UNIVERSITY OF CALIFORNIA SAN DIEGO

Spectral Properties of Neural Network Models

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics

by

Zhichao Wang

Committee in charge:

Professor Todd Kemp , Chair
Professor Ioana Dumitriu, Co-Chair
Professor Mikhail Belkin
Professor Alexander Cloninger
Professor Massimo Franceschetti

2024

The Dissertation of Zhichao Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

Dedicated to my parents and grandparents.

EPIGRAPH

Mathematics is the science of skillful operations with concepts and rules invented just for this purpose. The principal emphasis is on the invention of concepts. Mathematics would soon run out of interesting theorems if these had to be formulated in terms of the concepts which already appear in the axioms.

–Eugene Wigner

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to my advisors, Ioana Dumitriu and Todd Kemp, for their invaluable guidance and support throughout my PhD studies. Their expertise led me to explore the fascinating fields of random matrix theory and free probability theory, deepening my understanding of their extensive applications in machine learning. One of the key elements that has shaped my research journey has been the considerable freedom and support from Ioana and Todd. This has not only allowed me to dive deeper into deep learning theory but also to solidify my goal to pursue interdisciplinary research in mathematics and data science. This thesis would not have been possible without their assistance. Their support and encouragement have been instrumental in my development as an independent researcher. I am truly honored and fortunate to have had such outstanding mathematicians and, more importantly, such remarkable individuals as my advisors.

I also want to extend my deepest gratitude to my dissertation committee members, Mikhail Belkin, Alexander Cloninger, and Massimo Franceschetti, for their invaluable feedback and guidance throughout the course of this research. Special thanks to Mikhail Belkin for his mentorship and numerous insightful discussions that have significantly shaped both my research direction and career aspirations. Additionally, I am grateful to Michael Anshelevich for introducing me to free probability and random matrix theory. I am very fortunate to work with him. And thank you to Zhou Fan for sharing his wisdom in probability theory and statistics. Working with him has been a truly enlightening experience, and I consider myself very fortunate to have had this opportunity.

I thank my excellent collaborators: Zhou Fan, Yi Sun, Denny Wu, Yizhe Zhu, Andrew Engel, Natalie S. Frank, Ioana Dumitriu, Sutanay Choudhury, Anand Sarwate, Tony Chiang, Ethan Davis, David Jekel, Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, and Greg Yang. I have learned tremendously from many productive discussions with them.

I particularly want to thank my peers in mathematics, statistics, and machine learning: Yizhe Zhu, Haixiao Wang, Yubo Shuai, Dante Wu, Weiwei Wu, Yuyao Wang, Qingyuan Chen,

Sheng Qiao, Zhiyuan Jiang, Gregory Patchell, Collin Cranston, JD Flynn, Somak Maitra, Ryan Schneider, Libin Zhu, Daniel Beaglehole, Amirhesam Abedsoltan, Chaoyue Liu, Tianhao Wang, Simone Bombari, Alexander Wendel, Yaoqing Yang, Yilan Chen, Shuncheng Yuan, Wei Huang, Yiyun He, Yihan Zhang, and Denny Wu. I also want to thank researchers in probability and machine learning: Lucas Benigni, Courtney Paquette, Elliot Paquette, Xiucai Ding, Parthe Pandit, Sam Buchanan, Xiuyuan Cheng, Edgar Dobriban, Hamed Hassani, Boris Hanin, Murat A. Erdogdu, and Zhou Fan. I am grateful for their encouragement and help during my PhD study.

Finally, I thank my parents for their love and support.

Chapter 2 is extracted from "Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems* 33 (2020): 7710-7721". The thesis author is the co-author of this paper.

Chapter 3 is from the preprint "Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. arXiv:2402.10127 (2024)". The thesis author is the co-author of this paper.

Chapter 4 is extracted from "Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability* 34, no. 2 (2024): 1896-1947". The thesis author is the co-author of this paper.

Chapter 5 is extracted from a combination of "Zhichao Wang, Andrew Engel, Anand D. Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in Neural Information Processing Systems* 36 (2024)" and the preprint "Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. arXiv:2402.10127 (2024)". The thesis author is the co-author of these two papers.

VITA

2017	Bachelor of Science, Beihang University
2018	Master of Science, Texas A& M University
2019–2024	Teaching Assistant, Department of Mathematics Engineering University of California, San Diego
2021–2023	Research Assistant, Department of Mathematics Engineering University of California, San Diego
2021–2022	Research Intern, Pacific Northwest National Laboratory
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Zhichao Wang, Denny Wu, and Zhou Fan."Nonlinear spiked covariance matrices and signal propagation in deep neural networks." *arXiv:2402.10127* (2024), accepted by *The Thirty-Seventh Annual Conference on Learning Theory*.

Andrew Engel, Zhichao Wang, Natalie S. Frank, Ioana Dumitriu, Sutanay Choudhury, Anand Sarwate, and Tony Chiang. "Faithful and Efficient Explanations for Neural Networks via Neural Tangent Kernel Surrogate Models." *The Twelfth International Conference on Learning Representations* (2024).

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. "Learning in the presence of low-dimensional structure: a spiked random matrix perspective." *Advances in Neural Information Processing Systems* 36 (2024).

Zhichao Wang, Andrew Engel, Anand D. Sarwate, Ioana Dumitriu, and Tony Chiang. "Spectral evolution and invariance in linear-width neural networks." *Advances in Neural Information Processing Systems* 36 (2024).

Zhichao Wang and Yizhe Zhu. "Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks." *The Annals of Applied Probability* 34, no. 2 (2024): 1896-1947.

Zhichao Wang and Yizhe Zhu. "Overparameterized random feature regression with nearly orthogonal data." *In International Conference on Artificial Intelligence and Statistics*, pp. 8463-8493. PMLR, 2023.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. "High-dimensional asymptotics of feature learning: How one gradient step improves the representation." *Advances in Neural Information Processing* Systems 35 (2022): 37932-37946.

Ethan Davis, David Jekel, and Zhichao Wang. "Tree convolution for probability distributions with unbounded support." *Latin American Journal of Probability and Mathematical Statistics* (ALEA) 18.2 (2021), pp. 1585-1623.

Zhou Fan, Yi Sun, and Zhichao Wang. "Principal components in linear mixed models with general bulk." *The Annals of Statistics* 49, no. 3 (2021): 1489-1513.

Zhou Fan and Zhichao Wang. "Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks." *Advances in neural information processing systems* 33 (2020): 7710-7721.

Michael Anshelevich and Zhichao Wang. "Higher variations for free Lévy processes." *Studia Math.* 252 (2020), pp. 49-81.

ABSTRACT OF THE DISSERTATION

Spectral Properties of Neural Network Models

by

Zhichao Wang

Doctor of Philosophy in Mathematics

University of California San Diego, 2024

Professor Todd Kemp , Chair
Professor Ioana Dumitriu, Co-Chair

This thesis contributes to the emerging field of nonlinear random matrix theory and deep learning theory. The main contributions are summarized as follows:

In the linear-width regime, where the network widths scale proportionally with the sample size, we include proof of the global laws of the neural tangent kernel (NTK) and conjugate kernel (CK) matrices across layers. For datasets with low-dimensional signal structures, we characterize the outlier eigenvalues and eigenvector alignments of the CK matrices, extending recent results on spiked covariance models. In the ultra-wide regime, where the first layer's width is much larger than the sample size, we show that spectra of both CK and NTK matrices converge to a

deformed semicircle law.

Going beyond random initialization, we investigate the spectral properties of trained weight, CK, and NTK matrices through empirical and theoretical analyses. Empirically, it demonstrates the invariance of bulk spectra under small learning rates, the emergence of outliers with large learning rates, and the heavy-tailed distributions after adaptive gradient training, correlating them with feature learning. Theoretically, we prove the invariance of bulk spectra under small constant learning rates and characterize the feature learning phenomenon, where gradient descent optimizes the first-layer weights, leading to a rank-one spiked structure in the weight and CK matrices, with spike eigenvectors aligning with test labels.

Chapter 1

Introduction

Random matrix theory (RMT), free probability, and high-dimensional statistics have been recently employed successfully to study deep learning in a variety of cases. In this work, we aim to build a spectral analysis of neural networks [FW20, WZ24, WES$^+$23, BES$^+$22, WWF24], which may have important practical and theoretical consequences in deep learning. Generally speaking, there are three pillars of deep learning: neural network architectures, optimization and training dynamics, and data structures. In this thesis, we use high-dimensional probability theory to more realistically understand how these three pillars efficiently support neural networks (NNs) to achieve learning tasks:

(i) For **architectures**, we explore the impact of both the depth and width of fully connected neural network models. We consider different asymptotic scaling of the width of NNs.

(ii) For **data structures**, we go beyond isotropic Gaussian datasets to establish theory on datasets with anisotropic structures or even deterministic datasets with certain orthogonal properties.

(iii) For **optimization**, we analyze various training processes of NNs and how they learn features from training datasets and improve the representation learning in the neural network.

Modern datasets in areas like genomics, finance, and image processing usually have extremely large sizes in high dimensions which makes them difficult to process using traditional data processing techniques. In the context of machine learning and artificial intelligence, *big data* provides the extensive information needed to train artificial intelligence, improve accuracy,

and make data-driven decisions. Random matrix theory helps in studying the properties of these high-dimensional big data structures, by considering the sample size and feature dimension grow to infinity proportionally and then analyzing the eigenvalues and eigenvectors of matrices that represent the data. These properties can reveal important insights about the latent structure of high-dimensional data in large sizes.

Moreover, *overparameterization* (i.e., a large number of training parameters; [BHMM19, BHX20]) has become an indispensable component of artificial intelligence due to its presence in popular machine learning models, yet many properties of overparameterized estimators remain mysterious despite having been a major focus of recent research. To study the role of overparameterization in NNs, high-dimensional probability tools should be introduced to analyze the performances of NNs when sample size and width go to infinity with various rates. High-dimensional probability emerges as a potent tool in analyzing the asymptotic performances of overparameterized machine learning models.

Below, we will first summarize the background in deep learning theory and RMT, and then outline how our results harness the methodologies from random matrix theory to scrutinize spectral phenomena across diverse statistical and machine learning applications.

1.1 Neural Network Model and Kernel Matrices

In this section, we define the neural network model and its corresponding kernel matrices that will be discussed in this thesis.

Fully connected neural networks (NNs)

We define a fully-connected, feedforward neural network with input dimension d_0, hidden layers of dimensions $d_1, \ldots, d_L$, and a scalar output. For an input $\boldsymbol{x} \in \mathbb{R}^{d_0}$, we parametrize the network as

$$f_\theta(\boldsymbol{x}) = \boldsymbol{w}^\top \frac{1}{\sqrt{d_L}} \sigma\left(\boldsymbol{W}_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\ldots \frac{1}{\sqrt{d_2}} \sigma\Big(\boldsymbol{W}_2 \frac{1}{\sqrt{d_1}} \sigma(\boldsymbol{W}_1 \boldsymbol{x}) \Big) \Big) \right) \in \mathbb{R}. \tag{1.1.1}$$

Here, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function (applied entrywise to matrices or vectors) and

$$\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \quad \text{for } 1 \le \ell \le L, \qquad \boldsymbol{w} \in \mathbb{R}^{d_L}$$

are the network weights. We denote the collection of all training parameters by

$$\theta = (\mathrm{vec}(\boldsymbol{W}_1), \ldots, \mathrm{vec}(\boldsymbol{W}_L), \boldsymbol{w}),$$

where $\mathrm{vec}(\cdot)$ means vectorization of the weight matrix. We call $L \in \mathbb{N}$ the number of layers (the *depth*) of this NN and d_ℓ the number of neurons (the *width*) in ℓ-th layer for $1 \le \ell \le L$. Then, $\boldsymbol{W}_\ell$ is the weight matrix at the ℓ-th layer for $1 \le \ell \le L$.

The scalings by $1/\sqrt{d_\ell}$ in (1.1.1) reflect the "NTK-parametrization" of the network [JGH18]. We will discuss alternative scalings and an extension to multi-dimensional outputs in Sections 2.2.4 and 5.4.2.

Empirical kernel matrices in NN models

Kernel matrices associated with the nonlinear feature map of deep neural networks (NNs) provide insight into the optimization dynamics [JGH18, MZ20, FDP+20] and predictive performance [LBN+17, ADH+19a, OJMDF21]; consequently, properties of these kernel matrices can guide the design of network architecture [XBSD+18, MBD+21, LNR22] and learning algorithms [KO20, ZNB22]. Particular emphasis has been placed on the *spectral properties* of kernel matrices, due to their connection with the training and test performance of the underlying NN [BCP20, LGC+21a, WHS22]. This thesis will focus on two empirical kernel matrices on n training data points.

Given n training samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^{d_0}$, we denote the input data matrix by

$$\boldsymbol{X} \equiv \boldsymbol{X}_0 = \begin{pmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n \end{pmatrix} \in \mathbb{R}^{d_0 \times n}. \tag{1.1.2}$$

When passing through these n samples in (1.1.1), we can define the output matrix of post-activation at each layer by

$$\boldsymbol{X}_\ell = \frac{1}{\sqrt{d_\ell}}\sigma(\boldsymbol{W}_\ell \boldsymbol{X}_{\ell-1}) \in \mathbb{R}^{d_\ell \times n}, \qquad \text{for } \ell = 1\ldots, L, \tag{1.1.3}$$

with weight matrices $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, $\boldsymbol{X}_0 \equiv \boldsymbol{X}$ and $d_0 \equiv d$, and a nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ applied entrywise. The in-sample predictions of the network in (1.1.1) are given by

$$f_\theta(\boldsymbol{X}) = (f_\theta(\boldsymbol{x}_1), \ldots, f_\theta(\boldsymbol{x}_n)) = \boldsymbol{w}^\top \boldsymbol{X}_L \in \mathbb{R}^{1\times n}.$$

Definition 1. The **Conjugate Kernel (CK)** at ℓ-th layer of NN in (1.1.1) with n sample dataset $\boldsymbol{X}_0$ in (1.1.2) is defined by

$$\boldsymbol{K}_\ell^{\mathrm{CK}} = \boldsymbol{X}_\ell^\top \boldsymbol{X}_\ell \in \mathbb{R}^{n\times n}, \tag{1.1.4}$$

for $\ell = 1, \ldots, L$. We denote the last layer CK as $\boldsymbol{K}^{\mathrm{CK}}$ for simplicity.

This $\boldsymbol{K}^{\mathrm{CK}}$ is also called the equivalent Gaussian process kernel [Nea95, Wil97, CS09, DFS16, PLR$^+$16, SGGSD17, LBN$^+$18, MHR$^+$18]. Fixing the matrix $\boldsymbol{X}_L$, this $\boldsymbol{K}^{\mathrm{CK}}$ governs training and generalization in the linear regression model $f_\theta(\boldsymbol{X}) = \boldsymbol{w}^\top \boldsymbol{X}_L$. For very wide networks, $\boldsymbol{K}^{\mathrm{CK}}$ may be viewed as an approximation of its infinite-width limit, and regression using $\boldsymbol{X}_L$ is an approximation of regression in the Reproducing Kernel Hilbert Space (RKHS) defined by this limit kernel [RR08]. In this thesis, we use "conjugate kernel" and "neural tangent kernel" to refer to these matrices for a finite-width network, rather than their infinite-width limits. In the high-dimensional asymptotic setting where the width of the NN and the number of training samples diverge at the same rate, prior works employed random matrix theory to analyze the limit eigenvalue distribution of the CK matrix at random initialization [PW17, LLC18, Péc19, FW20]. These and related characterizations of the CK resolvent enable precise computations of various errors for NNs with random first-layer weights, known as *random features models* [MM22, TAP21, HJ22].

Definition 2. The **Neural Tangent Kernel (NTK)** is the Gram matrix of the Jacobian of in-sample predictions with respect to the network weights

$$\boldsymbol{K}^{\mathrm{NTK}} = J^\top J = \big(\nabla_\theta f_\theta(\boldsymbol{X})\big)^\top \big(\nabla_\theta f_\theta(\boldsymbol{X})\big) \in \mathbb{R}^{n\times n}, \tag{1.1.5}$$

where we denote the Jacobian matrix of the network predictions with respect to the weights θ as

$$J = \nabla_\theta f_\theta(\boldsymbol{X}) = \begin{pmatrix} \nabla_\theta f(\boldsymbol{x}_1) & \cdots & \nabla_\theta f(\boldsymbol{x}_n) \end{pmatrix} \in \mathbb{R}^{\dim(\theta)\times n}.$$

This K^{NTK} was introduced to study network training [JGH18, DZPS19a, AZLS19]. Under gradient-flow training dynamics, the in-sample predictions follow a differential equation governed by the NTK. Under gradient-flow training of the network with weights θ and training loss given by (1.1.7), the time evolutions of residual errors and in-sample predictions are given respectively by

$$\frac{d}{dt}\Big(\boldsymbol{y} - f_{\theta(t)}(\boldsymbol{X})\Big) = -\boldsymbol{K}^{\mathrm{NTK}}(t)\cdot\Big(\boldsymbol{y} - f_{\theta(t)}(\boldsymbol{X})\Big), \quad \frac{d}{dt} f_{\theta(t)}(\boldsymbol{X}) = \boldsymbol{K}^{\mathrm{NTK}}(t)\cdot\Big(\boldsymbol{y} - f_{\theta(t)}(\boldsymbol{X})\Big)$$

where $\theta(t)$ and $K^{\mathrm{NTK}}(t)$ are the parameters and NTK at training time t [JGH18, DZPS19a]. Denoting the eigenvalues and eigenvectors of $\boldsymbol{K}^{\mathrm{NTK}}(t)$ by $(\lambda_\alpha(t), \boldsymbol{v}_\alpha(t))_{\alpha=1}^n$, and the spectral components of the residual error by $r_\alpha(t) = \boldsymbol{v}_\alpha(t)^\top(\boldsymbol{y} - f_{\theta(t)}(\boldsymbol{X}))$, these training dynamics are expressed spectrally as

$$\boldsymbol{v}_\alpha(t)^\top \frac{d}{dt}\Big(\boldsymbol{y} - f_{\theta(t)}(\boldsymbol{X})\Big) = -\lambda_\alpha(t) r_\alpha(t), \qquad \frac{d}{dt} f_{\theta(t)}(\boldsymbol{X}) = \sum_{\alpha=1}^{n} \lambda_\alpha(t) r_\alpha(t)\cdot \boldsymbol{v}_\alpha(t).$$

Note that these relations hold instantaneously at each training time t, regardless of whether $\boldsymbol{K}^{\mathrm{NTK}}(t)$ evolves or remains approximately constant over training. Hence, $\lambda_\alpha(t)$ controls the instantaneous rate of decay of the residual error in the direction of $\boldsymbol{v}_\alpha(t)$.

For very wide networks, $\boldsymbol{K}^{\mathrm{NTK}}$, λ_α, and $\boldsymbol{v}_\alpha$ will all approximately stay close to the initial

values over the entirety of training [JGH18, DZPS19a, DLL$^+$19b, AZLS19, COB19]. This yields the closed-form solution $r_\alpha(t)\alpha \approx r_\alpha(0)e^{-t\lambda_\alpha}$, so that the in-sample predictions $f_{\theta(t)}(\boldsymbol{X})$ converge exponentially fast to the observed training labels $\mathbf{y}$, with a different exponential rate λ_α along each eigenvector $\boldsymbol{v}_\alpha$ of $\boldsymbol{K}^{\mathrm{NTK}}$.

The spectral decompositions of these kernel matrices are related to the training and generalization properties of the underlying network. Training occurs most rapidly along the eigenvectors of the largest eigenvalues [ASS20], and the eigenvalue distribution may determine the trainability of the model and the extent of implicit bias towards simpler functions [XPS19, YS19a]. It is thus of interest to understand the spectral properties of these matrices, using random matrix theory, both at random initialization and throughout training.

As an example, when $L = 2$, let $d_1 = h$ and $d_0 = d$ be the widths of the output and input layer. The CK is defined as $\boldsymbol{K}^{\mathrm{CK}} := \boldsymbol{X}_1^\top \boldsymbol{X}_1 \in \mathbb{R}^{n\times n}$, where $\boldsymbol{X}_1 := \frac{1}{\sqrt{h}}\sigma\left(\boldsymbol{W}\boldsymbol{X}/\sqrt{d}\right)$. Specifically, the empirical NTK of two-layer NN can be explicitly written as

$$\boldsymbol{K}^{\mathrm{NTK}} = \frac{1}{d}\boldsymbol{X}^\top\boldsymbol{X} \odot \frac{1}{h}\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X}\right)^\top \mathrm{diag}(\boldsymbol{v})^2\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X}\right) + \boldsymbol{K}^{\mathrm{CK}}. \tag{1.1.6}$$

Here we train both layers, so we have two parts in the NTK expression. If we only train the first-hidden layer, we can simply remove the second CK part.

Training processes of NNs.

We focus on the NTK parameterization and consider the kernel machine (1.1.9) induced by the initial NTK of the NN. We aim to seek the cases when the NN outperforms this kernel during the training process. For this purpose, we adopt different optimizers of training this NN to obtain different testing performances and spectral properties of trained weights and empirical kernels.

The loss function for training is a mean squared error (MSE)

$$\mathscr{L}(\theta) := \frac{1}{2n}\|\boldsymbol{y} - f_\theta(\boldsymbol{X})\|^2. \tag{1.1.7}$$

NNs are usually trained by gradient-based methods such as full-batch gradient descent (GD), mini-batch stochastic gradient descent (SGD), Adaptive Gradients (AdaGrad), and Adam [KB14]. We can represent GD by

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L}(\theta_t), \tag{1.1.8}$$

where η is the learning rate (step size) and $\nabla_\theta \mathscr{L}(\theta_t)$ is the gradient of the training loss w.r.t. trainable parameters θ at step $t \geq 0$. We will prove the global convergence of GD in some special (overparameterized) cases ensuring the convergence to the NN that interpolates the data. We will also show that the hyper-parameters (e.g. learning rate η) affect the spectral properties of NNs during training.

Lazy training regime.

Lazy training [COB19] can be viewed as a linear approximation of the NN, i.e.

$$f_\theta(\boldsymbol{x}) \approx f_{\theta_0}(\boldsymbol{x}) + (\theta - \theta_0)^\top \nabla_\theta f_{\theta_0}(\boldsymbol{x}),$$

defined by minimum-norm interpolation

$$\hat{\theta} := \arg\min \Big\{ \|\theta - \theta_0\| : (\theta - \theta_0)^\top \nabla_\theta f_{\theta_0}(\boldsymbol{X}) = \boldsymbol{y} - f_{\theta_0}(\boldsymbol{X}) \Big\}.$$

Then, lazy training also represents a kernel machine

$$\hat{f}(\boldsymbol{x}) = f_{\theta_0}(\boldsymbol{x}) + \big(\boldsymbol{y} - f_{\theta_0}(\boldsymbol{X})\big) \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}) \tag{1.1.9}$$

where $\hat{f}(\boldsymbol{x})$ is the unregularized regression prediction on test data $\boldsymbol{x} \in \mathbb{R}^d$, the kernel $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ is the initial $\boldsymbol{K}^{\mathrm{NTK}}$ on training data, and $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{x}) = (\nabla_\theta f_{\theta_0}(\boldsymbol{X}))^\top (\nabla_\theta f_{\theta_0}(\boldsymbol{x}))$. The asymptotic performance of $\hat{f}(\boldsymbol{x})$ has been analyzed by [AP20] under the LWR. We view this regime as a *benchmark*: [COB19, BMR21] prove that NN through gradient flow is close to lazy training if $h \gg n$; [HXAP20] shows NN can go beyond lazy training under a non-proportional regime.

1.2 An Overview of Random Matrix Theory

1.2.1 Empirical Spectral Distributions and Stieltjes Transforms

We will derive almost sure weak limits for the empirical spectral distributions (ESDs) of random symmetric kernel matrices $K \in \mathbb{R}^{n\times n}$ as $n \to \infty$. Throughout this thesis, we will denote this as

$$\lim \operatorname{spec} K = \mu$$

where μ is the limit probability distribution on $\mathbb{R}$. Letting $\{\lambda_\alpha\}_{\alpha=1}^n$ be the eigenvalues of K, this means

$$\frac{1}{n}\sum_{\alpha=1}^{n} f(\lambda_\alpha) \to \mathbb{E}_{\lambda\sim\mu}[f(\lambda)] \tag{1.2.1}$$

a.s. as $n \to \infty$, for any continuous bounded function $f : \mathbb{R} \to \mathbb{R}$. Intuitively, this may be understood as the convergence of the "bulk" of the spectral distribution of K. we call this convergence as global law of K. To be clear, this does not imply convergence of the extreme eigenvalues of K to the support of μ, which is a stronger notion of convergence for global law. We will also show that $\|K\| \le C$ a.s., for a constant $C > 0$ and all large n. Then (1.2.1) in fact holds for any continuous function $f : \mathbb{R} \to \mathbb{R}$, as such a function must be bounded on $[-C, C]$.

We will characterize the probability distribution μ and the ESD of $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ by their Stieltjes transforms. These are defined, respectively, for $z \in \mathbb{C}^+$, as

$$m_\mu(z) = \int \frac{1}{x-z} d\mu(x), \qquad m_K(z) = \frac{1}{n}\sum_{\alpha=1}^{n} \frac{1}{\lambda_\alpha - z} = \frac{1}{n}\operatorname{Tr}(K - z\,\mathrm{Id})^{-1}.$$

Here Id represents the identity matrix. The pointwise convergence $m_K(z) \to m_\mu(z)$ a.s. as $n \to \infty$ over $z \in \mathbb{C}^+$ implies $\lim \operatorname{spec} K = \mu$. For $z = x + i\eta \in \mathbb{C}^+$, the value $\pi^{-1} \operatorname{Im} m_\mu(z)$ is the density function of the convolution of μ with the distribution Cauchy$(0, \eta)$ at $x \in \mathbb{R}$. Hence, the function $m_\mu(z)$ uniquely defines μ, and evaluating $\pi^{-1} \operatorname{Im} m_\mu(x + i\eta)$ for small $\eta > 0$ yields an approximation for the density function of μ (provided this density exists at x).

For any $n \times n$ Hermitian matrix A_n, the Stieltjes transform of the empirical spectral distribution of A_n can be written as $\operatorname{tr}(A_n - z\,\mathrm{Id})^{-1}$. We call $(A_n - z\,\mathrm{Id})^{-1}$ the resolvent of A_n.

1.2.2 Deformed Marčenko-Pastur Law

An example of limiting eigenvalue distribution is given by the *Marčenko-Pastur law*, which describes the spectra of sample covariance matrices [MP67a]. Back in 1928, [Wis28] first studied this kind of random matrix model in statistical analysis.

Let $X \in \mathbb{R}^{d\times n}$ have i.i.d. $\mathcal{N}(0,1/d)$ entries, let $\Phi \in \mathbb{R}^{n\times n}$ be deterministic and positive semi-definite, and let $n \to \infty$ such that $\lim \operatorname{spec} \Phi = \nu$ and $n/d \to \gamma \in (0,\infty)$. Then the sample covariance matrix $\Phi^{1/2} X^\top X \Phi^{1/2}$ has an almost sure spectral limit,

$$\lim \operatorname{spec} \Phi^{1/2} X^\top X \Phi^{1/2} = \rho_\gamma^{\mathrm{MP}} \boxtimes \nu. \tag{1.2.2}$$

We call this limit $\rho_\gamma^{\mathrm{MP}} \boxtimes \nu$ the Marčenko-Pastur map of ν with aspect ratio γ, where $\boxtimes$ is an analog of the classical notion of multiplicative convolution of probability measures, called *free multiplicative convolution* in free probability theory. This distribution $\rho_\gamma^{\mathrm{MP}} \boxtimes \nu$ can be defined by its Stieltjes transform $m(z)$, which solves the Marčenko-Pastur fixed point equation [MP67a]. We define this fixed point equation below.

This free multiplicative convolution is from free harmonic analysis. For full descriptions of free independence and free multiplicative convolution, see [NS06, Lecture 18] and [AGZ10, Section 5.3.3]. The free multiplicative convolution $\boxtimes$ was first introduced in [Voi87], which later has many applications for products of asymptotic free random matrices. The main tool for computing free multiplicative convolution is the S-transform, invented by [Voi87]. S-transform was recently utilized to study the dynamical isometry of deep neural networks [PSG17, PSG18, XBSD$^+$18, HK21, CH23]. Some basic properties and intriguing examples for free multiplicative convolution with μ_s can also be found in [BZ10, Theorems 1.2, 1.3].

For a probability measure ν supported on $[0,\infty)$ and an aspect ratio parameter $\gamma > 0$,

consider the deformed Marčenko–Pastur measure

$$\mu = \rho_\gamma^{\mathrm{MP}} \boxtimes \nu$$

and its "companion" probability measure

$$\tilde{\mu} = \gamma\mu + (1-\gamma)\delta_0.$$

In general, μ and $\tilde{\mu}$ represent the limit eigenvalue distributions of

$$\boldsymbol{G}^\top \boldsymbol{G} \in \mathbb{R}^{n\times n}, \quad \text{and} \quad \boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{N\times N}$$

respectively, when $\boldsymbol{G} = \frac{1}{\sqrt{N}}[\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N] \in \mathbb{R}^{N\times n}$ has i.i.d. rows with mean 0 and covariance $\boldsymbol{\Sigma}$, and $n, N \to \infty$ with $n/N \to \gamma$ and $\frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\boldsymbol{\Sigma})} \to \nu$ weakly.

These measures $\mu, \tilde{\mu}$ can be defined by their Stieltjes transforms

$$m(z) = \int \frac{1}{x-z} d\mu(x), \qquad \tilde{m}(z) = \int \frac{1}{x-z} d\tilde{\mu}(x) \tag{1.2.3}$$

where $\tilde{m}(z) = \gamma m(z) + (1-\gamma)(-1/z)$. By the results of [MP67a, SB95], for any $z \in \mathbb{C}^+$, $m(z)$ and $\tilde{m}(z)$ are the unique solutions in $\{m \in \mathbb{C} : \gamma m + (1-\gamma)(-1/z) \in \mathbb{C}^+\}$ and $\mathbb{C}^+$, respectively, to the *Marčenko-Pastur equations*

$$m(z) = \int \frac{1}{\lambda(1-\gamma-\gamma z m(z)) - z} d\nu(\lambda), \quad z = -\frac{1}{\tilde{m}(z)} + \gamma \int \frac{\lambda}{1+\lambda\tilde{m}(z)} d\nu(\lambda). \tag{1.2.4}$$

We define $m(z)$ and $\tilde{m}(z)$ via (1.2.3) also on the full domains $\mathbb{C} \setminus \mathrm{supp}(\mu)$ and $\mathbb{C} \setminus \mathrm{supp}(\tilde{\mu})$ respectively. Notice that the support sets $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\tilde{\mu})$ may differ only at the single point $\{0\}$.

In the setting $\boldsymbol{\Sigma} = \mathrm{Id}$ (and $\nu = \delta_1$), the law $\mu = \rho_\gamma^{\mathrm{MP}}$ is the standard Marčenko-Pastur law

[MP67a], with explicit density function with respect to Lebesgue measure

$$d\rho_{\gamma}^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\gamma\lambda} \cdot \mathbf{1}_{\lambda\in[\lambda_-,\lambda_+]}\, d\lambda, \qquad \lambda_\pm := (1 \pm \sqrt{\gamma})^2$$

for $\gamma \leq 1$, and an additional point mass $(1-1/\gamma)$ at 0 when $\gamma > 1$. Here $\mathbf{1}_{\lambda\in[\lambda_-,\lambda_+]}$ is the indicator function on the subset $[\lambda_-, \lambda_+] \subset \mathbb{R}$.

In general, μ and $\tilde{\mu}$ may not have analytically explicit densities. However, $\mathrm{supp}(\tilde{\mu})$ is explicitly characterized in [SC95], and we review this characterization here: Define

$$\mathscr{T} = \{0\} \cup \{-1/\lambda : \lambda \in \mathrm{supp}(\nu)\}. \tag{1.2.5}$$

For $\tilde{m} \in \mathbb{C} \setminus \mathscr{T}$, define

$$z(\tilde{m}) = -\frac{1}{\tilde{m}} + \gamma \int \frac{\lambda}{1+\lambda\tilde{m}} d\nu(\lambda). \tag{1.2.6}$$

In light of the second equation of (1.2.4), this may be understood as a formal inverse of $\tilde{m}(z)$. From [SC95, Theorems 4.1 and 4.2], we have the following properties.

Proposition 3. $\tilde{m}(\cdot)$ defines a bijection from $\{z \in \mathbb{R} \setminus \mathrm{supp}(\tilde{\mu})\}$ to $\{\tilde{m} \in \mathbb{R} \setminus \mathscr{T} : z'(\tilde{m}) > 0\}$, whose inverse function is $z(\cdot)$. In particular, $x \in \mathbb{R}$ does not belong to $\mathrm{supp}(\tilde{\mu})$ if and only if there exists $\tilde{m} \in \mathbb{R} \setminus \mathscr{T}$ such that $z'(\tilde{m}) > 0$ and $z(\tilde{m}) = x$.

1.2.3 Deformed Semicircle Law

The *semicircle law* is another fundamental distribution in random matrix theory. It describes the limiting distribution of eigenvalues for many types of random symmetric matrices as the matrix size grows to infinity. The Wigner semicircle law Theorem [Wig55] states that the empirical distribution of the eigenvalues of a symmetric $n \times n$ random matrix with independent and identically distributed (i.i.d.) entries (up to the symmetry constraint) converges to a semicircle distribution as n approaches infinity. This type of random matrix is called the Wigner matrix in random matrix literature. If the entries of the matrix are centered (mean zero) with variance

σ^2/n, the density of the eigenvalue distribution in the limit is given by the density function:

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}\cdot \mathbf{1}_{\{|\lambda|\leq 2\sigma\}}$$

where $\mathbf{1}$ is the indicator function, restricting λ to the interval $[-2\sigma, 2\sigma]$. We denote the standard semicircle law as μ_s with density function $\rho(\lambda)$ and $\sigma = 1$. We refer to [Gui09, BS10, AGZ10, Tao12] for more details on the Wigner matrix, semicircle law μ_s and its relation with free probability.

Similarly with deformed Marčenko-Pastur Law, we can define the deformed semicircle law $\mu = \mu_s \boxtimes \mu_\Phi$ for a probability measure μ_Φ supported on $[0, +\infty)$. This probability measure μ can be characterized by its Stieltjes transform $m(z)$ governed by the following equation

$$m(z) + \int \frac{d\mu_\Phi(x)}{z + \beta(z)x} = 0 \tag{1.2.7}$$

for each $z \in \mathbb{C}^+$, where $\beta(z) \in \mathbb{C}^+$ is the unique solution to

$$\beta(z) + \int \frac{x d\mu_\Phi(x)}{z + \beta(z)x} = 0. \tag{1.2.8}$$

When $\mu_\Phi = \delta_1$, μ is exactly the *standard* semicircle law. For more details on this deformed semicircle law, see Chapter 4.

1.3 Summary of Contribution

Recently, nonlinear random matrix theory has emerged in neural networks (1.1.1) at random initialization, e.g. [PW17, LLC18, BP21]. This thesis contributes to this field by deriving the limiting spectral distributions of neural network kernel matrices in (1.1.4) and (1.1.5) for different regimes and general dataset X. Furthermore, we analyze spectral properties of weight matrices in (1.1.1) and neural network kernel matrices in (1.1.4) and (1.1.5) during the training

processes of neural network functions. We summarize the main contribution of this thesis below.

Global law for the linear-width regime in Chapter 2.

First, we considered the *linear-width* (proportional) regime where all the widths are scaled with the sample size n, i.e. $n/d_\ell \to \gamma_\ell$. In Chapter 2, we prove the limiting spectral distribution of CK_L can be obtained by a recursive sequence of deformed Marčenko-Pastur distributions $\mu_\ell = \rho_{\gamma_\ell}^{\mathrm{MP}} \boxtimes ((1-b_\sigma^2) + b_\sigma^2 \mu_{\ell-1})$ where $b_\sigma = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma'(\xi)]$ and $\rho_{\gamma_\ell}^{\mathrm{MP}}$ is the Marčenko-Pastur distribution with aspect ratio γ_ℓ, for $1 \le \ell \le L$. Here μ_0 is the limiting law of the initial data matrix $X^\top X$. For this model, we proved the limiting spectrum of NTK is a linear combination of limiting spectra of CK_ℓ for all $1 \le \ell \le L$. Our analysis shows how spectra propagate through multiple layers of a random neural network under the linear-width regime. The dataset we considered is very general with only some approximate orthogonality; for example, the CIFAR-10 image dataset can be covered in this case. Moreover, our proofs provide new techniques to extend the analysis of a single hidden layer to multi-layer networks and deeper NTKs, in a systematic and recursive way.

Spike analysis for linear-width regime in Chapter 3.

Following the global law results in Chapter 2, we studied the outlier eigenvalues in CK_l when the dataset X has some spikes that often capture the low-dimensional signal structure of the learning problem. Following the setup of Chapter 2, we consider the CK matrix defined by (1.1.4) in an L-layer fully connected NN at random initialization under the linear-width regime. Given spiked input data, we compute the magnitude of the eigenvalues in CK_ℓ outside the bulk distribution μ_ℓ and the alignments between the corresponding CK eigenvectors with those of the input data, across network depth $1 \le \ell \le L$.

To quantify these spike eigenvalues in CK_ℓ, we develop a new result of the signal eigenvalues and eigenvectors of spiked covariance matrices with arbitrary and possibly nonlinear dependence across features, showing a "universality" with the quantitative spectral properties of linear spiked covariance models established by [BY12]. We prove a deterministic equivalent for

the Stieltjes transform and resolvent for any spectral argument z separated from the support of the limit spectral measure, extending recent results for spectral arguments bounded away from the positive real line [Cho22, Cho23, SCDL23]. Using this, we characterize the BBP-type phase transition [BAP05] and first-order limits of the eigenvalues and eigenvector alignments in the proportional asymptotics regime, for spike eigenvalues of bounded size.

