$ and the matrix

$$\Xi_j := C_j(L_{-j} - zI_{\ell-1})^{-1}C_j - \xi_j C_j(L_{-j} - zI_{\ell-1})^{-1}l_j l_j^T(L_{-j} - zI_{\ell-1})^{-1}C_j.$$

We have the following block inversion formula.

Lemma 4.3.1. *For every $j \in \{1, 2, \dots, d\}$ and $z \in \mathbb{H}$,*

$$(P_j L P_j - zI_\ell)^{-1} = \begin{bmatrix} -\xi_j & \xi_j l_{-j}^T(L_{-j} - zI_{\ell-1})^{-1}C_j \\ \xi_j C_j(L_{-j} - zI_{\ell-1})^{-1}l_{-j} & \Xi_j \end{bmatrix}.$$

Proof. The lemma follows directly from the observation

$$P_j L P_j = \begin{bmatrix} -1 & l_{-j}^T C_j^{-1} \\ C_j^{-1} l_{-j} & C_j^{-1} L_{-j} C_j^{-1} \end{bmatrix}$$

and an application of Lemma 2.2.11. □

For every $j \in \{1, 2, \dots, d\}$, let $q_j = l_{-j}^T R_{-j} l_{-j}$ and $R_{-j} := (L_{-j} - zI_\ell)^{-1}$. Concentration of bilinear forms is a central ingredient of many random matrix theory proof. We obtain a concentration result for q_j by adapting [LLC18, Lemma 4].

Lemma 4.3.2. *Under the settings of Theorem 4.1.1, $\lim_{n\to\infty} \mathbb{E}[\max_{1\le j\le d} |q_j - \mathbb{E}q_j|^2] = 0$ for every $z \in \mathbb{H}$.*

Proof. Adapting [LLC18, Lemma 4], there exists some absolute constants $c_1, c_2 \in \mathbb{R}_{>0}$ such that

$$\mathbb{P}\left(|q_j - \mathbb{E}q_j| > t\right) \le c_1 e^{-c_2 n \min\{t, t^2\}}$$

for every $t \in \mathbb{R}_{\ge 0}$. Then, $\mathbb{E}[\max_{1\le j\le d} |q_j - \mathbb{E}q_j|^2] \le n^{-\frac{1}{2}} + \int_{n^{-\frac{1}{2}}}^1 \mathbb{P}(\max_{1\le j\le d} |q_j - \mathbb{E}q_j|^2 > t)\mathrm{d}t + \int_1^\infty \mathbb{P}(\max_{1\le j\le d} |q_j - \mathbb{E}q_j|^2 > t)\mathrm{d}t$. Using a union bound,

$$\int_{n^{-\frac{1}{2}}}^1 \mathbb{P}\left(\max_{1\le j\le d} |q_j - \mathbb{E}q_j|^2 > t\right)\mathrm{d}t \le c_1 d \int_{n^{-\frac{1}{2}}}^1 e^{-c_2 n t}\mathrm{d}t = \frac{c_1 d}{c_2 n}\left(e^{-c_2\sqrt{n}} - e^{-c_2 n}\right)$$

Also,

$$\int_1^\infty \mathbb{P}\left(\max_{1\le j\le d}|q_j - \mathbb{E}q_j|^2 > t\right)\mathrm{d}t \le c_1 d\int_1^\infty e^{-c_2 n\sqrt{t}}\mathrm{d}t = 2c_1 d\int_1^\infty te^{-c_2 nt}\mathrm{d}t$$
$$= \frac{2c_1 d}{c_2 n}e^{-c_2 n}\left(1 + \frac{1}{c_2 n}\right)$$

Taking $n \to \infty$ and using the fact that $d \propto n$ concludes the proof. □

We need one additional tool in order to show universality, which we state here. We omit the proof, as it follows directly from Hölder's inequality.

Lemma 4.3.3. *If* $\limsup_{n\to\infty}\max\{\mathbb{E}[\|A\|^4], \mathbb{E}[\|\hat{A}\|^4]\} < \infty$ *then* $\limsup_{n\to\infty}\mathbb{E}[\|L - \mathbb{E}L\|^4] < \infty$*.*

We are ready to show universality using a leave-one-out approach.

Lemma 4.3.4. *Fix $z \in \mathbb{H}$. Let L be the linearization defined in* (4.3) *and $\Delta(L,\tau)$ be defined as in* (3.18)*. Under the settings of Theorem 4.1.1,* $\lim_{\ell\to\infty}\|\Delta(L,\tau)\| = 0$ *for every* $\tau \in \mathbb{R}_{>0}$*.*

Proof. For simplicity, we will demonstrate that $\lim_{n\to\infty}\|\mathbb{E}[(L - \mathbb{E}L)(L - zI_\ell)^{-1}] + \mathbb{E}[(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}]\| = 0$ for every $z \in \mathbb{H}$, where $\tilde{L}$ is an i.i.d. copy of L. This adjustment streamlines notation without altering any steps in the proof. For every $j \in \{1, 2, \ldots, d\}$,

$$P_j L_j P_j = \begin{bmatrix} 0 & l_{-j}^T C_j^{-1} \\ C_j^{-1} l_{-j} & 0 \end{bmatrix}$$

and, by Lemma 4.3.1,

$$\begin{aligned}
\mathbb{E}\left[(L - \mathbb{E}L)(L - zI_\ell)^{-1}\right] &= \sum_{j=1}^d P_j \mathbb{E}\left[P_j L_j P_j (P_j L P_j - zI_\ell)^{-1}\right] P_j \\
&= \sum_{j=1}^d P_j \mathbb{E}\begin{bmatrix} \xi_j l_{-j}^T R_{-j} l_{-j} & l_{-j}^T R_{-j} C_j - \xi_j l_{-j}^T R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \\ -\xi_j C_j^{-1} l_{-j} & \xi_j C_j^{-1} l_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix} P_j \\
&= \sum_{j=1}^d P_j \mathbb{E}\begin{bmatrix} \xi_j l_{-j}^T R_{-j} l_{-j} & -\xi_j l_{-j}^T R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \\ -\xi_j C_j^{-1} l_{-j} & \xi_j C_j^{-1} l_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix} P_j
\end{aligned}$$

where we recall that $R_{-j} = (L_{-j} - zI_{\ell-1})^{-1}$. On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}\right] \\
&= \sum_{j=1}^{d} P_j \mathbb{E}\left[P_j \tilde{L}_j P_j (P_j L P_j - zI_\ell)^{-1} P_j \tilde{L}_j P_j (P_j L P_j - zI_\ell)^{-1}\right] P_j \\
&= \sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} \xi_j \tilde{l}_{-j}^T R_{-j} l_{-j} & \tilde{l}_{-j}^T R_{-j} C_j - \xi_j \tilde{l}_{-j}^T R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \\ -\xi_j C_j^{-1} \tilde{l}_{-j} & \xi_j C_j^{-1} \tilde{l}_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix}^2 P_j.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[(L - \mathbb{E}L)(L - zI_\ell)^{-1}\right] + \mathbb{E}\left[(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}\right] = \sum_{j=1}^{d} P_j \mathbb{E}[\xi_j \Psi_j] P_j$$

where $q_j = l_{-j}^T R_{-j} l_{-j}$, $\tilde{q}_j = \tilde{l}_{-j}^T R_{-j} \tilde{l}_{-j}$, $r_j = \tilde{l}_{-j}^T R_{-j} l_{-j}$ and

$$\Psi_j = \begin{bmatrix} q_j - \tilde{q}_j + 2\xi_j r_j^2 & r_j \tilde{l}_{-j}^T R_{-j} C_j - 2\xi_j r_j^2 l_{-j}^T R_{-j} C_j + (\tilde{q}_j - q_j) l_{-j}^T R_{-j} C_j \\ -2\xi_j r_j C_j^{-1} \tilde{l}_{-j} - C_j^{-1} l_{-j} & C_j^{-1}(l_{-j} l_{-j}^T - \tilde{l}_{-j} \tilde{l}_{-j}^T) R_{-j} C_j + 2\xi_j r_j C_j^{-1} \tilde{l}_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix}.$$

We will consider the upper blocks of $\sum_{j=1}^{d} P_j \mathbb{E}[\xi_j \Psi_j] P_j$ separately.

First, for the upper-left corner, the sum along with the permutation matrices P_j are simply tiling the diagonal. Furthermore, by Lemma 2.2.5, both $|\xi_j|$ and $\|R_j\|$ are bounded by $(\Im[z])^{-1}$ for every $j \in \{1, 2, \dots, d\}$. Then,

$$\begin{aligned}
\left\| \sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} \xi_j(q_j - \tilde{q}_j) + 2\xi_j^2 r_j^2 & 0 \\ 0 & 0 \end{bmatrix} P_j \right\| &\leq \max_{1 \leq j \leq d} |\mathbb{E}[\xi_j(q_j - \tilde{q}_j) + 2\xi_j^2 r_j^2]| \\
&\leq \frac{1}{\Im[z]} \mathbb{E}[|q - \tilde{q}|] + 2|\mathbb{E}[\xi_1^2 l_{-1}^T R_{-1} K R_{-1} l_{-1}]| \\
&\leq \frac{2}{\Im[z]} \mathbb{E}[|q - \mathbb{E}q|] + \frac{2\mathbb{E}[\|L - \mathbb{E}L\|^2]\|K\|}{(\Im[z])^4}.
\end{aligned}$$

Here, we introduced the correlation matrix $K = \mathbb{E}[l_{-1} l_{-1}^T]$. Using Jensen's inequality and

Cauchy Schwarz,

$$\begin{aligned}\|K\| &\le \|\mathbb{E}[a_1a_1^T]\| + \|\mathbb{E}[\hat{a}_1 a_1^T]\| + \|\mathbb{E}[a_1\hat{a}_1^T]\| + \|\mathbb{E}[\hat{a}_1\hat{a}_1^T]\| \\ &= d^{-1}(\|\mathbb{E}[AA^T]\| + \|\mathbb{E}[\hat{A} A^T]\| + \|\mathbb{E}[A\hat{A}^T]\| + \|\mathbb{E}[\hat{A}\hat{A}^T]\|) \\ &\le d^{-1}(\mathbb{E}[\|A\|^2] + 2\sqrt{\mathbb{E}[\|\hat{A}\|^2]\mathbb{E}[\|A\|^2]} + \mathbb{E}[\|\hat{A}\|^2]) \lesssim d^{-1}. \end{aligned} \tag{4.5}$$

Here, we used the fact that $\mathbb{E}[\|A\|^2]$ and $\mathbb{E}[\|\hat{A}\|^2]$ are bounded by assumption. Additionally, by Lemma 4.3.2, it is clear that $\mathbb{E}[|q - \mathbb{E}q|] \to 0$ as $n \to \infty$.