Global law for the ultra-wide regime in Chapter 4.

In Chapter 4, we focus on the *ultra-wide* regime for $L = 1$, where the width d_1 of the first layer is much larger than the sample size n. Under similar assumptions on X as [FW20], we showed that as $d_1/n \to \infty$ and $n \to \infty$, a deformed semicircle law $\mu_s \boxtimes ((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_0)$ emerges for both normalized CK and NTK, where μ_s is a semicircular distribution. In the proof, we first established necessary random matrix results for generalized sample covariance matrices with some dependency and a nonlinear Hanson-Wright inequality that is suitable for neural networks with random weights and Lipschitz activation functions. We also demonstrated non-asymptotic concentrations of CK and NTK around their limiting kernels in the spectral norm, along with sharper lower bounds on their smallest eigenvalues, which are useful in deep learning theory [DLL$^+$19a, OS20, MZ20, BMR21]. As an application, we showed that random feature regression induced by CK and NTK achieves the same asymptotic performance as its limiting kernel regression under the ultra-wide regime. This allows us to calculate the precise asymptotic training and test errors for random feature regression using the corresponding kernel.

Empirical and theoretical analysis of trained weight and kernel matrices in Chapter 5.

Going beyond the random initialization of NNs, we study the spectral properties of NNs in (1.1.1) with $L = 1$ after certain training processes under the linear-width regime, empirically and theoretically, respectively.

Empirically, we show that the spectra of the weight matrix, CK, and NTK are invariant when trained by gradient descent for small constant learning rates. We demonstrate similar characteristics when training with stochastic gradient descent with small learning rates. When the

learning rate is large, we exhibit the emergence of an outlier whose corresponding eigenvector is aligned with the training data structure. We also show that, where a lower test error and feature learning emerge after adaptive gradient training, both weight and kernel matrices exhibit heavy tail behavior. Simple examples are provided to empirically explain when heavy tails can have better generalizations. We exhibit different spectral properties such as invariant bulk, spike, and heavy-tailed distribution from a two-layer neural network using different training strategies, and then correlate them to the feature learning. Analogous phenomena also appear when we train conventional neural networks with real-world data. We conclude that monitoring the evolution of the spectra during training is an essential step toward understanding the training dynamics and feature learning.

Theoretically, we prove the invariance of the bulk spectra for both CK and NTK when the NN is trained by gradient descent for small constant learning rates. However, these invariant spectra impede the NN from learning informative features from the dataset during the training process. Therefore, we further investigate the phenomenon of emergence of an outlier we empirically observed. With proper scaling of NN and learning rate in [BES$^+$22], we prove the feature learning of NN, where the first-layer weights in a two-layer NN are optimized by gradient descent, and the learned weight matrix exhibits a rank-one spiked structure. Applying the spiked sample covariance result in Chapter 3, we characterize the spiked eigenstructure of the corresponding CK matrix for independent test data, and the alignment of spike eigenvectors with the test labels. This provides a quantitative description of how gradient descent improves the NN representation.

1.4 Notation

Let $[n] = \{i = 1, \ldots, n\}$ for any $n \in \mathbb{N}$. We use $\mathrm{tr}(A) = \frac{1}{n}\sum_i A_{ii}$ as the normalized trace of a matrix $A \in \mathbb{R}^{n\times n}$ and $\mathrm{Tr}(A) = \sum_i A_{ii}$. Denote vectors by lowercase boldface. Throughout this thesis, $\boldsymbol{v}^*$ and M^* denote the conjugate transpose, $\|\cdot\|$ denotes the ℓ_2 norm for vectors (Euclidean

norm), $\ell_2 \to \ell_2$ is the operator norm for matrices, i.e. $\|M\| = \sup_{\boldsymbol{v}\in\mathbb{C}^n:\|\boldsymbol{v}\|=1} \|M\boldsymbol{v}\|$, while $\|\cdot\|_F$ is the Frobenius norm, i.e. $\|M\|_F = (\mathrm{Tr}M^*M)^{1/2} = (\sum_{\alpha,\beta} |M_{\alpha\beta}|^2)^{1/2}$. We define the entry-wise 2-∞ matrix norm as

$$\|M\|_{2,\infty} := \max_{1\le i\le N} \|\mathbf{m}_i\|,$$

for any matrix $M \in \mathbb{R}^{N\times d}$ with the i-th row $\mathbf{m}_i \in \mathbb{R}^d$ and $1 \le i \le N$. We denote $\|\cdot\|_\infty$ as the ℓ_∞ norm for vectors. Note that we have

$$|\mathrm{tr}M| \le \|M\| \le \|M\|_F, \qquad \|M\|_F \le \sqrt{n}\|M\|, \qquad |\mathrm{tr}AB| \le n^{-1}\|A\|_F\|B\|_F, \tag{1.4.1}$$

for any matrices $M, A, B \in \mathbb{R}^{n\times n}$. Denote that $A \odot B$ is the Hadamard product of two matrices, i.e., $(A\odot B)_{ij} = A_{ij}B_{ij}$ for any $i, j \in [n]$. Given any vector $\boldsymbol{v}$, $\mathrm{diag}(\boldsymbol{v})$ is a diagonal matrix where the main diagonal elements are given by $\boldsymbol{v}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of any Hermitian matrix A, respectively. $s_{\max}(A)$ is the largest singular value of any matrix A. Let $\boldsymbol{I}_n$ be the $n\times n$ identity matrix. For simplicity, we may ignore the dimension n and denote $\boldsymbol{I}$ as the identity matrix. We will also use Id to denote the identity matrix.

Let $\mathbb{E}_{\boldsymbol{w}}[\cdot]$ and $\mathrm{Var}_{\boldsymbol{w}}[\cdot]$ be the expectation and variance only with respect to random vector $\boldsymbol{w}$. Also, let $o_{d,\mathbb{P}}(\cdot)$ represent little-o in probability as $d \to \infty$. For a probability measure μ, its support is the closed set

$$\mathrm{supp}(\mu) = \{x \in \mathbb{R} : \mu(O) > 0 \text{ for any open neighborhood } O \ni x\}.$$

We write $\mathrm{dist}(x, A) = \inf\{|x-y| : y \in A\}$ and define the ε-neighborhood

$$\mathrm{supp}(\mu) + (-\varepsilon, \varepsilon) = \{x \in \mathbb{R} : \mathrm{dist}(x, \mathrm{supp}(\mu)) < \varepsilon\}.$$

We write δ_x for the probability measure given by a point mass at $x \in \mathbb{R}$, and the linear transformation $a \otimes \mu \oplus b$ for the law of $a\mathsf{X} + b$ when $\mathsf{X} \sim \mu$ and $a, b \in \mathbb{R}$.

Chapter 2

Deformed Marčenko-Pastur Law for Linear-Width Multi-Layer NNs

In this Chapter, we apply techniques of random matrix theory to derive an exact asymptotic characterization of the eigenvalue distributions of the CK and NTK at random initialization, in a multi-layer feedforward network architecture. We study a "linear-width" asymptotic regime, where each hidden layer has a width proportional to the training sample size. We impose an assumption of approximate pairwise orthogonality for the training samples, which encompasses general settings of independent samples that need not have independent entries.

We show that the eigenvalue distributions for both the CK and the NTK converge to deterministic limits, depending on the limiting eigenvalue distribution of the training data. The limit distribution for the CK at each intermediate hidden layer is a Marčenko-Pastur map of a linear transformation of that of the previous layer. The NTK can be approximated by a linear combination of CK matrices, and its limiting eigenvalue distribution can be described by a recursively defined sequence of fixed point equations that extend this Marčenko-Pastur map. We demonstrate the agreement of these asymptotic limits with the observed spectra on both synthetic and CIFAR-10 training data of moderate size.

In this linear-width asymptotic regime, feature learning occurs, and both the CK and NTK evolve over training. Although our theory pertains only to their spectra at random initialization of the weights, we conclude with an empirical examination of their spectral evolutions during

training, on simple examples of learning a single neuron and learning a binary classifier for two classes in CIFAR-10. In these examples, the bulk eigenvalue distributions of the CK and NTK undergo elongations, and isolated principal components emerge that are highly predictive of the training labels. Recent theoretical work has studied the evolution of the NTK in an entrywise sense [HY19, DGA20], and we believe it is an interesting open question to translate this understanding to a more spectral perspective.

2.1 Related Work

Many properties of the CK and NTK have been established in the limit of infinite width and fixed sample size n. In this limit, both the CK [Nea95, Wil97, DFS16, LBN+18, MHR+18] and the NTK [JGH18, LXS+19, Yan19] at random initialization converge to fixed $n \times n$ kernel matrices. The associated random features regression models converge to kernel linear regression in the RKHS of these limit kernels. Furthermore, network training occurs in the lazy regime [COB19], where the NTK remains constant throughout training [JGH18, DZPS19a, DLL+19b, AZLS19, LXS+19, ADH+19b]. Spectral properties of the CK, NTK, and Hessian matrix of the training loss have been previously studied in this infinite-width limit in [PLR+16, SEG+17, XPS19, KAA19, GSd+19, JGH19]. Limitations of lazy training and these equivalent kernel regression models have been studied theoretically and empirically in [COB19, ADH+19b, YS19b, GMMM19, GMMM21, LRZ19], suggesting that trained neural networks of practical width are not fully described by this type of infinite-width kernel equivalence. The asymptotic behavior in the linear-width regime is different from the infinite-width regime: For example, for a linear activation $\sigma(x) = x$, the infinite-width limit of the CK for random weights is the input Gram matrix $X^\top X$, whereas its limit spectrum under linear-width asymptotics has an additional noise component from iterating the Marčenko-Pastur map.

Under linear-width asymptotics, the limit CK spectrum for one hidden layer was characterized in [PW17] for training data with i.i.d. Gaussian entries. For activations satisfying

$\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma'(\xi)] = 0$, [PW17] conjectured that this limit is a Marčenko-Pastur law also in multi-layer networks, and this was proven under a subgaussian assumption as part of the results of [BP21]. [LLC18] studied the one-hidden-layer CK with general training data, and [LC18b] specialized this to Gaussian mixture models. These works [LLC18, LC18b] showed that the limit spectrum is a Marčenko-Pastur map of the inter-neuron covariance. We build on this insight by analyzing this covariance across multiple layers, under approximate orthogonality of the training samples. This orthogonality condition is similar to that of [ALP22], which recently studied the one-hidden-layer CK with a bias term. This condition is also more general than the assumption of i.i.d. entries, and [FW20] describes the reduction to the one-hidden-layer result of [PW17] for i.i.d. Gaussian inputs, as this reduction is not immediately clear. [Péc19] provides another form of the limit distribution in [PW17], which is equivalent to our form via the relation described in [BG10].

The limit NTK spectrum for a one-hidden-layer network with i.i.d. Gaussian inputs was recently characterized in parallel work of [AP20]. In particular, [AP20] applied the same idea as in Lemma 7 below to study the Hadamard product arising in the NTK. [PB17, PW18] previously studied the equivalent spectrum of a sample covariance matrix derived from the network Jacobian, which is one of two components of the Hessian matrix of the training loss, in a slightly different setting and also for one hidden layer.

The spectra of the kernel matrices $X^\top X$ that we study are equivalent (up to the addition/removal of 0's) to the spectra of the sample covariance matrices in linear regression using the features X. As developed in a line of recent literature including [Dic16, PW17, DW18, LLC18, LC18a, HMRT22, MM22, AP20, dRBK20], this spectrum and the associated Stieltjes transform and resolvent are closely related to the training and generalization errors in this linear regression model. These works have collectively provided an asymptotic understanding of training and generalization errors for random features regression models derived from the CK and NTK of one-hidden-layer neural networks, and related qualitative phenomena of double and multiple descent in the generalization error curves.

2.2 Main Results

2.2.1 Additional Notation and Assumptions

In this Chapter, we use Greek indices α, β, etc. for samples in $\{1,\dots,n\}$, and Roman indices i, j, etc. for neurons in $\{1,\dots,d\}$. For a matrix $X \in \mathbb{R}^{d\times n}$, we denote by $\mathbf{x}_\alpha$ its α^{th} column and by $\mathbf{x}_i^\top$ its i^{th} row.

Definition 4. Let $\varepsilon, B > 0$. A matrix $X \in \mathbb{R}^{d\times n}$ is (ε, B)**-orthonormal** if its columns satisfy, for every $\alpha \neq \beta \in \{1,\dots,n\}$,

$$\big|\|\mathbf{x}_\alpha\|^2 - 1\big| \le \varepsilon, \qquad \big|\mathbf{x}_\alpha^\top \mathbf{x}_\beta\big| \le \varepsilon, \qquad \|X\| \le B, \qquad \sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|^2 - 1)^2 \le B^2.$$

Assumption 1. The number of layers $L \ge 1$ is fixed, and $n, d_0, d_1, \dots, d_L \to \infty$ with $n/d_\ell \to \gamma_\ell$ for constants $\gamma_\ell \in (0,\infty)$ and each $\ell = 1,2,\dots,L$, such that

(a) The weights $\theta = (\mathrm{vec}(W_1),\dots,\mathrm{vec}(W_L), \boldsymbol{w})$ are i.i.d. and distributed as $\mathcal{N}(0,1)$.

(b) The activation $\sigma(x)$ is twice differentiable, with $\sup_{x\in\mathbb{R}} |\sigma'(x)|, |\sigma''(x)| \le \lambda_\sigma$ for some $\lambda_\sigma < \infty$. For $\xi \sim \mathcal{N}(0,1)$, we have $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma^2(\xi)] = 1$.

(c) The input $X \in \mathbb{R}^{d_0\times n}$ is (ε_n, B)-orthonormal in the sense of Definition 4, where B is a constant, and $\varepsilon_n n^{1/4} \to 0$ as $n \to \infty$.

(d) As $n \to \infty$, $\lim \mathrm{spec}\, X^\top X = \mu_0$ for a probability distribution μ_0 on $[0,\infty)$.

Part (c) quantifies our assumption of approximate pairwise orthogonality of the training samples. Although not completely general, it encompasses many settings of independent samples with input dimension $d_0 \asymp n$, including:

- Gaussian inputs $\mathbf{x}_\alpha \sim \mathcal{N}(0,\Sigma)$, for any Σ satisfying $\mathrm{Tr}\,\Sigma = 1$ and $\|\Sigma\| \lesssim 1/n$.
- Inputs $\mathbf{x}_\alpha$ drawn from certain multi-class Gaussian mixture models, in the high-dimensional asymptotic regimes that were studied in [CBG16, LLC18, LC18b, LC18a, LC19].

- Inputs that may be expressed as $\sqrt{d_0} \cdot \mathbf{x}_\alpha = f(\mathbf{z}_\alpha)$, where $\mathbf{z}_\alpha \in \mathbb{R}^m$ has independent entries satisfying a log-Sobolev inequality, and $f : \mathbb{R}^m \to \mathbb{R}^{d_0}$ is any Lipschitz function.

In particular, the limit spectral law μ_0 in Assumption 1(d) can be very different from the Marčenko-Pastur spectrum that would correspond to X having i.i.d. entries. This approximate orthogonality is implied by the following more technical convex concentration property, which is discussed further in [VW15, Ada15]. We prove this result in Section 2.3.

Proposition 5. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent training samples satisfying $\mathbb{E}[\mathbf{x}_\alpha] = 0$ and $\mathbb{E}[\|\mathbf{x}_\alpha\|^2] = 1$. Suppose, for some constant $c_0 > 0$, that $d_0 \geq c_0 n$, and each vector $\sqrt{d_0} \cdot \mathbf{x}_\alpha$ satisfies the convex concentration property

$$\mathbb{P}\Big[\big|\varphi(\sqrt{d_0} \cdot \mathbf{x}_\alpha) - \mathbb{E}\varphi(\sqrt{d_0} \cdot \mathbf{x}_\alpha)\big| \geq t\Big] \leq 2e^{-c_0 t^2}$$

for every $t > 0$ and every 1-Lipschitz convex function $\varphi : \mathbb{R}^{d_0} \to \mathbb{R}$. Then for any $k > 0$, with probability $1 - n^{-k}$, X is $(\sqrt{\frac{C \log n}{d_0}}, B)$-orthonormal for some $C, B > 0$ depending only on c_0, k.

In Assumptions 1(a) and (b), the scaling of θ and the conditions $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma^2(\xi)] = 1$, together with the parametrization (1.1.1), ensure that all pre-activations have approximate mean 0 and variance 1. This may be achieved in practice by batch normalization [IS15]. For $\xi \sim \mathcal{N}(0,1)$, we define the following constants associated to $\sigma(x)$:

$$b_\sigma = \mathbb{E}[\sigma'(\xi)], \quad a_\sigma = \mathbb{E}[\sigma'(\xi)^2], \quad q_\ell = (b_\sigma^2)^{L-\ell}, \quad r_\ell = a_\sigma^{L-\ell}, \quad r_+ = \sum_{\ell=0}^{L-1} r_\ell - q_\ell. \quad (2.2.1)$$

We verify in Proposition 11 that under Assumption 1(b), we have $b_\sigma^2 \leq 1 \leq a_\sigma$. These parameters in (2.2.1) related to activation function σ will be employed to characterize the limiting spectra of empirical CK and NTK matrices in the following sections.

2.2.2 Spectrum of the Conjugate Kernel

Recall the Marčenko-Pastur map (1.2.2). Let $\mu_1, \mu_2, \mu_3, \ldots$ be the sequence of probability distributions on $[0, \infty)$ defined recursively by

$$\mu_\ell = \rho_{\gamma_\ell}^{\mathrm{MP}} \boxtimes \Big((1 - b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_{\ell-1} \Big). \tag{2.2.2}$$

Here, μ_0 is the input limit spectrum in Assumption 1(d), b_σ is defined in (2.2.1), and $(1 - b_\sigma^2) \oplus b_\sigma^2 \otimes \mu$ denotes the translation and rescaling of μ that is the distribution of $(1 - b_\sigma^2) + b_\sigma^2 \mathsf{X}$ when $\mathsf{X} \sim \mu$.

The following theorem shows that these distributions $\mu_1, \mu_2, \mu_3, \ldots$ are the asymptotic limits of the empirical eigenvalue distributions of the CK across the layers. Thus, the limit distribution for each layer ℓ is a Marčenko-Pastur map of a translation and rescaling of that of the preceding layer $\ell - 1$.

Theorem 6. *Suppose Assumption 1 holds, and define $\mu_1, \ldots, \mu_L$ by (2.2.2). Then (marginally) for each $\ell = 1, \ldots, L$, we have* $\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell$. *In particular,*

$$\lim \operatorname{spec} \boldsymbol{K}^{CK} = \mu_L.$$

Furthermore, $\|\boldsymbol{K}^{CK}\| \leq C$ a.s. for a constant $C > 0$ and all large n.

If $\sigma(x)$ is such that $b_\sigma = 0$, then each distribution μ_ℓ is simply the Marčenko-Pastur law $\rho_{\gamma_\ell}^{\mathrm{MP}}$. This special case was previously conjectured in [PW17] and proven in [BP21], for input data X with i.i.d. entries. Note that for such non-linearities, the limiting CK spectrum does not depend on the spectrum μ_0 of the input data, and furthermore $\mu_1 = \ldots = \mu_L$ if the layers have the same width $d_1 = \ldots = d_L$. Implications of this for the network discrimination ability in classification tasks and for learning performance have been discussed previously in [CBG16, PW17, LLC18, LC19, ALP22].

To connect Theorem 6 to our next result on the NTK, let us describe the iteration (2.2.2) more explicitly using a recursive sequence of fixed point equations derived from the Marčenko-Pastur equation (1.2.4): Let $m_\ell(z)$ be the Stieltjes transform of μ_ℓ, and define

$$\tilde{t}_\ell(z_{-1}, z_\ell) = \lim_{n\to\infty} \frac{1}{n} \operatorname{Tr}(z_{-1}\operatorname{Id} + z_\ell X_\ell^\top X_\ell)^{-1} = \frac{1}{z_\ell} m_\ell\left(-\frac{z_{-1}}{z_\ell}\right).$$

Applying the Marčenko-Pastur equation (1.2.4) to $m_\ell(-z_{-1}/z_\ell)$, and introducing $\tilde{s}_\ell(z_{-1}, z_\ell) = [z_\ell(1-\gamma_\ell+\gamma_\ell z_{-1}\tilde{t}_\ell(z_{-1}, z_\ell))]^{-1}$, one may check that (2.2.2) may be written as a pair of equations

$$\tilde{t}_\ell(z_{-1}, z_\ell) = \tilde{t}_{\ell-1}\left(z_{-1} + \frac{1-b_\sigma^2}{\tilde{s}_\ell(z_{-1}, z_\ell)},\ \frac{b_\sigma^2}{\tilde{s}_\ell(z_{-1}, z_\ell)}\right), \tag{2.2.3}$$

$$\tilde{s}_\ell(z_{-1}, z_\ell) = (1/z_\ell) + \gamma_\ell\Big(\tilde{s}_\ell(z_{-1}, z_\ell) - z_{-1}\tilde{s}_\ell(z_{-1}, z_\ell)\tilde{t}_\ell(z_{-1}, z_\ell)\Big), \tag{2.2.4}$$

where (2.2.4) is a rearrangement of the definition of $\tilde{s}_\ell$. Applying (2.2.3) to substitute $\tilde{t}_\ell(z_{-1}, z_\ell)$ in (2.2.4), the equation (2.2.4) is a fixed point equation that defines $\tilde{s}_\ell$ in terms of $\tilde{t}_{\ell-1}$. Then (2.2.3) defines $\tilde{t}_\ell$ in terms of $\tilde{s}_\ell$ and $\tilde{t}_{\ell-1}$. The limit Stieltjes transform for $\boldsymbol{K}^{\mathrm{CK}}$ is the specialization $m_{\mathrm{CK}}(z) = \tilde{t}_L(-z, 1)$.

2.2.3 Spectrum of the Neural Tangent Kernel

In the neural network model (1.1.1), an application of the chain rule yields an explicit form for NTK in (1.1.5)

$$\boldsymbol{K}^{\mathrm{NTK}} = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1})$$

for certain matrices $S_\ell \in \mathbb{R}^{d_\ell \times n}$, where $\odot$ is the Hadamard (entrywise) product. We refer to Section 2.8.1 for the exact expression; see also [HY19, Eq. (1.7)]. Our spectral analysis of $\boldsymbol{K}^{\mathrm{NTK}}$ relies on the following approximation, which shows that the limit spectrum of $\boldsymbol{K}^{\mathrm{NTK}}$ is equivalent to a linear combination of the CK matrices $X_0^\top X_0, \ldots, X_L^\top X_L$ and Id. We prove this in

Section 2.8.1.

Lemma 7. *Under Assumption 1, letting r_+ and q_ℓ be as defined in (2.2.1),*

$$\lim \operatorname{spec} \boldsymbol{K}^{NTK} = \lim \operatorname{spec} \Big(r_+ \operatorname{Id} + X_L^\top X_L + \sum_{\ell=0}^{L-1} q_\ell X_\ell^\top X_\ell \Big).$$

By this lemma, if $b_\sigma = 0$, then $q_0 = \ldots = q_{L-1} = 0$ and the limit spectrum of $\boldsymbol{K}^{\mathrm{NTK}}$ reduces to the limit spectrum of $r_+ \operatorname{Id} + X_L^\top X_L$ which is a translation of $\rho_{\gamma_L}^{\mathrm{MP}}$ described in Theorem 6. In the following, Thus we assume that $b_\sigma \neq 0$. Our next result provides an analytic description of the limit spectrum of $\boldsymbol{K}^{\mathrm{NTK}}$, by extending (2.2.3) and (2.2.4) to characterize the trace of rational functions of $X_0^\top X_0, \ldots, X_L^\top X_L$ and Id.

Denote the closed lower-half complex plane with 0 removed as $\mathbb{C}^* = \overline{\mathbb{C}^-} \setminus \{0\}$. For $\ell = 0, 1, 2, \ldots$, we define recursively two sequences of functions

$$\begin{aligned} t_\ell &: (\mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*) \times \mathbb{C}^{\ell+2} \to \mathbb{C}, && (\mathbf{z}, \boldsymbol{w}) \mapsto t_\ell(\mathbf{z}, \boldsymbol{w}), \\ s_\ell &: \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^* \to \mathbb{C}^+, && \mathbf{z} \mapsto s_\ell(\mathbf{z}). \end{aligned}$$

where $\mathbf{z} = (z_{-1}, z_0, \ldots, z_\ell) \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*$ and $\boldsymbol{w} = (w_{-1}, w_0, \ldots, w_\ell) \in \mathbb{C}^{\ell+2}$. We will define these functions such that $t_\ell(\mathbf{z}, \boldsymbol{w})$ will be the value of

$$\lim_{n \to \infty} n^{-1} \operatorname{Tr}(z_{-1} \operatorname{Id} + z_0 X_0^\top X_0 + \ldots + z_\ell X_\ell^\top X_\ell)^{-1} (w_{-1} \operatorname{Id} + w_0 X_0^\top X_0 + \ldots + w_\ell X_\ell^\top X_\ell).$$

For $\ell = 0$, we define the first function t_0 by

$$t_0\Big((z_{-1}, z_0), (w_{-1}, w_0)\Big) = \int \frac{w_{-1} + w_0 x}{z_{-1} + z_0 x} d\mu_0(x) \tag{2.2.5}$$

For $\ell \geq 1$, we then define the functions s_ℓ and t_ℓ recursively by

$$s_\ell(\mathbf{z}) = (1/z_\ell) + \gamma_\ell t_{\ell-1}\big(\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}),\mathbf{z}),\, (1-b_\sigma^2, 0, \ldots, 0, b_\sigma^2)\big), \tag{2.2.6}$$

$$t_\ell(\mathbf{z},\boldsymbol{w}) = (w_\ell/z_\ell) + t_{\ell-1}\big(\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}),\mathbf{z}),\, \boldsymbol{w}_{\text{prev}}\big) \tag{2.2.7}$$

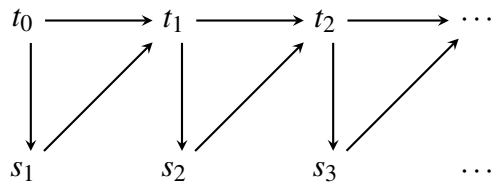
where we write as shorthand

$$\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}),\mathbf{z}) \equiv \left(z_{-1} + \frac{1-b_\sigma^2}{s_\ell(\mathbf{z})}, z_0, \ldots, z_{\ell-2}, z_{\ell-1} + \frac{b_\sigma^2}{s_\ell(\mathbf{z})}\right) \in \mathbb{C}^- \times \mathbb{R}^{\ell-1} \times \mathbb{C}^*, \tag{2.2.8}$$

$$\boldsymbol{w}_{\text{prev}} \equiv (w_{-1}, \ldots, w_{\ell-1}) - (w_\ell/z_\ell)\cdot(z_{-1}, \ldots, z_{\ell-1}) \in \mathbb{C}^{\ell+1}. \tag{2.2.9}$$

Proposition 8. Suppose $b_\sigma \neq 0$. For each $\ell \geq 1$ and any $\mathbf{z} \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*$, there is a unique solution $s_\ell(\mathbf{z}) \in \mathbb{C}^+$ to the fixed point equation (2.2.6).

Hence, (2.2.6) defines the function s_ℓ in terms of the function $t_{\ell-1}$, and this is then used in (2.2.7) to define t_ℓ. This is illustrated diagrammatically as



Specializing the function t_L for the last layer L to the values $(z_{-1}, z_0, \ldots, z_{L-1}, z_L) = (r_+, q_0, \ldots, q_{L-1}, 1)$ and $(w_{-1}, w_0, \ldots, w_L) = (1, 0, \ldots, 0)$, we obtain an analytic description for the limit spectrum of $\boldsymbol{K}^{\text{NTK}}$ via its Stieltjes transform.

Theorem 9. *Suppose $b_\sigma \neq 0$. Under Assumption 1, for any fixed values $z_{-1}, z_0, \ldots, z_L \in \mathbb{R}$ where $z_L \neq 0$, we have* $\lim \operatorname{spec}(z_{-1}\,\mathrm{Id} + z_0 X_0^\top X_0 + \ldots + z_L X_L^\top X_L) = \nu$ *where ν is the probability distribution with Stieltjes transform $m_\nu(z) = t_L((-z + z_{-1}, z_0, \ldots, z_L), (1, 0, \ldots, 0))$.*

In particular, $\lim \operatorname{spec} \boldsymbol{K}^{NTK}$ *is the probability distribution with Stieltjes transform*

$$m_{NTK}(z) = t_L\Big((-z+r_+, q_0, \ldots, q_{L-1}, 1), (1,0,\ldots,0)\Big).$$

Furthermore, $\|\boldsymbol{K}^{NTK}\| \leq C$ *a.s. for a constant* $C > 0$ *and all large* n*.*

We remark that Theorem 9 encompasses the previous result in Theorem 6 for $\boldsymbol{K}^{\mathrm{CK}} = X_L^\top X_L$, by specializing to $(z_0, \ldots, z_{L-1}, z_L) = (0, \ldots, 0, 1)$. Under this specialization,

$$s_\ell(z_{-1}, 0, \ldots, 0, z_\ell) = \tilde{s}_\ell(z_{-1}, z_\ell),$$
$$t_\ell((z_{-1}, 0, \ldots, 0, z_\ell), (1, 0, \ldots, 0)) = \tilde{t}_\ell(z_{-1}, z_\ell),$$

and (2.2.6) and (2.2.7) reduce to (2.2.3) and (2.2.4), respectively.

2.2.4 Multi-dimensional Outputs and Rescaled Parametrizations

Theorem 9 pertains to a network with scalar outputs, under the "NTK-parametrization" of network weights in (1.1.1). As neural network models used in practice often have multi-dimensional outputs and may be parametrized differently for backpropagation, we state here the extension of the preceding result to a network with k-dimensional output and a general scaling of the weights.

Consider the model

$$f_\theta(\mathbf{x}) = W_{L+1}^\top \frac{1}{\sqrt{d_L}} \sigma\bigg(W_L \frac{1}{\sqrt{d_{L-1}}} \sigma\Big(\ldots \frac{1}{\sqrt{d_2}} \sigma\Big(W_2 \frac{1}{\sqrt{d_1}} \sigma(W_1 \mathbf{x})\Big)\Big)\bigg) \in \mathbb{R}^k \tag{2.2.10}$$

where $W_{L+1}^\top \in \mathbb{R}^{k \times d_L}$. We write the coordinates of f_θ as $(f_\theta^1, \ldots, f_\theta^k)$, and the vectorized output for all training samples $X \in \mathbb{R}^{d_0 \times n}$ as $f_\theta(X) = (f_\theta^1(X), \ldots, f_\theta^k(X)) \in \mathbb{R}^{nk}$. We consider the NTK

$$\boldsymbol{K}^{\mathrm{NTK}} = \sum_{\ell=1}^{L+1} \tau_\ell \Big(\nabla_{W_\ell} f_\theta(X)\Big)^\top \Big(\nabla_{W_\ell} f_\theta(X)\Big) \in \mathbb{R}^{nk \times nk}. \tag{2.2.11}$$

For $\tau_1 = \ldots = \tau_{L+1} = 1$, this is a flattening of the NTK defined in [JGH18], and we recall briefly its derivation from gradient-flow training in Section 2.9.1. We consider general constants $\tau_1, \ldots, \tau_{L+1} > 0$ to allow for a different learning rate for each weight matrix W_ℓ, which may arise from backpropagation in the model (2.2.10) using a parametrization with different scalings of the weights.

Theorem 10. *Fix any $k \geq 1$. Suppose Assumption 1 holds, and $b_\sigma \neq 0$. Then $\|\boldsymbol{K}^{NTK}\| \leq C$ a.s. for a constant $C > 0$ and all large n, and $\operatorname{lim\,spec} \boldsymbol{K}^{NTK}$ is the probability distribution with Stieltjes transform*

$$m_{NTK}(z) = t_L\Big((-z + \tau \cdot r_+,\ \tau_1 q_0, \ldots, \tau_L q_{L-1}, \tau_{L+1}), (1, 0, \ldots, 0)\Big), \quad \tau \cdot r_+ \equiv \sum_{\ell=0}^{L-1} \tau_{\ell+1}(r_\ell - q_\ell).$$

2.3 Proof of Proposition 5

We prove Proposition 5. For convenience, in this section, we denote the input dimension d_0 simply as d, and we denote the rescaled input by $\widetilde{X} = \sqrt{d}X$, with columns $\tilde{\mathbf{x}}_\alpha = \sqrt{d} \cdot \mathbf{x}_\alpha$.

Bound for $\|\tilde{\mathbf{x}}_\alpha\|^2$: Note that $\mathbb{E}[\|\tilde{\mathbf{x}}_\alpha\|^2] = d$. Applying the convex concentration property and [Ada15, Theorem 2.5] with $A = \mathrm{Id}$, we have for any $t > 0$ that

$$\mathbb{P}\Big[\big|\|\tilde{\mathbf{x}}_\alpha\|^2 - d\big| > t\Big] \leq 2\exp\left(-c\min\left(\frac{t^2}{d}, t\right)\right) \tag{2.3.1}$$

for a constant c depending only on c_0. Applying this for $t = \sqrt{Kd\log n}$ and a union bound, with probability $1 - 2ne^{-cK\log n}$,

$$\Big|\|\tilde{\mathbf{x}}_\alpha\|^2 - d\Big| \leq \sqrt{Kd\log n} \quad \text{for all } \alpha \in [n]. \tag{2.3.2}$$

Rescaling, this shows $|\|\mathbf{x}_\alpha\|^2 - 1| \leq \sqrt{(K\log n)/d}$.

Bound for $\tilde{\mathbf{x}}_\alpha^\top \tilde{\mathbf{x}}_\beta$: Since $\tilde{\mathbf{x}}_\alpha$ and $\tilde{\mathbf{x}}_\beta$ are independent, conditional on $\tilde{\mathbf{x}}_\beta$, we have $\mathbb{E}[\tilde{\mathbf{x}}_\alpha^\top \tilde{\mathbf{x}}_\beta \mid \tilde{\mathbf{x}}_\beta] = 0$, and the map $\tilde{\mathbf{x}}_\alpha \mapsto \tilde{\mathbf{x}}_\alpha^\top \tilde{\mathbf{x}}_\beta$ is convex and $\|\tilde{\mathbf{x}}_\beta\|$-Lipschitz. Then the convex concentration

property implies, for any $t > 0$,

$$\mathbb{P}\left[|\tilde{\mathbf{x}}_\alpha^\top \tilde{\mathbf{x}}_\beta| > t \middle| \tilde{\mathbf{x}}_\beta\right] \le 2e^{-c_0 t^2/\|\tilde{\mathbf{x}}_\beta\|^2}.$$

On the event (2.3.2), applying this for $t = \sqrt{Kd\log n}$, this probability is at most $2e^{-cK\log n}$. Taking a union bound, with probability $1 - 2n^2 e^{-cK\log n}$,

$$\left|\tilde{\mathbf{x}}_\alpha^\top \tilde{\mathbf{x}}_\beta\right| \le \sqrt{Kd\log n} \quad \text{for all } \alpha \ne \beta \in [n].$$

Rescaling, this shows $|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \le \sqrt{(K\log n)/d}$.

Bound for $\|\widetilde{X}\|$: Fix any unit vector $\boldsymbol{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$. By [KR19, Lemma C.11], the random vector $\widetilde{X}\boldsymbol{v}$ also satisfies the convex concentration property, with a modified constant c_0'. Note that $\mathbb{E}[\|\widetilde{X}\boldsymbol{v}\|^2] = d\|\boldsymbol{v}\|^2 = d$. Then, as in (2.3.1), we have

$$\mathbb{P}\left[|\|\widetilde{X}\boldsymbol{v}\|^2 - d| > t\right] \le 2\exp\left(-c\min\left(\frac{t^2}{d}, t\right)\right).$$

Applying this with $t = (B^2/4 - 1)d$, and taking a union bound over a $1/2$-net $\mathbb{N}$ of the unit ball $\{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\| = 1\}$ with cardinality 5^n, we have with probability at least $1 - 5^n \cdot 2e^{-cB^2 d}$ that

$$\|\widetilde{X}\boldsymbol{v}\| \le (B/2)\sqrt{d} \quad \text{for all } \boldsymbol{v} \in \mathbb{N}.$$

Since

$$\|\widetilde{X}\| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \|\widetilde{X}\boldsymbol{v}\| \le \sup_{\boldsymbol{v}\in\mathbb{N}} \|\widetilde{X}\boldsymbol{v}\| + \|\widetilde{X}\|/2,$$

we have $\|\widetilde{X}\| \le B\sqrt{d}$ on this event. Rescaling, this shows $\|X\| \le B$.

Bound for $\sum_{\alpha=1}^n (\|\tilde{\mathbf{x}}_\alpha\|^2 - d)^2$: Define $\mathbf{z} = (z_1, \dots, z_n)$ where $\mathbf{z}_\alpha = \|\tilde{\mathbf{x}}_\alpha\|^2 - d$. Fixing any unit vector $\boldsymbol{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$, let us first bound $\boldsymbol{v}^\top \mathbf{z}$: We have

$$\boldsymbol{v}^\top \mathbf{z} = \sum_{\alpha=1}^n v_\alpha (\|\tilde{\mathbf{x}}_\alpha\|^2 - d),$$

which has mean 0. Note that integrating the tail bound (2.3.1) yields the sub-exponential condition

$$\mathbb{E}\big[\exp\big(\lambda(\|\tilde{\mathbf{x}}_\alpha\|^2 - d)\big)\big] \le \exp(Cd\lambda^2) \quad \text{for all } |\lambda| \le c'$$

and some constants $C, c' > 0$. (See e.g. [BLM13, Theorem 2.3], applied with $(v, c) = (C'd, C')$ and a large enough constant $C' > 0$.) Then, as $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n$ are independent and $\|\boldsymbol{v}\|^2 = 1$, also

$$\mathbb{E}[e^{\lambda \boldsymbol{v}^\top \mathbf{z}}] = \mathbb{E}\left[\exp\left(\lambda \sum_{\alpha=1}^n v_\alpha (\|\tilde{\mathbf{x}}_\alpha\|^2 - d)\right)\right] \le \exp(Cd\lambda^2) \quad \text{for all } |\lambda| \le c'.$$

For any $t > 0$, applying this with $\lambda = \min(t/(2Cd), c')$ yields the sub-exponential tail bound

$$\mathbb{P}[\boldsymbol{v}^\top \mathbf{z} \ge t] \le e^{-\lambda t} \mathbb{E}[e^{\lambda \boldsymbol{v}^\top \mathbf{z}}] \le \exp\left(-c \min\left(\frac{t^2}{d}, t\right)\right).$$

Now applying this for $t = (B/2)d$, and again taking a union bound over a $1/2$-net $\mathbb{N}$ of the unit ball, we have with probability $1 - 5^n \cdot e^{-cBd}$ that

$$\boldsymbol{v}^\top \mathbf{z} \le (B/2)d \quad \text{for all } \boldsymbol{v} \in \mathbb{N}.$$

On this event, we have as above that $\|\mathbf{z}\| \le Bd$, so $\|\mathbf{z}\|^2 \le B^2 d^2$. Rescaling, this shows $\sum_{\alpha=1}^n (\|\tilde{\mathbf{x}}_\alpha\|^2 - 1)^2 \le B^2$.