We now turn our attention to the upper-right $1 \times (\ell - 1)$ corner of $\sum_{j=1}^d P_j \mathbb{E}[\xi_j \Psi_j] P_j$. For every unit vector $x \in \mathbb{C}^\ell$,

$$\begin{aligned}&\left\|\sum_{j=1}^d P_j \mathbb{E}\begin{bmatrix}0 & \xi_j r_j \tilde{l}_{-j}^T R_{-j} C_j - 2\xi_j^2 r_j^2 l_{-j}^T R_{-j} C_j + \xi_j(\tilde{q}_j - q_j) l_{-j}^T R_{-j} C_j \\ 0 & 0\end{bmatrix} P_j x\right\|_2 \\ &= \left\|\mathbb{E}\begin{pmatrix}\left(0 \quad \xi_1 r_1 \tilde{l}_{-1}^T R_{-1} C_1 - 2\xi_1^2 r_1^2 l_{-1}^T R_{-1} C_1 + \xi_1(\tilde{q}_1 - q_1) l_{-1}^T R_{-1} C_1\right) P_1 x \\ \vdots \\ \left(0 \quad \xi_d r_d \tilde{l}_{-d}^T R_{-d} C_d - 2\xi_d^2 r_d^2 l_{-d}^T R_{-d} C_d + \xi_d(\tilde{q}_d - q_d) l_{-d}^T R_{-d} C_d\right) P_d x\end{pmatrix}\right\|_2 \\ &\le \sqrt{\ell} \max_{1 \le j \le d} \|\mathbb{E}[\xi_j r_j \tilde{l}_{-j}^T R_{-j} - 2\xi_j^2 r_j^2 l_{-j}^T R_{-j} + \xi_j(\tilde{q}_j - q_j) l_{-j}^T R_{-j}]\|_2.\end{aligned}$$

On one hand,

$$\mathbb{E}[\xi_j r_j \tilde{l}_{-j}^T R_{-j} - 2\xi_j^2 r_j^2 l_{-j}^T R_{-j}] = \mathbb{E}[\xi_j l_{-j}^T R_{-j} K R_{-j} - 2\xi_j^2 l_{-j}^T R_{-j} K R_{-j} l_{-j} l_{-j}^T R_{-j}]$$

and, since $|\xi_j| \le (\Im[z])^{-1}$,

$$\max_{1 \le j \le d} \|\mathbb{E}[\xi_j r_j \tilde{l}_{-j}^T R_{-j} - 2\xi_j^2 r_j^2 l_{-j}^T R_{-j}]\| \le \frac{\mathbb{E}[\|l_{-1}\|]\|K\|}{(\Im[z])^3} + \frac{2\mathbb{E}[\|l_{-1}\|^3]\|K\|}{(\Im[z])^5}.$$

Furthermore, by Cauchy-Schwarz for complex random variables,

$$\begin{aligned}\|\mathbb{E}[\xi_j(\tilde{q}_j - q_j)l_{-j}^T R_{-j}]\|_2 &= \sup_{\|y\|\leq 1} |\mathbb{E}[\xi_j(\tilde{q}_j - q_j)l_{-j}^T R_{-j}y]| \\ &\leq (\Im[z])^{-1} \sup_{\|y\|\leq 1} \sqrt{\mathbb{E}[|q-\tilde{q}|^2]\mathbb{E}[|l_{-j}^T R_{-j}y|^2]} \\ &= (\Im[z])^{-1} \sup_{\|y\|\leq 1} \sqrt{\mathbb{E}[|q-\tilde{q}|^2]\mathbb{E}[y^* R_{-j}^* K R_{-j} y]} \leq \frac{\sqrt{\mathbb{E}[|q-\tilde{q}|^2]\|K\|}}{(\Im[z])^2}.\end{aligned}$$

Combining everything, we obtain that the upper-right $1\times(\ell-1)$ corner of $\sum_{j=1}^d P_j\mathbb{E}[\xi_j\Psi_j]P_j$ is bounded, in norm, by

$$\frac{\sqrt{\ell}\mathbb{E}[\|l_{-1}\|]\|K\|}{(\Im[z])^3} + \frac{2\sqrt{\ell}\mathbb{E}[\|l_{-1}\|^3]\|K\|}{(\Im[z])^5} + \frac{\sqrt{\ell\mathbb{E}[|q-\tilde{q}|^2]\|K\|}}{(\Im[z])^2}.$$

We conclude that this bound vanishes as n increases using (4.5), $\mathbb{E}[\|l_{-1}\|] \leq \mathbb{E}[\|L - \mathbb{E}L\|]$ as well as lemmas 4.3.2 and 4.3.3.

We consider the two lower blocks together. For notational convenience, let

$$\underline{\Psi}_j = \begin{bmatrix} 0 & 0 \\ -2\xi_j r_j C_j^{-1}\tilde{l}_{-j} - C_j^{-1}l_{-j} & C_j^{-1}(l_{-j}l_{-j}^T - \tilde{l}_{-j}\tilde{l}_{-j}^T)R_{-j}C_j + 2\xi_j r_j C_j^{-1}\tilde{l}_{-j}l_{-j}^T R_{-j}C_j \end{bmatrix}$$

for every $j \in \{1,2,\ldots,d\}$. Since we expect q_j to concentrate around its mean, we write $\xi_j = (1+z+q_j)^{-1} = (1+z+\mathbb{E}q_j)^{-1} + \frac{\mathbb{E}q_j - q_j}{(1+z+\mathbb{E}q_j)}\xi_j$ and

$$\sum_{j=1}^d P_j\mathbb{E}[\xi_j\underline{\Psi}_j]P_j = (1+z+\mathbb{E}q)^{-1}\sum_{j=1}^d P_j\mathbb{E}[\underline{\Psi}_j]P_j - (1+z+\mathbb{E}q)^{-1}\sum_{j=1}^d P_j\mathbb{E}[(q_j - \mathbb{E}q_j)\xi_j\underline{\Psi}_j]P_j.$$

Using independence of R_{-j}, l_{-j} and $\tilde{l}_{-j}$,

$$\mathbb{E}\underline{\Psi}_j = \begin{bmatrix} 0 & 0 \\ -2\xi_j C_j^{-1}KR_{-j}l_{-j} & 2\xi_j C_j^{-1}KR_{-j}l_{-j}l_{-j}^T R_{-j}C_j \end{bmatrix}.$$

Using a similar argument as above,

$$\left\|\sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ -2\xi_j C_j^{-1} K R_{-j} l_{-j} & 0 \end{bmatrix} P_j\right\| \leq \frac{2\sqrt{\ell}\mathbb{E}[\|l_{-1}\|]\|K\|}{(\Im[z])^2} \xrightarrow{n\to\infty} 0.$$

Moreover, further decomposing the lower-right corner,

$$\begin{aligned}
&\sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ 0 & 2\xi_j C_j^{-1} K R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix} P_j \\
&\qquad = (1+z+\mathbb{E}q)^{-1} \sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ 0 & 2C_j^{-1} K R_{-j} K R_{-j} C_j \end{bmatrix} P_j \\
&\qquad - (1+z+\mathbb{E}q)^{-1} \sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ 0 & 2(q_j - \mathbb{E}q_j)\xi_j C_j^{-1} K R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix} P_j
\end{aligned}$$

with $|(1+z+\mathbb{E}q)^{-1}| \leq (\Im[z])^{-1}$,

$$\left\|\sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ 0 & 2C_j^{-1} K R_{-j} K R_{-j} C_j \end{bmatrix} P_j\right\| \leq \frac{2d\|K\|^2}{(\Im[z])^2}$$

and

$$\left\|\sum_{j=1}^{d} P_j \mathbb{E}\begin{bmatrix} 0 & 0 \\ 0 & 2(q_j - \mathbb{E}q_j)\xi_j C_j^{-1} K R_{-j} l_{-j} l_{-j}^T R_{-j} C_j \end{bmatrix} P_j\right\| \leq \frac{2d\|K\|\sqrt{\mathbb{E}[|q-\mathbb{E}q|^2]\mathbb{E}[\|l_{-1}\|^4]}}{(\Im[z])^3}.$$

In particular, $\|(1+z+\mathbb{E}q)^{-1}\sum_{j=1}^{d} P_j \mathbb{E}[\underline{\Psi}_j] P_j\| \xrightarrow{n\to\infty} 0$. It only remains to show that $\|(1+z+\mathbb{E}q)^{-1}\sum_{j=1}^{d} P_j \mathbb{E}[(q_j - \mathbb{E}q_j)\xi_j \underline{\Psi}_j] P_j\|$ vanishes. To this end, we undo the decomposition and notice that

$$\begin{aligned}
\sum_{j=1}^{d} P_j \mathbb{E}[(q_j - \mathbb{E}q_j)\xi_j \underline{\Psi}_j] P_j &= \mathbb{E}\left[(\underline{L} - \mathbb{E}\underline{L})\Omega(L - zI_\ell)^{-1}\right] \\
&\qquad + \mathbb{E}\left[(\tilde{\underline{L}} - \mathbb{E}\underline{L})\Omega(L - zI_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}\right]
\end{aligned}$$

where

$$\underline{L} = \begin{bmatrix} \delta I_{n_{\text{train}}} & A & 0 & 0 \\ 0 & -I_{d\times d} & 0 & 0 \\ 0 & 0 & 0 & -I_{n_{\text{test}}} \\ 0 & \hat{A} & -I_{n_{\text{test}}} & 0 \end{bmatrix}$$

and

$$\Omega = \text{diag}\{0_{n_{\text{train}}\times n_{\text{train}}}, \text{diag}\{q_j - \mathbb{E}q_j\}_{j=1}^{d}, 0_{2n_{\text{test}}\times 2n_{\text{test}}}\}.$$

Using the bound $\|(L - zI_\ell)^{-1}\| \leq (\Im[z])^{-1}$, it follows from Jensen's and Cauchy-Schwarz inequalities that

$$\left\|\mathbb{E}\left[(\underline{L} - \mathbb{E}\underline{L})\Omega(L - zI_\ell)^{-1}\right]\right\| \leq \frac{\sqrt{\mathbb{E}[\|L - \mathbb{E}L\|^2]\mathbb{E}[\max_{1\leq j\leq d}|q_j - \mathbb{E}q_j|^2]}}{\Im[z]}$$

and

$$\left\|\mathbb{E}\left[(\tilde{\underline{L}} - \mathbb{E}\underline{L})\Omega(L - zI_\ell)^{-1}(\tilde{L} - \mathbb{E}L)(L - zI_\ell)^{-1}\right]\right\| \leq \frac{\sqrt{\mathbb{E}[\|L - \mathbb{E}L\|^4]\mathbb{E}[\max_{1\leq j\leq d}|q_j - \mathbb{E}q_j|^2]}}{(\Im[z])^2}.$$

This term gives us the bottleneck conditions on the norm of the matrix $L - \mathbb{E}L$ and the concentration of q around its mean. By lemmas 4.3.2 and 4.3.3, both of the RHS bounds vanish as n diverges to infinity. □

4.3.2 First Deterministic Equivalent

With the computational leave-one-out argument out of the way, we now focus on establishing a deterministic equivalent for the random matrix $\hat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}$. As mentioned above, this will be established by showing that the solution to the MDE given in (4.4) is a deterministic equivalent for the pseudo-resolvent $(L - z\Lambda)^{-1}$ in a neighborhood of $z = 0$. Hence, the key lies in establishing control over $M(z)$ in the proximity of $z = 0$. This control is secured through the insights provided by the following lemma.