Applying all of the above bounds for sufficiently large constants $K, B > 0$, we obtain that these bounds hold with probability at least $1 - n^{-k}$, which yields Proposition 5.

2.4 Overview of Proofs and Preliminary Lemmas

The proofs of Theorems 6, 9, and 10 are contained in the subsequent Appendices 2.5–2.9. We provide here an outline of the argument.

We will apply induction across the layers $\ell = 1, \dots, L$, analyzing the post-activation matrix X_ℓ of each layer conditional on the previous post-activations $X_0, \dots, X_{\ell-1}$ (i.e. with respect to only the randomness of W_ℓ). For the Conjugate Kernel, this will entail analyzing the Stieltjes transform

$$\frac{1}{n}\operatorname{Tr}(X_L^\top X_L - z\,\mathrm{Id})^{-1}$$

conditional on the previous layers. For the Neural Tangent Kernel, given the approximation in Lemma 7, this will entail analyzing the Stieltjes transform

$$\frac{1}{n}\operatorname{Tr}(A + X_L^\top X_L - z\,\mathrm{Id})^{-1}$$

conditional on the previous layers, where A is a linear combination of $X_0^\top X_0, \dots, X_{L-1}^\top X_{L-1}$, and Id. Note that this matrix A is deterministic conditional on the previous layers.

In Section 2.5, we carry out a non-asymptotic analysis of (ε, B)-orthonormality. In particular, we show that if the deterministic input $X \equiv X_0$ is (ε, B)-orthonormal, then X_1 is $(C\varepsilon, CB)$-orthonormal with high probability, for a constant $C > 0$ depending only on λ_σ. Note that we require the fourth technical condition

$$\sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|^2 - 1)^2 \le B^2$$

in Definition 4 to ensure that the operator norm $\|X_1\|$ remains of constant order, as otherwise X_1 may have a rank-one component whose norm grows slowly with n. Applying this result conditionally for every layer, Assumption 1 then implies that $X_0, \dots, X_L$ are all $(\tilde{\varepsilon}_n, \tilde{B})$-orthonormal for modified parameters $(\tilde{\varepsilon}_n, \tilde{B})$ with high probability.

In Section 2.6, we carry out the analysis of the trace

$$\frac{1}{n}\mathrm{Tr}(A+\alpha X_1^\top X_1 - z\,\mathrm{Id})^{-1}$$

in a single layer, for a deterministic (ε_n, B)-orthonormal input X_0, symmetric matrix $A \in \mathbb{R}^{n\times n}$, and spectral parameters $\alpha \in \mathbb{C}^* \equiv \overline{\mathbb{C}^-} \setminus 0$ and $z \in \mathbb{C}^+$. We allow $\alpha \in \mathbb{C}^*$ (rather than fixing $\alpha = 1$), as the subsequent induction argument for the NTK will require this extension. When $A = 0$ and $\alpha = 1$, this reduces to the analysis in [LLC18], and also mirrors the proof of the Marčenko-Pastur equation (1.2.4). For $A \neq 0$, this trace will depend jointly on A and the second-moment matrix $\Phi_1 \in \mathbb{R}^{n\times n}$ for the rows of X_1. We derive a fixed point equation in terms of A and Φ_1, which approximates this trace in the $n \to \infty$ limit.

In Section 2.7, we prove Theorem 6 on the CK, by specializing this analysis to the setting $A = 0$ and $\alpha = 1$. The inductive loop is closed via an entrywise approximation of the second-moment matrix Φ_ℓ in each layer by a linear combination of $X_{\ell-1}^\top X_{\ell-1}$ and Id in the previous layer. The main argument for this approximation has been carried out in Section 2.5.

In Section 2.8, we prove Theorem 9 on the NTK. Our analysis reduces the trace of any linear combination of $X_0^\top X_0, \ldots, X_L^\top X_L$ and Id to the trace of a more general rational function of $X_0^\top X_0, \ldots, X_{L-1}^\top X_{L-1}$ and Id in the previous layer. In order to close the inductive loop, we analyze the trace of such a rational function across layers and show that it may be characterized by the recursive fixed point equations (2.2.6) and (2.2.7). In Section 2.8, we also establish the approximation in Lemma 7 and the existence and uniqueness of the fixed point to (2.2.6).

Finally, in Section 2.9, we prove Theorem 10, which is a minor extension of Theorem 9.

Before presenting all the proofs, let us first collect here a few basic results, which we will use in the subsequent sections.

Proposition 11. Under Assumption 1(b), the constants a_σ and b_σ in (2.2.1) satisfy

$$|b_\sigma| \le 1 \le \sqrt{a_\sigma} \le \lambda_\sigma.$$

For a universal constant $C > 0$, the activation function σ satisfies

$$|\sigma(x)| \le \lambda_\sigma(|x| + C) \qquad \text{for all } x \in \mathbb{R}. \tag{2.4.1}$$

Proof. It is clear from the definition that $a_\sigma \le \lambda_\sigma^2$. By the Gaussian Poincaré inequality,

$$1 = \mathbb{E}[\sigma(\xi)^2] = \mathrm{Var}[\sigma(\xi)] \le \mathbb{E}[\sigma'(\xi)^2] = a_\sigma.$$

By Gaussian integration-by-parts and Cauchy-Schwarz,

$$|b_\sigma| = |\mathbb{E}[\sigma'(\xi)]| = |\mathbb{E}[\xi \cdot \sigma(\xi)]| \le \mathbb{E}[\xi^2]^{1/2}\mathbb{E}[\sigma(\xi)^2]^{1/2} = 1.$$

We have

$$|\sigma(0)| \le \mathbb{E}[|\sigma(0) - \sigma(\xi)|] + \mathbb{E}[|\sigma(\xi)|] \le \lambda_\sigma \mathbb{E}[|\xi|] + \mathbb{E}[\sigma(\xi)^2]^{1/2} \le C\lambda_\sigma \tag{2.4.2}$$

(the last inequality applying $\lambda_\sigma \ge 1$). Then $|\sigma(x)| \le |\sigma(0)| + \lambda_\sigma |x| \le \lambda_\sigma(|x| + C)$. □

Proposition 12. Suppose $M = U + iV \in \mathbb{C}^{n\times n}$, where the real and imaginary parts $U, V \in \mathbb{R}^{n\times n}$ are symmetric, and V is invertible with either $V \succeq c_0 \,\mathrm{Id}$ or $V \preceq -c_0 \,\mathrm{Id}$ for a value $c_0 > 0$. Then M is invertible, and $\|M^{-1}\| \le 1/c_0$.

Proof. For any unit vector $\boldsymbol{v} \in \mathbb{C}^n$,

$$\|M\boldsymbol{v}\| = \|M\boldsymbol{v}\| \cdot \|\boldsymbol{v}\| \ge |\boldsymbol{v}^* M \boldsymbol{v}| = |\boldsymbol{v}^* U \boldsymbol{v} + i \cdot \boldsymbol{v}^* V \boldsymbol{v}| \ge |\boldsymbol{v}^* V \boldsymbol{v}|,$$

the last step holding because U, V are real-symmetric so that $\boldsymbol{v}^* U \boldsymbol{v}$ and $\boldsymbol{v}^* V \boldsymbol{v}$ are both real. By the given assumption on V, we have $|\boldsymbol{v}^* V \boldsymbol{v}| \ge c_0$, so $\|M\boldsymbol{v}\| \ge c_0$ for every unit vector $\boldsymbol{v} \in \mathbb{C}^n$. Then M is invertible, and $\|M^{-1}\| \le 1/c_0$. □

Proposition 13. Let $M, \widetilde{M} \in \mathbb{R}^{n\times n}$ be any two symmetric matrices satisfying

$$\frac{1}{n}\|M-\widetilde{M}\|_F^2 \to 0$$

a.s. as $n\to\infty$. If $\lim \operatorname{spec} M = \nu$ for a probability distribution ν on $\mathbb{R}$, then also $\lim \operatorname{spec} \widetilde{M} = \nu$.

Proof. For fixed $z \in \mathbb{C}^+$, let $m(z) = \operatorname{tr}(M - z\mathrm{Id})^{-1}$ and $\tilde{m}(z) = \operatorname{tr}(\widetilde{M} - z\mathrm{Id})^{-1}$ be the Stieltjes transforms. Then applying the identity

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}, \tag{2.4.3}$$

we may bound their difference by

$$\begin{aligned}
|m(z)-\tilde{m}(z)|^2 &= \frac{1}{n^2}\left|\operatorname{Tr}[(M-z\mathrm{Id})^{-1} - (\widetilde{M}-z\mathrm{Id})^{-1}]\right|^2 \\
&= \frac{1}{n^2}\left|\operatorname{Tr}(M-z\mathrm{Id})^{-1}(\widetilde{M}-M)(\widetilde{M}-z\mathrm{Id})^{-1}\right|^2 \\
&\le \frac{1}{n^2}\|\widetilde{M}-M\|_F^2\|(M-z\mathrm{Id})^{-1}(\widetilde{M}-z\mathrm{Id})^{-1}\|_F^2 \\
&\le \frac{1}{n}\|\widetilde{M}-M\|_F^2\|(M-z\mathrm{Id})^{-1}\|^2\|(\widetilde{M}-z\mathrm{Id})^{-1}\|^2
\end{aligned}$$

Applying $\|(M-z\mathrm{Id})^{-1}\| \le 1/\operatorname{Im} z$ by Proposition 12, and similarly for $\widetilde{M}$, the given condition shows that $m(z)-\tilde{m}(z)\to 0$ a.s., pointwise over $z\in\mathbb{C}^+$. If $\lim \operatorname{spec} M = \nu$, then $m(z)\to m_\nu(z) \equiv \int (x-z)^{-1} d\nu(x)$ a.s., and hence also $\tilde{m}(z)\to m_\nu(z)$ a.s. and $\lim \operatorname{spec}\widetilde{M} = \nu$. $\square$

2.5 Propagation of Approximate Pairwise Orthogonality

In this section, we work in the following (non-asymptotic) setting of a single layer: Consider any deterministic matrix $X \in \mathbb{R}^{d\times n}$, let $W \in \mathbb{R}^{d\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, and set

$$\boldsymbol{X} = \frac{1}{\sqrt{d}}\sigma(WX) \in \mathbb{R}^{d\times n}. \tag{2.5.1}$$

Note that $\boldsymbol{X}$ has i.i.d. rows with distribution $\sigma(\boldsymbol{w}^\top X)/\sqrt{d}$, where $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$. Define the second-moment matrix of $\boldsymbol{X}$ by

$$\Phi = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] = \mathbb{E}[\sigma(\boldsymbol{w}^\top X)^\top \sigma(\boldsymbol{w}^\top X)] \in \mathbb{R}^{n\times n} \tag{2.5.2}$$

where the expectations are over the standard Gaussian matrix W and standard Gaussian vector $\boldsymbol{w}$. Let $\Phi_{\alpha\beta}$ denote the (α,β) entry of Φ for any $\alpha,\beta \in [n]$. We show in this section the following result.

Lemma 14. *Suppose X is (ε, B)-orthonormal where $\varepsilon < 1/\lambda_\sigma$. Then for universal constants $C, c > 0$, with probability at least $1 - 2n^2 e^{-cd\varepsilon^2} - 3e^{-cn}$, the matrix $\boldsymbol{X}$ remains (ε, B)-orthonormal with*

$$\varepsilon = C\lambda_\sigma^2 \varepsilon, \qquad B = C\Big(1 + n/d\Big)\lambda_\sigma^2 B.$$

Corollary 15. *Under Assumption 1, there exist parameters $(\tilde{\varepsilon}_n, \tilde{B})$ still satisfying $\tilde{\varepsilon}_n n^{1/4} \to 0$, such that a.s. for all large n, every matrix $X_0, \ldots, X_L$ is $(\tilde{\varepsilon}_n, \tilde{B})$-orthonormal.*

Proof. Note that increasing ε_n represents a weaker assumption, so we may assume without loss of generality that $\varepsilon_n \geq n^{-0.49}$. Then by Lemma 14, there is a constant $C_0 \geq 1$ depending on $\lambda_\sigma, \gamma_1, \ldots, \gamma_L$, such that if $X_{\ell-1}$ is $(C_0^{\ell-1}\varepsilon_n, C_0^{\ell-1}B)$-orthonormal, then conditional on this event, X_ℓ is $(C_0^\ell \varepsilon_n, C_0^\ell B)$-orthonormal with probability at least $1 - e^{-n^{0.01}}$ for all large n. Thus, setting $\tilde{\varepsilon}_n = C_0^L \varepsilon_n$ and $\tilde{B} = C_0^L B$, with probability at least $1 - Le^{-n^{0.01}}$, every matrix $X_0, \ldots, X_L$ is $(\tilde{\varepsilon}_n, \tilde{B})$-orthonormal. The almost sure statement then follows from the Borel-Cantelli Lemma. □

In the remainder of this section, we prove Lemma 14. We divide the proof into Lemmas 16, 17, and 18 below, which check the individual requirements for (ε, B)-orthonormality of $\boldsymbol{X}$. We denote by $C, C', c, c' > 0$ universal constants that may change from instance to instance.

Lemma 16. *If X is (ε, B)-orthonormal where $\varepsilon < 1/\lambda_\sigma$, then for universal constants $C, c > 0$:*

(a) For all $\alpha \neq \beta \in [n]$,

$$|\Phi_{\alpha\beta} - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq C\lambda_\sigma^2 \varepsilon^2 \tag{2.5.3}$$

$$\left|\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(0,\mathrm{Id})}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]\right| \leq C\lambda_\sigma \left|\|\mathbf{x}_\alpha\|^2 - 1\right| \leq C\lambda_\sigma \varepsilon \tag{2.5.4}$$

$$|\Phi_{\alpha\alpha} - 1| \leq C\lambda_\sigma \left|\|\mathbf{x}_\alpha\|^2 - 1\right| \leq C\lambda_\sigma \varepsilon \tag{2.5.5}$$

(b) With probability at least $1 - 2n^2 e^{-cd\varepsilon^2}$, simultaneously for all $\alpha \neq \beta \in [n]$, the columns of $\boldsymbol{X}$ satisfy

$$\left|\|\boldsymbol{x}_\alpha\|^2 - 1\right| \leq C\lambda_\sigma^2 \varepsilon, \qquad \left|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\right| \leq C\lambda_\sigma^2 \varepsilon.$$

Note that (2.5.3) establishes an approximation of off-diagonal entries for Φ which is second-order in ε—this will be important in our later arguments which approximate Φ in both Frobenius norm and operator norm.

Proof of Lemma 16. For part (a), observe that $(\zeta_\alpha, \zeta_\beta) \equiv (\boldsymbol{w}^\top \mathbf{x}_\alpha, \boldsymbol{w}^\top \mathbf{x}_\beta)$ is bivariate Gaussian, with mean 0 and covariance

$$\Sigma = \begin{pmatrix} \|\mathbf{x}_\alpha\|^2 & \mathbf{x}_\alpha^\top \mathbf{x}_\beta \\ \mathbf{x}_\alpha^\top \mathbf{x}_\beta & \|\mathbf{x}_\beta\|^2 \end{pmatrix} = \mathrm{Id} + \Delta$$

where Δ is entrywise bounded by ε. Then performing a Gram-Schmidt orthogonalization

procedure, for some independent standard Gaussian variables $\xi_\alpha, \xi_\beta \sim \mathcal{N}(0,1)$, we have

$$\zeta_\alpha = u_\alpha \xi_\alpha, \qquad \zeta_\beta = u_\beta \xi_\beta + v_\beta \xi_\alpha \tag{2.5.6}$$

where $u_\alpha, u_\beta > 0$ and $v_\beta \in \mathbb{R}$ satisfy $|u_\alpha - 1|, |u_\beta - 1|, |v_\beta| \le C\varepsilon$ for a universal constant $C > 0$.

By a Taylor expansion of $\sigma(\zeta)$ around $\zeta = \xi$, there exists a random variable η between ζ and ξ such that

$$\sigma(\zeta) = \sigma(\xi) + \sigma'(\xi)(\zeta - \xi) + \frac{1}{2}\sigma''(\eta)(\zeta - \xi)^2. \tag{2.5.7}$$

For $\alpha \neq \beta$, applying this for both ζ_α and ζ_β, noting that the product of leading terms satisfies $\mathbb{E}[\sigma(\xi_\alpha)\sigma(\xi_\beta)] = 0$, and applying also the bounds $|\sigma'(x)|, |\sigma''(x)| \le \lambda_\sigma$ where $\lambda_\sigma \ge 1$, it is easy to check that

$$\Phi_{\alpha\beta} = \mathbb{E}[\sigma(\zeta_\alpha)\sigma(\zeta_\beta)] = \mathbb{E}\Big[\sigma(\xi_\alpha)\cdot\sigma'(\xi_\beta)(\zeta_\beta - \xi_\beta) + \sigma(\xi_\beta)\cdot\sigma'(\xi_\alpha)(\zeta_\alpha - \xi_\alpha)\Big] + \text{remainder}$$

where this remainder has magnitude at most $C\lambda_\sigma^2\varepsilon^2$. For the first term, substituting (2.5.6) and applying independence of ξ_α and ξ_β, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sigma(\xi_\alpha)\cdot\sigma'(\xi_\beta)(\zeta_\beta - \xi_\beta) + \sigma(\xi_\beta)\cdot\sigma'(\xi_\alpha)(\zeta_\alpha - \xi_\alpha)\Big] \\
&= (u_\beta - 1)\mathbb{E}[\sigma(\xi_\alpha)]\cdot\mathbb{E}[\sigma'(\xi_\beta)\xi_\beta] + v_\beta\mathbb{E}[\sigma(\xi_\alpha)\xi_\alpha]\cdot\mathbb{E}[\sigma'(\xi_\beta)] \\
&\quad + (u_\alpha - 1)\mathbb{E}[\sigma(\xi_\beta)]\cdot\mathbb{E}[\sigma'(\xi_\alpha)\xi_\alpha].
\end{aligned}$$

Applying $\mathbb{E}[\sigma(\xi)] = 0$ and the integration-by-parts identity $\mathbb{E}[\sigma(\xi)\xi] = \mathbb{E}[\sigma'(\xi)] = b_\sigma$, this term equals $v_\beta b_\sigma^2$. From (2.5.6), we have $u_\alpha v_\beta = \mathbb{E}[\zeta_\alpha\zeta_\beta] = \mathbf{x}_\alpha^\top \mathbf{x}_\beta$. Since $|u_\alpha - 1| \le C\varepsilon$ and $|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \le \varepsilon$, this implies $|v_\beta b_\sigma^2 - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta| \le C b_\sigma^2 \varepsilon^2 \le C\lambda_\sigma^2\varepsilon^2$. Combining these yields (2.5.3).

Similarly, from a first-order Taylor expansion analogous to (2.5.7),

$$\Big|\mathbb{E}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]\Big| = \Big|\mathbb{E}[\sigma(\zeta_\alpha)] - \mathbb{E}[\sigma(\xi_\alpha)]\Big| \le C\lambda_\sigma \cdot |u_\alpha - 1|,$$
$$|\Phi_{\alpha\alpha} - 1| = \Big|\mathbb{E}[\sigma(\zeta_\alpha)^2] - \mathbb{E}[\sigma(\xi_\alpha)^2]\Big| \le C\max\Big(\lambda_\sigma \cdot |u_\alpha - 1|,\ \lambda_\sigma^2 \cdot |u_\alpha - 1|^2\Big).$$

The bounds (2.5.4) and (2.5.5) follows from the observations $u_\alpha^2 = \mathbb{E}[\zeta_\alpha^2] = \|\mathbf{x}_\alpha\|^2$ and $|u_\alpha - 1| \le |u_\alpha - 1| \cdot |u_\alpha + 1| = |u_\alpha^2 - 1| \le \varepsilon$.

For part (b), let $\boldsymbol{w}_k^\top$ be the k^{th} row of W. Then by definition of $\boldsymbol{X}$, for any $\alpha, \beta \in [n]$ (including $\alpha = \beta$),

$$\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta = \frac{1}{d}\sum_{k=1}^{d} \sigma\Big(\boldsymbol{w}_k^\top \mathbf{x}_\alpha\Big)\sigma\Big(\boldsymbol{w}_k^\top \mathbf{x}_\beta\Big).$$

We next apply Bernstein's inequality. Denote by $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ the sub-Gaussian and sub-exponential norms of a random variable. For any deterministic vector $\mathbf{x} \in \mathbb{R}^d$, the function $\boldsymbol{w} \mapsto \sigma(\boldsymbol{w}^\top \mathbf{x})$ is $\lambda_\sigma\|\mathbf{x}\|$-Lipschitz. Then for $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$ and a universal constant $C > 0$, we have by Gaussian concentration-of-measure

$$\|\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha) - \mathbb{E}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]\|_{\psi_2} \le C\lambda_\sigma \|\mathbf{x}_\alpha\|.$$

From (2.5.4), $|\mathbb{E}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]| \le C\lambda_\sigma\varepsilon$. Thus (recalling that $|\|\mathbf{x}_\alpha\| - 1| \le \varepsilon$), we have

$$\|\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)\|_{\psi_2} \le C\lambda_\sigma$$

for a constant $C > 0$, and similarly for $\mathbf{x}_\beta$. So

$$\|\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)\sigma(\boldsymbol{w}^\top \mathbf{x}_\beta)\|_{\psi_1} \le \|\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)\|_{\psi_2}\|\sigma(\boldsymbol{w}^\top \mathbf{x}_\beta)\|_{\psi_2} \le C\lambda_\sigma^2. \tag{2.5.8}$$

Applying Bernstein's inequality (see [Ver18, Theorem 2.8.1]), for a universal constant $c > 0$ and

any $t > 0$,

$$\mathbb{P}\Big[\big|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta - \mathbb{E}\big[\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\big]\big| > t\Big] \le 2\exp\bigg(-cd\min\bigg(\frac{t^2}{\lambda_\sigma^4}, \frac{t}{\lambda_\sigma^2}\bigg)\bigg).$$

Applying this for $t = \lambda_\sigma^2 \varepsilon$ and taking a union bound over all $\alpha, \beta \in [n]$, we get

$$\mathbb{P}\Big[\big|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta - \mathbb{E}\big[\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\big]\big| \le \lambda_\sigma^2 \varepsilon \text{ for all } \alpha, \beta \in [n]\Big] \ge 1 - 2n^2 \exp\big(-cd \cdot \varepsilon^2\big). \tag{2.5.9}$$

Since $\mathbb{E}[\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta] = \Phi_{\alpha\beta}$, part (b) now follows from part (a). □

Lemma 17. *If X is (ε, B)-orthonormal where $\varepsilon < 1/\lambda_\sigma$, then for universal constants $C, c > 0$:*

(a) $\|\Phi\| \le C\lambda_\sigma^2 B^2$.

(b) With probability at least $1 - 2e^{-cn}$, $\|\boldsymbol{X}\| \le C\Big(1 + \sqrt{n/d}\Big)\lambda_\sigma B$.

Proof. For part (a), define

$$\Sigma = \mathbb{E}\Big[\sigma(\boldsymbol{w}^\top X)^\top \sigma(\boldsymbol{w}^\top X)\Big] - \mathbb{E}[\sigma(\boldsymbol{w}^\top X)]^\top \mathbb{E}[\sigma(\boldsymbol{w}^\top X)] \tag{2.5.10}$$

where the first term on the right is Φ. Then

$$\|\Sigma\| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \Sigma \boldsymbol{v} = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \Big|\mathbb{E}\Big[\big(\sigma(\boldsymbol{w}^\top X)\boldsymbol{v}\big)^2\Big] - \mathbb{E}\Big[\sigma(\boldsymbol{w}^\top X)\boldsymbol{v}\Big]^2\Big| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \mathrm{Var}\big[\sigma(\boldsymbol{w}^\top X)\boldsymbol{v}\big].$$

We bound this variance using the Gaussian Poincaré inequality: Let us fix $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| = 1$ and define

$$F(\boldsymbol{w}) = \sigma(\boldsymbol{w}^\top X)\boldsymbol{v} = \sum_{\alpha=1}^n v_\alpha \sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha).$$

Then, letting $\boldsymbol{u} \in \mathbb{R}^n$ be the vector with entries $u_\alpha = v_\alpha \sigma'(\boldsymbol{w}^\top \mathbf{x}_\alpha)$,

$$\nabla F(\boldsymbol{w}) = \sum_{\alpha=1}^n v_\alpha \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_\alpha) \cdot \boldsymbol{x}_\alpha = \boldsymbol{X}\boldsymbol{u}, \qquad \|\nabla F(\boldsymbol{w})\| \le \|\boldsymbol{X}\| \cdot \|\boldsymbol{u}\| \le \lambda_\sigma B. \tag{2.5.11}$$

Then by the Gaussian Poincaré inequality, $\mathrm{Var}[F(\boldsymbol{w})] \le \mathbb{E}[\|\nabla F(\boldsymbol{w})\|^2] \le \lambda_\sigma^2 B^2$, so $\|\Sigma\| \le \lambda_\sigma^2 B^2$. In addition, by (2.5.4), the difference between Φ and Σ is a rank-one perturbation controlled by

$$\|\Phi - \Sigma\| = \|\mathbb{E}[\sigma(\boldsymbol{w}^\top X)]\|^2 = \sum_{\alpha=1}^{n} \mathbb{E}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]^2 \le C\lambda_\sigma^2 \sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|^2 - 1)^2 \le C\lambda_\sigma^2 B^2, \quad (2.5.12)$$

the last inequality using the final condition of (ε, B)-orthonormality in Definition 4. This establishes part (a).

For part (b), we apply the concentration result of [Ver10, Eq. (5.26)] for matrices with independent sub-Gaussian rows. For any fixed unit vector $\boldsymbol{v} \in \mathbb{R}^n$, recall from (2.5.11) that $F(\boldsymbol{w}) = \sigma(\boldsymbol{w}^\top X)\boldsymbol{v}$ is $\lambda_\sigma B$-Lipschitz. Then by Gaussian concentration of measure,

$$\|F(\boldsymbol{w}) - \mathbb{E}[F(\boldsymbol{w})]\|_{\psi_2} \le C\lambda_\sigma B.$$

We have $|\mathbb{E}[F(\boldsymbol{w})]| \le \|\mathbb{E}[\sigma(\boldsymbol{w}^\top X)]\| \le C\lambda_\sigma B$ by (2.5.12), so also $\|F(\boldsymbol{w})\|_{\psi_2} \le C\lambda_\sigma B$. This holds for any unit vector $\boldsymbol{v} \in \mathbb{R}^n$, hence $\|\sigma(\boldsymbol{w}^\top X)\|_{\psi_2} \le C\lambda_\sigma B$ for the vector sub-Gaussian norm. Thus, $\sqrt{d}\boldsymbol{X}/(\lambda_\sigma B)$ has i.i.d. rows whose sub-Gaussian norm is at most a universal constant. Recalling $\Phi = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}]$ and applying [Ver10, Eq. (5.26)] with $A = \sqrt{d}\boldsymbol{X}/(\lambda_\sigma B)$, we obtain for some universal constants $C, c > 0$ that

$$\mathbb{P}\Big[\|\boldsymbol{X}^\top \boldsymbol{X} - \Phi\| > \max(\delta, \delta^2)\|\Phi\|\Big] \le 2e^{-ct^2}, \qquad \delta = C\sqrt{n/d} + t/\sqrt{d}.$$

Note that the complementary event $\|\boldsymbol{X}^\top \boldsymbol{X} - \Phi\| \le \max(\delta, \delta^2)\|\Phi\|$ implies

$$\|\boldsymbol{X}\| \le \sqrt{(1 + \max(\delta, \delta^2))\|\Phi\|} \le (1 + C'\delta)\sqrt{\|\Phi\|}$$

for a constant $C' > 0$. Then choosing $t = \sqrt{n}$ and applying part (a) yields part (b). □

Lemma 18. *If X is (ε, B)-orthonormal where $\varepsilon < 1/\lambda_\sigma$, then for universal constants $C, c > 0$,*

with probability at least $1 - e^{-cn}$*, the columns of* $\boldsymbol{X}$ *satisfy*

$$\sum_{\alpha=1}^{n} (\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \le C\Big(1 + n^2/d^2\Big)\lambda_\sigma^4 B^2.$$

Let us remark that in settings where $\varepsilon \gg 1/\sqrt{n}$, applying Lemma 16(b) to bound each term $(\|\boldsymbol{x}_\alpha\|^2 - 1)^2$ separately would not yield a constant-order bound for this sum. The proof below performs a more careful analysis of the combined fluctuations of $(\|\boldsymbol{x}_\alpha\|^2 - 1)^2$.

Proof. Let $\mathbf{z} = (z_1, \dots, z_n) \in \mathbb{R}^n$ and $\mathbf{r} = (r_1, \dots, r_n) \in \mathbb{R}^n$ be defined as

$$z_\alpha = \|\boldsymbol{x}_\alpha\|^2 - \mathbb{E}[\|\boldsymbol{x}_\alpha\|^2], \qquad r_\alpha = \mathbb{E}[\|\boldsymbol{x}_\alpha\|^2] - 1.$$

The quantity to be bounded is $\|\mathbf{z} + \mathbf{r}\|^2$. Note that $\|\mathbf{z} + \mathbf{r}\|^2 \le 2\|\mathbf{z}\|^2 + 2\|\mathbf{r}\|^2$. We have

$$\mathbb{E}[\|\boldsymbol{x}_\alpha\|^2] = \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} \sigma(\boldsymbol{w}_i^\top \mathbf{x}_\alpha)^2\right] = \Phi_{\alpha\alpha},$$

so applying (2.5.5) from Lemma 16,

$$\|\mathbf{r}\|^2 = \sum_{\alpha=1}^{n} (\Phi_{\alpha\alpha} - 1)^2 \le C\lambda_\sigma^2 \sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|^2 - 1)^2 \le C\lambda_\sigma^2 B^2. \tag{2.5.13}$$

Thus it remains to bound $\|\mathbf{z}\|^2$.

Let $\mathscr{N}$ be a $1/2$-net of the unit ball $\{\boldsymbol{w} \in \mathbb{R}^n : \|\boldsymbol{w}\| = 1\}$, of cardinality $|\mathscr{N}| \le 5^n$. Then

$$\|\mathbf{z}\| = \sup_{\boldsymbol{w}:\|\boldsymbol{w}\|\le 1} \boldsymbol{w}^\top \mathbf{z} \le \sup_{\boldsymbol{v}\in\mathscr{N}} \boldsymbol{v}^\top \mathbf{z} + \|\mathbf{z}\|/2,$$

so $\|\mathbf{z}\| \le 2\sup_{\boldsymbol{v}\in\mathcal{N}} \boldsymbol{v}^\top \mathbf{z}$. For each fixed vector $\boldsymbol{v} = (v_1,\dots,v_n) \in \mathcal{N}$, we have

$$\begin{aligned}\boldsymbol{v}^\top \mathbf{z} &= \sum_{\alpha=1}^{n} v_\alpha \cdot \frac{1}{d}\sum_{i=1}^{d}\Big(\sigma(\boldsymbol{w}_i^\top \mathbf{x}_\alpha)^2 - \mathbb{E}[\sigma(\boldsymbol{w}_i^\top \mathbf{x}_\alpha)^2]\Big) \\ &= \frac{1}{d}\sum_{i=1}^{d}\Bigg(\sum_{\alpha=1}^{n}\Big(\sigma(\boldsymbol{w}_i^\top \mathbf{x}_\alpha)^2 - \mathbb{E}[\sigma(\boldsymbol{w}_i^\top \mathbf{x}_\alpha)^2]\Big)v_\alpha\Bigg). \end{aligned} \tag{2.5.14}$$

We will bound the sub-exponential norm of each summand $i = 1,\dots,d$ and apply Bernstein's inequality.

For standard normal random vector $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$, denote

$$\mathbf{q} \equiv \mathbf{q}(\boldsymbol{w}) = (q_1,\dots,q_n) = (\boldsymbol{w}^\top \mathbf{x}_1,\dots,\boldsymbol{w}^\top \mathbf{x}_n), \qquad F(\mathbf{q}) = \sum_{\alpha=1}^{n}\Big(\sigma(q_\alpha)^2 - \mathbb{E}[\sigma(q_\alpha)^2]\Big)v_\alpha.$$

Observe that $\mathbf{q}(\boldsymbol{w}) = X^\top \boldsymbol{w}$. Thus we wish to bound the sub-exponential norm of $F(\mathbf{q}(\boldsymbol{w}))$ when $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$. By the Gaussian Sobolev inequality (see [AW15, Eq. (3)]), for any $p \ge 2$,

$$\|F(\mathbf{q}(\boldsymbol{w}))\|_{L^p} \le \sqrt{p}\cdot \Big\| \|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\| \Big\|_{L^p} \tag{2.5.15}$$

where $\|Y\|_{L^p} = \mathbb{E}[|Y|^p]^{1/p}$ denotes the L^p-norm of a random variable (and $\|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\|$ is the usual ℓ_2 vector norm of the gradient of $F(\mathbf{q}(\boldsymbol{w}))$ in $\boldsymbol{w}$). By the chain rule,

$$\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w})) = X \cdot \nabla_{\mathbf{q}} F(\mathbf{q}),$$

so

$$\|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\|^2 \le \|X\|^2 \|\nabla_{\mathbf{q}} F(\mathbf{q})\|^2 \le B^2 \|\nabla_{\mathbf{q}} F(\mathbf{q})\|^2.$$

We have $(\partial/\partial q_\alpha) F(\mathbf{q}) = 2\sigma(q_\alpha)\sigma'(q_\alpha)v_\alpha$. Hence,

$$\|\nabla_{\mathbf{q}} F(\mathbf{q})\|^2 = \sum_{\alpha=1}^{n} 4\sigma(q_\alpha)^2\sigma'(q_\alpha)^2 v_\alpha^2 \le 4\lambda_\sigma^2 \sum_{\alpha=1}^{n} \sigma(q_\alpha)^2 v_\alpha^2.$$

Recalling (2.5.8), we have $\|\sigma(q_\alpha)^2\|_{\psi_1} = \|\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)^2\|_{\psi_1} \le C\lambda_\sigma^2$. Then

$$\left\| \sum_{\alpha=1}^n \sigma(q_\alpha)^2 v_\alpha^2 \right\|_{\psi_1} \le C\lambda_\sigma^2 \sum_{\alpha=1}^n v_\alpha^2 = C\lambda_\sigma^2,$$

and, hence,

$$\left\| \|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\|^2 \right\|_{\psi_1} \le C\lambda_\sigma^4 B^2.$$

This implies the bound (see [Ver18, Proposition 2.7.1]), for any $p \ge 1$,

$$\left\| \|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\| \right\|_{L^{2p}}^{2p} = \mathbb{E}\left[\|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\|^{2p} \right] = \left\| \|\nabla_{\boldsymbol{w}} F(\mathbf{q}(\boldsymbol{w}))\|^2 \right\|_{L^p}^p \le (C'\lambda_\sigma^4 B^2 \cdot p)^p$$

for a universal constant $C' > 0$. Thus, applying this to (2.5.15), we obtain for any $p \ge 2$

$$\|F(\mathbf{q}(\boldsymbol{w}))\|_{L^p} \le \sqrt{p} \cdot C\lambda_\sigma^2 B\sqrt{p} = C\lambda_\sigma^2 B \cdot p.$$

Finally, this implies (see again [Ver18, Proposition 2.7.1]) $\|F(\mathbf{q}(\boldsymbol{w}))\|_{\psi_1} \le C'\lambda_\sigma^2 B$ for a universal constant $C' > 0$, which is our desired bound on the sub-exponential norm of $F(\mathbf{q}(\boldsymbol{w}))$.

Applying this and Bernstein's inequality to (2.5.14), for any $t > 0$,

$$\mathbb{P}[\boldsymbol{v}^\top \mathbf{z} > t] \le \exp\left(-cd \min\left(\frac{t^2}{\lambda_\sigma^4 B^2}, \frac{t}{\lambda_\sigma^2 B} \right) \right).$$

Setting

$$t = C_0 \lambda_\sigma^2 B \cdot \max(\delta, \delta^2), \qquad \delta = \sqrt{n/d}$$

for a large enough constant $C_0 > 0$, and taking the union bound over all 5^n vectors $\boldsymbol{v} \in \mathbb{N}$, we get

$$\mathbb{P}[\|\mathbf{z}\| > 2t] \le \mathbb{P}\left[\sup_{\boldsymbol{v} \in \mathbb{N}} \boldsymbol{v}^\top \mathbf{z} > t \right] \le e^{-cn}$$

for a constant $c > 0$. Combining with the bound on $\|\mathbf{r}\|^2$ in (2.5.13), we obtain the lemma. □

2.6 Resolvent Analysis for Single Layers

We consider the same setting of a single layer as in the preceding section. Let $\boldsymbol{X}$ and Φ be defined by the deterministic input $X \in \mathbb{R}^{d\times n}$ and Gaussian matrix $W \in \mathbb{R}^{d\times d}$ as in (2.5.1) and (2.5.2), and define the (n-dependent) aspect ratio

$$\gamma = n/d.$$

Consider a deterministic real-symmetric matrix $A \in \mathbb{R}^{n\times n}$, and two (possibly n-dependent) spectral arguments $\alpha \in \mathbb{C}^*$ and $z \in \mathbb{C}^+$, where $\mathbb{C}^* = \overline{\mathbb{C}^-} \setminus \{0\}$. We study the matrix

$$A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id}\,.$$

We collect here the set of assumptions that we will use in this section.