Lemma 4.3.5. *Let $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$ and $M \in \mathscr{M}$ be the unique solution to* (4.4)*. Then, $\Re[M_{1,1}(z)] \succ 0$ and $\Re[M_{2,2}(z)] \prec 0$. Additionally, $\|M_{1,1}(z)\| \leq (\delta - \Re[z])^{-1}$ and $\|M_{2,2}(z)\| \leq (1 + \Re[z])^{-1}$.*

Proof. Let N be any $(n_{\text{train}}+d)\times(n_{\text{train}}+d)$ matrix-valued analytic function on $\mathbb{H}$ such that $\Im[N(z)] \succ 0$, $N_{1,2}(z) = N_{2,1}(z) = 0$ for every $z \in \mathbb{H}$. Further assume that $\Re[N_{1,1}(z)] \succeq 0$ and $\Re[N_{2,2}(z)] \preceq 0$ for every $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$. By Lemma 2.2.1,

$$\Re[\mathcal{F}^{(\text{sub})}(N)] = \mathcal{F}^{(\text{sub})}(N) \begin{bmatrix} \substack{(\delta-\Re[z])I_{n_{\text{train}}} \\ -\operatorname{tr}(\Re[N_{2,2}])K_{AA^T}} & 0 \\ 0 & {\scriptstyle -(1+\Re[z]+\operatorname{tr}(K_{AA^T}\Re[N_{1,1}]))} \end{bmatrix} (\mathcal{F}^{(\text{sub})}(N))^* \tag{4.6}$$

where we omit the dependence of N on z. Thus, $\Re[\mathcal{F}^{(\text{sub})}_{1,1}(N)] \succ 0$ and $\Re[\mathcal{F}^{(\text{sub})}_{2,2}(N)] \prec 0$ for every $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$. Additionally, $\mathcal{F}^{(\text{sub})}_{1,2}(N) = \mathcal{F}^{(\text{sub})}_{2,1}(N) = 0$. Since the iterates $N_{k+1} = \mathcal{F}^{(\text{sub})}(N_k)$ converges to the unique solution to the sub-MDE, it must be the case that $\Re[M^{(\text{sub})}_{1,1}(z)] \succ 0$ and $\Re[M^{(\text{sub})}_{2,2}(z)] \prec 0$ for every $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$. In fact, by uniqueness of the solution to (3.2), $\Re[M_{1,1}(z)] \succ 0$ and $\Re[M_{2,2}(z)] \prec 0$ for every $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$. Using the fact that $\mathcal{F}^{(\text{sub})}(M^{(\text{sub})}) = M^{(\text{sub})}$ and (4.6), we get

$$\Re[M_{1,1}] \succeq (\delta - \Re[z])M_{1,1}(M_{1,1})^* \quad \text{and} \quad \Re[M_{2,2}] \preceq -(1+\Re[z])M_{2,2}(M_{2,2})^*$$

for every $z \in \mathbb{H}$ with $|z| \leq 1 \wedge \delta$. Since the spectral norm maintains the Loewner partial ordering, it follows from Lemma 2.2.4 that $\|M_{1,1}\| \geq (\delta - \Re[z])\|M_{1,1}(M_{1,1})^*\| = (\delta - \Re[z])\|M_{1,1}\|^2$ and $\|M_{2,2}\| \geq (1+\Re[z])\|M_{2,2}\|^2$. Rearranging yields the desired result. □

It will be useful later to not only have a deterministic equivalent for $(L - z\Lambda)^{-1}$, but also for $(L^{(\text{sub})} - zI_{n_{\text{train}}})^{-1}$.

Lemma 4.3.6. *Let $z \in \mathbb{H}$ with $|z| < \delta \wedge 1$ and $M \in \mathscr{M}$ be the unique solution to the sub-MDE. Under the settings of Theorem 4.1.1, $\operatorname{tr}(U((L^{(\text{sub})} - zI_{n_{\text{train}}+d})^{-1} - M^{(\text{sub})}(z))) \to 0$ almost surely as $n \to \infty$ for every sequence $U \in \mathbb{C}^{(n_{\text{train}}+d)\times(n_{\text{train}}+d)}$ with $\|U\|_* \leq 1$.*

Proof. We first check that assumptions 1 to 3 are satisfied in this context. Indeed, it follows from Hölder's inequality that $\limsup_{\ell\to\infty} \mathbb{E}\|(L - z\Lambda)^{-1}\|^2 < \infty$. In particular, Assumption 1 is satisfied. Since there is a spectral parameter spanning the entire diagonal in the sub-MDE, it follows from the stability properties of the MDE or [Alt+19, Corollary 3.8] that Assumption 2 is satisfied for the sub-MDE. Because both A and $\hat{A}$ are centered, and their norm have bounded fourth moments, it is clear that $\limsup_{\ell\to\infty}(\|\mathcal{S}\| \vee \|\mathbb{E}L\|) < \infty$. Finally, Assumption 3 is evidently satisfied. Hence, the assumptions are satisfied.

Let $D^{(\mathrm{sub})} = \mathbb{E}[(\mathbb{E}L^{(\mathrm{sub})} - L^{(\mathrm{sub})} - \mathcal{S}^{(\mathrm{sub})}(\mathbb{E}(L^{(\mathrm{sub})} - zI_{n_{\mathrm{train}}+d})^{-1}))(L^{(\mathrm{sub})} - zI_{n_{\mathrm{train}}+d})^{-1}]$ be the perturbation matrix, as defined in (3.17), associated with the linearization $L^{(\mathrm{sub})}$. In particular, $(\mathbb{E}L^{(\mathrm{sub})} - \mathcal{S}(\mathbb{E}(L^{(\mathrm{sub})} - zI_{n_{\mathrm{train}}+d})^{-1}) - zI_{n_{\mathrm{train}}+d})\mathbb{E}(L^{(\mathrm{sub})} - zI_{n_{\mathrm{train}}+d})^{-1} = I_{n_{\mathrm{train}}+d} + D^{(\mathrm{sub})}$. As a consequence of Lemma 4.3.4 with $\hat{A} = 0$, $\|\Delta(L^{(\mathrm{sub})}, \tau; z)\| \to 0$ as $n \to \infty$ for every $\tau \in \mathbb{R}_{>0}$. Hence, to show that the perturbation matrix is vanishing as the dimension increases, we only have to establish that the term involving the Lipschitz constant and the term involving the norm of $\tilde{\mathcal{S}}^{(\mathrm{sub})}$ in Lemma 3.4.8 are asymptotically negligible.

We derive some useful norm bounds. Recall that $R = ((1+z)^{-1}AA^T + (\delta - z)I_{n_{\mathrm{train}}})^{-1}$. For $|z| < 1 \wedge \delta$, $\Re[(1+z)^{-1}] \geq |1+z|^{-2}(1-|z|) \geq (1-|z|)/4 > 0$ and $\Re[\delta - z] \geq \delta - |z|$. Hence, $\Re[R] \geq (\delta - |z|)RR^*$ which implies that $\|R\| \leq (\delta - |z|)^{-1}$. A similar argument applied to $\bar{R} = -((1+z)I_d + (\delta - z)^{-1}A^TA)^{-1}$ gives $\|\bar{R}\| \leq (1-|z|)^{-1}$. Furthermore, we know that $\|RA\|^2 = \|RAA^TR^*\| \leq \|RAA^T\|\|R\|$. By definition, $RAA^T = (1+z)I_{n_{\mathrm{train}}} - (1+z)(\delta - z)R$. Thus, $\|RAA^T\| \leq 2(2+\delta)(\delta - |z|)^{-1}$ and $\|RA\| \leq \sqrt{2(2+\delta)}(\delta - |z|)^{-1}$.

Based on Assumption 3, write $L^{(\mathrm{sub})} \equiv L^{(\mathrm{sub})}(Z) = \mathcal{C}(Z) + \mathbb{E}L^{(\mathrm{sub})}$ for $Z \in \mathbb{R}^{n_0 \times d}$ a matrix of i.i.d. standard normal entries and let λ be the Lipschitz constant associated to the function $Z \in (\mathbb{R}^{n_0 \times d}, \|\cdot\|_F) \mapsto \mathcal{S}^{(\mathrm{sub})}((L^{(\mathrm{sub})}(Z) - zI_{n_{\mathrm{train}}+d})^{-1}) \in (\mathbb{C}^{(n_{\mathrm{train}}+d)\times(n_{\mathrm{train}}+d)}, \|\cdot\|_2)$. As mentioned in Remark 3.4.2, $\lambda \leq (\Im[z])^{-2}\|\mathcal{S}^{(\mathrm{sub})}\|_{F\mapsto 2}\lambda_{\mathcal{C}}$ where $\lambda_{\mathcal{C}}$ is the Lipschitz constant associated with map $\mathcal{C} : Z \in (\mathbb{R}^{n_0\times d}, \|\cdot\|_F) \mapsto \mathcal{C}(Z) \in (\mathbb{R}^{(n_{\mathrm{train}}+d)\times(n_{\mathrm{train}}+d)}, \|\cdot\|_F)$. For every $N \in \mathbb{C}^{(n_{\mathrm{train}}+d)\times(n_{\mathrm{train}}+d)}$, we can use Cauchy-Schwarz inequality to obtain

$$\|\mathcal{S}^{(\mathrm{sub})}(N)\| \leq \|K_{AA^T}\||\operatorname{tr}(N_{2,2})| + |\operatorname{tr}(K_{AA^T}N_{1,1})| \leq (\sqrt{d} + \sqrt{n_{\mathrm{train}}})\|K_{AA^T}\|\|N\|_F.$$

By Jensen's inequality, $\|K_{AA^T}\| = \|d^{-1}\mathbb{E}[AA^T]\| \leq d^{-1}\mathbb{E}\|A\|^2$. In fact, by a similar argument, $\|K_{AA^T}\| \vee \|K_{\hat{A}A^T}\| \vee \|K_{A\hat{A}^T}\| \vee \|K_{\hat{A}\hat{A}^T}\| \lesssim n^{-1}$. Since we assumed that $\mathbb{E}\|A\|^4$ is bounded, and we are working in the proportional limit, $\|\mathcal{S}^{(\mathrm{sub})}\|_{F\mapsto 2} \lesssim n^{-1/2}$. Next, let $Z_1, Z_2 \in \mathbb{R}^{n_0\times d}$ and notice that $\|\mathcal{C}(Z_1) - \mathcal{C}(Z_2)\|_F \leq n^{-1/2}\lambda_\sigma\lambda_\varphi\|X\|\|Z_1 - Z_2\|_F$. Since λ_σ, λ_φ and $\|X\|$ are all bounded by assumption, $\lambda \lesssim (\Im[z])^{-2}n^{-1}$. Finally, for every $N \in \mathbb{C}^{(n_{\mathrm{train}}+d)\times(n_{\mathrm{train}}+d)}$,

$$\tilde{\mathcal{S}}^{(\mathrm{sub})}(N) = \mathbb{E}\begin{bmatrix} 0 & K_{AA^T}N_{2,1}^T + K_{A\hat{A}^T}N_{2,4}^T \\ N_{1,2}^TK_{AA^T} + N_{4,2}^TK_{\hat{A}A^T} & 0 \end{bmatrix}.$$

Since $\|K_{AA^T}\| \vee \|K_{\hat{A}A^T}\| \vee \|K_{A\hat{A}^T}\| \vee \|K_{\hat{A}\hat{A}^T}\| \lesssim n^{-1}$, $\|\tilde{\mathcal{S}}^{(\mathrm{sub})}\| \lesssim n^{-1}$. Combining everything along with concentration of linear functional of the resolvent around their mean, the result follows from Theorem 3.4.2. □

The final prerequisite needed to establish that the solution to (4.4) acts as a deterministic equivalent for $(L - z\Lambda)^{-1}$, where L is defined in (4.3), for every $z \in \mathbb{H}$ within a neighborhood of the origin, is to confirm that Assumption 2 holds. We demonstrate this in the following lemma.