Assumption 2. There are constants $B, C_0, c_0 > 0$ such that

(a) $\alpha \in \mathbb{C}^*$ and $z \in \mathbb{C}^+$, and $\gamma, |\alpha|, |z|, \mathrm{Im}\, z \in [c_0, C_0]$.

(b) X is (ε_n, B)-orthonormal, where $\varepsilon_n < n^{-0.01}$.

(c) $A \in \mathbb{R}^{n\times n}$ is deterministic and symmetric, satisfying $\|A\| \le C_0$.

(d) W has i.i.d. $\mathcal{N}(0,1)$ entries, and $\sigma(x)$ satisfies Assumption 1(b).

Throughout this section, $C, C', c, c', n_0 > 0$ denote constants changing from instance to instance that may depend on λ_σ and the above values B, C_0, c_0.

Proposition 12 ensures that $A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id}$ is invertible. Define the resolvent

$$R = (A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id})^{-1} \in \mathbb{C}^{n\times n} \tag{2.6.1}$$

and the deterministic (n-dependent) parameter

$$\bar{s} = \alpha^{-1} + \gamma \cdot \mathbb{E}[\operatorname{tr} R\Phi]. \tag{2.6.2}$$

The goal of this section is to prove the following result, which approximates this resolvent R by replacing the random matrix $\alpha \boldsymbol{X}^\top \boldsymbol{X}$ with a deterministic matrix $\bar{s}^{-1}\Phi$, and provides an approximate fixed point equation that defines this parameter $\bar{s}$.

For $A = 0$ and $\alpha = 1$, we will verify in Section 2.7 that this result reduces to the Marčenko-Pastur equation (1.2.4).

Lemma 19. *Under Assumption 2, there are constants $C, c, c', n_0 > 0$ such that for all $n \geq n_0$, any deterministic matrix $M \in \mathbb{C}^{n \times n}$, and any $t \in (n^{-1}, c')$,*

(a) $\mathbb{P}\left[\left|\operatorname{tr} RM - \operatorname{tr}\left(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id}\right)^{-1} M\right| > \|M\| t\right] \leq Cne^{-cnt^2}$

(b) $\mathbb{P}\left[\left|\bar{s} - \left(\alpha^{-1} + \gamma \operatorname{tr}\left(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id}\right)^{-1}\Phi\right)\right| > t\right] \leq Cne^{-cnt^2}$

2.6.1 Basic Bounds

Proposition 20. Recall the notation in the above section. Under Assumption 2, deterministically for some constants $C, c, n_0 > 0$ and all $n \geq n_0$,

$$\|R\| \leq C, \qquad \|\Phi\| \leq C, \qquad |\bar{s}| \leq C, \qquad \operatorname{Im} \bar{s} \geq c.$$

Furthermore, with probability at least $1 - 2e^{-c'n}$ for a constant $c' > 0$,

$$\operatorname{Im} \operatorname{tr} R\Phi \geq c.$$

Proof. We may write $A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id} = U + iV$ where $U = A + (\operatorname{Re}\alpha)\boldsymbol{X}^\top \boldsymbol{X} - (\operatorname{Re} z)\,\mathrm{Id}$ and $V = (\operatorname{Im}\alpha)\boldsymbol{X}^\top \boldsymbol{X}^\top - (\operatorname{Im} z)\,\mathrm{Id}$. Both U and V are symmetric, and $V \preceq (-\operatorname{Im} z)\,\mathrm{Id}$ because $\operatorname{Im}\alpha \leq 0$ and $\operatorname{Im} z > 0$. Then $\|R\| \leq 1/\operatorname{Im} z \leq C$ by Proposition 12.

The bound $\|\Phi\| \le C$ comes from Lemma 17(a) and the (ε_n, B)-orthonormality assumption for X. Then from the definition of $\bar{s}$ in (2.6.2) and the bounds $\|R\|, \|\Phi\| \le C$, we have also $|\bar{s}| \le C$. For the lower bound for $\operatorname{Im} \bar{s}$ and $\operatorname{Im} \operatorname{tr} R\Phi$, let us write

$$\operatorname{tr} R\Phi = \operatorname{tr}\left(\frac{R+R^*}{2}\right)\Phi + \operatorname{tr}\left(\frac{R-R^*}{2}\right)\Phi.$$

The first trace is real because $R+R^*$ is Hermitian, so

$$\operatorname{Im} \operatorname{tr} R\Phi = \operatorname{Im} \operatorname{tr}\left(\frac{R-R^*}{2}\right)\Phi.$$

Denoting $Y = A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id}$ and applying the identity (2.4.3), we have

$$R - R^* = Y^{-1} - (Y^*)^{-1} = Y^{-1}(Y^* - Y)(Y^*)^{-1} = R(Y^* - Y)R^*.$$

Then, writing $Y = U + iV$ as above and applying $Y^* - Y = -2iV$, we get

$$\begin{aligned} \operatorname{Im} \operatorname{tr} R\Phi &= \operatorname{Im}(-i \cdot \operatorname{tr} RVR^*\Phi) \\ &= \operatorname{Re}\Big(-(\operatorname{Im}\alpha) \cdot \operatorname{tr} R\boldsymbol{X}^\top \boldsymbol{X} R^*\Phi + (\operatorname{Im} z) \cdot \operatorname{tr} RR^*\Phi\Big). \end{aligned}$$

Since $\operatorname{tr} R\boldsymbol{X}^\top \boldsymbol{X} R^*\Phi = \operatorname{tr} \Phi^{1/2} R\boldsymbol{X}^\top \boldsymbol{X} R^* \Phi^{1/2}$, where this matrix is positive semi-definite, this trace is real and non-negative. Similarly, $\operatorname{tr} RR^*\Phi$ is real and non-negative. Then the above yields the lower bound

$$\operatorname{Im} \operatorname{tr} R\Phi \ge \operatorname{Im} z \cdot \operatorname{tr} RR^*\Phi \ge \operatorname{Im} z \cdot \lambda_{\min}(RR^*) \cdot \operatorname{tr} \Phi,$$

where $\lambda_{\min}(RR^*)$ is the smallest eigenvalue of RR^*. By (2.5.5) and the condition $\varepsilon_n < n^{-0.01}$, we have $\operatorname{tr} \Phi \ge c$ for a constant $c > 0$ and large enough n_0. Observe that $\lambda_{\min}(RR^*) = 1/\|Y\|^2$, and $\|Y\| \le \|A\| + |\alpha| \cdot \|\boldsymbol{X}\|^2 + |z|$. By Lemma 17(b), with probability $1 - 2e^{-c'n}$, we have $\|\boldsymbol{X}\| \le C$, so putting this together yields $\operatorname{Im} \operatorname{tr} R\Phi \ge c$ with this probability. Finally, for the deterministic

bound $\operatorname{Im} \bar{s} \geq c$, we may apply $\operatorname{Im} \operatorname{tr} R\Phi \geq c$ on the event where $\|\boldsymbol{X}\| \leq C$ holds, and $\operatorname{Im} \operatorname{tr} R\Phi \geq 0$ on the complementary event. Taking an expectation and applying the definition (2.6.2) yields $\operatorname{Im} \bar{s} \geq c$. □

2.6.2 Resolvent Approximation

We recall the result of [LLC18, Lemma 1], which establishes a concentration of quadratic forms in the rows of $\boldsymbol{X}$. The following is its specialization to standard Gaussian matrices W, and stated in our notation.

Lemma 21 ([LLC18])**.** *Suppose $\sigma(x)$ is λ_σ-Lipschitz, and let $\boldsymbol{x}_i^\top$ be a row of $\boldsymbol{X}$. Then for any deterministic matrix $Y \in \mathbb{R}^{n\times n}$ with $\|Y\| \leq 1$, for some constants $C, c > 0$ (depending on λ_σ), and for any $t > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{\gamma}\boldsymbol{x}_i^\top Y \boldsymbol{x}_i - \operatorname{tr} Y\Phi\right| > t\right) \leq C\exp\left(-\frac{cn}{\|X\|^2}\min\left(\frac{t^2}{t_0^2}, t\right)\right) \tag{2.6.3}$$

where $t_0 = |\sigma(0)| + \lambda_\sigma \|X\|\sqrt{1/\gamma}$.

Using this result, we establish the following approximation for the resolvent R in (2.6.1).

Lemma 22. *Consider any deterministic matrix $M \in \mathbb{C}^{n\times n}$, and set*

$$\delta_n = \operatorname{tr} M - \operatorname{tr} R\left(A + \frac{1}{\alpha^{-1} + \gamma \operatorname{tr} R\Phi}\Phi - z\,\mathrm{Id}\right)M.$$

Under Assumption 2, there exist constants $C, c, c', n_0 > 0$ such that for all $n \geq n_0$ and $t \in (n^{-1}, c')$,

$$\mathbb{P}[|\delta_n| > \|M\| t] \leq Cne^{-cnt^2}.$$

Proof. By rescaling M, we may assume that $\|M\| \leq 1$. We have $\mathrm{Id} = R(A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id}) = RA + \alpha R\boldsymbol{X}^\top \boldsymbol{X} - zR$. Writing $\boldsymbol{X}^\top \boldsymbol{X} = \sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$ (where $\boldsymbol{x}_i^\top$ is the i^{th} row of $\boldsymbol{X}$), multiplying by M,

and taking the normalized trace $\mathrm{tr} = n^{-1}\mathrm{Tr}$,

$$\begin{aligned}\mathrm{tr}\, M &= \mathrm{tr}\, RAM + \alpha\, \mathrm{tr}\, R\boldsymbol{X}^\top \boldsymbol{X} M - z\,\mathrm{tr}\, RM \\ &= \mathrm{tr}\, RAM + \frac{\alpha}{n}\sum_{i=1}^{d} \boldsymbol{x}_i^\top MR\boldsymbol{x}_i - z\,\mathrm{tr}\, RM.\end{aligned}$$

Hence

$$\delta_n = \frac{\alpha}{n}\sum_{i=1}^{d} \boldsymbol{x}_i^\top MR\boldsymbol{x}_i - \frac{\mathrm{tr}\, R\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R\Phi}.$$

Let us define the leave-one-out resolvent, for each $1 \le i \le d$,

$$R^{(i)} = \left(A + \alpha \sum_{j:j\neq i} \boldsymbol{x}_j \boldsymbol{x}_j^\top - z\,\mathrm{Id}\right)^{-1}.$$

We may then decompose δ_n as $\delta_n = J_1 + \gamma J_2$ where (recalling $\gamma = n/d$)

$$\begin{aligned}J_1 &= \frac{1}{n}\sum_{i=1}^{d}\left(\alpha \boldsymbol{x}_i^\top MR\boldsymbol{x}_i - \frac{\gamma\,\mathrm{tr}\, R^{(i)}\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R^{(i)}\Phi}\right), \\ J_2 &= \frac{1}{n}\sum_{i=1}^{d}\left(\frac{\mathrm{tr}\, R^{(i)}\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R^{(i)}\Phi} - \frac{\mathrm{tr}\, R\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R\Phi}\right).\end{aligned}$$

Let us denote these summands as

$$J_1^{(i)} = \alpha \boldsymbol{x}_i^\top MR\boldsymbol{x}_i - \frac{\gamma\,\mathrm{tr}\, R^{(i)}\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R^{(i)}\Phi} \quad \text{and} \quad J_2^{(i)} = \frac{\mathrm{tr}\, R^{(i)}\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R^{(i)}\Phi} - \frac{\mathrm{tr}\, R\Phi M}{\alpha^{-1} + \gamma\,\mathrm{tr}\, R\Phi}.$$

Bound for J_1. Momentarily fix the index $i \in \{1, \dots, d\}$. Applying the Sherman-Morrison-Woodbury identity, we have

$$R = R^{(i)} - \frac{\alpha R^{(i)} \boldsymbol{x}_i \boldsymbol{x}_i^\top R^{(i)}}{1 + \alpha \boldsymbol{x}_i^\top R^{(i)} \boldsymbol{x}_i}. \tag{2.6.4}$$

Then, introducing $A_1 = \boldsymbol{x}_i^\top M R^{(i)} \boldsymbol{x}_i$ and $A_2 = \boldsymbol{x}_i^\top R^{(i)} \boldsymbol{x}_i$,

$$\alpha \boldsymbol{x}_i^\top MR\boldsymbol{x}_i = \alpha A_1 - \frac{\alpha^2 A_1 A_2}{1 + \alpha A_2} = \frac{A_1}{\alpha^{-1} + A_2}.$$

Recall that the rows of $\boldsymbol{X}$ are i.i.d. Let $\boldsymbol{X}^{(i)}$ be the matrix $\boldsymbol{X}$ with the i^{th} row $\boldsymbol{x}_i$ removed, and let $\mathbb{E}_{\boldsymbol{x}_i}[\cdot]$ be the expectation over only $\boldsymbol{x}_i$ (i.e. conditional on $\boldsymbol{X}^{(i)}$). Observe that $R^{(i)}$ is a function of $\boldsymbol{X}^{(i)}$. Applying Proposition 20 with $\boldsymbol{X}^{(i)}$ in place of $\boldsymbol{X}$, we see that $\|R^{(i)}\|$ and $\|MR^{(i)}\|$ are both bounded by a constant. Then applying Lemma 21 conditional on $\boldsymbol{X}^{(i)}$, and recalling the bound (2.4.1) for $\sigma(0)$, there are constants $C, c > 0$ for which

$$\mathbb{P}[|A_k - \mathbb{E}_{\boldsymbol{x}_i}[A_k]| > t] \le Ce^{-cn\min(t^2,t)} \qquad \text{for } k = 1,2.$$

Note that

$$\mathbb{E}_{\boldsymbol{x}_i}[A_1] = \operatorname{Tr} MR^{(i)}\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top] = \frac{1}{d}\operatorname{Tr} MR^{(i)}\Phi = \gamma \operatorname{tr} R^{(i)}\Phi M.$$

Similarly, $\mathbb{E}_{\boldsymbol{x}_i}[A_2] = \gamma \operatorname{tr} R^{(i)}\Phi$, so

$$J_1^{(i)} = \frac{A_1}{\alpha^{-1} + A_2} - \frac{\mathbb{E}_{\boldsymbol{x}_i}[A_1]}{\alpha^{-1} + \mathbb{E}_{\boldsymbol{x}_i}[A_2]}.$$

Applying Proposition 20, we have for some constants $C, c, c' > 0$, on an event $\mathscr{E}(\boldsymbol{X}^{(i)})$ of probability $1 - 2e^{-c'n}$, that

$$|\mathbb{E}_{\boldsymbol{x}_i}[A_1]| \le C, \qquad |\alpha^{-1} + \mathbb{E}_{\boldsymbol{x}_i}[A_2]| \ge \operatorname{Im}(\alpha^{-1} + \mathbb{E}_{\boldsymbol{x}_i}[A_2]) \ge c.$$

Then, for any t such that $t < c/2$, on the event where $|A_1 - \mathbb{E}_{\boldsymbol{x}_i}[A_1]| \le t$, $|A_2 - \mathbb{E}_{\boldsymbol{x}_i}[A_2]| \le t$, and $\mathscr{E}(\boldsymbol{X}^{(i)})$ all hold,

$$\left|J_1^{(i)}\right| \le \frac{|A_1 - \mathbb{E}_{\boldsymbol{x}_i}[A_1]|}{|\alpha^{-1} + A_2|} + |\mathbb{E}_{\boldsymbol{x}_i}[A_1]| \cdot \frac{|A_2 - \mathbb{E}_{\boldsymbol{x}_i}[A_2]|}{|\alpha^{-1} + A_2| \cdot |\alpha^{-1} + \mathbb{E}_{\boldsymbol{x}_i}[A_2]|} \le Ct. \tag{2.6.5}$$

Thus, for $t < c'$ and a sufficiently small constant $c' > 0$, we have $\mathbb{P}[|J_1^{(i)}| \ge t] \le Ce^{-cnt^2}$. Applying a union bound over $i \in \{1, \dots, d\}$, this yields $\mathbb{P}[|J_1| \ge t] \le Cne^{-cnt^2}$.

Bound for J_2. Applying the identity (2.4.3),

$$R^{(i)} - R = R^{(i)}(R^{-1} - (R^{(i)})^{-1})R = \alpha R^{(i)} \boldsymbol{x}_i \boldsymbol{x}_i^\top R.$$

Then, applying also the bounds $\|R\|, \|R^{(i)}\| \le C$ from Proposition 20,

$$|\operatorname{tr}(R^{(i)} - R)\Phi M| = \frac{1}{n}|\alpha \boldsymbol{x}_i^\top R\Phi M R^{(i)} \boldsymbol{x}_i| \le \frac{C\|\boldsymbol{X}\|^2}{n}.$$

Applying Lemma 17(b), with probability $1 - 2e^{-cn}$, this is at most C/n for every $i \in \{1, \dots, d\}$. Similarly, $|\operatorname{tr}(R^{(i)} - R)\Phi| \le C/n$ with this probability. Applying again $|\operatorname{tr} R\Phi M| \le C$, $|\alpha^{-1} + \gamma \operatorname{tr} R\Phi| \ge c$, and an argument similar to (2.6.5), we obtain $|J_2^{(i)}| \le C'/n$ for a constant $C' > 0$. Taking a union bound over $i \in \{1, \dots, d\}$, this yields $\mathbb{P}[|J_2| > C/n] \le C' n e^{-cn}$. Combining these bounds for J_1 and J_2, choosing $t > cn^{-1}$, and re-adjusting the constants yields the lemma. □

2.6.3 Proof of Lemma 19

We now prove Lemma 19 using Lemma 22. Define the random n-dependent parameter

$$s = \alpha^{-1} + \gamma \operatorname{tr} R\Phi,$$

so that $\bar{s} = \mathbb{E}[s]$. The following establishes a concentration of s around $\bar{s}$.

Lemma 23. *Under Assumption 2, for some constants $c, n_0 > 0$, all $n \ge n_0$, and any $t > 0$,*

$$\mathbb{P}[|s - \bar{s}| > t] \le 2e^{-cnt^2}.$$

Proof. Define $F(W) = \gamma \operatorname{tr} R\Phi$, where R and $\boldsymbol{X}$ are considered as a function of W. Fix any matrices $W, \Delta \in \mathbb{R}^{d \times n}$ where $\|\Delta\|_F = 1$, and define $W_t = W + t\Delta$. Then, applying $\partial R = -R(\partial(R^{-1}))R$

and $R = R^\top$,

$$\begin{aligned}\operatorname{vec}(\Delta)^\top(\nabla F(W)) = \frac{d}{dt}\Big|_{t=0} F(W_t) &= -\gamma \operatorname{tr} R\left(\frac{d}{dt}\Big|_{t=0} R^{-1}\right) R\Phi \\ &= -2\gamma\alpha \operatorname{tr} R\left(\boldsymbol{X}^\top \cdot \frac{d}{dt}\Big|_{t=0} \boldsymbol{X}\right) R\Phi \\ &= -\frac{2\gamma\alpha}{\sqrt{d}} \operatorname{tr} R\Big(\boldsymbol{X}^\top \cdot \big(\sigma'(WX) \odot (\Delta X)\big)\Big) R\Phi,\end{aligned}$$

where $\odot$ is the Hadamard product, and σ' is applied entrywise. Applying Proposition 20,

$$\Big|\operatorname{vec}(\Delta)^\top(\nabla F(W))\Big| \le \frac{C}{\sqrt{d}} \cdot \Big\|R\boldsymbol{X}^\top \cdot (\sigma'(WX) \odot (\Delta X)) \cdot R\Big\| \le \frac{C'}{\sqrt{d}} \cdot \|R\boldsymbol{X}^\top\| \cdot \|\sigma'(WX) \odot (\Delta X)\|.$$

For the first term,

$$\begin{aligned}\|R\boldsymbol{X}^\top\|^2 = \frac{1}{|\alpha|}\|R(\alpha\boldsymbol{X}^\top\boldsymbol{X})R^*\| &\le \frac{1}{|\alpha|}\Big(\|R(A + \alpha\boldsymbol{X}^\top\boldsymbol{X} - z\,\mathrm{Id})R^*\| + \|R(A - z\,\mathrm{Id})R^*\|\Big) \\ &\le \frac{1}{|\alpha|}(\|R\| + \|R\|^2(\|A\| + |z|)) \le C.\end{aligned}$$

For the second term,

$$\|\sigma'(WX) \odot (\Delta X)\| \le \|\sigma'(WX) \odot (\Delta X)\|_F \le \lambda_\sigma \|\Delta X\|_F \le \lambda_\sigma \|\Delta\|_F \cdot \|X\| \le C.$$

Thus $|\operatorname{vec}(\Delta)^\top(\nabla F(W))| \le C/\sqrt{n}$. This holds for every Δ such that $\|\Delta\|_F = 1$, so $F(W)$ is $C/\sqrt{n}$-Lipschitz in W with respect to the Frobenius norm. Then the result follows from the Gaussian concentration of measure. □

To conclude the proof of Lemma 19, we may again assume $\|M\| \le 1$ by rescaling M. Set

$$\widetilde{M} = \big(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id}\big)^{-1} M.$$

Note that $\bar{s}^{-1} \in \mathbb{C}^-$, so $\|\widetilde{M}\| \le \|(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id})^{-1}\| \le C$ by Proposition 12. Applying Lemma 22 with $\widetilde{M}$,

$$\mathbb{P}\Big[\Big|\operatorname{tr}\widetilde{M} - \operatorname{tr} R\big(A + s^{-1}\Phi - z\,\mathrm{Id}\big)\widetilde{M}\Big| > t\Big] \le Cne^{-cnt^2} \tag{2.6.6}$$

for all $t \in (n^{-1}, c')$. Furthermore, applying the definition of $\widetilde{M}$,

$$\begin{aligned}|\operatorname{tr} R\big(A + s^{-1}\Phi - z\,\mathrm{Id}\big)\widetilde{M} - \operatorname{tr} RM| &= \Big|\operatorname{tr} R\Big(\big(A + s^{-1}\Phi - z\,\mathrm{Id}\big) - \big(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id}\big)\Big)\widetilde{M}\Big| \\ &= |s^{-1} - \bar{s}^{-1}| \cdot |\operatorname{tr} R\Phi\widetilde{M}| \le C|s^{-1} - \bar{s}^{-1}|.\end{aligned}$$

Recall that $|\bar{s}| \ge \operatorname{Im}\bar{s} \ge c$. Then, on the event where $|s - \bar{s}| \le t$ and $t < c/2$, we have $|s^{-1} - \bar{s}^{-1}| \le Ct$. Then applying Lemma 23, for some constants $c, c' > 0$ and all $t \in (0, c')$,

$$\mathbb{P}\Big[|\operatorname{tr} R\big(A + s^{-1}\Phi - z\,\mathrm{Id}\big)\widetilde{M} - \operatorname{tr} RM| > t\Big] \le 2e^{-cnt^2}.$$

Combining this with (2.6.6) yields Lemma 19(a). Specializing Lemma 19(a) to $M = \Phi$, we obtain

$$\mathbb{P}\big[\big|s - \big(\alpha^{-1} + \gamma\operatorname{tr}(A + \bar{s}^{-1}\Phi - z\,\mathrm{Id})^{-1}\Phi\big)\big| > t\big] \le Cne^{-cnt^2}.$$

Applying again Lemma 23 to bound $|s - \bar{s}|$, we obtain Lemma 19(b).

2.7 Analysis for the Conjugate Kernel

Theorem 6 is a special case of Theorem 9, but let us provide here a simpler argument. Define, for each layer, the $n \times n$ matrices

$$\Phi_\ell = \mathbb{E}_{\boldsymbol{w}}\Big[\sigma(\boldsymbol{w}^\top X_{\ell-1})^\top \sigma(\boldsymbol{w}^\top X_{\ell-1})\Big] \tag{2.7.1}$$

$$\tilde{\Phi}_\ell = b_\sigma^2 X_{\ell-1}^\top X_{\ell-1} + (1 - b_\sigma^2)\,\mathrm{Id} \tag{2.7.2}$$

where $\mathbb{E}_{\boldsymbol{w}}$ denotes the expectation over only the random vector $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$. Here, Φ_ℓ and $\tilde{\Phi}_\ell$ are deterministic conditional on $X_{\ell-1}$, but are random unconditionally for $\ell \geq 2$. For each fixed $\ell = 1, \ldots, L$, we will show that as $n \to \infty$,

$$\lim \text{spec } \Phi_\ell = \lim \text{spec } \tilde{\Phi}_\ell. \tag{2.7.3}$$

Conditional on $X_{\ell-1}$, the spectral limit of $X_\ell^\top X_\ell$ was shown in [LLC18] to be a Marčenko-Pastur map of the spectral limit of Φ_ℓ—we reproduce a short proof below under our assumptions, by specializing Lemma 19 to $\alpha = 1$ and $A = 0$. Combining with (2.7.3) and iterating from $\ell = 1, \ldots, L$ yields Theorem 6.

Lemma 24. *Under Assumption 1, for each $\ell = 1, \ldots, L$, almost surely as $n \to \infty$,*

$$\frac{1}{n} \|\Phi_\ell - \tilde{\Phi}_\ell\|_F^2 \to 0.$$

Proof. By Corollary 15, increasing (ε_n, B) as needed, we may assume that each matrix $X_0, \ldots, X_L$ is (ε_n, B)-orthonormal. Denote by $\Phi_\ell[\alpha, \beta]$ and $\tilde{\Phi}_\ell[\alpha, \beta]$ the (α, β) entries of these matrices. Then Lemma 16(a) shows for $\alpha \neq \beta$ that

$$|\Phi_\ell[\alpha, \beta] - \tilde{\Phi}_\ell[\alpha, \beta]| \leq C\varepsilon_n^2.$$

For $\alpha = \beta$, applying $\tilde{\Phi}_\ell[\alpha, \alpha] = 1 - b_\sigma^2 + b_\sigma^2 \|\mathbf{x}_\alpha^{\ell-1}\|^2$, we have

$$|\Phi_\ell[\alpha, \alpha] - \tilde{\Phi}_\ell[\alpha, \alpha]| \leq |\Phi_\ell[\alpha, \alpha] - 1| + b_\sigma^2 |\|\mathbf{x}_\alpha^{\ell-1}\|^2 - 1| \leq C\varepsilon_n.$$

Then

$$\|\Phi_\ell - \tilde{\Phi}_\ell\|_F^2 \leq Cn(n-1)\varepsilon_n^4 + Cn\varepsilon_n^2,$$

and the result follows from the condition $\varepsilon_n n^{1/4} \to 0$. □

Proof of Theorem 6. By Corollary 15, we may assume that each matrix $X_0, \dots, X_L$ is (ε_n, B)-orthonormal. This implies the bounds $\|X_\ell\| \le C$ and $\|\boldsymbol{K}^{\mathrm{CK}}\| \le C$ for all large n.

For the spectral convergence, suppose by induction that $\lim \operatorname{spec} X_{\ell-1}^\top X_{\ell-1} = \mu_{\ell-1}$, where the base case $\lim \operatorname{spec} X_0^\top X_0 = \mu_0$ holds by assumption. Defining

$$\nu_\ell = (1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1},$$

Proposition 13 and Lemma 24 together show that

$$\lim \operatorname{spec} \Phi_\ell = \lim \operatorname{spec} \tilde{\Phi}_\ell = \nu_\ell.$$

Specializing Lemma 19(b) to the setting $A = 0$, $\alpha = 1$, $X = X_{\ell-1}$, and $\boldsymbol{X} = X_\ell$, and choosing $t \equiv t_n$ such that $t_n \to 0$ and $n t_n^2 \gg \log n$, we obtain

$$\left| \bar{s} - 1 - (n/d_\ell) \operatorname{tr}(\bar{s}^{-1} \Phi_\ell - z \operatorname{Id})^{-1} \Phi_\ell \right| \to 0 \tag{2.7.4}$$

a.s. as $n \to \infty$, where

$$\bar{s} = 1 + \frac{n}{d_\ell} \mathbb{E}_{W_\ell}[\operatorname{tr}(X_\ell^\top X_\ell - z \operatorname{Id})^{-1} \Phi_\ell].$$

Here, this expectation is taken over only W_ℓ (i.e. conditional on $X_0, \dots, X_{\ell-1}$).

Proposition 20 verifies that $\bar{s}$ is bounded as $n \to \infty$, so for any subsequence in n, there is a further sub-subsequence along which $\bar{s} \to s_0$ for a limit $s_0 \equiv s_0(z) \in \mathbb{C}^+$. Applying (2.4.3) and

Propositions 12 and 20,

$$\begin{aligned}
&\left|\operatorname{tr}(\bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell - \operatorname{tr}(s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell\right| \\
&= |s_0^{-1} - s^{-1}| \cdot \operatorname{tr}\left|(s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell(\bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell\right| \\
&\leq |s_0^{-1} - s^{-1}| \cdot \|(s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\| \cdot \|(\bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\| \cdot \|\Phi_\ell\|^2 \\
&\leq C|s_0^{-1} - s^{-1}|.
\end{aligned}$$

Thus, along the sub-subsequence where $\bar{s} \to s_0$, we get

$$\operatorname{tr}(\bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell - \operatorname{tr}(s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell \to 0. \tag{2.7.5}$$

We have also

$$\operatorname{tr}(s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell \to \int \frac{x}{s_0^{-1}x - z} d\nu_\ell(x), \tag{2.7.6}$$

since the function $x \mapsto x/(s_0^{-1}x - z)$ is continuous and bounded over $\mathbb{R}$, and $\lim \operatorname{spec} \Phi_\ell = \nu_\ell$. Thus, taking the limit of (2.7.4) along this sub-subsequence, the value s_0 must satisfy

$$s_0 - 1 - \gamma_\ell \int \frac{x}{s_0^{-1}x - z}\, d\nu_\ell(x) = 0. \tag{2.7.7}$$

Now applying Lemma 19(a) with $M = \mathrm{Id}$, and taking the limit along this sub-subsequence, by a similar argument we obtain that

$$\operatorname{tr}(X_\ell^\top X_\ell - z\,\mathrm{Id})^{-1} \to \int \frac{1}{s_0^{-1}x - z} d\nu_\ell(x). \tag{2.7.8}$$

Denoting this limit by $m_\ell(z)$, and rewriting (2.7.7) by applying

$$\int \frac{x}{s_0^{-1}x - z} d\nu_\ell(x) = s_0 \int \left(1 + \frac{z}{s_0^{-1}x - z}\right) d\nu_\ell(x) = s_0(1 + z m_\ell(z)),$$

we get $s_0^{-1} = 1 - \gamma_\ell - \gamma_\ell z m_\ell(z)$. Applying this back to the definition of $m_\ell(z)$ in (2.7.8), this shows that $m_\ell(z)$ satisfies the Marčenko-Pastur equation

$$m(z) = \int \frac{1}{x(1-\gamma_\ell - \gamma_\ell z m(z)) - z} d\nu_\ell(x),$$

so $m_\ell(z)$ is the Stieltjes transform of $\mu_\ell = \rho_{\gamma_\ell}^{\mathrm{MP}} \boxtimes \nu_\ell = \rho_{\gamma_\ell}^{\mathrm{MP}} \boxtimes ((1-b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_{\ell-1})$.

We have shown that $\mathrm{tr}(X_\ell^\top X_\ell - z\,\mathrm{Id})^{-1} \to m_\ell(z)$ almost surely along this sub-subsequence in n. Since, for every subsequence in n, there exists such a sub-subsequence, this implies $\lim_{n\to\infty} \mathrm{tr}(X_\ell^\top X_\ell - z\,\mathrm{Id})^{-1} = m_\ell(z)$ almost surely. Thus $\lim \mathrm{spec}\, X_\ell^\top X_\ell = \mu_\ell$, which completes the induction. □

2.8 Analysis for the Neural Tangent Kernel

2.8.1 Spectral Approximation and Operator Norm Bound

We first prove the spectral approximation stated in Lemma 7, as well as the operator norm bound $\|\boldsymbol{K}^{\mathrm{NTK}}\| \le C$. The following form of $\boldsymbol{K}^{\mathrm{NTK}}$ is derived also in [HY19, Eq. (1.7)]: Denote by $\mathbf{x}_\alpha^\ell$ the α^{th} column of X_ℓ. For each $\ell = 1, \ldots, L$, define the matrix $S_\ell \in \mathbb{R}^{d_\ell \times n}$ whose α^{th} column is given by

$$\mathbf{s}_\alpha^\ell = D_\alpha^\ell \frac{W_{\ell+1}^\top}{\sqrt{d_\ell}} D_\alpha^{\ell+1} \frac{W_{\ell+2}^\top}{\sqrt{d_{\ell+1}}} D_\alpha^{\ell+2} \cdots \frac{W_L^\top}{\sqrt{d_{L-1}}} D_\alpha^L \frac{\boldsymbol{w}}{\sqrt{d_L}}, \tag{2.8.1}$$

where we define diagonal matrices indexed by $\alpha \in [n]$ and $k \in [L]$ as

$$D_\alpha^k \equiv \mathrm{diag}\left(\sigma'(W_k \mathbf{x}_\alpha^{k-1})\right) \in \mathbb{R}^{d_k \times d_k}.$$

Applying the chain rule, we may verify for each input sample $\mathbf{x}_\alpha$ that

$$\nabla_{\boldsymbol{w}} f_\theta(\boldsymbol{x}_\alpha) = \mathbf{x}_\alpha^L \in \mathbb{R}^{d_L}, \quad \nabla_{W_\ell} f_\theta(\mathbf{x}_\alpha) = \mathbf{s}_\alpha^\ell \otimes \mathbf{x}_\alpha^{\ell-1} \in \mathbb{R}^{d_\ell d_{\ell-1}}.$$

Then, we can write

$$\begin{aligned}
\big(\nabla_{\boldsymbol{w}} f_\theta(X)\big)^\top \big(\nabla_{\boldsymbol{w}} f_\theta(X)\big) &= X_L^\top X_L, \\
\big(\nabla_{W_\ell} f_\theta(X)\big)^\top \big(\nabla_{W_\ell} f_\theta(X)\big) &= (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}),
\end{aligned}$$

where $\odot$ is the Hadamard product. Thus, the NTK is given by

$$K^{\mathrm{NTK}} = \Big(\nabla_\theta f_\theta(X)\Big)^\top \Big(\nabla_\theta f_\theta(X)\Big) = X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}). \tag{2.8.2}$$

Lemma 25. *Let $X \in \mathbb{R}^{d\times n}$ be (ε, B)-orthonormal, let $W \in \mathbb{R}^{d\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, and let $\mathbf{x}_\alpha, \mathbf{x}_\beta$ be two columns of X where $\alpha \neq \beta$. Then for universal constants $C, c > 0$ and any $t > 0$:*

(a) With probability at least $1 - 2e^{-cdt^2}$,

$$\left| \frac{1}{d} \operatorname{Tr}\Big(\operatorname{diag}\big(\sigma'(W\mathbf{x}_\alpha)\big) \operatorname{diag}\big(\sigma'(W\mathbf{x}_\beta)\big)\Big) - b_\sigma^2 \right| \le C\lambda_\sigma^2(\varepsilon + t).$$

(b) Let $M \in \mathbb{R}^{d\times d}$ be any deterministic symmetric matrix, and denote

$$T(\mathbf{x}_\alpha, \mathbf{x}_\beta) = \frac{1}{d} \operatorname{Tr}\Big(\operatorname{diag}\big(\sigma'(W\mathbf{x}_\alpha)\big) W M W^\top \operatorname{diag}\big(\sigma'(W\mathbf{x}_\beta)\big)\Big).$$

With probability at least $1 - (2d+2)e^{-c\min(t^2 d, t\sqrt{d})}$,

$$\big|T(\mathbf{x}_\alpha, \mathbf{x}_\beta) - b_\sigma^2 \operatorname{Tr} M\big| \le C\lambda_\sigma^2 \Big(\varepsilon\sqrt{d} + t\sqrt{d} + t\sqrt{d}\Big) \|M\|_F.$$

Furthermore, both (a) and (b) hold with $(\mathbf{x}_\alpha, \mathbf{x}_\alpha)$ *in place of* $(\mathbf{x}_\alpha, \mathbf{x}_\beta)$*, upon replacing* b_σ^2 *by* a_σ*.*

Proof. Write $\boldsymbol{w}_k^\top \in \mathbb{R}^d$ for the k^{th} row of W. Then

$$\frac{1}{d}\operatorname{Tr}\Big(\operatorname{diag}\big(\sigma'(W\mathbf{x}_\alpha)\big)\operatorname{diag}\big(\sigma'(W\mathbf{x}_\beta)\big)\Big) = \frac{1}{d}\sum_{k=1}^{d}\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\alpha)\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\beta).$$

Applying $\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\alpha)\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\beta) \in [-\lambda_\sigma^2, \lambda_\sigma^2]$ and Hoeffding's inequality,

$$\mathbb{P}\left[\left|\frac{1}{d}\sum_{k=1}^{d}\Big(\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\alpha)\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\beta) - \mathbb{E}[\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\alpha)\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\beta)]\Big)\right| > \lambda_\sigma^2 t\right] \le 2e^{-cdt^2}.$$

To bound the mean, recall that $(\zeta_\alpha, \zeta_\beta) \equiv (\boldsymbol{w}_k^\top\mathbf{x}_\alpha, \boldsymbol{w}_k^\top\mathbf{x}_\beta)$ is bivariate Gaussian, which we may write as

$$\zeta_\alpha = u_\alpha\xi_\alpha, \qquad \zeta_\beta = u_\beta\xi_\beta + v_\beta\xi_\alpha$$

as in (2.5.6). Here, $\xi_\alpha, \xi_\beta \sim \mathcal{N}(0,1)$ are independent, $u_\alpha, u_\beta > 0$ and $v_\beta \in \mathbb{R}$, and these satisfy $|u_\alpha - 1|, |u_\beta - 1|, |v_\beta| \le C\varepsilon$. Applying the Taylor expansion

$$\sigma'(\zeta) = \sigma'(\xi) + \sigma''(\eta)(\zeta - \xi)$$

for some η between ζ and ξ, and the conditions $\mathbb{E}[\sigma'(\xi)] = b_\sigma$ and $|\sigma''(x)| \le \lambda_\sigma$, it is easy to check that $|\mathbb{E}[\sigma'(\zeta_\alpha)\sigma'(\zeta_\beta)] - b_\sigma^2| \le C\lambda_\sigma^2\varepsilon$. Then part (a) follows. The statement with $(\mathbf{x}_\alpha, \mathbf{x}_\alpha)$ and a_σ follows similarly from this Taylor expansion and the bound $|\mathbb{E}[\sigma'(\zeta_\alpha)^2] - a_\sigma| \le C\lambda_\sigma^2\varepsilon$.