Lemma 4.3.7. *Suppose that $M \in \mathscr{M}$ is the unique solution to* (4.4) *and $M^{(\tau)}$ is the unique solution to the regularized version of the same equation. Then, M and $M^{(\tau)}$ satisfy Assumption 2 for all $z \in \mathbb{H}$ with $|z| < \delta \wedge 1$.*

Proof. Fix $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$ and let $\tau \in \mathbb{R}_{>0}$. Expanding (3.6), we have $((\delta - z - i\tau)I_{n_{\text{train}}} - \mathrm{tr}(M^{(\tau)}_{2,2})K_{AA^T})M^{(\tau)}_{1,4} = \mathrm{tr}(M^{(\tau)}_{2,2})K_{A\hat{A}^T}M^{(\tau)}_{4,4}$, $-i\tau M^{(\tau)}_{3,4} - M^{(\tau)}_{4,4} = 0$ and $-M^{(\tau)}_{3,4} - (\mathrm{tr}(M^{(\tau)}_{2,2})K_{\hat{A}\hat{A}^T} + i\tau I_{n_{\text{train}}})M^{(\tau)}_{4,4} - \mathrm{tr}(M^{(\tau)}_{2,2})K_{\hat{A}A^T}M^{(\tau)}_{1,4} = I_{n_{\text{test}}}$. Using those equations, we may solve for $M^{(\tau)}_{4,4}$ and $M^{(\tau)}_{1,4}$. In particular, we have

$$M^{(\tau)}_{1,4} = \mathrm{tr}(M^{(\tau)}_{2,2})((\delta - z - i\tau)I_{n_{\text{train}}} - \mathrm{tr}(M^{(\tau)}_{2,2})K_{AA^T})^{-1}K_{A\hat{A}^T}M^{(\tau)}_{4,4}$$

and $(I_{n_{\text{test}}} - i\tau\Xi)M^{(\tau)}_{4,4} = i\tau I_{n_{\text{test}}}$ with $\Xi = (\mathrm{tr}(M^{(\tau)}_{2,2})K_{\hat{A}\hat{A}^T} + i\tau I_{n_{\text{train}}}) - (\mathrm{tr}(M^{(\tau)}_{2,2}))^2 K_{\hat{A}A^T}((\delta - z - i\tau)I_{n_{\text{train}}} - \mathrm{tr}(M^{(\tau)}_{2,2})K_{AA^T})^{-1}K_{A\hat{A}^T}$. Since $\|M^{(\tau)}_{2,2}\| \leq (\Im[z])^{-1}$ and $\Im[M^{(\tau)}_{2,2}] \succeq 0$ for every $\tau \in \mathbb{R}_{>0}$, it follows that $\|\Xi\|$ is bounded as $\tau \to 0$. Consequently, for every τ small enough, $\|i\tau\Xi\| < 1$ and $I_{n_{\text{test}}} - i\tau\Xi$ is invertible. In particular, using the fact that $\|K_{AA^T}\| \vee \|K_{\hat{A}A^T}\| \vee \|K_{A\hat{A}^T}\| \vee \|K_{\hat{A}\hat{A}^T}\| \lesssim n^{-1}$, $M^{(\tau)}_{4,4}$ approaches 0 as $\tau \to 0$ uniformly in n. The same argument can be applied to $M^{(\tau)}_{1,4}$.

We turn our attention to (3.6) again. Expanding, it is straightforward to see that $M^{(\tau)}_{1,2} = M^{(\tau)}_{2,1} = M^{(\tau)}_{2,4} = M^{(\tau)}_{4,2} = 0$. Furthermore, we can write

$$(\mathbb{E}L^{(\text{sub})} - \mathcal{S}^{(\text{sub})}(\mathrm{diag}\{M^{(\tau)}_{1,1}, M^{(\tau)}_{2,2}\}) - (z + i\tau)I_{n_{\text{train}}+d})\,\mathrm{diag}\{M^{(\tau)}_{1,1}, M^{(\tau)}_{2,2}\} = I_{n_{\text{train}}+d} + D^{(\text{sub})}$$

with $D^{(\text{sub})} = \mathrm{diag}\{\mathrm{tr}(M^{(\tau)}_{2,2})K_{A\hat{A}^T}M^{(\tau)}_{4,1}, \mathrm{tr}(K_{A\hat{A}^T}M^{(\tau)}_{4,1} + K_{\hat{A}A^T}M^{(\tau)}_{1,4} + K_{\hat{A}\hat{A}^T}M^{(\tau)}_{4,4})M^{(\tau)}_{2,2}\}$. In particular, $\mathrm{diag}\{M^{(\tau)}_{1,1}, M^{(\tau)}_{2,2}\}$ almost solves the sub-MDE up to an additive perturbation term $D^{(\text{sub})}$ which vanishes as $\tau \to 0$ uniformly in n. Using the stability properties of the sub-MDE, it follows that $\|M_{1,1}^{(\tau)} - M_{1,1}\| \to 0$ and $\|M_{2,2}^{(\tau)} - M_{2,2}\| \to 0$ as $\tau \to 0$ uniformly in n. Expanding (3.6), we can pass this convergence to the remaining blocks. Hence, $M^{(\tau)}$ converges to M as $\tau \to 0$ uniformly in n. This completes the proof. □

Lemma 4.3.8. *Let $z \in \mathbb{H}$ with $|z| < \delta \wedge 1$ and $M \in \mathscr{M}$ be the unique solution to* (4.4). *Under the settings of Theorem 4.1.1,* $\operatorname{tr}(U((L-z)^{-1} - M(z))) \to 0$ *almost surely as $n \to \infty$ for every $U \in \mathbb{C}^{\ell \times \ell}$ with $\|U\|_* \leq 1$.*

Proof. Utilizing a similar argument as in the proof of Lemma 4.3.6, we observe that Assumption 1 is satisfied, $\lim_{n\to\infty} \sqrt{n}\lambda = 0$, and $\limsup_{n\to\infty} \|\tilde{\mathcal{S}}\| = 0$, where λ is the Lipschitz constant defined in Lemma 3.4.8. In particular, since $\|\Delta(L, \tau; z)\| \to 0$ as $n \to \infty$ for every $\tau \in \mathbb{R}_{>0}$ by Lemma 4.3.4, it follows from Lemma 3.4.8 that $\|D^{(\tau)}\| \to 0$ as $n \to \infty$ for every $\tau \in \mathbb{R}_{>0}$. By Lemma 4.3.7, Assumption 2 holds. The application of Theorem 3.4.2 yields the desired result. □

Having established that the solution to (4.4) acts as a deterministic equivalent for the pseudo-resolvent $(L - z\Lambda)^{-1}$ linked to (4.3), we aim to retrieve the expression in (4.2) by taking the spectral parameter to 0. To accomplish this, we need further control over the MDE near the origin.

Lemma 4.3.9. *Let $z \in \mathbb{H}$ with $|z| < \delta \wedge 1$ and $M \in \mathscr{M}$ the unique solution to* (4.4). *Then, for every $\epsilon \in (0, 2^{-1}]$ with $(2(\delta - \Re[z])^{-2} d^2 \|K_{AA^T}\|^2 - 1 - \Re[z])\epsilon \leq 2^{-1}(1 + \Re[z])$,*

$$1 - d\|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}(z)\|^2 \geq \epsilon.$$

In particular, under the settings of Theorem 4.1.1, there exists $\epsilon \in (0, 2^{-1}]$ depending on $0 < \eta < 1 \wedge \delta$ such that

$$\limsup_{n\to\infty} d\|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}(z)\|^2 < 1 - \epsilon$$

for all $z \in \mathbb{H}$ with $|z| \leq \eta$.

Proof. Fix $z \in \mathbb{H}$ with $|z| < 1 \wedge \delta$ and write $M \equiv M(z)$. Let $\epsilon \in (0, 2^{-1}]$ such that

$$(2(\delta - \Re[z])^{-2} d^2 \|K_{AA^T}\|^2 - 1 - \Re[z])\epsilon \leq 2^{-1}(1 + \Re[z])$$

and assume, by contradiction, that $d\|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}(z)\|^2 > 1 - \epsilon$. Using the definition of $M_{1,1}$ and $M_{2,2}$ and repeatedly applying (4.6),

$$\begin{aligned}
\operatorname{tr}(K_{AA^T}\Re[M_{1,1}]) &= (\delta - \Re[z]) \operatorname{tr}(K_{AA^T} M_{1,1} M_{1,1}^*) - \operatorname{tr}(\Re[M_{2,2}]) \operatorname{tr}(K_{AA^T} M_{1,1} K_{AA^T} M_{1,1}^*) \\
&= (\delta - \Re[z]) \operatorname{tr}(K_{AA^T} M_{1,1} M_{1,1}^*) \\
&+ d\|M_{2,2}\|^2 \|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 (1 + \Re[z]) \\
&+ d\|M_{2,2}\|^2 \|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \operatorname{tr}(K_{AA^T}\Re[M_{1,1}]) \\
&\geq (\delta - \Re[z]) \operatorname{tr}(K_{AA^T} M_{1,1} M_{1,1}^*) + (1 - \epsilon)(1 + \Re[z]) \\
&+ (1 - \epsilon) \operatorname{tr}(K_{AA^T}\Re[M_{1,1}]).
\end{aligned}$$

Solving for $\operatorname{tr}(K_{AA^T}\Re[M_{1,1}])$, we obtain that $\operatorname{tr}(K_{AA^T}\Re[M_{1,1}]) = t\epsilon^{-1}$ with

$$t := (\delta - \Re[z]) \operatorname{tr}(K_{AA^T} M_{1,1} M_{1,1}^*) + (1 - \epsilon)(1 + \Re[z]).$$

In particular, taking the real part of $M_{2,2}$, we have

$$-\Re[M_{2,2}] = \|M_{2,2}\|^2 (1 + \Re[z] + \operatorname{tr}(K_{AA^T}\Re[M_{1,1}])) I_d \succeq \|M_{2,2}\|^2 (1 + \Re[z] + t\epsilon^{-1}) I_d.$$

By taking the norm on both sides and leveraging the properties that the spectral norm preserves the Loewner partial ordering and that the spectral norm of the real and imaginary parts of a complex matrix are bounded above by the spectral norm of the matrix itself, we obtain

$$\|M_{2,2}\|^2 (1 + \Re[z] + t\epsilon^{-1}) \leq \|\Re[M_{2,2}]\| \leq \|M_{2,2}\|.$$

Rearranging, this implies that $\|M_{2,2}\| \leq (1 + \Re[z] + t\epsilon^{-1})^{-1}$. By Lemma 4.3.5 and the definition of ϵ,

$$d\|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}(z)\|^2 \leq \frac{(\delta - \Re[z])^{-2} d^2 \|K_{AA^T}\|^2}{1 + \Re[z] + t\epsilon^{-1}} \leq 2^{-1}.$$

This is a contradiction. Thus, it must be the case that

$$1 - d\|K_{AA^T}^{\frac{1}{2}} M_{1,1}(z) K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}(z)\|^2 \geq \epsilon$$

for every $\epsilon \in (0, 2^{-1}]$ with $(2\delta^{-2}d^2\|X\|^4 - 1 - \Re[z])\epsilon \leq 2^{-1}(1 + \Re[z])$. The result follows. □

The statement of Lemma 4.3.9 is intricate because the left-hand side of the inequality $(2\delta^{-2}d^2\|K_{AA^T}\|^2 - 1 - \Re[z])\epsilon \leq 2^{-1}(1 + \Re[z])$ may be negative. Nevertheless, the essence of Lemma 4.3.9 lies in the fact that the quantity $1 - d\|K_{AA^T}^{1/2} M_{1,1}(z) K_{AA^T}^{1/2}\|_F^2 \|M_{2,2}(z)\|^2$ can be consistently bounded away from 0 regardless of the dimension.

Now that we have some control on the solution of the MDE when the spectral parameter is close to the origin, we still need to continuously extend the function M to its boundary point 0. To do so, we analytically extend M by reflection to the lower complex plane $\{z \in \mathbb{H} : \Im[z] < 0\}$ through an open interval containing the origin.

Lemma 4.3.10. *The unique solution M to* (3.2) *can be extended analytically to the lower-half complex plane through the open interval* $(-(1 \wedge \delta), 1 \wedge \delta)$.

Proof. Using the definition of matrix imaginary part and the resolvent identity, we obtain the system of equations $\Im[M_{1,1}] = M_{1,1}(\Im[z] + \operatorname{tr}(\Im[M_{2,2}])K_{AA^T})(M_{1,1})^*$ and $\operatorname{tr}(\Im[M_{2,2}]) = d\|M_{2,2}\|^2(\Im[z] + \operatorname{tr}(K_{AA^T}\Im[M_{1,1}]))$. Combining the two equalities, we get $d^{-1}\operatorname{tr}(\Im[M_{2,2}])(1 - \|\sqrt{d}K_{AA^T}^{1/2} M_{1,1} K_{AA^T}^{1/2}\|_F^2 \|M_{2,2}\|^2) = \|M_{2,2}\|^2 \Im[z](1 + \|\sqrt{d}K_{AA^T}^{1/2} M_{1,1}\|_F^2)$. By Lemma 4.3.9, $1 - d\|K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}\|^2 > 0$ uniformly on $\{z \in \mathbb{H} : |z| \leq \epsilon\}$ for every $0 < \epsilon < 1 \wedge \delta$. Using Lemma 4.3.5,

$$d^{-1}\operatorname{tr}(\Im[M_{2,2}]) \leq \frac{\Im[z](1 + \|\sqrt{d}(\delta - \Re[z])^{-1} K_{AA^T}^{\frac{1}{2}}\|_F^2)}{1 - \|\sqrt{d}K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2 \|M_{2,2}\|^2}.$$

Thus, we observe that $\Im[M_{2,2}(z)] \downarrow 0$ uniformly as $\Im[z] \to 0$ on $(-\epsilon, \epsilon)$ and similarly for $\|\Im[M_{1,1}]\|$. Since $M_{1,1}$ and $M_{2,2}$ fully define the solution of the MDE, $\|\Im[M(z)]\|$ vanishes uniformly for $\Re[z] \in (-\epsilon, \epsilon)$ as $\Im[z] \to 0$. By Lemma 2.2.9, the positive semidefinite measure in Theorem 3.3.1 has no support in $(-\epsilon, \epsilon)$. The result follows from Lemma 2.2.10. □

We may now remove the spectral parameter in Lemma 4.3.8.

Corollary 4.3.1. *Let $M \in \mathscr{M}$ be the unique solution to* (4.4)*. Under the settings of Theorem 4.1.1,* $\operatorname{tr}(U(L^{-1} - M(0))) \to 0$ *almost surely as $n \to \infty$ for every sequence $U \in \mathbb{C}^{\ell \times \ell}$ with $\|U\|_* \leq 1$.*

The solution to (4.4) is fully defined by the scalar $\operatorname{tr}(M_{2,2})$. By fully defined, we mean that we may explicitly construct the full solution of the MDE using only knowledge of $\operatorname{tr}(M_{2,2})$. Using the sub-MDE defined above, we get the following numerical result.

Lemma 4.3.11. *Suppose that $M(0)$ solves* (4.4) *when $z = 0$. Let*

$$T : x \in \mathbb{R}_{<0} \mapsto -(1 + \operatorname{tr}(K_{AA^T}(\delta I_{n_{\text{train}}} - dxK_{AA^T})^{-1}))^{-1} \in \mathbb{R}_{<0}$$

and consider the iterates $\{\alpha_k\}_{k \in \mathbb{N}_0}$ obtained via $\alpha_{k+1} = T(\alpha_k)$ for every $k \in \mathbb{N}$ with arbitrary $\alpha_0 \in \mathbb{R}_{\leq 0}$. Then,

$$M(0) = \begin{bmatrix} (\delta I_{n_{\text{train}}} - d\alpha K_{AA^T})^{-1} & 0 & -d\alpha M_{1,1}(0) K_{A\hat{A}^T} & 0 \\ 0 & \alpha I_d & 0 & 0 \\ -d\alpha K_{\hat{A}A^T} M_{1,1}(0) & 0 & (d\alpha)^2 K_{\hat{A}A^T} M_{1,1}(0) K_{A\hat{A}^T} + d\alpha K_{\hat{A}\hat{A}^T} & -I_{n_{\text{test}}} \\ 0 & 0 & -I_{n_{\text{test}}} & 0 \end{bmatrix}$$

where $\alpha := d^{-1} \operatorname{tr}(M_{2,2}(0)) = \lim_{k \to \infty} \alpha_k$.

Proof. In order to use Theorem 2.1.1, we consider the set $\mathscr{S}$ of complex matrices $N \in \mathbb{C}^{(n_{\text{train}}+d) \times (n_{\text{train}}+d)}$ with $N_{1,2} = N_{2,1} = 0$, $\Re[N_{1,1}] \succ 0$ and $\Re[N_{2,2}] \prec 0$ and denote

$$\mathcal{F}^{(\text{sub})} : N \in \mathscr{S} \mapsto (\mathbb{E}L^{(\text{sub})} - \mathcal{S}^{(\text{sub})}(N))^{-1} \in \mathscr{S}.$$

Using an argument analogous to the one in lemmas 3.3.1 and 4.3.5, we see that

$$\Re[\mathcal{F}^{(\text{sub})}_{1,1}(N)] \succeq \delta \mathcal{F}^{(\text{sub})}_{1,1}(N)(\mathcal{F}^{(\text{sub})}_{1,1}(N))^* \succeq \delta(\delta + d\|K_{AA^T}\|\|N\|)^{-2}$$

and

$$\Re[\mathcal{F}^{(\text{sub})}_{2,2}(N)] \preceq -\mathcal{F}^{(\text{sub})}_{2,2}(N)(\mathcal{F}^{(\text{sub})}_{2,2}(N))^* \preceq -(1 + d\|K_{AA^T}\|\|N\|)^{-2}$$

for every $N \in \mathscr{S}$. This implies that $\mathcal{F}^{(\text{sub})}$ is well-defined. Furthermore, by Lemma 2.2.4,

$\|\mathcal{F}^{(\mathrm{sub})}_{1,1}(N)\| \leq \delta^{-1}$ and $\|\Re[\mathcal{F}^{(\mathrm{sub})}_{2,2}(N)]\| \leq 1$ for every $N \in \mathscr{S}$. Let $\mathscr{S}_b = \mathscr{S} \cap \mathscr{B}_b(0)$. Then, for every $b > \delta \vee 1$, $\mathcal{F}^{(\mathrm{sub})}$ is strictly holomorphic on $\mathscr{S}_b$. By Theorem 2.1.1, there exists a unique $N \in \mathscr{S}_b$ such that $\mathcal{F}^{(\mathrm{sub})}(N) = N$. Furthermore, the sequence $\{N_k\}_{k\in\mathbb{N}_0}$ with $N_{k+1} = \mathcal{F}^{(\mathrm{sub})}(N_k)$ for every $k \in \mathbb{N}$ converges to N for every $N_0 \in \mathscr{S}_b$. Since $\mathscr{S} = \bigcup_{b>0} \mathscr{S}_b$, it follows that there exists a unique $N \in \mathscr{S}$ such that $\mathcal{F}^{(\mathrm{sub})}(N) = N$. Also, the sequence $\{N_k\}_{k\in\mathbb{N}_0}$ with $N_{k+1} = \mathcal{F}^{(\mathrm{sub})}(N_k)$ for every $k \in \mathbb{N}$ converges to N for every $N_0 \in \mathscr{S}$. Choosing $N_0 = \operatorname{diag}\{I_{n_{\mathrm{train}}}, \alpha_0 I_d\}$ gives the result. □

4.3.3 Second Deterministic Equivalent

We now consider the squared matrix $(AA^T + \delta I_{n_{\mathrm{train}}})^{-1} A\hat{A}^T \hat{A}A^T (AA^T + \delta I_{n_{\mathrm{train}}})^{-1}$. Notice that $(L^{-2})_{1,1} = R^2 + RA^TAR + RA\hat{A}^T\hat{A}AR$ and $(L^{(\mathrm{sub})})^{-2}_{1,1} = R^2 + RA^TAR$ for $R := (AA^T + \delta I_{n_{\mathrm{train}}})^{-1}$. Rearranging, we get that

$$(AA^T + \delta I_{n_{\mathrm{train}}})^{-1} A\hat{A}^T\hat{A}A(AA^T + \delta I_{n_{\mathrm{train}}})^{-1} = (L^{-2})_{1,1} - (L^{(\mathrm{sub})})^{-2}_{1,1}. \tag{4.7}$$

Therefore, it suffices to find deterministic equivalents for $(L^{(\mathrm{sub})})^{-2}$ and L^{-2} to obtain a deterministic equivalent for the random matrix $(AA^T + \delta I_{n_{\mathrm{train}}})^{-1} A\hat{A}^T\hat{A}A(AA^T + \delta I_{n_{\mathrm{train}}})^{-1}$.