For part (b), we write

$$T(\mathbf{x}_\alpha, \mathbf{x}_\beta) = \frac{1}{d}\sum_{k=1}^{d}\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\alpha)\sigma'(\boldsymbol{w}_k^\top\mathbf{x}_\beta)\cdot\boldsymbol{w}_k^\top M\boldsymbol{w}_k.$$

By the Hanson-Wright inequality (see [RV13, Theorem 1.1]),

$$\mathbb{P}\Big[|\boldsymbol{w}_k^\top M\boldsymbol{w}_k - \operatorname{Tr}M| > \|M\|_F\cdot t\sqrt{d}\Big] \le 2e^{-c\min(t^2d, t\sqrt{d})}$$

for a constant $c > 0$. Then, applying $|\sigma'(x)| \le \lambda_\sigma$ and a union bound over $k = 1, \dots, d$, with probability at least $1 - 2de^{-c\min(t^2 d, t\sqrt{d})}$,

$$\left| T(\mathbf{x}_\alpha, \mathbf{x}_\beta) - \operatorname{Tr} M \cdot \frac{1}{d} \sum_{k=1}^{d} \sigma'(\boldsymbol{w}_k^\top \mathbf{x}_\alpha) \sigma'(\boldsymbol{w}_k^\top \mathbf{x}_\beta) \right| \le \|M\|_F \cdot \lambda_\sigma^2 t \sqrt{d}.$$

Then part (b) follows by combining with part (a) and applying $\operatorname{Tr} M \le \sqrt{d}\|M\|_F$. □

Corollary 26. *Let $\mathbf{s}_\alpha^\ell$ be as defined in (2.8.1), and let q_ℓ, r_ℓ be the constants in (2.2.1). Under Assumption 1, for a constant $C > 0$, almost surely for all large n and for all $\ell \in [L]$ and $\alpha \ne \beta \in [n]$,*

$$\left| {\mathbf{s}_\alpha^\ell}^\top \mathbf{s}_\beta^\ell - q_{\ell-1} \right| \le C \max(\varepsilon_n, n^{-0.48}), \qquad \left| \|\mathbf{s}_\alpha^\ell\|^2 - r_{\ell-1} \right| \le C \max(\varepsilon_n, n^{-0.48}). \tag{2.8.3}$$

Proof. By Corollary 15, we may assume that each matrix $X_0, \dots, X_L$ is (ε_n, B)-orthonormal. Since a larger value of ε_n corresponds to a weaker assumption, we may assume without loss of generality that $\varepsilon_n \ge n^{-0.48}$.

Fix $\ell \in [L]$ and $\alpha, \beta \in [n]$, and define

$$\begin{aligned} M_\ell &= D_\alpha^\ell D_\beta^\ell \\ M_k &= D_\alpha^k \frac{W_k}{\sqrt{d_{k-1}}} \dots D_\alpha^{\ell+1} \frac{W_{\ell+1}}{\sqrt{d_\ell}} D_\alpha^\ell D_\beta^\ell \frac{W_{\ell+1}^\top}{\sqrt{d_\ell}} D_\beta^{\ell+1} \dots \frac{W_k^\top}{\sqrt{d_{k-1}}} D_\beta^k \end{aligned} \tag{2.8.4}$$

for $\ell + 1 \le k \le L$. Recalling the definition (2.8.1) and applying the Hanson-Wright inequality conditional on $W_1, \dots, W_L$,

$$\left| {\mathbf{s}_\alpha^\ell}^\top \mathbf{s}_\beta^\ell - \frac{1}{d_L} \operatorname{Tr} M_L \right| \le C\varepsilon_n \sqrt{n} \cdot \frac{1}{d_L} \|M_L\|_F \tag{2.8.5}$$

with probability $1 - e^{-c\min(\varepsilon_n^2 n, \varepsilon_n \sqrt{n})} \ge 1 - e^{-n^{0.01}}$. Next, for each $k = L, L-1, \dots, \ell+1$, we

apply Lemma 25(b) conditional on $W_1, \dots, W_{k-1}$, with $t = \varepsilon_n$, $M = M_{k-1}/d_{k-1}$, $d = d_{k-1}$, and $d = d_k$. Note that $k-1 \geq \ell \geq 1$, so that both d_{k-1} and d_k are proportional to n. Then

$$\left| \frac{1}{d_k} \operatorname{Tr} M_k - b_\sigma^2 \cdot \frac{1}{d_{k-1}} \operatorname{Tr} M_{k-1} \right| \leq C\varepsilon_n \sqrt{n} \cdot \frac{1}{d_{k-1}} \|M_{k-1}\|_F$$

with probability $1 - e^{-n^{0.01}}$. Finally, for $k = \ell$, applying Lemma 25(a) conditional on $W_1, \dots, W_{\ell-1}$ and with $t = \varepsilon_n$,

$$\left| \frac{1}{d_\ell} \operatorname{Tr} M_\ell - b_\sigma^2 \right| \leq C\varepsilon_n$$

with probability $1 - e^{-n^{0.01}}$. Combining these bounds, with probability $1 - C'e^{-n^{0.01}}$,

$$\left| {\mathbf{s}_\alpha^\ell}^\top \mathbf{s}_\beta^\ell - (b_\sigma^2)^{L-\ell+1} \right| \leq \frac{C\varepsilon_n}{\sqrt{n}} \big(\|M_L\|_F + \ldots + \|M_\ell\|_F + \sqrt{n} \big).$$

We also have $\|W_k/\sqrt{d_k}\| \leq C$ for each $k = 2, \dots, L$ with probability $1 - C'e^{-cn}$, see e.g. [Ver18, Theorem 4.4.5]. Then, applying $\|D_k\| \leq \lambda_\sigma$, we have $\|M_k\|_F \leq C\sqrt{n}\|M_k\| \leq C'\sqrt{n}$ for every $k = 1, \dots, L$. Then the first bound of (2.8.3) follows. The second bound of (2.8.3) is the same, applying Lemma 25 for $(\mathbf{x}_\alpha, \mathbf{x}_\alpha)$ instead of $(\mathbf{x}_\alpha, \mathbf{x}_\beta)$. The almost sure statement follows from the Borel-Cantelli Lemma. □

Lemma 27. *Under Assumption 1, almost surely as $n \to \infty$,*

$$\frac{1}{n} \left\| \boldsymbol{K}^{NTK} - \left(r_+ \operatorname{Id} + X_L^\top X_L + \sum_{\ell=0}^{L-1} q_\ell X_\ell^\top X_\ell \right) \right\|_F^2 \to 0.$$

Furthermore, for a constant $C > 0$, almost surely for all large n, $\|\boldsymbol{K}^{NTK}\| \leq C$.

Proof. By Corollary 15, we may assume that each matrix $X_0, \dots, X_L$ is (ε_n, B)-orthonormal. Then

$$\left| {\mathbf{x}_\alpha^{\ell-1}}^\top \mathbf{x}_\beta^{\ell-1} \right| \leq \varepsilon_n, \qquad \left| \|\mathbf{x}_\alpha^{\ell-1}\|^2 - 1 \right| \leq \varepsilon_n.$$

Increasing ε_n if necessary, we may assume $\varepsilon_n \geq n^{-0.48}$. Combining with (2.8.3), we have for the off-diagonal entries of the Hadamard product that

$$\left| \big((S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}) \big)[\alpha,\beta] - q_{\ell-1} X_{\ell-1}^\top X_{\ell-1}[\alpha,\beta] \right| \leq C\varepsilon_n^2,$$

and for the diagonal entries that

$$\begin{aligned} &\left| \big((S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1})[\alpha,\alpha] - q_{\ell-1} (X_{\ell-1}^\top X_{\ell-1})[\alpha,\alpha] - (r_{\ell-1} - q_{\ell-1}) \right| \\ \leq &\left| \big((S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1})[\alpha,\alpha] - r_{\ell-1} \right| + q_{\ell-1} \left| X_{\ell-1}^\top X_{\ell-1}[\alpha,\alpha] - 1 \right| \leq C\varepsilon_n. \end{aligned}$$

Then applying this to (2.8.2),

$$\left\| \boldsymbol{K}^{\mathrm{NTK}} - \left(r_+ \,\mathrm{Id} + X_L^\top X_L + \sum_{\ell=0}^{L-1} q_\ell X_\ell^\top X_\ell \right) \right\|_F^2 \leq Cn(n-1)\varepsilon_n^4 + Cn\varepsilon_n^2.$$

The first statement of the lemma then follows from the assumption $\varepsilon_n n^{1/4} \to 0$.

For the second statement on the operator norm, we have

$$\| (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1}) \| \leq \max_{1 \leq \alpha \leq n} \left| {\mathbf{s}_\alpha^\ell}^\top \mathbf{s}_\alpha^\ell \right| \cdot \| X_{\ell-1}^\top X_{\ell-1} \|.$$

See [Joh90, Eq. (3.7.9)], applied with $X = Y = S_\ell$. Then $\|\boldsymbol{K}^{\mathrm{NTK}}\| \leq C$ follows from (2.8.2), the (ε_n, B)-orthonormality of each matrix $X_{\ell-1}$, and the bound for $\|\mathbf{s}_\alpha^\ell\|^2$ in (2.8.3). □

Combining Lemma 27 and Proposition 13, this proves Lemma 7.

As a remark, Lemmas 27 and 7 imply $\lim \operatorname{spec} \boldsymbol{K}^{\mathrm{NTK}} = \lim \operatorname{spec} (r_+ \,\mathrm{Id} + X_L^\top X_L)$ when $b_\sigma = 0$, since every $q_\ell = 0$ in this case. Thus, the Stieltjes transform of $\lim \operatorname{spec} \boldsymbol{K}^{\mathrm{NTK}}$ is actually $m_{\mathrm{NTK}}(z) = m(-r_+ + z)$ defined by the Stieltjes transform of $\rho_\gamma^{\mathrm{MP}}$ in (1.2.4) with $\gamma = \gamma_L$. Thus in the following arguments for the limit spectrum of $\boldsymbol{K}^{\mathrm{NTK}}$, we restrict to the case $b_\sigma \neq 0$.

2.8.2 Unique Solution of the Fixed Point Equation

Let $A, \Phi \in \mathbb{R}^{n\times n}$ be symmetric matrices, where Φ is positive semi-definite. Let $z \in \mathbb{C}^+$, $\alpha \in \mathbb{C}^*$, and $\gamma > 0$. For $s \in \mathbb{C}^+$, define

$$S(s) = (A + s^{-1}\Phi - z\,\mathrm{Id})^{-1}, \quad f_n(s) = \alpha^{-1} + \gamma \mathrm{tr}\, S(s)\Phi.$$

Lemma 28. *(a) For any $s \in \mathbb{C}^+$, setting $S \equiv S(s)$,*

$$\mathrm{Im}\, f_n(s) \geq \mathrm{Im}\, z \cdot \gamma \mathrm{tr}\, S\Phi S^* \geq 0.$$

(b) For any $s_1, s_2 \in \mathbb{C}^+$, setting $S_1 \equiv S(s_1)$ and $S_2 \equiv S(s_2)$,

$$\begin{aligned} &|f_n(s_1) - f_n(s_2)| \\ &\leq |s_1 - s_2| \cdot \left(\frac{\mathrm{Im}\, f_n(s_1) - \mathrm{Im}\, z \cdot \gamma \mathrm{tr}\, S_1 \Phi S_1^*}{\mathrm{Im}\, s_1} \right)^{1/2} \left(\frac{\mathrm{Im}\, f_n(s_2) - \mathrm{Im}\, z \cdot \gamma \mathrm{tr}\, S_2 \Phi S_2^*}{\mathrm{Im}\, s_2} \right)^{1/2} \end{aligned}$$

Proof. For part (a), let us write

$$S\Phi = S\Phi S^*(A + s^{-1}\Phi - z\,\mathrm{Id})^* = S\Phi S^* A + (1/s^*) S\Phi S^* \Phi - z^* S\Phi S^*.$$

Since $S\Phi S^*$ is Hermitian and positive semi-definite, the quantities $\mathrm{tr}\, S\Phi S^* A$, $\mathrm{tr}\, S\Phi S^* \Phi$, and $\mathrm{tr}\, S\Phi S^*$ are all real, and the latter two are nonnegative. Then

$$\mathrm{Im}\, f_n(s) = \mathrm{Im}\, \alpha^{-1} + \gamma \mathrm{Im}\, \mathrm{tr}\, S\Phi = \mathrm{Im}\, \alpha^{-1} + \frac{\mathrm{Im}\, s}{|s|^2} \cdot \gamma \mathrm{tr}\, S\Phi S^* \Phi + \mathrm{Im}\, z \cdot \gamma \mathrm{tr}\, S\Phi S^*. \tag{2.8.6}$$

Each term on the right side of (2.8.6) is nonnegative, and dropping the first two of these terms yields (a).

For part (b), applying the identity (2.4.3), we have

$$S_1 - S_2 = S_1(s_2^{-1}\Phi - s_1^{-1}\Phi)S_2 = \frac{s_1 - s_2}{s_1 s_2} S_1 \Phi S_2,$$

so

$$f_n(s_1) - f_n(s_2) = \gamma \mathrm{tr}\, S_1\Phi - \gamma \mathrm{tr}\, S_2\Phi = \frac{\gamma(s_1 - s_2)}{s_1 s_2} \mathrm{tr}\, S_1\Phi S_2\Phi.$$

Applying Cauchy-Schwartz to the inner-product $\langle S_1, S_2\rangle_\Phi = \mathrm{tr}\, S_1\Phi S_2^*\Phi$,

$$|\mathrm{tr}\, S_1\Phi S_2\Phi|^2 = |\langle S_1, S_2^*\rangle_\Phi|^2 \le \langle S_1, S_1\rangle_\Phi \cdot \langle S_2^*, S_2^*\rangle_\Phi = \mathrm{tr}\, S_1\Phi S_1^*\Phi \cdot \mathrm{tr}\, S_2\Phi S_2^*\Phi.$$

Then

$$|f_n(s_1) - f_n(s_2)| \le |s_1 - s_2| \cdot \left(\frac{\gamma \mathrm{tr}\, S_1\Phi S_1^*\Phi}{|s_1|^2}\right)^{1/2} \left(\frac{\gamma \mathrm{tr}\, S_2\Phi S_2^*\Phi}{|s_2|^2}\right)^{1/2}.$$

Dropping $\mathrm{Im}\,\alpha^{-1}$ in (2.8.6) and applying this to upper-bound $\gamma \mathrm{tr}\, S\Phi S^*\Phi/|s|^2$, part (b) follows. □

Corollary 29. *As $n \to \infty$, suppose that $f_n(s) \to f(s)$ pointwise for each $s \in \mathbb{C}^+$, the empirical spectral distributions of Φ and A converge weakly to deterministic limits, and the limit for Φ is not the point distribution at 0. Then the fixed point equation $s = f(s)$ has at most one solution $s \in \mathbb{C}^+$.*

Proof. Let us first show that for each $s \in \mathbb{C}^+$ and a value $c_0(s) > 0$ independent of n,

$$\liminf_{n\to\infty} \mathrm{tr}\, S(s)\Phi S(s)^* \ge c_0(s) > 0. \tag{2.8.7}$$

Denoting $S \equiv S(s)$ and applying the von Neumann trace inequality,

$$\mathrm{tr}\, S\Phi S^* = \frac{1}{n}\mathrm{Tr}\,\Phi S^* S \ge \frac{1}{n}\sum_{\alpha=1}^{n} \lambda_\alpha(\Phi)\lambda_{n+1-\alpha}(S^*S),$$

where $\lambda_1(\cdot) \geq \ldots \geq \lambda_n(\cdot)$ denote the sorted eigenvalues. Since Φ has a non-degenerate limit spectrum, there is a constant $\varepsilon > 0$ for which $\lambda_{\varepsilon n}(\Phi) > \varepsilon$ for all large n. (Throughout the proof, εn, $\varepsilon n/2$, etc. should be understood as their roundings to the nearest integer.) Then

$$\operatorname{tr} S\Phi S^* \geq \varepsilon \cdot \frac{1}{n} \sum_{\alpha=1}^{\varepsilon n} \lambda_{n+1-\alpha}(S^*S).$$

Denoting by $\sigma_\alpha(\cdot)$ the α^{th} largest singular value, observe that

$$\lambda_{n+1-\alpha}(S^*S) = \sigma_{n+1-\alpha}(S)^2 = \sigma_\alpha(A + s^{-1}\Phi - z\,\mathrm{Id})^{-2}.$$

Applying $\sigma_{\alpha+\beta-1}(A+B) \leq \sigma_\alpha(A) + \sigma_\beta(B)$, we have

$$\sigma_\alpha(A + s^{-1}\Phi - z\,\mathrm{Id}) \leq \sigma_{\alpha/2}(A) + |s|^{-1}\sigma_{\alpha/2+1}(\Phi) + |z|.$$

Since the spectra of A and Φ converge to deterministic limits, this implies that there is a constant $C(s) > 0$ (also depending on z and ε) such that $\sigma_\alpha(A + s^{-1}\Phi - z\,\mathrm{Id}) \leq C(s)$ for every $\alpha \in [\varepsilon n/2, \varepsilon n]$ and all large n. Thus

$$\operatorname{tr} S\Phi S^* \geq \varepsilon \cdot \frac{\varepsilon n - \varepsilon n/2}{n} \cdot C(s)^{-2}$$

for all large n, and this shows the claim (2.8.7).

Then, taking the limit $n \to \infty$ in Lemma 28(b), we get

$$|f(s_1) - f(s_2)| \leq |s_1 - s_2| \cdot \left(\frac{\operatorname{Im} f(s_1) - \operatorname{Im} z \cdot \gamma c_0(s_1)}{\operatorname{Im} s_1}\right)^{1/2} \left(\frac{\operatorname{Im} f(s_2) - \operatorname{Im} z \cdot \gamma c_0(s_2)}{\operatorname{Im} s_2}\right)^{1/2}.$$

If $s_1 = f(s_1)$ and $s_2 = f(s_2)$, then this yields $|s_1 - s_2| \leq |s_1 - s_2| \cdot h(s_1, s_2)$ for some quantity $h(s_1, s_2) \in [0, 1)$, where $h(s_1, s_2) < 1$ strictly because $c_0(s_1), c_0(s_2) > 0$. This contradiction implies $s_1 = s_2$, so the equation $s = f(s)$ has at most one solution $s \in \mathbb{C}^+$. □

2.8.3 Proof of Proposition 8 and Theorem 9

The operator norm bound in Theorem 9 was shown in Lemma 27. For the spectral convergence, note that by Lemma 7, the limit Stieltjes transform of $\boldsymbol{K}^{\mathrm{NTK}}$ at any $z \in \mathbb{C}^+$ is given by

$$m_{\mathrm{NTK}}(z) = \lim_{n\to\infty} \operatorname{tr}\left((-z + r_+)\operatorname{Id} + X_L^\top X_L + \sum_{\ell=0}^{L-1} q_\ell X_\ell^\top X_\ell \right)^{-1},$$

provided that this limit exists and defines the Stieltjes transform of a probability measure. For

$$\mathbf{z} = (z_{-1}, \dots, z_\ell) \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*, \qquad \boldsymbol{w} = (w_{-1}, \dots, w_\ell) \in \mathbb{C}^{\ell+2},$$

recall the functions

$$\mathbf{z} \mapsto s_\ell(\mathbf{z}), \quad (\mathbf{z}, \boldsymbol{w}) \mapsto t_\ell(\mathbf{z}, \boldsymbol{w})$$

defined recursively by (2.2.6) and (2.2.7). Proposition 8 and Theorem 9 are immediate consequences of the following extended result.

Lemma 30. *Suppose $b_\sigma \neq 0$. Under Assumption 1, for each $\ell = 1, \dots, L$:*

(a) For every $\mathbf{z} \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*$, the equation (2.2.6) has a unique fixed point $s_\ell(\mathbf{z}) \in \mathbb{C}^+$.

(b) For every $(\mathbf{z}, \boldsymbol{w}) \in (\mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*) \times \mathbb{C}^{\ell+2}$, almost surely

$$\begin{aligned} & t_\ell(\mathbf{z}, \boldsymbol{w}) \\ &= \lim_{n\to\infty} \operatorname{tr}\Big(z_{-1}\operatorname{Id} + z_0 X_0^\top X_0 + \dots + z_\ell X_\ell^\top X_\ell \Big)^{-1} \Big(w_{-1}\operatorname{Id} + w_0 X_0^\top X_0 + \dots + w_\ell X_\ell^\top X_\ell \Big). \end{aligned} \tag{2.8.8}$$

In particular, for any $z_{-1}, \dots, z_\ell \in \mathbb{R}$ where $z_\ell \neq 0$,

$$\lim \operatorname{spec}\Big(z_{-1}\operatorname{Id} + z_0 X_0^\top X_0 + \dots + z_\ell X_\ell^\top X_\ell \Big) = \boldsymbol{\nu}$$

where ν is a probability measure on $\mathbb{R}$ with Stieltjes transform

$$m(z) = t_\ell\Big((-z+z_{-1}, z_0, \ldots, z_\ell), (1,0,\ldots,0)\Big).$$

Proof. By Corollary 15, we may assume that each matrix $X_0, \ldots, X_L$ is (ε_n, B)-orthonormal.

Define $\Phi_\ell, \tilde{\Phi}_\ell$ by (2.7.1) and (2.7.2). For $\mathbf{z} = (z_{-1}, \ldots, z_\ell)$, let us write as shorthand

$$\mathbf{z} \cdot \mathbf{X}^\top \mathbf{X}(\ell) = z_{-1}\,\mathrm{Id} + z_0 X_0^\top X_0 + \ldots + z_\ell X_\ell^\top X_\ell,$$

where the parenthetical (ℓ) signifies the index of the last term in this sum. Let us define similarly $\boldsymbol{w} \cdot \mathbf{X}^\top \mathbf{X}(\ell)$.

Note that part (b) holds for $\ell = 0$, by the assumption $\lim \mathrm{spec}\, X_0^\top X_0 = \mu_0$, the definition of $t_0((z_{-1}, z_0), (w_{-1}, w_0))$ in (2.2.5), and the fact that the function $x \mapsto (w_{-1} + w_0 x)/(z_{-1} + z_0 x)$ is continuous and bounded over the non-negative real line when $z_{-1} \in \mathbb{C}^-$ and $z_0 \in \mathbb{C}^*$.

We induct on ℓ. Suppose that part (b) holds for $\ell - 1$. To show part (a) for ℓ, fix any $\mathbf{z} = (z_{-1}, \ldots, z_\ell) \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*$ (not depending on n) and consider the matrix

$$R = \Big(\mathbf{z} \cdot \mathbf{X}^\top \mathbf{X}(\ell)\Big)^{-1}. \tag{2.8.9}$$

We apply the analysis of Section 2.6, conditional on $X_0, \ldots, X_{\ell-1}$, and with the identifications

$$\boldsymbol{X} = X_\ell, \qquad X = X_{\ell-1}, \qquad d = d_\ell, \qquad d = d_{\ell-1},$$

$$A = z_0 X_0^\top X_0 + \ldots + z_{\ell-1} X_{\ell-1}^\top X_{\ell-1}, \qquad \alpha = z_\ell, \qquad z = -z_{-1}.$$

Observe that $\alpha \in \mathbb{C}^*$ and $z \in \mathbb{C}^-$. The matrix R in (2.8.9) is exactly

$$R = (A + \alpha \boldsymbol{X}^\top \boldsymbol{X} - z\,\mathrm{Id})^{-1}.$$

Since each $X_0, \dots, X_{\ell-1}$ is (ε_n, B)-orthonormal, we have $\|A\| \le C$ for some constant $C > 0$ (depending on $z_{-1}, \dots, z_\ell, \lambda_\sigma$). Thus Assumption 2 holds, conditional on $X_0, \dots, X_{\ell-1}$. Let us define the n-dependent parameter

$$\bar{s} = \frac{1}{\alpha} + \frac{n}{d_\ell} \operatorname{tr} \mathbb{E}_{W_\ell}[R\Phi_\ell]$$

where this expectation is over only the weights W_ℓ. Then, applying Lemma 19(b) with a value $t \equiv t_n$ such that $t \to 0$ and $nt^2 \gg \log n$, we obtain

$$\left| \bar{s} - \frac{1}{\alpha} - \frac{n}{d_\ell} \operatorname{tr}(A + \bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell \right| \to 0 \tag{2.8.10}$$

almost surely as $n \to \infty$.

Proposition 20 shows that $|\bar{s}|$ is bounded, so for any subsequence in n, there is a further sub-subsequence where $\bar{s} \to s_0$ for a limit $s_0 \equiv s_0(\mathbf{z}) \in \mathbb{C}^+$. Let us now replace $\bar{s}$ and Φ_ℓ above by s_0 and $\tilde{\Phi}_\ell$: First we have

$$\operatorname{tr}\big(A + \bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell - \operatorname{tr}\big(A + s_0^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell \to 0$$

by the same argument as (2.7.5). Then, we have

$$\begin{aligned}
&\left| \operatorname{tr}\big(A + s_0^{-1}\Phi_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell - \operatorname{tr}\big(A + s_0^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell \right| \\
&= \left| s_0^{-1} \operatorname{tr}\big(A + s_0^{-1}\Phi_\ell - z\,\mathrm{Id}\big)^{-1}(\tilde{\Phi}_\ell - \Phi_\ell)\big(A + s_0^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell \right| \\
&\le \frac{C}{n}\|\tilde{\Phi}_\ell - \Phi_\ell\|_F \cdot \left\| (A + s_0^{-1}\tilde{\Phi} - z\,\mathrm{Id})^{-1}\Phi(A + s_0^{-1}\Phi - z\,\mathrm{Id})^{-1} \right\|_F \\
&\le \frac{C}{\sqrt{n}}\|\tilde{\Phi}_\ell - \Phi_\ell\|_F \cdot \|(A + s_0^{-1}\tilde{\Phi} - z\,\mathrm{Id})^{-1}\| \cdot \|\Phi\| \cdot \|(A + s_0^{-1}\Phi - z\,\mathrm{Id})^{-1}\| \to 0,
\end{aligned}$$

where the convergence to 0 follows from Lemma 27. Finally, we have

$$\begin{aligned}&\left|\operatorname{tr}\big(A+s_0^{-1}\Phi_\ell - z\,\mathrm{Id}\big)^{-1}\Phi_\ell - \operatorname{tr}\big(A+s_0^{-1}\Phi_\ell - z\,\mathrm{Id}\big)^{-1}\tilde{\Phi}_\ell\right| \\ &\le \frac{1}{n}\|(A+s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\|_F \cdot \|\Phi_\ell - \tilde{\Phi}_\ell\|_F \le \frac{1}{\sqrt{n}}\|(A+s_0^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\| \cdot \|\Phi_\ell - \tilde{\Phi}_\ell\|_F \to 0.\end{aligned}$$

Applying these approximations to (2.8.10), we have almost surely along this sub-subsequence that

$$\left|s_0 - \frac{1}{\alpha} - \gamma_\ell \operatorname{tr}(A+s_0^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id})^{-1}\tilde{\Phi}_\ell\right| \to 0. \tag{2.8.11}$$

Now observe from the definitions of A, $\tilde{\Phi}_\ell$, and z that

$$\begin{aligned}A+s_0^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id} &= \Big(z_{-1} + \frac{1-b_\sigma^2}{s_0}\Big)\,\mathrm{Id} + \sum_{k=0}^{\ell-2} z_k X_k^\top X_k + \Big(z_{\ell-1} + \frac{b_\sigma^2}{s_0}\Big)X_{\ell-1}^\top X_{\ell-1}, \\ \tilde{\Phi}_\ell &= (1-b_\sigma^2)\,\mathrm{Id} + b_\sigma^2 X_{\ell-1}^\top X_{\ell-1}.\end{aligned}$$

Then, applying (2.8.11) and the induction hypothesis that part (b) holds for $\ell-1$, we obtain that the value s_0 must satisfy

$$s_0 = \frac{1}{\alpha} + \gamma_\ell t_{\ell-1}\Big(\mathbf{z}_{\mathrm{prev}}(s_0,\mathbf{z}), (1-b_\sigma^2, 0, \ldots, 0, b_\sigma^2)\Big),$$

where $\mathbf{z}_{\mathrm{prev}}$ is defined in (2.2.8). This shows the existence of a solution (in $\mathbb{C}^+$) to the fixed point equation (2.2.6). Notice that because $b_\sigma \neq 0$ and $s_0 \in \mathbb{C}^+$, the last entry of $\mathbf{z}_{\mathrm{prev}}(s_0,\mathbf{z})$ is in $\mathbb{C}^*$ and $(\mathbf{z}_{\mathrm{prev}}(s_0,\mathbf{z}), (1-b_\sigma^2, 0, \ldots, 0, b_\sigma^2))$ is in the domain of function $t_{\ell-1}$.

To show uniqueness, we apply Corollary 29: For any fixed $s \in \mathbb{C}^+$, defining

$$f_n(s) = \frac{1}{\alpha} + (n/d_\ell)\operatorname{tr}(A+s^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\Phi_\ell,$$

the same arguments as above establish that

$$\lim_{n\to\infty} f_n(s) = f(s) \equiv \frac{1}{\alpha} + \gamma_\ell t_{\ell-1}\Big(\mathbf{z}_{\text{prev}}(s,\mathbf{z}), (1-b_\sigma^2, 0, \ldots, 0, b_\sigma^2)\Big).$$

Part (b) holding for $\ell - 1$ implies that both A and Φ_ℓ have deterministic spectral limits, where

$$\lim \operatorname{spec} \Phi_\ell = \lim \operatorname{spec} \tilde{\Phi}_\ell$$

by (2.7.3). This cannot be the point distribution at 0, because (2.5.5) implies that $\operatorname{tr} \Phi_\ell \geq 1/2$ for all large n, and $\|\Phi_\ell\| \leq C$ so at least $n/(2C)$ eigenvalues of Φ_ℓ exceed $1/2$ for every n. Thus, Corollary 29 implies that the fixed point $s = f(s)$ is unique. So the fixed point $s_\ell(\mathbf{z}) \in \mathbb{C}^+$ is uniquely defined by (2.2.6), and this shows part (a) for ℓ.

By the uniqueness of this fixed point, we have also shown that $s_0 = s_\ell(\mathbf{z})$, where s_0 is the limit of $\bar{s}$ along the above sub-subsequence. Since for any subsequence in n, there exists a sub-subsequence for this which holds, this shows that $\lim_{n\to\infty} \bar{s} = s_\ell(\mathbf{z})$ almost surely.

Now, to show that part (b) holds for ℓ, let us also fix any $\boldsymbol{w} = (w_{-1}, \ldots, w_\ell) \in \mathbb{C}^{\ell+2}$. Using that $z_\ell \neq 0$, we may write

$$\boldsymbol{w} \cdot \mathbf{X}^\top \mathbf{X}(\ell) = \frac{w_\ell}{z_\ell} \cdot \mathbf{z} \cdot \mathbf{X}^\top \mathbf{X}(\ell) + \boldsymbol{w}_{\text{prev}} \cdot \mathbf{X}^\top \mathbf{X}(\ell-1),$$

where $\boldsymbol{w}_{\text{prev}}$ is as defined in (2.2.9). Then

$$\Big(\mathbf{z} \cdot \mathbf{X}^\top \mathbf{X}(\ell)\Big)^{-1} \Big(\boldsymbol{w} \cdot \mathbf{X}^\top \mathbf{X}(\ell)\Big) = \frac{w_\ell}{z_\ell} \operatorname{Id} + \Big(\mathbf{z} \cdot \mathbf{X}^\top \mathbf{X}(\ell)\Big)^{-1} \Big(\boldsymbol{w}_{\text{prev}} \cdot \mathbf{X}^\top \mathbf{X}(\ell-1)\Big). \tag{2.8.12}$$

We now apply Lemma 19(a) conditional on $X_0, \ldots, X_{\ell-1}$, with the same identifications as above and with

$$M = \boldsymbol{w}_{\text{prev}} \cdot \mathbf{X}^\top \mathbf{X}(\ell-1).$$

Note that M is indeed deterministic conditional on $X_0,\dots,X_{\ell-1}$, and $\|M\| \le C$ for a constant $C > 0$ (depending on $\mathbf{z}$ and $\boldsymbol{w}$) since $X_0,\dots,X_{\ell-1}$ are (ε_n, B)-orthonormal. Then, we can apply Lemma 19(a) to conclude that

$$\operatorname{tr}\left[\left(\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\right)^{-1}\left(\boldsymbol{w}_{\text{prev}}\cdot\mathbf{X}^\top\mathbf{X}(\ell-1)\right)\right] - \operatorname{tr}\left[(A+\bar{s}^{-1}\Phi_\ell - z\,\mathrm{Id})^{-1}\left(\boldsymbol{w}_{\text{prev}}\cdot\mathbf{X}^\top\mathbf{X}(\ell-1)\right)\right] \to 0.$$

By the same arguments as above, we may replace $\bar{s}$ by $s_0 = s_\ell(\mathbf{z})$ and Φ_ℓ by $\tilde{\Phi}_\ell$. Then, applying this to (2.8.12),

$$\begin{aligned}\operatorname{tr}\left[\left(\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\right)^{-1}\left(\boldsymbol{w}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\right)\right] - \frac{w_\ell}{z_\ell} \\ - \operatorname{tr}\left[(A+s_\ell(\mathbf{z})^{-1}\tilde{\Phi}_\ell - z\,\mathrm{Id})^{-1}\left(\boldsymbol{w}_{\text{prev}}\cdot\mathbf{X}^\top\mathbf{X}(\ell-1)\right)\right] \to 0.\end{aligned}$$

Finally, applying that part (b) holds for $\ell-1$, this yields

$$\lim_{n\to\infty}\operatorname{tr}\left[\left(\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\right)^{-1}\left(\boldsymbol{w}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\right)\right] = \frac{w_\ell}{z_\ell} + t_{\ell-1}(\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}),\mathbf{z}),\boldsymbol{w}_{\text{prev}}),$$

which is the definition of $t_\ell(\mathbf{z},\boldsymbol{w})$. This establishes (2.8.8).

For any fixed $z_{-1},\dots,z_\ell \in \mathbb{R}$ where $z_\ell \neq 0$, and any fixed $z \in \mathbb{C}^+$, this implies that the Stieltjes transform of $\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)$ has the almost sure limit

$$m(z) = t_\ell\Big((-z+z_{-1},z_0,\dots,z_\ell),(1,0,\dots,0)\Big).$$

So $m(z)$ defines the Stieltjes transform of a sub-probability distribution ν, and the empirical eigenvalue distribution of $\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)$ converges vaguely a.s. to ν. Since $\|\mathbf{z}\cdot\mathbf{X}^\top\mathbf{X}(\ell)\|$ is bounded because $X_0,\dots,X_L$ are (ε_n,B)-orthonormal, this limit ν must, in fact, be a probability distribution, and the ESD converges weakly to ν. This concludes the induction and the proof. $\square$

2.9 Multi-dimensional Outputs and Rescaling

In this section, we provide some motivation for the form of the NTK in (2.2.11) for networks with a k-dimensional output, and we prove Theorem 10 regarding its spectrum.

2.9.1 Derivation of (2.2.11) from Gradient Flow Training

Consider the training process of the network (2.2.10) based on a gradient flow with training samples $\{\mathbf{x}_\alpha, \mathbf{y}_\alpha\}_{\alpha=1}^n$ where $\mathbf{x}_\alpha \in \mathbb{R}^{d_0}$ and $\mathbf{y}_\alpha \in \mathbb{R}^k$, under the general training loss

$$F(\theta) = \sum_{\alpha=1}^{n} \mathscr{L}(f_\theta(\mathbf{x}_\alpha), \mathbf{y}_\alpha).$$

Here, $\mathscr{L} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is the loss function. We denote by $\nabla \mathscr{L}(f_\theta(\mathbf{x}_\alpha), \mathbf{y}_\alpha) \in \mathbb{R}^k$ the gradient of $\mathscr{L}$ with respect to its first argument, and by $\nabla_{W_\ell} f_\theta(\mathbf{x}_\alpha) \in \mathbb{R}^{\dim(W_\ell) \times k}$ the Jacobian of $f_\theta(\mathbf{x}_\alpha)$ with respect to the weights W_ℓ.