Lemma 4.3.12. *Under the settings of Theorem 4.1.1, let* $\alpha = d^{-1}\operatorname{tr}(M_{2,2}(0))$ *as in Lemma 4.3.11,* $R = (AA^T + \delta I_{n_{\mathrm{train}}})^{-1}$ *and define*

$$\beta = \frac{\alpha^2 \operatorname{tr}\left(K_{\hat{A}\hat{A}^T} + d\alpha K_{\hat{A}A^T} M_{1,1}(0)(I_{n_{\mathrm{train}}} + \delta M_{1,1}(0))K_{A\hat{A}^T}\right)}{1 - \|\sqrt{d}\alpha K^{\frac{1}{2}}_{AA^T} M_{1,1}(0) K^{\frac{1}{2}}_{AA^T}\|_F^2} \in \mathbb{R}_{\geq 0}.$$

Then, $\operatorname{tr} U(RA\hat{A}^T\hat{A}A^TR - d\beta M_{1,1}(0)K_{AA^T}M_{1,1}(0) - M_{1,3}(0)M_{3,1}(0)) \to 0$ *almost surely as* $n \to \infty$ *for every sequence* $U \in \mathbb{C}^{n_{\mathrm{train}} \times n_{\mathrm{train}}}$ *with* $\|U\|_* \leq 1$.

Proof. First, as stated in Lemma 3.3.3, we note that $i\tau \mapsto (L - i\tau)^{-1}$ is an analytic function with $\partial_{i\tau}(L - i\tau)^{-1} = (L - i\tau)^{-2}$. Overloading notation, let $M^{(\zeta)} \in \mathscr{M}_+$ be the unique solution to the MDE $(\mathbb{E}L - \mathcal{S}(M^{(\zeta)}) - \zeta I_\ell)M^{(\zeta)} = I_\ell$ where $\zeta \in \mathbb{H}$. By the proof of [EKN20, Theorem 2.14], the function $\zeta \mapsto M^{(\zeta)}$ is analytic on $\mathbb{H}$. Adapting a general argument resembling to the one in [Sch+23, equation (174)], it follows from Cauchy's integral formula that

$$(L-i\tau)^{-2}-\partial_{i\tau}M^{(\tau)}(0)=\partial_{i\tau}\left((L-i\tau)^{-1}-M^{(\tau)}(0)\right)=\frac{1}{2\pi}\oint_{\gamma}\frac{(L-\zeta)^{-1}-M^{(\zeta)}}{(\zeta-i\tau)^2}\mathrm{d}\zeta$$

where γ forms a counterclockwise circle of radius $\tau/2$ around $i\tau$. We know that $M^{(\zeta)}$ is a deterministic equivalent for $(L-\zeta)^{-1}$ for every fixed $\zeta\in\mathbb{H}$. By the resolvent identity, $\zeta\mapsto(L-\zeta I_\ell)^{-1}$ is $4/\tau^2$-Lipschitz on $\{z\in\mathbb{H}:\Im[z]\geq\tau/2\}$. Similarly, by the proof of [EKN20, Theorem 2.14], the function $\zeta\mapsto M^{(\zeta)}$ is $(2/\tau)^{12}$-Lipschitz on $\{z\in\mathbb{H}:\Im[z]\geq\tau/2\}$. Therefore, we obtain $\operatorname{tr}(U((L-i\tau)^{-2}-\partial_{i\tau}M^{(\tau)}(0)))\to 0$ almost surely as $n\to\infty$ for every $\tau\in\mathbb{R}_{>0}$ and $U\in\mathbb{C}^{\ell\times\ell}$ with $\|U\|_*\leq 1$. Taking the derivative of (3.6), we obtain $\partial_{i\tau}M^{(\tau)}(0)=M^{(\tau)}(0)(\mathcal{S}(\partial_{i\tau}M^{(\tau)}(0))+I_\ell)M^{(\tau)}(0)$ or, relating this equation to the stability operator, $\mathcal{L}^{(\tau)}(\partial_{i\tau}M^{(\tau)}(0))=(M^{(\tau)}(0))^2$ with $\mathcal{L}^{(\tau)}:N\in\mathbb{C}^{\ell\times\ell}\mapsto N-M^{(\tau)}(0)\mathcal{S}(N)M^{(\tau)}(0)$.

In what follows, we omit the argument of $M^{(\tau)}$ and write $M^{(\tau)}\equiv M^{(\tau)}(0)$. Using simple but tedious computations, we decompose $\partial_{i\tau}M^{(\tau)}_{j,k}=C_{j,k}+D_{j,k}\operatorname{tr}(\partial_{i\tau}M^{(\tau)}_{2,2})$ for every $(j,k)\in\{(1,1),(1,4),(4,4)\}$ with

$$\begin{aligned}
C_{1,1}&:=M^{(\tau)}_{1,4}M^{(\tau)}_{4,1}+(M^{(\tau)}_{1,1})^2+M^{(\tau)}_{1,3}M^{(\tau)}_{3,1},\\
D_{1,1}&:=M^{(\tau)}_{1,1}K_{AA^T}M^{(\tau)}_{1,1}+M^{(\tau)}_{1,1}K_{A\hat{A}^T}M^{(\tau)}_{4,1}+M^{(\tau)}_{1,4}K_{\hat{A}A^T}M^{(\tau)}_{1,1}+M^{(\tau)}_{1,4}K_{\hat{A}\hat{A}^T}M^{(\tau)}_{4,1},\\
C_{4,4}&:=M^{(\tau)}_{4,1}M^{(\tau)}_{1,4}+(M^{(\tau)}_{4,4})^2+M^{(\tau)}_{4,3}M^{(\tau)}_{3,4},\\
D_{4,4}&:=M^{(\tau)}_{4,1}K_{AA^T}M^{(\tau)}_{1,4}+M^{(\tau)}_{4,1}K_{A\hat{A}^T}M^{(\tau)}_{4,4}+M^{(\tau)}_{4,4}K_{\hat{A}A^T}M^{(\tau)}_{1,4}+M^{(\tau)}_{4,4}K_{\hat{A}\hat{A}^T}M^{(\tau)}_{4,4},\\
C_{1,4}&:=M^{(\tau)}_{1,1}M^{(\tau)}_{1,4}+M^{(\tau)}_{1,4}M^{(\tau)}_{4,4}+M^{(\tau)}_{1,3}M^{(\tau)}_{3,4},\ \text{and}\\
D_{1,4}&:=M^{(\tau)}_{1,1}K_{AA^T}M^{(\tau)}_{1,4}+M^{(\tau)}_{1,1}K_{A\hat{A}^T}M^{(\tau)}_{4,4}+M^{(\tau)}_{1,4}K_{\hat{A}A^T}M^{(\tau)}_{1,4}+M^{(\tau)}_{1,4}K_{AA^T}M^{(\tau)}_{4,4}.
\end{aligned}$$

Taking the trace of the $2,2$ block of $\partial_{i\tau}M^{(\tau)}(0)$, we get

$$\begin{aligned}
\operatorname{tr}(\partial_{i\tau}M^{(\tau)}_{2,2})&=\operatorname{tr}((M^{(\tau)}_{2,2})^2)(\rho(\partial_{i\tau}M^{(\tau)})+1)\\
&=\operatorname{tr}((M^{(\tau)}_{2,2})^2)(\operatorname{tr}(K_{AA^T}C_{1,1}+K_{A\hat{A}^T}C^T_{1,4}+K_{\hat{A}A^T}C_{1,4}+K_{\hat{A}\hat{A}^T}C_{4,4})+1)\\
&\quad+\operatorname{tr}(\partial_{i\tau}M^{(\tau)}_{2,2})\operatorname{tr}((M^{(\tau)}_{2,2})^2)\operatorname{tr}(K_{AA^T}D_{1,1}+K_{A\hat{A}^T}D^T_{1,4}+K_{\hat{A}A^T}D_{1,4}+K_{\hat{A}\hat{A}^T}D_{4,4}).
\end{aligned}$$

By the proof of Lemma 4.3.8, we observe that there exists a function $f:\mathbb{R}_{>0}\mapsto\mathbb{R}_{\geq 0}$ with

$\lim_{\tau\downarrow 0} f(\tau) = 0$ such that $\|D_{4,4}\| \leq f(\tau) + o_n(1)$, $\|D_{1,1} - M_{1,1}K_{AA^T}M_{1,1}\| \leq f(\tau) + o_n(1)$, $\|D_{1,4}\| = \|D_{4,1}^T\| \leq f(\tau) + o_n(1)$ and $\|M_{2,2}^{(\tau)} - M_{2,2}\| \leq f(\tau) + o_n(1)$. By Lemma 4.3.9,

$$|1 - \operatorname{tr}((M_{2,2}^{(\tau)})^2) \operatorname{tr}(K_{AA^T}D_{1,1} + K_{A\hat{A}^T}D_{4,1} + K_{\hat{A}A^T}D_{1,4} + K_{\hat{A}\hat{A}^T}D_{4,4})|$$

is bounded away from 0 for every $n \in \mathbb{N}$ large enough and $\tau \in \mathbb{R}_{>0}$ small enough. In particular, the limit $\lim_{\tau\to 0} \partial_{i\tau} M^{(\tau)}$ exists and satisfies

$$\begin{aligned}\lim_{\tau\to 0} d^{-1} \operatorname{tr}(\partial_{i\tau} M_{2,2}^{(\tau)}) &= \frac{\alpha^2(\operatorname{tr}(K_{AA^T}M_{1,1}^2 + K_{AA^T}M_{1,3}M_{3,1} - K_{A\hat{A}^T}M_{3,1} - K_{\hat{A}A^T}M_{1,3} + K_{\hat{A}\hat{A}^T}) + 1)}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2} \\ &= \beta + \frac{\alpha^2(1 + \operatorname{tr}(K_{AA^T}M_{1,1}^2))}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2}\end{aligned}$$

where we recall that $\alpha = d^{-1}\operatorname{tr}(M_{2,2})$ as defined in Lemma 4.3.11. Plugging this into the expression for $\partial_{i\tau} M_{1,1}^{(\tau)}$ and taking the limit as $\tau \downarrow 0$, we get that

$$d\beta M_{1,1}K_{AA^T}M_{1,1} + M_{1,3}M_{3,1} + M_{1,1}^2 + \frac{d\alpha^2(1 + \operatorname{tr}(K_{AA^T}M_{1,1}^2))}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2} M_{1,1}K_{AA^T}M_{1,1}$$

is an asymptotic deterministic equivalent for $(L^{-2})_{1,1}$.

Using a similar argument, we note that $\partial_z M^{(\mathrm{sub})}(0)$ is a deterministic equivalent for $(L^{(\mathrm{sub})})^{-2}$ and

$$\partial_z M_{1,1}^{(\mathrm{sub})} = M_{1,1}^2 + \frac{d\alpha^2(1 + \operatorname{tr}(K_{AA^T}M_{1,1}^2))}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M_{1,1} K_{AA^T}^{\frac{1}{2}}\|_F^2} M_{1,1}K_{AA^T}M_{1,1}.$$

We obtain the result by (4.7). □

5

Conclusions and Future Work

In this thesis, our objective was to develop analytical tools to better understand the behavior of machine learning models. Given the inherent complexity of these models, characterized by their large size and non-linearities, direct analysis proves challenging. To circumvent these obstacles, we adopt the strategy of modeling a portion of the machine learning process using random matrices. By doing so, we can study various aspects of machine learning through the analysis of rational expressions involving random matrices. This approach allows us to use tools from random matrix theory to analyze machine learning models. In that regard, our particular focus lies in extending the matrix Dyson equation to correlated linearizations. One key advantage of the matrix Dyson equation framework, as opposed to alternative techniques for finding deterministic equivalents, is its ability to streamline analysis by offering a candidate deterministic equivalent and identifying natural quantities that must be bounded to establish deterministic equivalence. We have expanded this framework by demonstrating the existence of a unique solution to the matrix Dyson equation for linearizations under general settings. Additionally, we have established multiple properties such as effective support bounds and shown that the solution to the fixed-point equation serves as an asymptotic deterministic equivalent for a pseudo-resolvent. Our methodology is novel, offering potential applications to a range of other problems. For instance, the utilization of the Carathéodory-Riffen-Finsler pseudometric to establish stability properties of the matrix Dyson equation represents a methodological innovation with broader applicability.

Our work diverges from existing literature on the matrix Dyson equation in two significant ways. Firstly, recognizing that global information about the spectral properties of random objects often suffices to draw conclusions about machine learning problems, we concentrate solely on the properties of the matrix Dyson equation and pseudo-resolvents on a macroscale. This contrasts with much of the existing literature concerning the matrix Dyson equation, which primarily focuses on the mesoscale and microscale. While local convergence results on the mesoscale and microscale offer greater precision, they are often unnecessary for understanding machine learning model behavior. By operating on the macroscale, we are able to relax distributional assumptions on the random matrices and consider linearizations with more general correlation structures, thereby improving the fidelity of our modeling of real-world datasets. Nevertheless, there remain numerous avenues for further exploration in this field. For example, we believe that the implicit assumptions on the matrices B and Q, imposed by the requirement of $\|\tilde{\mathcal{S}}\|$ vanishing as $n \to \infty$, could be relaxed. This would significantly expand the applicability of the matrix Dyson equation framework to even more general linearizations. Additionally, several questions regarding the matrix Dyson equation itself remain open. In particular, it is unclear what minimal assumptions are necessary to ensure the validity of the matrix Dyson equation, a question of considerable importance as it relates to a form of universality. These areas represent promising directions for future research.

To demonstrate the potential of our framework, we applied our theory to derive a deterministic equivalent for the empirical test error of random features ridge regression. The random features model, incorporating a non-linear activation and the potential to be overparameterized, is of particular interest due to its simplicity for theoretical analysis while capturing key aspects of more complex machine learning models. We considered the random features model with a general Lipschitz activation function and last-layer weights trained by fitting a regularized linear model. Our main result establishes that the norm squared error of the trained random features model on a test dataset concentrates around a deterministic quantity. We provide a simple characterization of this deterministic quantity, which involves solving a single scalar fixed-point equation. Additionally, we offer a numerical result guaranteeing the convergence of an algorithm to solve this fixed-point equation, enabling efficient computation of the deterministic equivalent. Having a deterministic equivalent enables us to simplify the analysis of the generalization error of the random features model by focusing on the analysis of the conjugate kernel. This approach allows us to characterize implicit regularization, which is crucial for understanding the generalization properties of a model. Additionally, we offer a Gaussian equivalence theorem, extending previous similar results to the empirical test error. This theorem simplifies the analysis of more intricate networks by reducing them to the analysis of a simpler surrogate Gaussian network.

Importantly, our approach does not rely on any specific distributional assumptions about the data, unlike many other results studying the test error of random features ridge regression. This generality broadens the applicability of our result to a wider range of datasets, including those like the MNIST dataset [Den12], which do not conform to simple isotropic Gaussian assumptions as shown in Figure 1.2. This flexibility opens up promising avenues for future research in privacy, where the distribution of the test data could be chosen adversarially. Furthermore, our result accommodates a wide range of activation functions, although theoretical proof is lacking for non-Lipschitz activation functions, warranting further investigation. Future research directions could include extending our result to more general architectures, such as deep random features. While the random features model has served as a proxy for the theoretical analysis of machine learning models, its exact relationship with neural networks remains unclear. For example, it is uncertain whether deep structured random features networks are as expressive as deep neural networks. Addressing this question could offer valuable insights and directions for future research endeavors. Moreover, one of the primary motivations behind studying the empirical test error of random features ridge regression is the necessity of data anisotropy to surpass the kernel lower bound. This requirement acknowledges that data is typically non-isotropic, contrary to the common assumption in many theoretical results in machine learning. It also reflects the process of feature learning, which involves learning a meaningful representation of the data, a fundamental aspect of useful machine learning models. Another avenue to surpass the kernel lower bound is to move beyond the linear scaling regime, which presents another intriguing direction for future research.

In conclusion, this thesis adds to the growing body of literature dedicated to elucidating the behavior of machine learning models. However, there is still much work to be done to achieve a full understanding of the inner workings of those models. Bridging the gap between theory and practice will require collaborative efforts across diverse communities. This understanding holds crucial importance in safeguarding fairness and privacy, comprehending when the output of machine learning models can be trusted. This pursuit will become increasingly important as machine learning models become increasingly intertwined with our daily routines.

Bibliography

[AEK17] Oskari H. Ajanki, László Erdős, and Torben Krüger. “Universality for general Wigner-type matrices”. In: *Probability Theory and Related Fields* 169 (2017).

[AEK18] Johannes Alt, László Erdős, and Torben Krüger. “The Dyson equation with linear self-energy: spectral bands, edges and cusps”. In: *Documenta Mathematica* 25 (2018), pp. 1421–1539.

[AEK19a] Oskari H. Ajanki, László Erdős, and Torben Krüger. “Quadratic Vector Equations On Complex Upper Half-Plane”. In: *Memoirs of the American Mathematical Society* 261.1261 (2019).

[AEK19b] Oskari H. Ajanki, László Erdős, and Torben Krüger. “Stability of the Matrix Dyson Equation and Random Matrices with Correlations”. In: *Probability Theory and Related Fields* 173.1-2 (2019), pp. 293–373.

[AKE17] Oskari H. Ajanki, Torben Krüger, and László Erdős. “Singularities of Solutions to Quadratic Vector Equations on the Complex Upper Half-Plane”. In: *Communications on Pure and Applied Mathematics* 70.9 (2017), pp. 1672–1705.

[ALP22] Ben Adlam, Jake A. Levinson, and Jeffrey Pennington. “A Random Matrix Perspective on Mixtures of Nonlinearities in High Dimensions”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics.* Vol. 151. Proceedings of Machine Learning Research. PMLR, 2022, pp. 3434–3457.

[Alt+19] Johannes Alt et al. “Location of the spectrum of Kronecker random matrices”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 55.2 (2019), pp. 661–696.

[Alt+20] Johannes Alt et al. “Correlated random matrices: Band rigidity and edge universality”. In: *The Annals of Probability* 48.2 (2020).

[Alt18] Johannes Alt. “Dyson equation and eigenvalue statistics of random matrices”. PhD thesis. IST Austria, 2018.

[And13] Greg W. Anderson. “Convergence of the largest singular value of a polynomial in independent Wigner matrices”. In: *The Annals of Probability* 41.3B (2013).

[And15] Greg W. Anderson. “A local limit law for the empirical spectral distribution of the anticommutator of independent Wigner matrices”. In: *Annales de l’I.H.P. Probabilités et statistiques* 51.3 (2015), pp. 809–841.

[AP20a] Ben Adlam and Jeffrey Pennington. “The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization”. In: *Proceedings of the 37th International Conference on Machine Learning.* Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 74–84.

[AP20b] Ben Adlam and Jeffrey Pennington. “Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition”. In: *Advances in Neural Information Processing Systems.* Vol. 33. Curran Associates, Inc., 2020, pp. 11022–11032.

[Ba+22] Jimmy Ba et al. “High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation”. In: *Advances in Neural Information Processing Systems.* Vol. 35. Curran Associates, Inc., 2022, pp. 37932–37946.

[Bah+21] Yasaman Bahri et al. *Explaining Neural Scaling Laws.* 2021. arXiv: `2102.06701 [cs.LG]`.

[Bel+19] Mikhail Belkin et al. “Reconciling modern machine-learning practice and the classical bias-variance trade-off”. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[Bel10] Richard E. Bellman. *Dynamic Programming.* Princeton: Princeton University Press, 2010.

[BH23] Tatiana Brailovskaya and Ramon van Handel. *Universality and sharp matrix concentration inequalities.* 2023. arXiv: `2201.05142 [math.PR]`.

[BMS17] Serban T. Belinschi, Tobias Mai, and Roland Speicher. "Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 2017.732 (2017), pp. 21–53.

[BP21] Lucas Benigni and Sandrine Péché. "Eigenvalue distribution of some nonlinear models of random matrices". In: *Electronic Journal of Probability* 26 (2021), pp. 1–37.

[BP22] Lucas Benigni and Sandrine Péché. *Largest Eigenvalues of the Conjugate Kernel of Single-Layered Neural Networks*. 2022. arXiv: `2201.04753 [math.PR]`.

[BPH23] David Bosch, Ashkan Panahi, and Babak Hassibi. *Precise Asymptotic Analysis of Deep Random Feature Models*. 2023. arXiv: `2302.06210 [stat.ML]`.

[Bro+20] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[BS10] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer, 2010.

[BZ08] Zhidong Bai and Wang Zhou. "Large Sample Covariance Matrices Without Independence Structures in Columns". In: *Statistica Sinica* 18.2 (2008), pp. 425–442.

[Cho22] Clément Chouard. *Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure*. 2022. arXiv: `2211.13044 [math.PR]`.