Consider a possibly reweighted gradient-flow training of θ, where the evolution of weights W_ℓ is given by

$$\frac{d}{dt} W_\ell(t) = -\tau_\ell \cdot \nabla_{W_\ell} F(\theta(t)) = -\tau_\ell \sum_{\alpha=1}^{n} \nabla_{W_\ell} f_{\theta(t)}(\mathbf{x}_\alpha) \cdot \nabla \mathscr{L}(f_{\theta(t)}(\mathbf{x}_\alpha), \mathbf{y}_\alpha).$$

The learning rate for each weight matrix W_ℓ is scaled by a constant τ_ℓ—this may arise, for example, from reparametrizing the network (2.2.10) using $\widetilde{W}_\ell = \tau_\ell^{-1} \cdot W_\ell$ and considering gradient flow training for $\widetilde{W}_\ell$. Denoting the vectorization of all training predictions and their Jacobian by

$$f_\theta(X) = (f_\theta^1(X), \ldots, f_\theta^k(X)) \in \mathbb{R}^{nk}, \qquad \nabla_{W_\ell} f_\theta(X) \in \mathbb{R}^{\dim(W_\ell) \times nk},$$

and the corresponding vectorization of $(\nabla \mathscr{L}(f_\theta(\mathbf{x}_\alpha), \mathbf{y}_\alpha))_{\alpha=1}^n$ by $\nabla \mathscr{L}(f_\theta(X), \mathbf{y}) \in \mathbb{R}^{nk}$, this may

be written succinctly as

$$\frac{d}{dt}W_\ell(t) = -\tau_\ell \cdot \nabla_{W_\ell} f_{\theta(t)}(X) \cdot \nabla \mathscr{L}(f_{\theta(t)}(X), \mathbf{y}).$$

Then the time evolution of in-sample predictions is given by

$$\begin{aligned}
\frac{d}{dt}f_{\theta(t)}(X) &= \Big(\nabla_\theta f_{\theta(t)}(X)\Big)^\top \cdot \frac{d}{dt}\theta(t) \\
&= -\sum_{\ell=1}^{L+1} \tau_\ell \Big(\nabla_{W_\ell} f_{\theta(t)}(X)\Big)^\top \Big(\nabla_{W_\ell} f_{\theta(t)}(X)\Big) \cdot \nabla \mathscr{L}(f_{\theta(t)}(X), \mathbf{y}) \\
&= -\boldsymbol{K}^{\mathrm{NTK}}(t) \cdot \nabla \mathscr{L}(f_{\theta(t)}(X), \mathbf{y}),
\end{aligned}$$

where $\boldsymbol{K}^{\mathrm{NTK}}$ is the matrix defined in (2.2.11). For $\tau_1 = \ldots = \tau_{L+1} = 1$, this matrix is simply

$$\boldsymbol{K}^{\mathrm{NTK}} = \Big(\nabla_\theta f_\theta(X)\Big)^\top \Big(\nabla_\theta f_\theta(X)\Big) \in \mathbb{R}^{nk \times nk},$$

which is a flattening of the neural tangent kernel $K \in \mathbb{R}^{n\times n\times k\times k}$ (identified as a map $K : \mathbb{R}^{n\times n} \to \mathbb{R}^{k\times k}$) that is defined in [JGH18].

2.9.2 Proof of Theorem 10

The matrix $\boldsymbol{K}^{\mathrm{NTK}}$ in (2.2.11) admits a $k \times k$ block decomposition

$$\boldsymbol{K}^{\mathrm{NTK}} = \begin{pmatrix} K_{11}^{\mathrm{NTK}} & \cdots & K_{1k}^{\mathrm{NTK}} \\ \vdots & \ddots & \vdots \\ K_{k1}^{\mathrm{NTK}} & \cdots & K_{kk}^{\mathrm{NTK}} \end{pmatrix}$$

where

$$K_{ij}^{\mathrm{NTK}} = \sum_{\ell=1}^{L+1} \tau_\ell \Big(\nabla_{W_\ell} f_\theta^i(X)\Big)^\top \Big(\nabla_{W_\ell} f_\theta^j(X)\Big) \in \mathbb{R}^{n\times n},$$

for general constants $\tau_1, \dots, \tau_{L+1} > 0$. Considering

$$W_{L+1} = \begin{pmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_k^\top \end{pmatrix},$$

we can use the chain rule similar to (2.8.2) to verify

$$K_{ij}^{\mathrm{NTK}} = \mathbf{1}\{i=j\}\tau_{L+1} X_L^\top X_L + \sum_{\ell=1}^{L} \tau_\ell (S_\ell^{i\top} S_\ell^j) \odot (X_{\ell-1}^\top X_{\ell-1})$$

where $S_\ell^i \in \mathbb{R}^{d_\ell \times n}$ is the matrix with the same column-wise definition as in (2.8.1), replacing $\boldsymbol{w}$ by $\boldsymbol{w}_i$.

Lemma 31. *Under the assumptions of Theorem 10, for any indices $i \neq j \in [k]$, almost surely as $n \to \infty$,*

$$\frac{1}{n}\|K_{ij}^{NTK}\|_F^2 \to 0.$$

Furthermore, for a constant $C > 0$, almost surely for all large n, $\|K_{ij}^{NTK}\| \leq C$.

Proof. By Corollary 15, we may assume that each $X_0, \dots, X_L$ is (ε_n, B)-orthonormal.

Let us fix i, j, ℓ and denote the columns of S_ℓ^i and S_ℓ^j by $\mathbf{s}_\alpha^{\ell,i}$ and $\mathbf{s}_\beta^{\ell,j}$ for $\alpha, \beta \in [n]$. We apply the Hanson-Wright inequality conditional on $W_1, \dots, W_L$, which is similar to (2.8.5). However, since $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are independent, there is no trace term, and we obtain instead

$$\left|\mathbf{s}_\alpha^{\ell,i\top} \mathbf{s}_\beta^{\ell,j}\right| \leq C\varepsilon_n \sqrt{n} \frac{1}{d_L} \|M_L\|_F$$

for both $\alpha = \beta$ and $\alpha \neq \beta$ with probability $1 - e^{-n^{0.01}}$, where M_L is the same matrix as defined in (2.8.4). Applying the bound $\|M_L\|_F \leq C\sqrt{n}$ as in the proof of Corollary 26, this yields

$$\left|\mathbf{s}_\alpha^{\ell,i\top} \mathbf{s}_\beta^{\ell,j}\right| \leq C\varepsilon_n$$

almost surely for all $\alpha, \beta \in [n]$ and all large n. Combining with the (ε_n, B)-orthonormality of $X_{\ell-1}$, we get for $\alpha \neq \beta$ that

$$\left|({S_\ell^i}^\top S_\ell^j) \odot (X_{\ell-1}^\top X_{\ell-1})[\alpha,\beta]\right| \le C\varepsilon_n^2, \qquad \left|({S_\ell^i}^\top S_\ell^j) \odot (X_{\ell-1}^\top X_{\ell-1})[\alpha,\alpha]\right| \le C\varepsilon_n.$$

Then

$$\|({S_\ell^i}^\top S_\ell^j) \odot (X_{\ell-1}^\top X_{\ell-1})\|_F^2 \le Cn(n-1)\varepsilon_n^4 + Cn\varepsilon_n^2,$$

and the first statement follows from the assumption $\varepsilon_n n^{1/4} \to 0$. The second statement on the operator norm follows from the bound

$$\|({S_\ell^i}^\top S_\ell^j) \odot (X_{\ell-1}^\top X_{\ell-1})\| \le \left(\max_{1\le\alpha\le n}\left|\mathbf{s}_\alpha^{\ell,i\top}\mathbf{s}_\alpha^{\ell,i}\right|\right)^{1/2}\left(\max_{1\le\alpha\le n}\left|\mathbf{s}_\alpha^{\ell,j\top}\mathbf{s}_\alpha^{\ell,j}\right|\right)^{1/2}\cdot\|X_{\ell-1}^\top X_{\ell-1}\|.$$

See [Joh90, Eq. (3.7.9)] applied with $X = S_\ell^i$ and $Y = S_\ell^j$. The bound $\|K_{ij}^{\mathrm{NTK}}\| \le C$ then follows from the (ε_n, B)-orthonormality of $X_{\ell-1}$ and Corollary 26, applied to S_ℓ^i and S_ℓ^j. □

Applying this lemma together with Proposition 13, we obtain

$$\operatorname{lim\,spec} \boldsymbol{K}^{\mathrm{NTK}} = \operatorname{lim\,spec} \begin{pmatrix} K_{11}^{\mathrm{NTK}} & & \\ & \ddots & \\ & & K_{kk}^{\mathrm{NTK}} \end{pmatrix}$$

where the off-diagonal blocks K_{ij}^{NTK} may be replaced by 0. Then the limit spectral distribution of $\boldsymbol{K}^{\mathrm{NTK}}$ is an equally weighted mixture of those of $K_{11}^{\mathrm{NTK}}, \dots, K_{kk}^{\mathrm{NTK}}$. For each diagonal block K_{ii}^{NTK}, the argument of Lemma 27 shows that

$$\operatorname{lim\,spec} K_{ii}^{\mathrm{NTK}} = \operatorname{lim\,spec}\left(\tau\cdot r_+\,\mathrm{Id} + \tau_{L+1}X_L^\top X_L + \sum_{\ell=0}^{L-1}\tau_{\ell+1}q_\ell X_\ell^\top X_\ell\right).$$

Then by Theorem 9, each diagonal block K_{ii}^{NTK} has the same limit spectral distribution, whose

Stieltjes transform is given by the function $m_{\mathrm{NTK}}(z)$ in Theorem 10. Furthermore, since $\|K_{ii}^{\mathrm{NTK}}\| \le C$ by Lemma 27 and $\|K_{ij}^{\mathrm{NTK}}\| \le C$ for $i \neq j$ by Lemma 31, this shows $\|\boldsymbol{K}^{\mathrm{NTK}}\| \le C$. This establishes Theorem 10. Again, when $b_\sigma = 0$, the limit spectrum of each K_{ii}^{NTK} reduces to $\lim \mathrm{spec}\,(\tau \cdot r_+ \,\mathrm{Id} + \tau_{L+1} X_L^\top X_L)$, which can be computed via the Stieltjes transform of $\rho_{\gamma_L}^{\mathrm{MP}}$.

2.10 Numerical Solution of the Fixed Point Equations

Theorem 9 characterizes the limit Stieltjes transform $m(z)$ of matrices such as $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$. By the discussion in Section 1.2.1, a numerical approximation to the density functions of the corresponding spectral distributions may be obtained by computing $m(z)$ for $z = x + i\eta$, across a fine grid of values $x \in \mathbb{R}$ and for a fixed small imaginary part $\eta > 0$. We describe here one possible approach for this computation.

To compute the limit spectrum for $z_{-1}\,\mathrm{Id} + z_0 X_0^\top X_0 + \ldots + z_L X_L^\top X_L$ and general values $z_{-1}, \ldots, z_L \in \mathbb{R}$, fix the spectral argument $z = x + i\eta$ and denote

$$\mathbf{z}_L = (-z + z_{-1}, z_0, \ldots, z_L),\ \mathbf{z}_{L-1} = \mathbf{z}_{\mathrm{prev}}(s_L(\mathbf{z}_L), \mathbf{z}_L),\ \mathbf{z}_{L-2} = \mathbf{z}_{\mathrm{prev}}(s_{L-1}(\mathbf{z}_{L-1}), \mathbf{z}_{L-1}),\ \text{etc.}$$

Here, for $s \in \mathbb{C}^+$ and $\mathbf{z} \in \mathbb{C}^- \times \mathbb{R}^\ell \times \mathbb{C}^*$, the quantity

$$\mathbf{z}_{\mathrm{prev}}(s, \mathbf{z}) = \left(z_{-1} + \frac{1 - b_\sigma^2}{s}, z_0, \ldots, z_{\ell-2}, z_{\ell-1} + \frac{b_\sigma^2}{s} \right) \in \mathbb{C}^- \times \mathbb{R}^{\ell-1} \times \mathbb{C}^*$$

is as defined in (2.2.8). Denote $s_\ell \equiv s_\ell(\mathbf{z}_\ell)$ for each $\ell = 1, \ldots, L$. Observe that, if we are given $s_1, \ldots, s_L$, then the value $t_\ell(\mathbf{z}_\ell, \boldsymbol{w})$ may be directly computed from (2.2.7), for any $\ell \in \{0, \ldots, L\}$ and any vector $\boldsymbol{w} \in \mathbb{C}^{\ell+2}$. This is because the fixed points needed to compute the arguments $\mathbf{z}_{\mathrm{prev}}(s_\ell(\mathbf{z}_\ell), \mathbf{z}_\ell)$, $\mathbf{z}_{\mathrm{prev}}(s_{\ell-1}(\mathbf{z}_{\ell-1}), \mathbf{z}_{\ell-1})$, etc. for the successive evaluations of t_ℓ, $t_{\ell-1}$, etc. are provided by this given sequence $s_1, \ldots, s_L$.

Thus, we apply an iterative procedure of initializing $s_1^{(0)}, \ldots, s_L^{(0)} \in \mathbb{C}^+$, and computing the *simultaneous* updates $s_1^{(t+1)}, \ldots, s_L^{(t+1)}$ using the previous values $s_1^{(t)}, \ldots, s_L^{(t)}$. That is, we

iterate the following two steps:

1. Set $\mathbf{z}_L^{(t)} = \mathbf{z}_L$, and compute $\mathbf{z}_{L-1}^{(t)} = \mathbf{z}_{\text{prev}}(s_L^{(t)}, \mathbf{z}_L^{(t)})$, $\mathbf{z}_{L-2}^{(t)} = \mathbf{z}_{\text{prev}}(s_{L-1}^{(t)}, \mathbf{z}_{L-1}^{(t)})$, etc.

2. Compute an update $s_\ell^{(t+1)}$ for the value of $s_\ell(\mathbf{z}_\ell)$ and each $\ell = 1, \ldots, L$, using the right side of (2.2.6) with $\mathbf{z}_\ell^{(t)}$ and $\mathbf{z}_{\ell-1}^{(t)} \equiv \mathbf{z}_{\text{prev}}(s_\ell^{(t)}, \mathbf{z}_\ell^{(t)})$ in place of $\mathbf{z}_\ell$ and $\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}_\ell), \mathbf{z}_\ell)$.

After this iteration converges to fixed points $s_1^*, \ldots, s_L^*$, we then compute

$$m(z) = t_L(\mathbf{z}_L, (1, 0, \ldots, 0))$$

using (2.2.7) and these fixed points. For each successive value $z = x + i\eta$ along the grid of values $x \in \mathbb{R}$, we initialize $s_1^{(0)}, \ldots, s_L^{(0)}$ by linear interpolation from the computed fixed points at the preceding two values of x along this grid, for faster computation.

Note that for each value $z = x + i\eta$, if the above iteration converges to fixed points $s_1^*, \ldots, s_L^* \in \mathbb{C}^+$, then this procedure computes the correct value for $m(z)$: This is because, denoting

$$\mathbf{z}_{L-1}^* = \mathbf{z}_{\text{prev}}(s_L^*, \mathbf{z}_L), \quad \mathbf{z}_{L-2}^* = \mathbf{z}_{\text{prev}}(s_{L-1}^*, \mathbf{z}_{L-1}^*), \quad \ldots, \qquad \mathbf{z}_1^* = \mathbf{z}_{\text{prev}}(s_2^*, \mathbf{z}_2^*),$$

it can be checked iteratively from (2.2.6), (2.2.7), and the uniqueness guarantee of Proposition 8 that $s_1^* = s_1(\mathbf{z}_1^*)$, then $s_2^* = s_2(\mathbf{z}_2^*)$, etc., and finally that $s_L^* = s_L(\mathbf{z}_L)$. This then means that $\mathbf{z}_{L-1}^* = \mathbf{z}_{\text{prev}}(s_L(\mathbf{z}_L), \mathbf{z}_L) = \mathbf{z}_{L-1}$, then $\mathbf{z}_{L-2}^* = \mathbf{z}_{\text{prev}}(s_{L-1}(\mathbf{z}_{L-1}), \mathbf{z}_{L-1}) = \mathbf{z}_{L-2}$, etc., and so $s_\ell^* = s_\ell(\mathbf{z}_\ell)$ for each ℓ. Then this method computes the correct value for $m(z) = t_L(\mathbf{z}_L, (1, 0, \ldots, 0))$.

We have found in practice that the above iteration occasionally converges to fixed points $s_1, \ldots, s_L$ not belonging to $\mathbb{C}^+$ (i.e. this is not a mapping from $(\mathbb{C}^+)^L$ to $(\mathbb{C}^+)^L$). If this occurs, we randomly re-initialize $s_1^{(0)}, \ldots, s_L^{(0)} \in \mathbb{C}^+$, and we have found that the method reaches the correct fixed point within a small number of random initialization.

To clarify this approach, let us illustrate this computation in a simple example: Consider $L = 2$. Fix any grid value $x \in \mathbb{R}$ and $\eta > 0$. An approximate density function for the limit

spectrum of $X_2^\top X_2$ at x is given by $\frac{1}{\pi}\operatorname{Im} t_2((-z,0,0,1),(1,0,0,0))$, where $z = x + i\eta$. Based on (2.2.5), (2.2.6), and (2.2.7), we can get

$$\begin{aligned} t_2((-z,0,0,1),(1,0,0,0)) &= t_1\left(\left(-z+\frac{1-b_\sigma^2}{s_2},0,\frac{b_\sigma^2}{s_2}\right),(1,0,0)\right) \\ &= t_0\left(\left(-z+\frac{1-b_\sigma^2}{s_2}+\frac{1-b_\sigma^2}{s_1},\frac{b_\sigma^2}{s_1}\right),(1,0)\right) \\ &= \int\left(-z+\frac{1-b_\sigma^2}{s_2}+\frac{1-b_\sigma^2}{s_1}+\frac{b_\sigma^2}{s_1}x\right)^{-1} d\mu_0(x), \end{aligned}$$

where $s_1, s_2 \in \mathbb{C}^+$ satisfy the fixed point equations

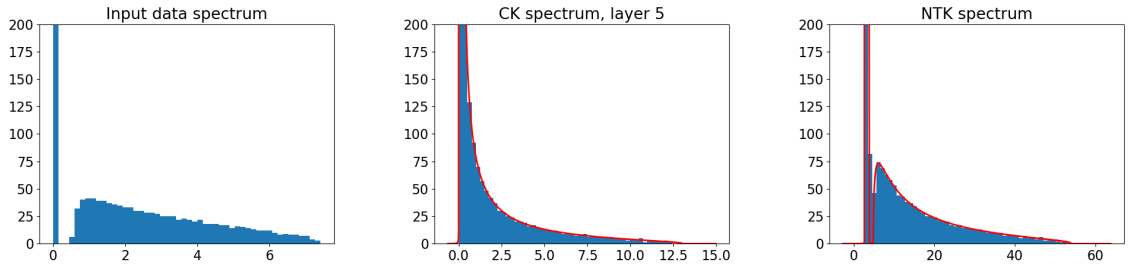
$$s_2 = 1 + \gamma_2 s_2 + \gamma_2 t_0\left(\left(-z+\frac{1-b_\sigma^2}{s_2}+\frac{1-b_\sigma^2}{s_1},\frac{b_\sigma^2}{s_1}\right),(s_2 z,0)\right) \tag{2.10.1}$$

$$s_1 = \frac{s_2}{b_\sigma^2} + \gamma_1 t_0\left(\left(-z+\frac{1-b_\sigma^2}{s_2}+\frac{1-b_\sigma^2}{s_1},\frac{b_\sigma^2}{s_1}\right),(1-b_\sigma^2,b_\sigma^2)\right). \tag{2.10.2}$$

We randomly initialize $s_1^{(0)}, s_2^{(0)} \in \mathbb{C}^+$, and update $s_1^{(t+1)}, s_2^{(t+1)}$ simultaneously by substituting $s_1 = s_1^{(t)}$ and $s_2 = s_2^{(t)}$ into the right side of (2.10.1) and (2.10.2). We iterate this until convergence, and then substitute into the above expression for $t_2((-z,0,0,1),(1,0,0,0))$ to approximate the limit spectral density of $X_2^\top X_2$ at x.

2.11 Experiments

We describe in Section 2.10 an algorithm to numerically compute the limit spectral densities of Theorem 9. The computational cost is independent of the dimensions $(n, d_0, \ldots, d_L)$, and each limit density below was computed within a few seconds on our laptop computer. Using this procedure, in this section, we investigate the accuracy of the theoretical predictions of Theorems 6 and 9.



(a) (b) (c)

Figure 2.1. Simulated spectra at initialization for i.i.d. Gaussian training samples in a 5-layer network, for (a) the input gram matrix $X_0^\top X_0$, (b) $\boldsymbol{K}^{\mathrm{CK}} = X_5^\top X_5$, and (c) $\boldsymbol{K}^{\mathrm{NTK}}$. Numerical computations of the limit spectra in Theorems 6 and 9 are superimposed in red.

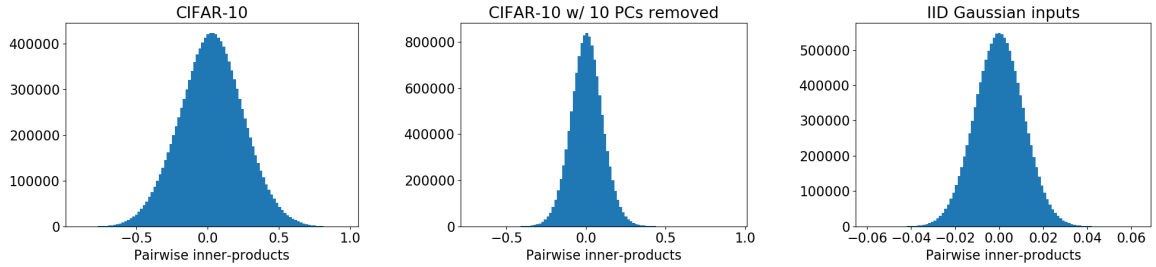
2.11.1 Simulated Gaussian Training Data

We consider $n = 3000$ training samples with i.i.d. $\mathcal{N}(0, 1/d_0)$ entries, input dimension $d_0 = 1000$, and $L = 5$ hidden layers of dimensions $d_1 = \ldots = d_5 = 6000$. We take $\sigma(x) \propto \tan^{-1}(x)$, normalized so that $\mathbb{E}[\sigma(\xi)^2] = 1$. A close agreement between the observed and limit spectra is displayed in Figure 2.1, for both $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ at initialization.

We highlight two qualitative phenomena: The spectral distribution of the NTK (at initialization) is separated from 0, as explained by the Id component in Lemma 7. Across layers $\ell = 1, \ldots, L$, there is a merging of the spectral bulk components of the CK, and an extension of its spectral support.

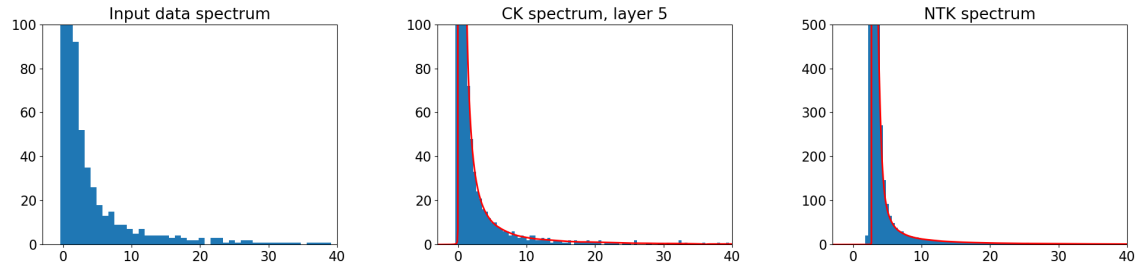
2.11.2 CIFAR-10 Training Data

We consider $n = 5000$ samples randomly selected from the CIFAR-10 training set [Kri09], with input dimension $d_0 = 3072$, and $L = 5$ hidden layers of dimensions $d_1 = \ldots = d_5 = 10000$. Strong principal component structure may cause the training samples to have large pairwise inner products, which is shown in Figure 2.2. CIFAR-10 training samples were mean-centered and normalized to satisfy $\mathbf{x}_\alpha^\top 1 = 0$ and $\|\mathbf{x}_\alpha\|^2 = 1$ in Figure 2.2(a) and (b). The pairwise inner-products in Figure 2.2(a) span a typical range of $[-0.5, 0.5]$. Those in Figure 2.2(b) span a range of about $[-0.2, 0.2]$, and those in Figure 2.2(c) about $[-0.02, 0.02]$. Thus, with 10 PCs removed,



(a) (b) (c)

Figure 2.2. All pairwise inner-products $\{\mathbf{x}_\alpha^\top \mathbf{x}_\beta : 1 \le \alpha < \beta \le n\}$, for (a) 5000 CIFAR-10 training samples, (b) 5000 CIFAR-10 training samples with the first 10 PCs removed, and (c) i.i.d. Gaussian training data of the same dimensions.



(a) (b) (c)

Figure 2.3. Same plots as Figure 2.1, for 5000 training samples from CIFAR-10 with 10 leading PCs removed.

these inner-products for CIFAR-10 are larger than for i.i.d. Gaussian inputs by a factor of 10. Thus, before computing the spectra of CK and NTK on the CIFAR-10 dataset, we pre-process the training samples by removing the leading 10 PCs—a few example images before and after this removal are depicted in the Appendix of [FW20]. A close agreement between the observed and limit spectra is displayed in Figure 2.3, for both $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$. Without removing these leading 10 PCs, there is still a close agreement for $\boldsymbol{K}^{\mathrm{CK}}$ but a deviation from the theoretical prediction for $\boldsymbol{K}^{\mathrm{NTK}}$. This suggests that the approximation in Lemma 7 is sensitive to large but low-rank perturbations of X.

In conclusion, we analyze the limiting eigenvalue distributions of CK and NTK for linear-width neural networks at random initialization. These results can be viewed as a benchmark for spectral analysis of trained neural network models. In the following chapters, we will extend

these results from different perspectives. Here, we derive the convergence of empirical eigenvalue distribution of CK to a deformed Marčenko-Pastur law, globally. Chapter 3 will address the possibility of spikes in CK, i.e. extreme eigenvalues of the CK which converge outside the bulk of deformed Marčenko-Pastur law. In Chapter 4, we will analyze the limiting eigenvalue distributions of CK and NTK when the width is much larger than the sample size. Additionally, in Section 2.11, we only present the spectra of CK and NTK at random initialization, but the spectra behavior for trained neural networks is more intriguing for deep learning theory. In Chapter 5, we will further explore the spectral behavior of CK and NTK matrices during the training process and compare with the results of limiting spectra at initialization from this Chapter.

2.12 Acknowledgment

Chapter 2 is extracted from "Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems* 33 (2020): 7710-7721". The thesis author is the co-author of this paper.

Chapter 3

Spike Analysis for Linear-Width Multi-Layer NNs

In this Chapter, we study the spike eigenvalues of the CK matrix at random initialization. It is worth noting that while Chapter 2 establishes the weak convergence of the empirical spectral measure, the precise behavior of "spike" eigenvalues that are separated from the spectral bulk remains largely unexplored. In learning applications, these spike eigenvalues and corresponding eigenvectors are often the primary spectral features (signal) of interest, because they pertain to low-rank structure of the underlying learning problem (e.g., class labels or the direction of the target function). For the linearly defined *spiked covariance model* $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n\times d}$, whose dependence across features is induced by a linear map $\boldsymbol{\Sigma}^{1/2}(\cdot)$ applied to $\boldsymbol{Z}$ having i.i.d. coordinates, classical work in random matrix theory provides a quantitative description of the spike eigenvalue/eigenvector behavior [Joh01, BS06, BGN12, BKYY16]. In this Chapter, we establish an analogous characterization of spiked spectral structure for the CK, motivated in part by the following applications:

Real data often contain **low-dimensional structure** despite the high ambient dimensionality [LV07, HTFF09, PZA$^+$21], and the leading eigenvectors of the input covariance matrix may be good predictors of the training labels. Common examples where the input features exhibit a low-dimensional spiked structure include Gaussian mixture models [LGC$^+$21a, RGKZ21, BAGJ23] and the block-covariance setting of [GMMM20, BES$^+$23, MHWSE23]. Assuming

that the input data $\boldsymbol{X}$ has *informative* spikes eigenvectors, we ask the natural question:

How does the low-dimensional signal propagate through nonlinear layers of the NN?

When do we observe a similar spiked structure in the CK matrix?

3.1 Related Work

Eigenvalues of nonlinear random matrices.

Global convergence of the empirical eigenvalue distribution of nonlinear kernel matrices has been studied in both proportional and polynomial scaling regimes [EK10, CS13, FM19, LY22, DLMY23]. Building upon related techniques, recent works characterized the spectrum of the CK matrix [PW17, LLC18, Péc19] and the neural tangent kernel (NTK) matrix [MZ20, AP20], with generalizations to deeper networks studied in [FW20] and [Cho23].

[BP22] gave a precise characterization of the largest eigenvalue in a one-hidden-layer CK matrix when the input data $\boldsymbol{X}$ and weight matrix $\boldsymbol{W}$ both have i.i.d. entries, identifying possible uninformative spike eigenvalues when the nonlinear activation is not an odd function. [GKK$^+$23] and [Fel23a] recently characterized spiked eigenstructure in models where an activation is applied to a spiked Wigner matrix or rectangular information-plus-noise matrix entrywise, for possibly growing spike sizes and activations having degenerate information/Hermite coefficients.

Precise error analysis of NNs.

An important application of spectral analyses of the CK matrix is the precise computation of generalization error of random features regression, first performed for two-layer models in proportional scaling regimes [LLC18, MM22] and later extended to deep random features models [SCDL23, BPH23] and polynomial scaling regimes [GMMM21, XHM$^+$22]. These risk analyses reveal a *Gaussian equivalence principle*, where generalization error coincides with that of a Gaussian covariates model, and this equivalence has been extended to other settings of nonlinear (regularized) empirical risk minimization [HL20, GLR$^+$21, MS22].

Going beyond random features, [BES$^+$22] derived the precise asymptotics of representation learning in a two-layer NN when the first-layer weights are trained by one (or finitely

many) gradient descent steps; see also [DLS22, BES+23, DKL+23]. The computation follows from an information-plus-noise characterization of the weight matrix due to a low-rank gradient update. [MLHD23] derived a corresponding information-plus-noise decomposition of the CK matrix defined by the resulting trained weights, in an asymptotic regime different from ours where the learning rate and spike eigenvalues diverge. [BAGHJ23] examined the emerging spike eigenstructure in the NN Hessian that arises during SGD training.

Eigenvalues of sample covariance matrices.

Asymptotic spectral analyses of sample covariance matrices have a long history in random matrix theory [MP67b, Sil95, SB95, BS98], with the strongest known results in the linearly defined model $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, see e.g. [BEK+14, KY17]. Outside of this linear setting, [SV13] and [CT18] develop sharp bounds for the extremal eigenvalues with isotropic population covariance, and [BX22] develop eigenvalue rigidity and Tracy-Widom fluctuation results for isotropic and log-concave distributions.

The spiked covariance model was introduced in [Joh01]. [BAP05, BS06, Pau07] initiated the study of spiked eigenstructure and phase transition phenomena for spiked covariance matrices with isotropic bulk covariance. [Péc06, BGN11, BGN12, Cap13, Cap18] studied spiked eigenstructure in related Wigner and information-plus-noise models. Closely related to our work are the results of [BY12] that characterize spike eigenvalues in linearly defined models $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ with general population covariance $\boldsymbol{\Sigma}$, and we extend this characterization to nonlinear settings.

3.2 Propagation of Signal Through Multi-Layer NNs

Consider input features $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ are independent samples. Define a L-hidden-layer feedforward neural network by (1.1.3). The Conjugate Kernel (CK) at each layer $\ell = 1, \ldots, L$ is given by the Gram matrix

$$\boldsymbol{K}_\ell = \boldsymbol{X}_\ell^\top \boldsymbol{X}_\ell \in \mathbb{R}^{n \times n}. \tag{3.2.1}$$

In the limit $n, d_0, \ldots, d_L \to \infty$ with $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each $\ell = 0, \ldots, L$, under deterministic conditions for the input data $\boldsymbol{X}$ and for random weight matrices $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ as specified below, Chapter 2 showed that the empirical eigenvalue distribution $\widehat{\mu}_\ell$ of $\boldsymbol{K}_\ell$ for each $\ell = 1, \ldots, L$ satisfies the weak convergence

$$\widehat{\mu}_\ell := \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{K}_\ell)} \to \mu_\ell \text{ a.s.} \tag{3.2.2}$$

for limit measures $\mu_1, \ldots, \mu_L$ defined as follows: Let μ_0 be the limit eigenvalue distribution of the input gram matrix $\boldsymbol{K}_0 = \boldsymbol{X}^\top \boldsymbol{X}$ (c.f. Assumption 4). Then, for $\ell = 1, \ldots, L$, let

$$\nu_{\ell-1} = b_\sigma^2 \otimes \mu_{\ell-1} \oplus (1 - b_\sigma^2) \tag{3.2.3}$$

denote the law of $b_\sigma^2 \mathsf{X} + (1 - b_\sigma^2)$ when $\mathsf{X} \sim \mu_{\ell-1}$ and $b_\sigma := \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$, and define

$$\mu_\ell = \rho_{\gamma_\ell}^{\mathrm{MP}} \boxtimes \nu_{\ell-1}. \tag{3.2.4}$$

Here, $\rho_\gamma^{\mathrm{MP}} \boxtimes \nu$ is *deformed Marčenko-Pastur law* defined in Chapter 1.

In this section, we provide a precise quantitative characterization of the spike eigenvalues and eigenvectors of $\boldsymbol{K}_\ell$ for each $\ell = 1, \ldots, L$ when the input data $\boldsymbol{X}$ has a fixed number of spike singular values of bounded magnitude. We assume the following conditions for the random weights, input data, and activation.

Assumption 3. The number of layers $L \geq 1$ is fixed, and $n, d_0, \ldots, d_L \to \infty$ such that

$$n/d_\ell \to \gamma_\ell \in (0, \infty) \text{ for each } \ell = 0, \ldots, L.$$

The weights $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ have entries $[\boldsymbol{W}_\ell]_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1)$, independent of each other and of $\boldsymbol{X}$.

Definition 32. A feature matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ is τ_n**-orthonormal** if

$$\big|\|\boldsymbol{x}_\alpha\| - 1\big| \le \tau_n, \qquad \big|\|\boldsymbol{x}_\beta\| - 1\big| \le \tau_n, \qquad \big|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\big| \le \tau_n$$

for all pairs $\alpha \ne \beta \in [n]$, where $\{\boldsymbol{x}_\alpha\}_{\alpha=1}^n$ are the columns of $\boldsymbol{X}$.

Assumption 4. For some $\tau_n > 0$ such that $\lim_{n\to\infty} \tau_n \cdot n^{1/3} = 0$, $\boldsymbol{X} \equiv \boldsymbol{X}_0$ is τ_n-orthonormal almost surely for all large n. Furthermore, $\boldsymbol{K}_0 = \boldsymbol{X}^\top \boldsymbol{X}$ has eigenvalues $\lambda_1(\boldsymbol{K}_0), \dots, \lambda_n(\boldsymbol{K}_0)$ (not necessarily ordered by magnitude) such that for some fixed $r \ge 0$, as $n, d \to \infty$,

(a) There exists a compactly supported probability measure μ_0 on $[0,\infty)$ such that

$$\frac{1}{n-r} \sum_{i=r+1}^{n} \delta_{\lambda_i(\boldsymbol{K}_0)} \to \mu_0 \text{ weakly a.s.}$$

and for any fixed $\varepsilon > 0$, almost surely for all large n,

$$\lambda_i(\boldsymbol{K}_0) \in \operatorname{supp}(\mu_0) + (-\varepsilon, \varepsilon) \text{ for all } i \ge r+1.$$

(b) There exist distinct values $\lambda_1, \dots, \lambda_r > 0$ with $\lambda_1, \dots, \lambda_r \notin \operatorname{supp}(\mu_0)$ such that

$$\lambda_i(\boldsymbol{K}_0) \to \lambda_i \quad \text{ a.s. for each } \quad i = 1, \dots, r.$$

Assumption 5. The activation $\sigma : \mathbb{R} \to \mathbb{R}$ is twice differentiable with $\sup_{x\in\mathbb{R}} |\sigma'(x)|, |\sigma''(x)| \le \lambda_\sigma$ for some $\lambda_\sigma \in (0,\infty)$. Under $\xi \sim \mathcal{N}(0,1)$, we have $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma^2(\xi)] = 1$. Furthermore,

$$b_\sigma := \mathbb{E}[\sigma'(\xi)] \ne 0, \qquad \mathbb{E}[\sigma''(\xi)] = 0. \tag{3.2.5}$$

Assumption 3 defines the linear-width asymptotic regime. Similarly to Assumption 1(c), Assumption 4 requires an orthogonality condition for the input features, and also codifies our spiked eigenstructure assumption for the input data. We briefly comment on (3.2.5) in

Assumption 5: The condition $b_\sigma \neq 0$ ensures that the linear component of $\sigma(\cdot)$ is non-degenerate; if $b_\sigma = 0$, then spiked eigenstructure does not propagate across the NN layers in our studied regime of bounded spike magnitudes. The condition $\mathbb{E}[\sigma''(\xi)] = 0$ ensures that $\boldsymbol{K}_\ell$ does not have uninformative spike eigenvalues; otherwise, as shown in [BP22], $\boldsymbol{K}_\ell$ may have spike eigenvalues even when the input $\boldsymbol{K}_0$ has no spiked structure. We assume $\mathbb{E}[\sigma''(\xi)] = 0$ for clarity, to avoid characterizing also such uninformative spikes across layers. This condition holds, in particular, for odd activation functions $\sigma(\cdot)$ such as tanh.

The following theorem first extends Theorem 6 by affirming that the weak convergence statement (3.2.2) holds under the above assumptions, and furthermore, each $\boldsymbol{K}_\ell$ has no outlier eigenvalues outside its limit spectral support when the input $\boldsymbol{K}_0$ has no spike eigenvalues.

Theorem 33. *Suppose Assumptions 3, 4, and 5 hold. Then for each $\ell = 1, \dots, L$,* (3.2.2) *holds weakly a.s. as $n \to \infty$. Furthermore, if the number of spikes is $r = 0$ in Assumption 4, then for any fixed $\varepsilon > 0$, almost surely for all large n,*

$$\boldsymbol{K}_\ell \text{ has no eigenvalues outside } \operatorname{supp}(\mu_\ell) + (-\varepsilon, \varepsilon).$$

The main result of this section characterizes the eigenvalues of $\boldsymbol{K}_\ell$ outside $\operatorname{supp}(\mu_\ell)$ when $r \geq 1$. To describe this characterization, define for each $\ell = 1, \dots, L$ the domain

$$\mathscr{T}_\ell = \{-1/\lambda : \lambda \in \operatorname{supp}(\nu_{\ell-1})\}$$

where $\nu_{\ell-1}$ is defined by (3.2.3), and define $z_\ell, \varphi_\ell : (0,\infty) \setminus \mathscr{T}_\ell \to \mathbb{R}$ by

$$z_\ell(s) = -\frac{1}{s} + \gamma_\ell \int \frac{\lambda}{1+\lambda s} \nu_{\ell-1}(d\lambda), \qquad \varphi_\ell(s) = -\frac{s z_\ell'(s)}{z_\ell(s)}. \tag{3.2.6}$$

It is known from the results of [BY12] and [YZB15, Chapter 11] that these are precisely the functions that characterize the spike eigenvalues and eigenvectors in linear spiked covariance

models. Set

$$\mathscr{I}_0 = \{1, \dots, r\}, \qquad s_{i,0} = -\frac{1}{b_\sigma^2 \lambda_i + (1 - b_\sigma^2)} \text{ for } i \in \mathscr{I}_0,$$

where λ_i and b_σ are defined in Assumptions 4 and 5 respectively. Here, $\mathscr{I}_0$ records the indices of the spike eigenvalues of the input Gram matrix $\boldsymbol{K}_0$. Then define recursively for $\ell = 1, \dots, L$

$$\mathscr{I}_\ell = \Big\{ i \in \mathscr{I}_{\ell-1} : z_\ell'(s_{i,\ell-1}) > 0 \Big\}, \qquad s_{i,\ell} = -\frac{1}{b_\sigma^2 z_\ell(s_{i,\ell-1}) + (1 - b_\sigma^2)} \text{ for } i \in \mathscr{I}_\ell. \tag{3.2.7}$$

The condition $z_\ell'(s_{i,\ell-1}) > 0$ describes the "phase transition" phenomenon for spike eigenvalues in this model, where spikes $i \in \mathscr{I}_{\ell-1}$ with $z_\ell'(s_{i,\ell-1}) > 0$ induce spike eigenvalues in the CK matrix $\boldsymbol{K}_\ell$ of the next layer, while spikes with $z_\ell'(s_{i,\ell-1}) \le 0$ are absorbed into the bulk spectrum of $\boldsymbol{K}_\ell$.

Theorem 34. *Suppose Assumptions 3, 4, and 5 hold. Then for each $\ell = 1, \dots, L$:*

(a) $s_{i,\ell-1} \in (0,\infty) \setminus \mathscr{T}_\ell$ for each $i \in \mathscr{I}_{\ell-1}$, so $z_\ell(s_{i,\ell-1})$ and $\mathscr{I}_\ell$ are well-defined. Furthermore, if $i \in \mathscr{I}_\ell$ (i.e. if $z_\ell'(s_{i,\ell-1}) > 0$) then $z_\ell(s_{i,\ell-1}) > 0$ and $\varphi_\ell(s_{i,\ell-1}) > 0$.

(b) For any fixed and sufficiently small $\varepsilon > 0$, almost surely for all large n, there is a 1-to-1 correspondence between the eigenvalues of $\boldsymbol{K}_\ell$ outside $\operatorname{supp}(\mu_\ell) + (-\varepsilon, \varepsilon)$ and $\{i : i \in \mathscr{I}_\ell\}$. Denoting these eigenvalues of $\boldsymbol{K}_\ell$ by $\{\widehat{\lambda}_{i,\ell} : i \in \mathscr{I}_\ell\}$, for each $i \in \mathscr{I}_\ell$ as $n \to \infty$,

$$\widehat{\lambda}_{i,\ell} \to z_\ell(s_{i,\ell-1}) \text{ a.s.}$$

(c) Let $\widehat{\boldsymbol{v}}_{i,\ell}$ be a unit-norm eigenvector of $\boldsymbol{K}_\ell$ corresponding to its eigenvalue $\widehat{\lambda}_{i,\ell}$, and let $\boldsymbol{v}_j$ be a unit-norm eigenvector of $\boldsymbol{K}_0$ corresponding to its spike eigenvalue $\lambda_j(\boldsymbol{K}_0)$. Then for each $i \in \mathscr{I}_\ell$ and $j \in \mathscr{I}_0$, as $n \to \infty$,

$$|\widehat{\boldsymbol{v}}_{i,\ell}^\top \boldsymbol{v}_j|^2 \to \prod_{k=1}^{\ell} \varphi_k(s_{i,k-1}) \cdot \mathbf{1}\{i = j\} \text{ a.s.}$$

Moreover, for each $i \in \mathscr{I}_\ell$ and any unit vector $\boldsymbol{v} \in \mathbb{R}^n$ independent of $\boldsymbol{W}_1, \dots, \boldsymbol{W}_\ell$,

$$|\widehat{\boldsymbol{v}}_{i,\ell}^\top \boldsymbol{v}|^2 - \prod_{k=1}^{\ell} \varphi_k(s_{i,k-1}) \cdot |\boldsymbol{v}_i^\top \boldsymbol{v}|^2 \to 0 \text{ a.s.}$$

We present the following corollary as a concrete example in which the assumptions of the theorem are satisfied. The corollary encompasses, for instance, Gaussian mixture models with a fixed number r of balanced classes, each class having $\Theta(n)$ samples.

Corollary 35. *Suppose the input data $\boldsymbol{X}$ is itself a low-rank signal-plus-noise matrix*

$$\boldsymbol{X} = \sum_{i=1}^{r} \theta_i \boldsymbol{a}_i \boldsymbol{b}_i^\top + \boldsymbol{Z} \in \mathbb{R}^{d \times n} \tag{3.2.8}$$

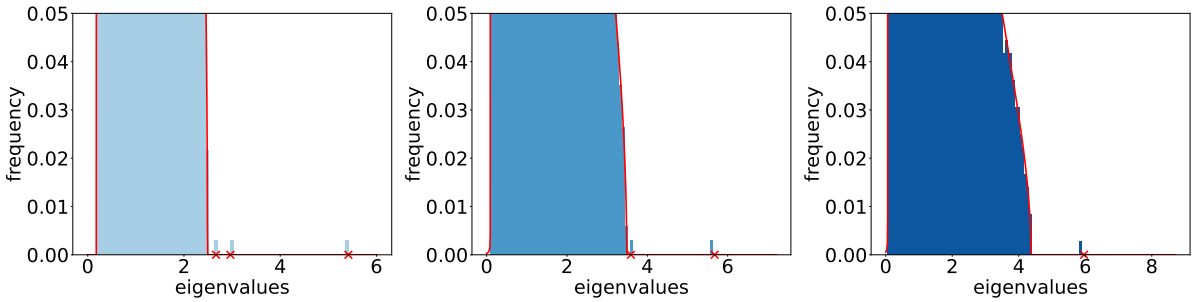
where $\theta_1, \dots, \theta_r > 0$ are fixed distinct signal strengths, $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r \in \mathbb{R}^d$ and $\boldsymbol{b}_1, \dots, \boldsymbol{b}_r \in \mathbb{R}^n$ are orthonormal sets of unit vectors, and $\boldsymbol{Z}$ has i.i.d. $\mathcal{N}(0, 1/d)$ entries. Assume that $\boldsymbol{b}_1, \dots, \boldsymbol{b}_r$ satisfy the ℓ_∞-delocalization condition: for any sufficiently small $\varepsilon > 0$ and all large n,

$$\max_{1 \le i \le r} \|\boldsymbol{b}_i\|_\infty < n^{-1/2+\varepsilon}.$$

Define $\varphi_\ell(\cdot)$ and $s_{i,\ell-1}$ by (3.2.6) *and* (3.2.7)*, with the initial measures $\mu_0 = \rho_{\gamma_0}^{\mathrm{MP}}$ and $\nu_0 = b_\sigma^2 \otimes \mu_0 \oplus (1 - b_\sigma^2)$ and initial spike values $\lambda_i = (1+\theta_i^2)(\gamma_0 + \theta_i^2)/\theta_i^2$ for $i \in \mathscr{I}_0$.*

Then for each $\ell = 1, \dots, L$, $\boldsymbol{K}_\ell$ has a spike eigenvalue corresponding to the input signal component θ_i if and only if $\theta_i > \gamma_0^{1/4}$ and $i \in \mathscr{I}_\ell$. In this case, its corresponding unit eigenvector $\widehat{\boldsymbol{v}}_{i,\ell}$ satisfies, as $n \to \infty$,

$$|\widehat{\boldsymbol{v}}_{i,\ell}^\top \boldsymbol{b}_i|^2 \to \prod_{k=1}^{\ell} \varphi_k(s_{i,k-1}) \cdot \left(1 - \frac{\gamma_0(1+\theta_i^2)}{\theta_i^2(\theta_i^2+\gamma_0)}\right) \text{ a.s.} \tag{3.2.9}$$



(a) Spectrum of $\boldsymbol{K}_0$ (b) Spectrum of $\boldsymbol{K}_1$ (c) Spectrum of $\boldsymbol{K}_2$

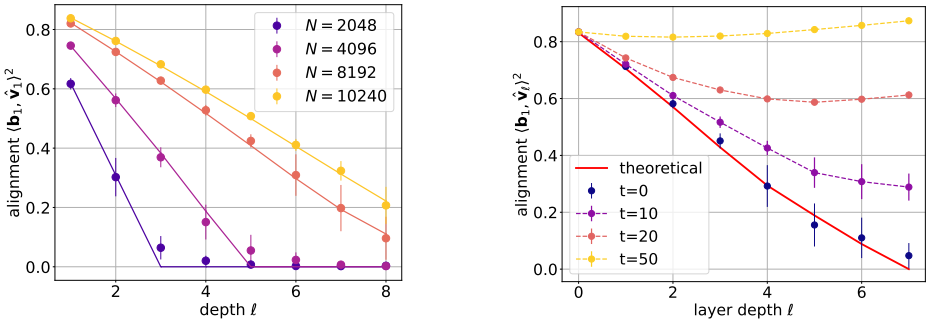
Figure 3.1. Spectra of three-layer CK matrices defined by (3.2.1) on GMM input data with $r = 3$ in (3.2.8). (a)-(c) are theoretically predicted (red) and empirical (blue) bulk distributions and spikes of $\boldsymbol{K}_\ell$ for $\ell = 0, 1, 2$.

Numerical illustration.

A simple illustration of this result for a 3-component Gaussian mixture model is provided in Figure 3.1. Here, we present the spectra of three-layer CK matrices defined by (3.2.1) with $n = 5000$, $d_0 = d_1 = d_2 = 15000$, and $\sigma \propto \arctan$. The input data is a GMM satisfying (3.2.8) with $r = 3$, $\theta_1 = 2.0$, $\theta_2 = 1.18$, and $\theta_3 = 1.0$. Observe that, in Figure 3.1, the number of informative spikes is non-increasing with respect to the depth. Theorem 34 shows that $\mathscr{I}_L \subseteq \cdots \subseteq \mathscr{I}_0$ and $\varphi_\ell(s_{i,\ell-1}) \in (0,1)$, so the number of spike eigenvalues of $\boldsymbol{K}_\ell$ induced by $\boldsymbol{K}_0$ and the alignment of the spike eigenvectors of $\boldsymbol{K}_\ell$ with the true class label vectors $\{\boldsymbol{b}_i\}_{i=1}^r$ are both non-increasing in the network depth, see also Figure 3.2. In other words, at random initialization, the input signal diminishes as the depth of the NN increases.

In Figure 3.2(a), we consider multiple-layer NNs at random initialization in (1.1.3) with varying hidden widths $N = 2048, 4096, 8192, 10240$, and the activation function $\sigma \propto \tanh$ satisfying Assumption 5. Here we propagate a Gaussian mixture data (3.2.8) with $r = 1$ and $\theta_1 = 2.5$. Figure 3.2(a) presents the eigenvector alignment between the largest eigenvector $\widehat{\boldsymbol{v}}_{1,\ell}$ of the CK matrix $\boldsymbol{K}_\ell$ with genuine signal $\boldsymbol{b}_1$ (class labels) for different layer $\ell = 1, \ldots, 8$.

Figure 3.2(b) considers a NN with $d_\ell = 6000$ for $\ell = 0, 1, \ldots, 7$ using $n = 2000$ training data points sampled from (3.2.8) with $r = 1$ and $\theta_1 = 1.8$. Figure 3.2(b) shows the eigenvector alignment between the largest eigenvector $\widehat{\boldsymbol{v}}_{1,\ell}$ of the CK matrix $\boldsymbol{K}_\ell$ with genuine signal $\boldsymbol{b}_1$



(a) Effect of width on alignment.

(b) Effect of GD training.

Figure 3.2. We consider multiple-layer NNs in (1.1.3) on Gaussian mixture data with $r = 1$, and compute the alignment between the largest eigenvector of the CK matrix $\boldsymbol{K}_\ell$ with genuine signal $\boldsymbol{b}_1$ (class labels) for different layer ℓ. (a) NNs at random initialization with varying hidden widths $N = 2048, 4096, 8192, 10240$. (b) NNs trained by gradient descent with learning rate $\eta = 0.1$ for varying steps $t = 0, 10, 20, 50$.

(class labels), for different layer $\ell = 0, \ldots, 7$. Here, we train the NN by gradient descent with learning rate $\eta = 0.1$ for varying steps $t = 0, 10, 20, 50$; we use the μ-parameterization [YH20] to encourage feature learning. When $t = 0$, Figure 3.2(b) presents the initial eigenvector alignment at different depths, which matches the theoretical solid curve in red from Corollary 35.

In Figure 3.2, all dots are empirical values (over 10 runs) and solid curves represent theoretical predictions at random initialization from Theorem 34. In summary, we can observe that Figure 3.2 highlights two remedies to this "curse of depth" at random initialization.

- In Figure 3.2(a) we observe that when the width of NN becomes larger, alignment between the leading eigenvector of $\boldsymbol{K}_\ell$ at random initialization and the signal can be preserved across a larger depth. This illustrates the benefit of overparameterization by increasing the network width.

- In Figure 3.2(b) we observe that gradient descent training on the weight matrices also restores and even amplifies the informative signal in the CK matrix of each layer; specifically, after 50 steps of GD training (yellow curve), the alignment between the class labels and the leading eigenvector of $\boldsymbol{K}_\ell$ may increase through depth. This demonstrates the benefit of gradient-based feature learning. In Section 5.4.2 we precisely quantify this improved alignment due to gradient descent in a simplified two-layer setting.

In addition, Figure 3.2(b) illustrates that gradient descent training on the parameters also restores and even amplifies the informative signal in the CK matrix of each layer, demonstrating the benefit of feature learning. In Section 5.4.2 we precisely quantify this improved alignment due to gradient descent in a simplified two-layer setting.

3.3 Results For the Nonlinear Spiked Covariance Model

In this section, we state a new random matrix result for nonlinear spiked sample covariance matrices. The proof will be presented in the next section. Theorem 34 for spikes in CK matrices can then be analyzed by this general random matrix result. In Chapter 5, we will employ this nonlinear spiked random matrix result again to investigate the trained CK matrix. Before stating the main results, we first introduce the following notation and proposition.

Stochastic domination notation.

We use the following standard notation for stochastic domination of random variables, see e.g. [EKY13, Definition 2.4]: For random variables $X \equiv X(u)$ and $Y \equiv Y(u) \geq 0$ depending implicitly on N and a parameter $u \in U_N$, as $N \to \infty$, we write

$$X \prec Y \text{ or } X = O_{\prec}(Y) \text{ uniformly over } u \in U_N$$

if, for any fixed $\varepsilon, D > 0$ and all large N,

$$\sup_{u \in U_N} \mathbb{P}\Big[|X(u)| > N^{\varepsilon} Y(u)\Big] < N^{-D}.$$

Throughout, "for all large N" means for all $N \geq N_0$ where N_0 may depend on ε, D, any quantities that are constant in the context of the statement, and convergence rates of the spike eigenvalues and empirical spectral measures in the given assumptions.

If $X = \mathbf{1}\{\mathscr{E}\}$ is the indicator of an event $\mathscr{E} \equiv \mathscr{E}_N$, then $\mathbf{1}\{\mathscr{E}\} \prec 0$ means $\mathbb{P}[\mathscr{E}] < N^{-D}$ for any fixed $D > 0$ and all large N. If X and Y are both deterministic, then $X \prec Y$ means $|X| \leq N^{\varepsilon} Y$

(deterministically) for any $\varepsilon > 0$ and all large N. For an event $\mathscr{E} \equiv \mathscr{E}_N$, we will write

$$X = O_{\prec}^{\mathscr{E}}(Y)$$

as shorthand for $X \cdot \mathbf{1}\{\mathscr{E}\} \prec Y$.

We will use the following basic properties often implicitly.

Proposition 36. Suppose $X \prec Y$ uniformly over $u \in U_N$.

(a) If $|U_N| \le N^C$ for a constant $C > 0$, then for any fixed $\varepsilon, D > 0$ and all large N,

$$\mathbb{P}\Big[\text{there exists } u \in U_N \text{ with } |X(u)| \ge N^{\varepsilon} Y(u)\Big] \le N^{-D}.$$

(b) If $|U_N| \le N^C$ for a constant $C > 0$, then $\sum_{u \in U_N} X(u) \prec \sum_{u \in U_N} Y(u)$.

(c) If $|U_N| \le C$ for a constant $C > 0$, then $\prod_{u \in U_N} X(u) \prec \prod_{u \in U_N} Y(u)$.

(d) If Y is deterministic, and $\mathbb{E}[X^2] \le N^C$ and $Y \ge N^{-C}$ for a constant $C > 0$, then also $\mathbb{E}[|X|] \prec Y$ uniformly over $u \in U_N$.

Proof. The first three statements follow from a union bound over U_N. For the last statement, for any fixed $\varepsilon > 0$, observe that

$$\mathbb{E}|X| \le N^{\varepsilon/2} Y + \mathbb{E}\Big[|X| \mathbf{1}\{|X| > N^{\varepsilon/2} Y\}\Big] \le N^{\varepsilon/2} Y + \mathbb{E}[X^2]^{1/2} \mathbb{P}[|X| > N^{\varepsilon/2} Y]^{1/2}.$$

Applying $\mathbb{E}[X^2] \le N^C$, $Y \ge N^{-C}$, and $\mathbb{P}[|X| > N^{\varepsilon/2} Y] < N^{-D}$ for sufficiently large $D > 0$ shows that the second term is less than $N^{\varepsilon/2} Y$ for all large N, hence $\mathbb{E}|X| < N^{\varepsilon} Y$. □

3.3.1 Deterministic Equivalent for the Resolvent

We consider the sample covariance and Gram matrix

$$\boldsymbol{K} = \boldsymbol{G}^\top \boldsymbol{G} \in \mathbb{R}^{n\times n}, \quad \widetilde{\boldsymbol{K}} = \boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{N\times N}, \quad \text{where} \quad \boldsymbol{G} = \frac{1}{\sqrt{N}}[\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N] \in \mathbb{R}^{N\times n}.$$

The following are our basic assumptions, where we recall that $\mathbf{1}\{\mathscr{E}\} \prec 0$ means $\mathbb{P}[\mathscr{E}] \le N^{-D}$ for any fixed $D > 0$ and all large N.

Assumption 6. The rows of $\boldsymbol{G}$ are independent and satisfy $\mathbb{E}[\boldsymbol{g}_i] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{g}_i \boldsymbol{g}_i^\top] = \boldsymbol{\Sigma}$ for all $i \in [N]$, such that:

(a) There exist constants $C, c > 0$ such that $c < n/N < C$ and $\|\boldsymbol{\Sigma}\| < C$.

(b) There exists a constant $B > 0$ such that $\mathbf{1}\{\|\boldsymbol{K}\| > B\} \prec 0$.

(c) Uniformly over deterministic matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and over $i \neq j \in [N]$,

$$\boldsymbol{g}_i^\top \boldsymbol{A}\boldsymbol{g}_i - \mathbb{E}[\boldsymbol{g}_i^\top \boldsymbol{A}\boldsymbol{g}_i] \prec \|\boldsymbol{A}\|_F, \qquad \boldsymbol{g}_i^\top \boldsymbol{A}\boldsymbol{g}_j \prec \|\boldsymbol{A}\|_F.$$

(d) For any integer $\alpha > 0$, there exists a constant $C = C(\alpha) > 0$ such that $\mathbb{E}[\|\boldsymbol{g}_i\|^\alpha] \le N^C$.

Denote the finite-N dimension ratio and empirical eigenvalue distribution of $\boldsymbol{\Sigma}$ by

$$\gamma_N = \frac{n}{N}, \qquad \nu_N = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\boldsymbol{\Sigma})}. \tag{3.3.1}$$

Let

$$\mu_N = \rho_{\gamma_N}^{\mathrm{MP}} \boxtimes \nu_N, \qquad \tilde{\mu}_N = \gamma_N \mu_N + (1-\gamma_N)\delta_0.$$

Denote the Stieltjes transforms of $\mu_N, \tilde{\mu}_N$ by $m_N(z), \tilde{m}_N(z)$. These are characterized exactly as in (1.2.4) with (γ_N, ν_N) in place of (γ, ν).

We first establish that with high probability, $\boldsymbol{K}$ and $\widetilde{\boldsymbol{K}}$ have no outlier eigenvalues far from the support set

$$\mathscr{S}_N = \mathrm{supp}(\mu_N) \cup \{0\} = \mathrm{supp}(\tilde{\mu}_N) \cup \{0\}. \tag{3.3.2}$$

Theorem 37. *Suppose Assumption 6 holds. Then for any fixed $\varepsilon > 0$,*

$$\mathbf{1}\Big\{\boldsymbol{K} \text{ has an eigenvalue outside } \mathscr{S}_N + (-\varepsilon, \varepsilon)\Big\} \prec 0.$$

In asymptotic settings where $\nu_N \to \nu$ and $\mu_N \to \mu$ weakly and $\boldsymbol{\Sigma}$ has no spike eigenvalues, this set $\mathscr{S}_N$ will converge to $\mathscr{S} := \mathrm{supp}(\mu) \cup \{0\}$. In general, $\mathscr{S}_N$ may contain intervals around spike eigenvalues of $\boldsymbol{K}$ that are separated from $\mathrm{supp}(\mu) \cup \{0\}$ if $\boldsymbol{\Sigma}$ has a spiked structure, and this will be clarified in the subsequent section.

Next, we establish a deterministic equivalent approximation for the resolvent of $\boldsymbol{K}$, for spectral arguments separated from this support set $\mathscr{S}_N$. Let us denote by

$$\boldsymbol{R}(z) = (\boldsymbol{K} - z\boldsymbol{I})^{-1}, \qquad m_{\boldsymbol{K}}(z) = \frac{1}{n}\operatorname{Tr}\boldsymbol{R}(z)$$

the resolvent and Stieltjes transform of $\boldsymbol{K}$ for $z \notin \mathrm{supp}(\mu_N)$. For any $\varepsilon > 0$, define the domain

$$U_N(\varepsilon) = \Big\{z \in \mathbb{C} : |z| \le \varepsilon^{-1}, \mathrm{dist}(z, \mathscr{S}_N) \ge \varepsilon\Big\}. \tag{3.3.3}$$

Theorem 38. *Suppose Assumption 6 holds. Then for any fixed $\varepsilon > 0$, uniformly over $z \in U_N(\varepsilon)$ and over deterministic matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, we have*

$$m_{\boldsymbol{K}}(z) - m_N(z) \prec \frac{1}{N}, \qquad \operatorname{Tr}\Big[\boldsymbol{R}(z)\boldsymbol{A} - (-z\tilde{m}_N(z)\boldsymbol{\Sigma} - z\boldsymbol{I})^{-1}\boldsymbol{A}\Big] \prec \frac{1}{\sqrt{N}}\|\boldsymbol{A}\|_F.$$

For spectral arguments $z \in \mathbb{C} \setminus \mathbb{R}_+$ separated from the positive real line, such a result has been shown recently in [Cho22, SCDL23] (using different proof techniques). We use Theorem

37 as an input to establish this approximation also for spectral arguments in $\mathbb{R}_+ \setminus \mathscr{S}_N$, as such a result (and its extension to a generalized resolvent) is needed for our analysis of spiked eigenstructure to follow.

3.3.2 Spike Eigenvalues and Eigenvectors

Now we consider an asymptotic setting with a specific spiked structure for the population covariance matrix $\mathbf{\Sigma}$, having a fixed number of spikes outside the support of the weak limit of its spectral law. This assumption is summarized as follows.

Assumption 7. $\mathbf{\Sigma}$ has eigenvalues $\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_n(\mathbf{\Sigma})$ (not necessarily ordered by magnitude) where, for a fixed integer $r \geq 0$, as $N \to \infty$:

(a) $n/N \to \gamma \in (0, \infty)$.

(b) There exists a probability measure ν with compact support in $(0, \infty)$, such that

$$\frac{1}{n-r} \sum_{i=r+1}^{n} \delta_{\lambda_i(\mathbf{\Sigma})} \to \nu \text{ weakly.}$$

Furthermore, for any fixed $\varepsilon > 0$ and all large N,

$$\lambda_i(\mathbf{\Sigma}) \in \operatorname{supp}(\nu) + (-\varepsilon, \varepsilon) \text{ for all } i \geq r+1.$$

(c) There exist distinct values $\lambda_1, \ldots, \lambda_r > 0$ with $\lambda_1, \ldots, \lambda_r \notin \operatorname{supp}(\nu)$ such that

$$\lambda_i(\mathbf{\Sigma}) \to \lambda_i \text{ for all } i = 1, \ldots, r.$$

Under this assumption, we analyze the outlier singular values of $\boldsymbol{G}$ and their corresponding singular vectors. Let

$$\gamma_{N,0} = \frac{n-r}{N}, \qquad \nu_{N,0} = \frac{1}{n-r} \sum_{i=r+1}^{n} \delta_{\lambda_i(\mathbf{\Sigma})}$$

be the finite-N aspect ratio and population spectral measure corresponding to the bulk component of $\boldsymbol{\Sigma}$. Define the laws

$$\mu_{N,0} = \rho^{\mathrm{MP}}_{\gamma_{N,0}} \boxtimes \nu_{N,0}, \qquad \tilde{\mu}_{N,0} = \gamma_{N,0}\mu_{N,0} + (1-\gamma_{N,0})\delta_0$$

and let $m_{N,0}(z), \tilde{m}_{N,0}(z)$ be their Stieltjes transforms. In the setting of Assumption 7, we note that $\mu_{N,0} \to \mu = \rho^{\mathrm{MP}}_{\gamma} \boxtimes \nu$ and $\tilde{\mu}_{N,0} \to \tilde{\mu} = \gamma\mu + (1-\gamma)\delta_0$ weakly as $N \to \infty$, where the Stieltjes transforms $m(z), \tilde{m}(z)$ of these limits $\mu, \tilde{\mu}$ are characterized by (1.2.4).

Denote the limit support set

$$\mathscr{S} = \mathrm{supp}(\mu) \cup \{0\} = \mathrm{supp}(\tilde{\mu}) \cup \{0\}. \tag{3.3.4}$$

Under Assumption 7 when $r = 0$, i.e. $\boldsymbol{\Sigma}$ does not have spike eigenvalues, the following is a corollary of Theorem 37. A similar "no outlier" statement has been shown for linearly defined sample covariance models in [BS98].

Corollary 39. *Suppose Assumptions 6 and 7 hold, where $r = 0$. Then for any fixed $\varepsilon > 0$,*

$$\mathbf{1}\Big\{ \boldsymbol{K} \textit{ has an eigenvalue outside } \mathscr{S} + (-\varepsilon, \varepsilon) \Big\} \prec 0.$$

We now give a more quantitative description of the spike eigenvalues of $\boldsymbol{K} = \boldsymbol{G}^\top \boldsymbol{G}$ and corresponding singular vectors of $\boldsymbol{G}$ when there are possibly spike eigenvalues in $\boldsymbol{\Sigma}$. Define the domain

$$\mathscr{T}_{N,0} = \{0\} \cup \{-1/\lambda : \lambda \in \mathrm{supp}(\nu_{N,0})\}.$$

For $\tilde{m} \in \mathbb{C} \setminus \mathscr{T}_{N,0}$, define the functions

$$z_{N,0}(\tilde{m}) = -\frac{1}{\tilde{m}} + \gamma_{N,0} \int \frac{\lambda}{1+\lambda\tilde{m}} d\nu_{N,0}(\lambda), \qquad \varphi_{N,0}(\tilde{m}) = -\frac{\tilde{m} z'_{N,0}(\tilde{m})}{z_{N,0}(\tilde{m})}. \tag{3.3.5}$$

We note that under Assumption 7, the domain $\mathscr{T}_{N,0}$ converges in Hausdorff distance to $\mathscr{T}$ as defined in (1.2.5). We will verify in the proof (c.f. Lemma 51) that $z_{N,0}(\tilde{m}) \to z(\tilde{m})$ and $z'_{N,0}(\tilde{m}) \to z'(\tilde{m})$ for each fixed $\tilde{m} \in \mathbb{C} \setminus \mathscr{T}$, where $z(\cdot)$ is as defined in (1.2.6). Then also $\varphi_{N,0}(\tilde{m}) \to \varphi(\tilde{m})$ for the limiting function

$$\varphi(\tilde{m}) = -\frac{\tilde{m} z'(\tilde{m})}{z(\tilde{m})}. \tag{3.3.6}$$

Theorem 40. *Suppose Assumptions 6 and 7 hold. Let*

$$\mathscr{I} = \big\{ i \in \{1, \dots, r\} : z'(-1/\lambda_i) > 0 \big\}.$$

(a) For any sufficiently small constant $\varepsilon > 0$ and all large N, on an event $\mathscr{E} \equiv \mathscr{E}_N$ satisfying $\mathbf{1}\{\mathscr{E}^c\} \prec 0$, there is a 1-to-1 correspondence between the eigenvalues of $\boldsymbol{K}$ outside $\mathscr{I} + (-\varepsilon, \varepsilon)$ and $\{\lambda_i : i \in \mathscr{I}\}$. Denoting these eigenvalues of $\boldsymbol{K}$ by $\{\widehat{\lambda}_i : i \in \mathscr{I}\}$, we have

$$\widehat{\lambda}_i - z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) = O^{\mathscr{E}}_{\prec}\bigg(\frac{1}{\sqrt{N}}\bigg)$$

for each $i \in \mathscr{I}$, where $z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) \to z(-1/\lambda_i) > 0$ as $N \to \infty$.

(b) On this event $\mathscr{E}$, for each $i \in \mathscr{I}$, let $\widehat{\boldsymbol{v}}_i \in \mathbb{R}^n$ be a unit-norm eigenvector of $\boldsymbol{K}$ (i.e. right singular vector of $\boldsymbol{G}$) corresponding to its eigenvalue $\widehat{\lambda}_i$, and let $\boldsymbol{v}_i$ be a unit-norm eigenvector of $\boldsymbol{\Sigma}$ corresponding to $\lambda_i(\boldsymbol{\Sigma})$. Then, uniformly over (deterministic) unit vectors $\boldsymbol{v} \in \mathbb{R}^n$,

$$|\boldsymbol{v}^\top \widehat{\boldsymbol{v}}_i| - \sqrt{\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))} \cdot |\boldsymbol{v}^\top \boldsymbol{v}_i| = O^{\mathscr{E}}_{\prec}\bigg(\frac{1}{\sqrt{N}}\bigg) \tag{3.3.7}$$

where $\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) \to \varphi(-1/\lambda_i) > 0$ as $N \to \infty$. In particular, for each $i \in \mathscr{I}$, $|\boldsymbol{v}_i^\top \widehat{\boldsymbol{v}}_i|^2 \to \varphi(-1/\lambda_i)$ and $\sup_{j \in [n]: j \neq i} |\boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i|^2 \to 0$ almost surely as $N \to \infty$.

(c) Let $\boldsymbol{u} = \frac{1}{\sqrt{N}}(u_1, \dots, u_N)^\top \in \mathbb{R}^N$ be a random vector such that $[\boldsymbol{u}, \boldsymbol{G}] \in \mathbb{R}^{N \times (n+1)}$ has inde-

pendent rows also satisfying Assumption 6. Denote by $\mathbb{E}[u\boldsymbol{g}] \in \mathbb{R}^n$ *the common value of* $\mathbb{E}[u_j \boldsymbol{g}_j]$ *for all* $j \in [N]$.

On this event $\mathscr{E}$*, for each* $i \in \mathscr{I}$*, let* $\widehat{\boldsymbol{u}}_i \in \mathbb{R}^N$ *be a unit-norm eigenvector of* $\widetilde{\boldsymbol{K}}$ *(i.e. left singular vector of* $\boldsymbol{G}$*) corresponding to its eigenvalue* $\widehat{\lambda}_i$*, and let* $\boldsymbol{v}_i$ *be the eigenvector of* $\boldsymbol{\Sigma}$ *as in part (b). Then*

$$|\boldsymbol{u}^\top \widehat{\boldsymbol{u}}_i| - \frac{\sqrt{z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))}}{\lambda_i(\boldsymbol{\Sigma})} \cdot \left|\mathbb{E}[u\boldsymbol{g}]^\top \boldsymbol{v}_i\right| = O_{\prec}^{\mathscr{E}}\left(\frac{1}{\sqrt{N}}\right). \tag{3.3.8}$$

3.4 Proof Ideas of Theorem 40

Statements (a–b) in Theorem 40 are known in a linear setting $\boldsymbol{g}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ when $\boldsymbol{z}_i$ has i.i.d. entries, see e.g. [BY12] and [YZB15, Theorems 11.3 and 11.5]. The above theorem thus verifies an exact asymptotic *equivalence* between spiked spectral phenomena in a nonlinear spiked covariance model with those of a linearly defined (possibly Gaussian) model.

In Section 3.2, each CK matrix $\boldsymbol{K}_\ell$ has (approximately) the structure of the above matrix $\boldsymbol{K}$ over the randomness of $\boldsymbol{W}_\ell$, conditional on the features $\boldsymbol{X}_{\ell-1}$ of the preceding layer, and Theorem 34 follows from Theorem 40(a,b). Additionally, in Section 5.4.2, the CK matrix $\boldsymbol{K}$ defined by trained weights has (approximately) this structure over the randomness of $\tilde{\boldsymbol{X}}$, conditional on $\boldsymbol{W}_{\text{trained}}$, and Theorem 105 follows from Theorem 40(a,c).

Proof ideas of Theorem 40.

Analyses in the linearly defined model $\boldsymbol{g}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ commonly stem from block matrix inversion identities with respect to the block decompositions

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_0 \end{pmatrix}, \qquad \boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_r & \boldsymbol{G}_0 \end{pmatrix}$$

where $\mathbf{\Sigma}_r$ contains the spike eigenvalues of $\mathbf{\Sigma}$, and $\boldsymbol{G}_r$ is independent of $\boldsymbol{G}_0$. This independence does *not* hold in our setting, and we develop a different "master equation" approach.

Let $\widehat{\lambda}^{1/2}$ be a spike singular value of $\boldsymbol{G}$ with corresponding unit singular vectors $(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})$. We consider the linearized equation

$$0 = \begin{pmatrix} -\widehat{\lambda}\boldsymbol{I} & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{v}} \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}} \end{pmatrix}. \tag{3.4.1}$$

Writing $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ for the r spike eigenvectors of $\mathbf{\Sigma}$, we define a generalized resolvent

$$\mathscr{R}(z,\alpha) = \begin{pmatrix} -z\boldsymbol{I} - \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I} \end{pmatrix}^{-1},$$

add to (3.4.1) the quantity $-\alpha \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}$ on both sides for some large $\alpha > 0$, and rewrite this as

$$\begin{pmatrix} \widehat{\boldsymbol{v}} \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}} \end{pmatrix} = -\alpha \mathscr{R}(\widehat{\lambda}, \alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}. \tag{3.4.2}$$

We will show that $\mathscr{R}(z,\alpha)$ exists and is bounded in operator norm for any z separated from the limit bulk spectral support of $\boldsymbol{K}$ and any large enough $\alpha > 0$. Then, multiplying (3.4.2) by $(\boldsymbol{V}_r^\top \ \boldsymbol{0})$ and applying a block matrix inversion identity,

$$\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = -\alpha \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix}^\top \mathscr{R}(\widehat{\lambda}, \alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = -\alpha \boldsymbol{V}_r^\top \left(\boldsymbol{G}^\top \boldsymbol{G} - \widehat{\lambda}\boldsymbol{I} - \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top \right)^{-1} \boldsymbol{V}_r \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}.$$

As a result, spike eigenvalues $\widehat{\lambda}$ are roots $z = \widehat{\lambda}$ of the master equation

$$\det\left(\boldsymbol{I}_r + \alpha \boldsymbol{V}_r^\top \left(\boldsymbol{G}^\top \boldsymbol{G} - z\boldsymbol{I} - \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top \right)^{-1} \boldsymbol{V}_r \right) = 0,$$

for any fixed and large $\alpha > 0$. Singular vector alignments may be characterized likewise from (3.4.2).

The core of the proof is an asymptotic analysis of this master equation via a deterministic equivalent approximation

$$\boldsymbol{v}_1^\top \boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{v}_2 := \boldsymbol{v}_1^\top (\boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{\Gamma})^{-1}\boldsymbol{v}_2 \approx -\boldsymbol{v}_1^\top (\boldsymbol{\Gamma} + z\tilde{m}(z)\boldsymbol{\Sigma})^{-1}\boldsymbol{v}_2 \tag{3.4.3}$$

for any deterministic unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$ and low-rank perturbations $\boldsymbol{\Gamma}$ of $z\boldsymbol{I}$, where $\tilde{m}(z)$ is the Stieltjes transform of the "companion" limit measure $\tilde{\mu}$ for the eigenvalue distribution of $\boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{N\times N}$. We extend results of [Cho22, SCDL23] by establishing this approximation not only for $\boldsymbol{\Gamma} = z\boldsymbol{I}$ but also perturbations thereof, and for spectral arguments $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$ that may belong to the positive real line. The latter extension requires showing, a priori, that all eigenvalues of $\boldsymbol{K} = \boldsymbol{G}^\top \boldsymbol{G}$ fall close to $\operatorname{supp}(\mu)$ in the absence of spiked structure. We show this by adapting an argument of [BS98] and using a fluctuation averaging lemma described below.