[CL22] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[Col+23] Elizabeth Collins-Woodfin et al. *Hitting the High-Dimensional Notes: An ODE for SGD learning dynamics on GLMs and multi-index models*. 2023. arXiv: `2308.08977 [math.OC]`.

[Cui+24] Hugo Cui et al. *Asymptotics of feature learning in two-layer networks after one gradient-step*. 2024. arXiv: `2402.04980 [stat.ML]`.

[Den12] Li Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[Dic16] Lee H. Dicker. “Ridge regression and asymptotic minimax estimation over spheres of growing dimension”. In: *Bernoulli* 22.1 (2016), pp. 1–37.

[DW18] Edgar Dobriban and Stefan Wager. “High-dimensional asymptotics of prediction: Ridge regression and classification”. In: *The Annals of Statistics* 46.1 (2018), pp. 247–279.

[EH70] Clifford J. Earle and Richard S. Hamilton. “A fixed point theorem for holomorphic mappings”. In: *Global Analysis (Proc. Sympos. Pure Math., Vol. XVI, Berkeley, 1968).* Vol. 16. Rhode Island: AMS, 1970, pp. 61–65.

[EKN20] László Erdős, Torben Krüger, and Yuriy Nemish. “Local laws for polynomials of Wigner matrices”. In: *Journal of Functional Analysis* 278.12 (2020), p. 108507.

[EKS19] László Erdős, Torben Krüger, and Dominik Schröder. “Random Matrices With Slow Correlation Decay”. In: *Forum of Mathematics, Sigma* 7 (2019).

[Epo22] Epoch. *Parameter, Compute and Data Trends in Machine Learning.* Accessed: 2024-02-02. 2022.

[Erd19] László Erdős. *The matrix Dyson equation and its applications for random matrices.* 2019. arXiv: `1903.10060 [math.PR]`.

[FKN23] Jacob Fronk, Torben Krüger, and Yuriy Nemish. *Norm Convergence Rate for Multivariate Quadratic Polynomials of Wigner Matrices.* 2023. arXiv: `2308.16778 [math.PR]`.

[FM19] Zhou Fan and Andrea Montanari. “The spectral norm of random inner-product kernel matrices”. In: *Probability Theory and Related Fields* 173.1 (2019), pp. 27–85.

[FW20] Zhou Fan and Zhichao Wang. “Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks”. In: *Advances in Neural Information Processing Systems.* Vol. 33. Curran Associates, Inc., 2020, pp. 7710–7721.

[Ger+20] Federica Gerace et al. "Generalisation error in learning with random features and the hidden manifold model". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3452–3462.

[Gol+22] Sebastian Goldt et al. "The gaussian equivalence of generative models for learning with shallow neural networks". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 426–471.

[GT00] Fritz Gesztesy and Eduard Tsekanovskii. "On Matrix-Valued Herglotz Functions". In: *Mathematische Nachrichten* 218.1 (2000), pp. 61–138.

[Gut+23] Florentin Guth et al. *A Rainbow in Deep Network Black Boxes*. 2023. arXiv: `2305.18512 [cs.LG]`.

[Har03] Lawrence A. Harris. "Fixed point of holomorphic mappings for domains in Banach spaces". In: *Abstract and Applied Analysis* 2003 (2003).

[Har79] Lawrence A. Harris. "Schwarz-Pick Systems of Pseudometrics for Domains in Normed Linear Spaces". In: *Advances in Holomorphy*. Ed. by Jorge Albedo Barroso. Vol. 34. North-Holland Mathematics Studies. North-Holland, 1979, pp. 345–406.

[Has+22] Trevor Hastie et al. "Surprises in high-dimensional ridgeless least squares interpolation". In: *The Annals of Statistics* 50.2 (2022), pp. 949–986.

[HFS07] J. William Helton, Reza Rashidi Far, and Roland Speicher. "Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints". In: *International Mathematics Research Notices* (2007).

[Hil48] Einar Hille. *Functional analysis and semi-groups.* English. Vol. 31. American Mathematical Society colloquium publications. New York: American Mathematical Society, 1948.

[HL23] Hong Hu and Yue M. Lu. "Universality Laws for High-Dimensional Learning With Random Features". In: *IEEE Transactions on Information Theory* 69.3 (2023), pp. 1932–1964.

[HLM24] Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. *Asymptotics of Random Feature Regression Beyond the Linear Scaling Regime*. 2024. arXiv: `2403.08160 [stat.ML]`.

[HMS18] J. William Helton, Tobias Mai, and Roland Speicher. "Applications of realizations (aka linearizations) to free probability". In: *Journal of Functional Analysis* 274.1 (2018), pp. 1–79.

[HMV06] J. William Helton, Scott A. McCullough, and Victor Vinnikov. "Noncommutative convexity arises from linear matrix inequalities". In: *Journal of Functional Analysis* 240.1 (2006), pp. 105–191.

[HT05] Uffe Haagerup and Steen Thorbjørnsen. "A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group". In: *Annals of Mathematics. Second Series* 2 (2005).

[Jac+20] Arthur Jacot et al. "Implicit Regularization of Random Feature Models". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4631–4640.

[Kri09] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009.

[LCM21] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. "A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (2021), p. 124006.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Soc., 2001. 196 pp.

[Lee+22] Kiwon Lee et al. "Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022, pp. 36944–36957.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. "A random matrix approach to neural networks". In: *The Annals of Applied Probability* 28.2 (2018), pp. 1190–1248.

[LM21] Zhenyu Liao and Michael W Mahoney. “Hessian Eigenspectra of More Realistic Nonlinear Models”. In: *Advances in Neural Information Processing Systems.* Vol. 34. Curran Associates, Inc., 2021, pp. 20104–20117.

[LN17] Annemarie Luger and Mitja Nedic. “A characterization of Herglotz–Nevanlinna functions in two variables via integral representations”. In: *Arkiv för Matematik* 55.1 (2017), pp. 199–216.

[Lou+22] Bruno Loureiro et al. “Learning curves of generic features maps for realistic datasets with a teacher-student model*”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2022.11 (2022), p. 114001.

[LP09] Anna Lytova and Leonid Pastur. “Central limit theorem for linear eigenvalue statistics of random matrices with independent entries”. In: *The Annals of Probability* 37.5 (2009).

[LP23] Hugo Latourelle-Vigeant and Elliot Paquette. *Matrix Dyson equation for correlated linearizations and test error of random features regression.* 2023. arXiv: `2312.09194 [math.ST]`.

[LS02] Tzon-Tzer Lu and Sheng-Hua Shiou. “Inverses of 2 × 2 block matrices”. In: *Computers & Mathematics with Applications* 43.1 (2002), pp. 119–129.

[MG21] Gabriel Mel and Surya Ganguli. “A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions”. In: *Proceedings of the 38th International Conference on Machine Learning.* Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 7578–7587.

[MM22] Song Mei and Andrea Montanari. “The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve”. In: *Communications on Pure and Applied Mathematics* 75.4 (2022), pp. 667–766.

[MMM22] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration”. In: *Applied and Computational Harmonic Analysis* 59 (2022). Special Issue on Harmonic Analysis and Machine Learning, pp. 3–84.

[MP22] Gabriel Mel and Jeffrey Pennington. "Anisotropic Random Feature Regression in High Dimensions". In: *International Conference on Learning Representations.* 2022.

[Paq+21] Courtney Paquette et al. "SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality". In: *Proceedings of Thirty Fourth Conference on Learning Theory.* Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3548–3626.

[Paq+23] Courtney Paquette et al. "Halting Time is Predictable for Large Models: A Universality Property and Average-Case Analysis". In: *Foundations of Computational Mathematics* 23.2 (2023), pp. 597–673.

[Pas05] Leonid Pastur. "A Simple Approach to the Global Regime of Gaussian Ensembles of Random Matrices". In: *Ukrainian Mathematical Journal* 57.6 (2005), pp. 936–966.

[PP21] Courtney Paquette and Elliot Paquette. "Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models". In: *Advances in Neural Information Processing Systems.* Vol. 34. Curran Associates, Inc., 2021, pp. 9229–9240.

[PS11] Leonid Pastur and Mariya Shcherbina. *Eigenvalue distribution of large random matrices.* 171. American Mathematical Soc., 2011.

[PS21] Vanessa Piccolo and Dominik Schröder. "Analysis of one-hidden-layer neural networks via the resolvent method". In: *Advances in Neural Information Processing Systems.* Vol. 34. Curran Associates, Inc., 2021, pp. 5225–5235.

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning". In: *Advances in Neural Information Processing Systems.* Vol. 30. Curran Associates, Inc., 2017.

[Ram+21] Aditya Ramesh et al. "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning.* Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8821–8831.

[Ram+22] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents.* 2022. arXiv: 2204.06125 [cs.CV].

[RR07] Ali Rahimi and Benjamin Recht. “Random Features for Large-Scale Kernel Machines”. In: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc., 2007.

[Sam59] Arthur L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229.

[Sch+23] Dominik Schröder et al. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 30285–30320.

[Sch+24] Dominik Schröder et al. *Asymptotics of Learning with Deep Structured (Random) Features*. 2024. arXiv: `2402.13999 [stat.ML]`.

[Ste81] Charles M. Stein. “Estimation of the Mean of a Multivariate Normal Distribution”. In: *The Annals of Statistics* 9.6 (1981), pp. 1135–1151.

[Tao12] Terence Tao. *Topics in Random Matrix Theory*. Vol. 132. Graduate Studies in Mathematics. American Mathematical Society, 2012.

[TAP21] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. “Overparameterization Improves Robustness to Covariate Shift in High Dimensions”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 13883–13897.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[WHS22] Alexander Wei, Wei Hu, and Jacob Steinhardt. “More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize”. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 23549–23588.

[Wig55] Eugene P. Wigner. “Characteristic Vectors of Bordered Matrices With Infinite Dimensions”. In: *Annals of Mathematics* 62.3 (1955), pp. 548–564.

[Wis28] John Wishart. “The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population”. In: *Biometrika* 20A (1928), pp. 32–52.

[WWF24] Zhichao Wang, Denny Wu, and Zhou Fan. *Nonlinear spiked covariance matrices and signal propagation in deep neural networks*. 2024. arXiv: `2402.10127 [stat.ML]`.

[WX20] Denny Wu and Ji Xu. "On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10112–10123.

[WZ23] Zhichao Wang and Yizhe Zhu. "Overparameterized Random Feature Regression with Nearly Orthogonal Data". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 8463–8493.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. arXiv: `1708.07747 [cs.LG]`.

[Zha+21] Chiyuan Zhang et al. "Understanding deep learning (still) requires rethinking generalization". In: *Commun. ACM* 64.3 (2021), pp. 107–115.

[ZP23] Jacob Zavatone-Veth and Cengiz Pehlevan. "Learning Curves for Deep Structured Gaussian Feature Models". In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 42866–42897.