Let us conclude with a brief discussion of our proof of (3.4.3): From manipulations of the identity

$$\operatorname{Tr}\boldsymbol{B} = \operatorname{Tr}(\boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{\Gamma})\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B} = -\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Gamma} + \frac{1}{N}\sum_{i=1}^N \boldsymbol{g}_i^\top \boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{g}_i$$

for appropriately chosen matrices $\boldsymbol{B} \in \mathbb{C}^{n\times n}$, the Sherman-Morrison (leave-one-out) formula for matrix inversion applied to $\boldsymbol{R}(\boldsymbol{\Gamma})$, and the concentration of bilinear forms in $\boldsymbol{g}_i$, one may show

$$\boldsymbol{v}_1^\top (\boldsymbol{\Gamma} + z\tilde{m}(z)\boldsymbol{\Sigma})^{-1}\boldsymbol{v}_2 \approx -\boldsymbol{v}_1^\top \boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{v}_2 + \frac{1}{1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})} \cdot \frac{1}{N}\sum_{i=1}^N (1 - \mathbb{E}_{\boldsymbol{g}_i})T_i \tag{3.4.4}$$

where $T_i = \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{v}_2 \cdot \boldsymbol{v}_1^\top (\boldsymbol{\Gamma} + z m_{\tilde{\boldsymbol{K}}}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i$. Here, $\boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})$ and $m_{\tilde{\boldsymbol{K}}}^{(i)}(\boldsymbol{\Gamma})$ are generalized leave-one-out resolvents and empirical Stieltjes transforms defined by $\{\boldsymbol{g}_j\}_{j\neq i}$, and $\mathbb{E}_{\boldsymbol{g}_i}$ is the partial expectation over only $\boldsymbol{g}_i$. Under our assumptions for $\boldsymbol{g}_i$, each error term $(1 - \mathbb{E}_{\boldsymbol{g}_i})T_i$ has

mean 0 and $O(1)$ fluctuations. In the next section, we develop a fluctuation-averaging lemma using recursive applications of the Sherman-Morrison-Woodbury identity to further resolve the dependence of $\boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})$ and $m^{(i)}_{\check{\boldsymbol{K}}}(\boldsymbol{\Gamma})$ on fixed subsets of rows $\{\boldsymbol{g}_j\}_{j\neq i}$, to show that the errors $(1-\mathbb{E}_{\boldsymbol{g}_i})T_i$ are weakly correlated across $i \in [N]$. Hence their average has a mean 0 and fluctuates on the asymptotically negligible scale of $O(N^{-1/2})$, and applying this to (3.4.4) shows (3.4.3).

3.5 Analysis of the Resolvent

We now prove the results of Section 3.3.1. Section 3.5.1 first develops a fluctuation averaging lemma for the sample covariance model. Section 3.5.2 applies this lemma within the arguments of [BS98], to prove the "no outliers" result of Theorem 37. Section 3.5.3 uses Theorem 37 and a second application of the fluctuation averaging lemma to prove the deterministic equivalent approximation of Theorem 38.

3.5.1 Fluctuation Averaging Lemma

Recall the definitions

$$\boldsymbol{K} = \boldsymbol{G}^\top \boldsymbol{G}, \qquad \widetilde{\boldsymbol{K}} = \boldsymbol{G}\boldsymbol{G}^\top.$$

For $S \subset [N]$, let $\boldsymbol{G}^{(S)} \in \mathbb{R}^{(N-|S|)\times n}$ be the matrix obtained by removing the rows of $\boldsymbol{G}$ corresponding to $i \in S$, and define

$$\boldsymbol{K}^{(S)} = {\boldsymbol{G}^{(S)}}^\top \boldsymbol{G}^{(S)} = \frac{1}{N}\sum_{i\in[N]\setminus S} \boldsymbol{g}_i\boldsymbol{g}_i^\top \in \mathbb{R}^{n\times n}.$$

Then, for $\boldsymbol{\Gamma} \in \mathbb{C}^{n\times n}$, define

$$\begin{aligned} \boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}) &= (\boldsymbol{K}^{(S)} - \boldsymbol{\Gamma})^{-1}, \qquad m^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma}) = \frac{1}{n}\operatorname{Tr}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}), \\ \tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma}) &= \gamma_N m^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma}) + (1-\gamma_N)\left(-\frac{1}{z}\right) = \frac{1}{N}\operatorname{Tr}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}) + \left(1-\frac{n}{N}\right)\left(-\frac{1}{z}\right). \end{aligned} \tag{3.5.1}$$

Importantly, these quantities are independent of $\{\boldsymbol{g}_i : i \in S\}$. We say that $\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})$ exists (and hence also $m_{\boldsymbol{K}}^{(S)}, \tilde{m}_{\boldsymbol{K}}^{(S)}$ exist) when $\boldsymbol{K}^{(S)} - \boldsymbol{\Gamma}$ is invertible. For simplicity, we write $\boldsymbol{R} = \boldsymbol{R}^{\emptyset}$, $\boldsymbol{R}^{(i)} = \boldsymbol{R}^{(\{i\})}$, $\boldsymbol{R}^{(Si)} = \boldsymbol{R}^{(S\cup\{i\})}$, and similarly for $m_{\boldsymbol{K}}$ and $\tilde{m}_{\boldsymbol{K}}$.

Lemma 41. *Suppose Assumption 6 holds. Suppose also that there are constants $C_0, c_0, \delta, \upsilon > 0$, N-dependent domains $U \subset \mathbb{C} \setminus \{0\}$ and $\mathscr{D}_\Gamma, \mathscr{D}_A \subseteq \mathbb{C}^{n\times n}$, and N-dependent maps $\Phi_N : \mathscr{D}_\Gamma \times \mathscr{D}_A \to (N^{-\upsilon}, N^{\upsilon})$ and $\Psi_N : \mathscr{D}_\Gamma \to (N^{-\upsilon}, N^{1-\delta})$, such that for any fixed $L \geq 1$, the events*

$$\begin{aligned}\mathscr{E}(z,\boldsymbol{\Gamma},\boldsymbol{A},S) = \Big\{ &\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}) \text{ exists, } \|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{A}\|_F \leq \Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}),\ \|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\|_F \leq \Psi_N(\boldsymbol{\Gamma}), \\ &\|(z^{-1}\boldsymbol{\Gamma} + \tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\| \leq C_0, \text{ and } |1 + N^{-1}\boldsymbol{g}_j^\top \boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{g}_j| \geq c_0 \text{ for all } j \in S\Big\}\end{aligned} \tag{3.5.2}$$

satisfy $\mathbf{1}\{\mathscr{E}(z,\boldsymbol{\Gamma},\boldsymbol{A},S)^c\} \prec 0$ uniformly over $z \in U$, $\boldsymbol{\Gamma} \in \mathscr{D}_\Gamma$, $\boldsymbol{A} \in \mathscr{D}_A$, and $S \subset [N]$ with $|S| \leq L$.

Then, denoting by $\mathbb{E}_{\boldsymbol{g}_i}$ the partial expectation over only $\boldsymbol{g}_i$ (i.e. conditional on $\{\boldsymbol{g}_j\}_{j\neq i}$), also uniformly over $z \in U$, $\boldsymbol{\Gamma} \in \mathscr{D}_\Gamma$, and $\boldsymbol{A} \in \mathscr{D}_A$,

$$\frac{1}{N}\sum_{i=1}^N (1 - \mathbb{E}_{\boldsymbol{g}_i})\big[\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + \tilde{m}_{\boldsymbol{K}}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i\big] \prec \max\left(\frac{\Psi_N(\boldsymbol{\Gamma})}{N}, \frac{1}{\sqrt{N}}\right)\cdot \Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}). \tag{3.5.3}$$

We remark that applying Assumption 6(c) and the conditions of $\mathscr{E}(z,\boldsymbol{\Gamma},\boldsymbol{A},i)$ separately to each summand of the left side of (3.5.3) gives the naive bound

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N (1 - \mathbb{E}_{\boldsymbol{g}_i})[\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + \tilde{m}_{\boldsymbol{K}}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i] \\ \prec &\max_{1\leq i\leq N} \|\boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}\|_F \cdot \|(z^{-1}\boldsymbol{\Gamma} + \tilde{m}_{\boldsymbol{K}}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\| \prec \Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}).\end{aligned}$$

The content of the lemma is to improve this by the additional factor of $\max(\frac{\Psi_N(\boldsymbol{\Gamma})}{N}, \frac{1}{\sqrt{N}}) \ll 1$.

In this work, we will apply Lemma 41 only to spectral arguments z with $O(1)$-separation from $\operatorname{supp}(\mu_N)$ (and matrices $\boldsymbol{\Gamma} = z\boldsymbol{I}$ or a finite-rank perturbation thereof), in which case we will take $\Psi_N(\boldsymbol{\Gamma}) = C/\sqrt{N}$ for a constant $C > 0$. For full-rank matrices $\boldsymbol{A}$ having bounded operator

norm, we will also take $\Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}) = C/\sqrt{N}$, whereas for finite-rank matrices $\boldsymbol{A}$ we will take $\Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}) = C$. We state the result here more abstractly, as it may be of independent interest to prove local laws in this nonlinear sample covariance model for spectral arguments z that approach $\mathrm{supp}(\mu_N)$.

In the remainder of this section, we prove Lemma 41. Fix $z \in U$, $\boldsymbol{\Gamma} \in \mathscr{D}_\Gamma$, and $\boldsymbol{A} \in \mathscr{D}_A$, and write as shorthand

$$\boldsymbol{R}^{(S)} = \boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}), \qquad \tilde{m}^{(S)} = \tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma}), \qquad \boldsymbol{\Omega}^{(S)} = (z^{-1}\boldsymbol{\Gamma} + \tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1},$$

$$\Phi_N = \Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}), \qquad \Psi_N = \Psi_N(\boldsymbol{\Gamma}), \qquad \mathscr{E}(S) = \mathscr{E}(z,\boldsymbol{\Gamma},\boldsymbol{A},S).$$

All subsequent instances of $\prec$ will be implicitly uniform over $z \in U$, $\boldsymbol{\Gamma} \in \mathscr{D}_\Gamma$, and $\boldsymbol{A} \in \mathscr{D}_A$. Define the quantities, for $i \in S$, $j,k \in S \setminus \{i\}$, and $d \geq 0$,

$$\begin{aligned}
Y_i^{(S)}[d] &= \mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{\Omega}^{(S)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(S)}]^d, \\
Z_{ijk}^{(S)}[d] &= N^{-1}\,\mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{g}_j\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{\Omega}^{(S)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(S)}]^d, \\
B_{jk}^{(S)} &= N^{-1}\boldsymbol{g}_j^\top\boldsymbol{R}^{(S)}\boldsymbol{g}_k, \\
C_{jk}^{(S)} &= N^{-2}\boldsymbol{g}_j^\top(\boldsymbol{R}^{(S)})^2\boldsymbol{g}_k, \\
Q_j^{(S)} &= (1 + N^{-1}\boldsymbol{g}_j^\top\boldsymbol{R}^{(S)}\boldsymbol{g}_j)^{-1}.
\end{aligned}$$

For each $L \geq 1$, define also the event

$$\mathscr{E}_L = \bigcap_{S \subset [N] : |S| \leq L} \mathscr{E}(S). \tag{3.5.4}$$

Lemma 42. *For any fixed $L, D \geq 1$, uniformly over $S \subset [N]$ with $|S| \leq L$, and over $i \in S$ and*

$j,k \in S \setminus \{i\}$ *and* $d \le D$,

$$\begin{aligned} Y_i^{(S)}[d] &= O_{\prec}^{\mathscr{E}(S)}(\Phi_N), \quad Z_{ijk}^{(S)}[d] = O_{\prec}^{\mathscr{E}(S)}\big(N^{-1}\Psi_N\Phi_N\big), \\ B_{jk}^{(S)} &= O_{\prec}^{\mathscr{E}(S)}\big(N^{-1}\Psi_N\big) \text{ for } j \neq k, \quad C_{jk}^{(S)} = O_{\prec}^{\mathscr{E}(S)}\big(N^{-2}\Psi_N^2\big), \quad Q_j^{(S)} = O_{\prec}^{\mathscr{E}(S)}(1). \end{aligned} \tag{3.5.5}$$

Furthermore, for any $\alpha > 0$*, there exists a constant* $C = C(\alpha, L, D) > 0$ *such that*

$$\begin{aligned} &\mathbb{E}\big[|Y_i^{(S)}[d]|^\alpha \mathbf{1}\{\mathscr{E}(S)\}\big] < N^C, \quad \mathbb{E}\big[|Z_{ijk}^{(S)}[d]|^\alpha \mathbf{1}\{\mathscr{E}(S)\}\big] < N^C, \\ &\mathbb{E}\big[|B_{jk}^{(S)}|^\alpha \mathbf{1}\{\mathscr{E}(S)\}\big] < N^C, \quad \mathbb{E}\big[|C_{jk}^{(S)}|^\alpha \mathbf{1}\{\mathscr{E}(S)\}\big] < N^C, \quad \mathbb{E}\big[|Q_j^{(S)}|^\alpha \mathbf{1}\{\mathscr{E}(S)\}\big] < N^C. \end{aligned} \tag{3.5.6}$$

Proof. On the event $\mathscr{E}(S)$, we have by definition $Q_j^{(S)} \le 1/c_0$, so the two statements for $Q_j^{(S)}$ hold immediately. The remaining statements of (3.5.6) follow easily from Holder's inequality, the moment bounds for $\|\boldsymbol{g}_i\|$ in Assumption 6(d), the bound $\|\boldsymbol{\Sigma}\| < C$ in Assumption 6(a), and the conditions $\|\boldsymbol{R}^{(S)}\boldsymbol{A}\| \le \Phi_N \le N^\upsilon$, $\|\boldsymbol{R}^{(S)}\|_F \le \Psi_N \le N$, and $\|\boldsymbol{\Omega}^{(S)}\| \le C_0$ defining $\mathscr{E}(S)$.

For the bounds for $B_{jk}^{(S)}$ and $C_{jk}^{(S)}$ in (3.5.5), note that when $j \neq k$, Assumption 6(c) implies $B_{jk}^{(S)} \prec N^{-1}\|\boldsymbol{R}^{(S)}\|_F$ and $C_{jk}^{(S)} \prec N^{-2}\|(\boldsymbol{R}^{(S)})^2\|_F \le N^{-2}\|\boldsymbol{R}^{(S)}\|_F^2$. When $j = k$, Assumption 6(c) implies also

$$\begin{aligned} C_{jj}^{(S)} &\prec N^{-2}|\operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{R}^{(S)})^2| + N^{-2}\|(\boldsymbol{R}^{(S)})^2\|_F \\ &\le N^{-2}\|\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}\|_F\|\boldsymbol{R}^{(S)}\|_F + N^{-2}\|\boldsymbol{R}^{(S)}\|_F^2 \le N^{-2}(\|\boldsymbol{\Sigma}\|+1)\|\boldsymbol{R}^{(S)}\|_F^2. \end{aligned}$$

Then these bounds in (3.5.5) follow from the condition $\|\boldsymbol{R}^{(S)}\|_F \le \Psi_N$ defining $\mathscr{E}(S)$.

Finally, for the bounds for $Y_i^{(S)}[d]$ and $Z_{ijk}^{(S)}[d]$ in (3.5.5), observe that for any matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ independent of $\boldsymbol{g}_i$, we have $\operatorname{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{A} \prec \|\boldsymbol{A}\|_F$ by Assumption 6(c). Then $Y_i^{(S)}[d] \prec \|\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{\Omega}^{(S)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(S)}]^d\|_F \le \|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F \cdot \|\boldsymbol{\Omega}^{(S)}\|^{d+1}\|\boldsymbol{\Sigma}\|^d$, so the bound for $Y_i^{(S)}[d]$ in (3.5.5) follows from the conditions $\|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F \le \Phi_N$ and $\|\boldsymbol{\Omega}^{(S)}\| \le C_0$ defining $\mathscr{E}(S)$. For $Z_{ijk}^{(S)}[d]$, similarly by

Assumption 6(c),

$$Z_{ijk}^{(S)} \prec N^{-1}\|\boldsymbol{R}^{(S)}\boldsymbol{g}_j\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{\Omega}^{(S)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(S)}]^d\|_F \le N^{-1}\|\boldsymbol{R}^{(S)}\boldsymbol{g}_j\|\cdot\|\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\|\cdot\|\boldsymbol{\Omega}^{(S)}\|^{d+1}\|\boldsymbol{\Sigma}\|^d.$$

Applying again Assumption 6(c), we have

$$\|\boldsymbol{R}^{(S)}\boldsymbol{g}_j\|^2 = \boldsymbol{g}_j^\top(\boldsymbol{R}^{(S)})^*\boldsymbol{R}^{(S)}\boldsymbol{g}_j \prec |\operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{R}^{(S)})^*\boldsymbol{R}^{(S)}| + \|(\boldsymbol{R}^{(S)})^*\boldsymbol{R}^{(S)}\|_F \prec \|\boldsymbol{R}^{(S)}\|_F^2$$

and similarly $\|\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\|^2 \prec \|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F^2$. Then the bound for $Z_{ijk}^{(S)}[d]$ in (3.5.5) follows from the conditions $\|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F \le \Phi_N$, $\|\boldsymbol{R}^{(S)}\|_F \le \Psi_N$, and $\|\boldsymbol{\Omega}^{(S)}\| \le C_0$ defining $\mathscr{E}(S)$. □

Lemma 43. *Fix any $L, D \ge 1$. Then there exist coefficients $\alpha(d,d',D) \in \mathbb{R}$ such that the following holds: Uniformly over $S \subset [N]$ with $|S| \le L-1$, and over $i \in S$, $j,k \in S\setminus\{i\}$, $l \in [N]\setminus S$, and $d \le D$, we have*

$$Y_i^{(S)}[d] = \sum_{d'=d}^{d+\lceil D/2\rceil}\alpha(d,d',D)\Big[C_{ll}^{(Sl)}Q_l^{(Sl)}\Big]^{d'-d}\Big(Y_i^{(Sl)}[d'] - Z_{ill}^{(Sl)}[d']Q_l^{(Sl)}\Big) + O_\prec^{\mathscr{E}_L}\big(N^{-D}\Psi_N^D\Phi_N\big) \tag{3.5.7}$$

$$\begin{aligned} Z_{ijk}^{(S)}[d] = \sum_{d'=d}^{d+\lceil D/2\rceil}\alpha(d,d',D)\Big[C_{ll}^{(Sl)}Q_l^{(Sl)}\Big]^{d'-d}\Big(&Z_{ijk}^{(Sl)}[d'] - Z_{ilk}^{(Sl)}[d']B_{lj}^{(Sl)}Q_l^{(Sl)} \\ &- Z_{ijl}^{(Sl)}[d']B_{kl}^{(Sl)}Q_l^{(Sl)} + Z_{ill}^{(Sl)}[d']B_{lj}^{(Sl)}B_{kl}^{(Sl)}(Q_l^{(Sl)})^2\Big) + O_\prec^{\mathscr{E}_L}\big(N^{-D}\Psi_N^D\Phi_N\big), \end{aligned} \tag{3.5.8}$$

$$B_{jk}^{(S)} = B_{jk}^{(Sl)} - B_{jl}^{(Sl)}B_{lk}^{(Sl)}Q_l^{(Sl)}, \tag{3.5.9}$$

$$C_{jk}^{(S)} = C_{jk}^{(Sl)} - B_{jl}^{(Sl)}C_{lk}^{(Sl)}Q_l^{(Sl)} - C_{jl}^{(Sl)}B_{lk}^{(Sl)}Q_l^{(Sl)} + B_{jl}^{(Sl)}C_{ll}^{(Sl)}B_{lk}^{(Sl)}(Q_l^{(Sl)})^2, \tag{3.5.10}$$

$$Q_j^{(S)} = \sum_{d=1}^{\lceil D/2\rceil}\Big(Q_j^{(Sl)}\Big)^d\Big[(B_{jl}^{(Sl)})^2Q_l^{(Sl)}\Big]^{d-1} + O_\prec^{\mathscr{E}_L}\big(N^{-D}\Psi_N^D\big). \tag{3.5.11}$$

Proof. By the Sherman-Morrison-Woodbury formula, on the event $\mathscr{E}_L$ where $\boldsymbol{R}^{(S)}$ and $\boldsymbol{R}^{(Sl)}$ both

exist, we have

$$\boldsymbol{R}^{(S)} = \boldsymbol{R}^{(Sl)} - N^{-1}\boldsymbol{R}^{(Sl)}\boldsymbol{g}_l\boldsymbol{g}_l^\top\boldsymbol{R}^{(Sl)} \cdot Q_l^{(Sl)}. \tag{3.5.12}$$

Applying this to each copy of $\boldsymbol{R}^{(S)}$ defining $B_{jk}^{(S)}$ and $C_{jk}^{(S)}$ yields immediately (3.5.9) and (3.5.10), as well as the identities

$$\begin{aligned} z^{-1}\boldsymbol{\Gamma} + \tilde{m}^{(S)}\boldsymbol{\Sigma} &= z^{-1}\boldsymbol{\Gamma} + \Big(N^{-1}\operatorname{Tr}\boldsymbol{R}^{(S)} + (1-\gamma_N)(-1/z)\Big)\boldsymbol{\Sigma} \\ &= (z^{-1}\boldsymbol{\Gamma} + \tilde{m}^{(Sl)}\boldsymbol{\Sigma}) - C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}, \\ 1 + B_{jj}^{(S)} &= 1 + B_{jj}^{(Sl)} - (B_{jl}^{(Sl)})^2 Q_l^{(Sl)}. \end{aligned}$$

Taking inverses and applying the expansion

$$(\boldsymbol{A}-\boldsymbol{\Delta})^{-1} = \sum_{d=1}^{\lceil D/2\rceil} \boldsymbol{A}^{-1}(\boldsymbol{\Delta}\boldsymbol{A}^{-1})^{d-1} + (\boldsymbol{A}-\boldsymbol{\Delta})^{-1}(\boldsymbol{\Delta}\boldsymbol{A}^{-1})^{\lceil D/2\rceil},$$

we obtain

$$\boldsymbol{\Omega}^{(S)} = \sum_{d=1}^{\lceil D/2\rceil} \boldsymbol{\Omega}^{(Sl)}[C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d-1} + \boldsymbol{E}, \tag{3.5.13}$$

$$Q_j^{(S)} = \sum_{d=1}^{\lceil D/2\rceil} Q_j^{(Sl)}[(B_{jl}^{(Sl)})^2 Q_l^{(Sl)}Q_j^{(Sl)}]^{d-1} + e, \tag{3.5.14}$$

for remainder terms $\boldsymbol{E} \in \mathbb{C}^{n\times n}$ and $e \in \mathbb{C}$ satisfying, by the bounds of Lemma 42,

$$\|\boldsymbol{E}\| = O_{\prec}^{\mathscr{E}_L}\Big(|C_{ll}^{(Sl)}|^{D/2}\Big) = O_{\prec}^{\mathscr{E}_L}\big((N^{-1}\Psi)^D\big), \quad |e| = O_{\prec}^{\mathscr{E}_L}\Big(|(B_{jl}^{(Sl)})^2|^{D/2}\Big) = O_{\prec}^{\mathscr{E}_L}\big((N^{-1}\Psi)^D\big).$$

In particular, (3.5.14) shows (3.5.11). Applying (3.5.13) to the definitions of $Y_i^{(S)}[d]$ and $Z_{ijk}^{(S)}[d]$,

we get

$$Y_i^{(S)}[d] = \mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{A}\left(\sum_{d'=1}^{\lceil D/2\rceil}\boldsymbol{\Omega}^{(Sl)}[C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d'-1} + \boldsymbol{E}\right)$$
$$\cdot\left(\boldsymbol{\Sigma}\left[\sum_{d'=1}^{\lceil D/2\rceil}\boldsymbol{\Omega}^{(Sl)}[C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d'-1} + \boldsymbol{E}\right]\right)^d,$$
$$Z_{ijk}^{(S)}[d] = \frac{1}{N}\mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{g}_j\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\left(\sum_{d'=1}^{\lceil D/2\rceil}\boldsymbol{\Omega}^{(Sl)}[C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d'-1} + \boldsymbol{E}\right)$$
$$\cdot\left(\boldsymbol{\Sigma}\left[\sum_{d'=1}^{\lceil D/2\rceil}\boldsymbol{\Omega}^{(Sl)}[C_{ll}^{(Sl)}Q_l^{(Sl)}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d'-1} + \boldsymbol{E}\right]\right)^d.$$

For any matrix $\boldsymbol{B}\in\mathbb{C}^{n\times n}$ independent of $\boldsymbol{g}_i$, observe that $\mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{B} = O_\prec^{\mathscr{E}(S)}(\Phi_N\|\boldsymbol{B}\|)$ and $\mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{g}_j\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\boldsymbol{B} = O_\prec^{\mathscr{E}(S)}(\Psi_N\Phi_N\|\boldsymbol{B}\|)$ by the same arguments as those bounding $Y_i^{(S)}[d]$ and $Z_{ijk}^{(S)}[d]$ in the proof of Lemma 42. Then, expanding the above and absorbing all terms containing $\boldsymbol{E}$ and all terms with combined power of $C_{ll}^{(Sl)}$ larger than $D/2$ into $O_\prec^{\mathscr{E}(S)}\big(N^{-D}\Psi_N^D\Phi_N\big)$ remainders, we obtain for some coefficients $\alpha(d,d',D)\in\mathbb{R}$ that

$$Y_i^{(S)}[d] = \mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{A}\sum_{d'=0}^{\lceil D/2\rceil}\alpha(d,d',D)[C_{ll}^{(Sl)}Q_l^{(Sl)}]^{d'}\boldsymbol{\Omega}^{(Sl)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d+d'}$$
$$+ O_\prec^{\mathscr{E}(S)}\big(N^{-D}\Psi_N^D\Phi_N\big),$$
$$Z_{ijk}^{(S)}[d] = \frac{1}{N}\mathrm{Tr}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{\Sigma})\boldsymbol{R}^{(S)}\boldsymbol{g}_j\boldsymbol{g}_k^\top\boldsymbol{R}^{(S)}\boldsymbol{A}\sum_{d'=0}^{\lceil D/2\rceil}\alpha(d,d',D)[C_{ll}^{(Sl)}Q_l^{(Sl)}]^{d'}\boldsymbol{\Omega}^{(Sl)}[\boldsymbol{\Sigma}\boldsymbol{\Omega}^{(Sl)}]^{d+d'}$$
$$+ O_\prec^{\mathscr{E}(S)}\big(N^{-D}\Psi_N^D\Phi_N\big).$$

Finally, applying the Sherman-Morrison-Woodbury formula (3.5.12) to expand each copy of $\boldsymbol{R}^{(S)}$, and re-indexing the summations by $d+d'\mapsto d'$, we get (3.5.7) and (3.5.8). □

Lemma 44. *Fix any $L,D\geq 1$. Uniformly over $S\subset[N]$ with $|S|\leq L$ and over $i\in S$, the following holds: Denote $\bar{S}=S\setminus\{i\}$. Then there exists a collection of monomials $\mathscr{M}_{i,S}$ such that $Y_i^{(i)}[0]$ can*

be approximated as

$$Y_i^{(i)}[0] = \sum_{q \in \mathscr{M}_{i,S}} q\Bigg(\{Y_i^{(S)}[d]\}_{d \le \lfloor D/2 \rfloor}, \{Z_{ijk}^{(S)}[d]\}_{j,k \in \bar{S}, d \le \lfloor D/2 \rfloor}, \{B_{jk}^{(S)}\}_{j \ne k \in \bar{S}}, \{C_{jk}^{(S)}\}_{j,k \in \bar{S}}, \{Q_j^{(S)}\}_{j \in \bar{S}} \Bigg) + O_{\prec}^{\mathscr{E}_L}\big(N^{-D}\Psi_N^D \Phi_N\big). \quad (3.5.15)$$

Each monomial $q \in \mathscr{M}_{i,S}$ is a product of a real-valued scalar coefficient and one or more factors of the form $Y_i^{(S)}[d]$, $Z_{ijk}^{(S)}[d]$, $B_{jk}^{(S)}$ with $j \ne k$, $C_{jk}^{(S)}$, $Q_j^{(S)}$ for $j,k \in \bar{S}$ and $d \le \lfloor D/2 \rfloor$. We have $q = O_{\prec}^{\mathscr{E}_L}(\Phi_N)$ uniformly over $q \in \mathscr{M}_{i,S}$, and the number of monomials $|\mathscr{M}_{i,S}|$ is most a constant depending on L, D. Furthermore:

(a) There is exactly one factor of the form $Y_i^{(S)}[d]$ or $Z_{ijk}^{(S)}[d]$ appearing in q.

(b) The number of factors $Z_{ijk}^{(S)}[d]$, $B_{jk}^{(S)}$, and $C_{jk}^{(S)}$ appearing in q is no less than the number of distinct indices of $\bar{S}$ (not including i) that appear as lower indices across all factors of q.

Proof. We arbitrarily order the indices of $\bar{S} = S \setminus \{i\}$ as $l_1, l_2, \ldots, l_{|S|-1}$. Beginning with the monomial $Y_i^{(i)}[0]$, iteratively for $j = 1, 2, \ldots, |S| - 1$, we replace all factors with superscript $(il_1 \ldots l_{j-1})$ by a sum of terms with superscript $(il_1 \ldots l_j)$, using the recursions (3.5.7)–(3.5.11). It is then direct to check that this gives a representation of the form (3.5.15), where:

- Each application of (3.5.7)–(3.5.8) replaces a factor $Y_i^{(\ldots)}[d]$ or $Z_{ijk}^{(\ldots)}[d]$ by terms having exactly one such factor. Thus, each monomial $q \in \mathscr{M}_{i,S}$ has exactly one factor $Y_i^{(S)}[d]$ or $Z_{ijk}^{(S)}[d]$.
- The number of total applications of (3.5.7)–(3.5.11) is bounded by a constant depending on L, D, so $|\mathscr{M}_{i,S}|$ and the scalar coefficient of each $q \in \mathscr{M}_{i,S}$ are both bounded by constants depending on L and D. Then, by the bounds of (3.5.5), each $q \in \mathscr{M}_{i,S}$ satisfies $q = O_{\prec}^{\mathscr{E}_L}(\Phi_N)$, and the remainder in (3.5.15) is at most $O_{\prec}^{\mathscr{E}_L}\big(N^{-D}\Psi_N^D \Phi_N\big)$. If q has the term $Y_i^{(S)}[d]$ or $Z_i^{(S)}[d]$, then it also has combined power of $\{C_{jk}^{(S)}\}_{j,k \in \bar{S}}$ equal to d, and hence may be absorbed into the remainder of (3.5.15) if $d > D/2$.

- Each term on the right side of (3.5.7)–(3.5.11) that contains the new lower index l has at least one more factor of the form $Z_{ijk}^{(\dots)}[d]$, $B_{jk}^{(\dots)}$, or $C_{jk}^{(\dots)}$ than the left side. Thus, each monomial $q \in \mathscr{M}_{i,S}$ is such that the number of distinct lower indices of $\bar{S}$ across all of its factors is no greater than the number of its factors of the form $Z_{ijk}^{(\dots)}[d]$, $B_{jk}^{(\dots)}$, or $C_{jk}^{(\dots)}$.

Combining these observations yields the lemma. □

Proof of Lemma 41. For each $\varepsilon, D > 0$, let us fix an even integer $L = L(\varepsilon, D) > D/\varepsilon$. The assumption of this lemma guarantees $\mathbf{1}\{\mathscr{E}(S)^c\} \prec 0$ uniformly over $S \subset [N]$ with $|S| \le L$. Since the number of such subsets is at most N^L, we may take a union bound (c.f. Proposition 36(a)) to obtain $\mathbf{1}\{\mathscr{E}_L^c\} \prec 0$ for the intersection event $\mathscr{E}_L$ of (3.5.4). Noting that $(1 - \mathbb{E}_{\boldsymbol{g}_i})[\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{g}_i] = Y_i^{(i)}[0]$, to prove the lemma, it suffices to show for any $\varepsilon, D > 0$ and all sufficiently large N that

$$\mathbb{P}\left[\left(\frac{1}{N}\sum_{i=1}^N Y_i^{(i)}[0]\right)\mathbf{1}\{\mathscr{E}_L\} > \max\left(\frac{\Psi_N}{N}, \frac{1}{\sqrt{N}}\right)\Phi_N \cdot N^{\varepsilon}\right] < N^{-D}. \tag{3.5.16}$$

In anticipation of applying Markov's inequality, we analyze

$$\mathbb{E}\left[\left(\sum_{i=1}^N Y_i^{(i)}[0]\right)^L \mathbf{1}\{\mathscr{E}_L\}\right] = \sum_{i_1,\dots,i_L=1}^N \underbrace{\mathbb{E}\left[\prod_{l=1}^L Y_{i_l}^{(i_l)}[0]\mathbf{1}\{\mathscr{E}_L\}\right]}_{:=\mathbb{E}[m(i_1,\dots,i_L)]}. \tag{3.5.17}$$

Fix any index tuple $(i_1, \dots, i_L)$. Letting $S = \{i_1, \dots, i_L\}$ be the set of distinct indices in this tuple, we apply Lemma 44 to each term $Y_{i_l}^{(i_l)}[0]$, with this set S and with $D = L$. This gives

$$m(i_1,\dots,i_L) = \sum_{q^{(1)} \in \mathscr{M}(i_1,S)} \cdots \sum_{q^{(l)} \in \mathscr{M}(i_l,S)} \prod_{l=1}^L q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\} + O_{\prec}\big((N^{-1}\Psi_N)^L \Phi_N^L\big), \tag{3.5.18}$$

where each $\mathscr{M}(i_l, S)$ is the collection of monomials arising in the approximation of $Y_{i_l}^{(i_l)}[0]$, and we have applied $q^{(l)} = O_{\prec}^{\mathscr{E}_L}(\Phi_N)$ to bound the remainder. Observe that by (3.5.6) and Holder's inequality, we have $\mathbb{E}[|m(i_1,\dots,i_L)|^2] \le N^C$ and $\mathbb{E}[|\prod_{l=1}^L q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}|^2] \le N^C$ for all $q^{(1)}, \dots, q^{(L)}$

and a constant $C > 0$. By this and the given condition $\Psi_N, \Phi_N \geq N^{-\upsilon}$, we may take expectations in (3.5.18) using Proposition 36(d) to get

$$\mathbb{E}[m(i_1,\ldots,i_L)] = \sum_{q^{(1)} \in \mathscr{M}(i_1,S)} \cdots \sum_{q^{(l)} \in \mathscr{M}(i_l,S)} \mathbb{E}\bigg[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}\bigg] + O_{\prec}\big((N^{-1}\Psi_N)^L \Phi_N^L\big). \quad (3.5.19)$$

Now to bound $\mathbb{E}[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}]$, we consider separately two cases, focusing on those indices i_l which appear exactly once in $(i_1,\ldots,i_L)$. In the first case, suppose there is some such index i_l that does not appear as a lower index of $q^{(l')}$ for any $l' \neq l$. Fixing this set $S = \{i_1,\ldots,i_L\}$ and index $i_l \in S$, let us introduce

$$\mathscr{E}' = \bigg\{ \boldsymbol{R}^{(S)} \text{ exists, } \|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F \leq \Phi_N,\ \|\boldsymbol{R}^{(S)}\|_F \leq \Psi_N,\ \|(z^{-1}\boldsymbol{\Gamma} + \tilde{m}^{(S)}\boldsymbol{\Sigma})^{-1}\| \leq C_0, \\ \text{and } |1 + N^{-1}\boldsymbol{g}_j^\top \boldsymbol{R}^{(S)} \boldsymbol{g}_j| \geq c_0 \text{ for all } j \in S \setminus \{i_l\} \bigg\}.$$

Comparing with the definition of $\mathscr{E}(S)$ from (3.5.2), observe that only the last condition defining $\mathscr{E}'$ is different (where we do not require the bound for $j = i_l$), so that this event $\mathscr{E}'$ is independent of $\boldsymbol{g}_{i_l}$. Then $\mathscr{E}_L \subseteq \mathscr{E}(S) \subseteq \mathscr{E}'$, and

$$\mathbb{E}\bigg[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}\bigg] = \mathbb{E}\bigg[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}'\}\bigg] - \mathbb{E}\bigg[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}'\}\mathbf{1}\{\mathscr{E}_L^c\}\bigg]. \quad (3.5.20)$$

For the first term of (3.5.20), observe that both $\{q^{(l')} : l' \neq l\}$ and $\mathscr{E}'$ are independent of $\boldsymbol{g}_{i_l}$, and only the one factor $Y_{i_l}^{(S)}[d]$ or $Z_{i_l jk}^{(S)}[d]$ in $q^{(l)}$ depends on $\boldsymbol{g}_{i_l}$. Then, noting that $\mathbb{E}_{\boldsymbol{g}_i}[Y_i^{(S)}[d]] = 0$ and $\mathbb{E}_{\boldsymbol{g}_i}[Z_{ijk}^{(S)}[d]] = 0$, the first term of (3.5.20) is 0. For the second term of (3.5.20), observe that all statements of (3.5.6) continue to hold with $\mathscr{E}(S)$ replaced by $\mathscr{E}'$, except for the bound on $Q_{i_l}^{(S)}$. But $Q_{i_l}^{(S)}$ appears neither in $\{q^{(l')} : l' \neq l\}$ nor in $q^{(l)}$, so we may apply Holder's inequality to get $\mathbb{E}[|\prod_{l=1}^{L} q^{(l)}|^2 \mathbf{1}\{\mathscr{E}'\}] \leq N^C$ for a constant $C > 0$. Then, applying Cauchy-Schwarz and $\mathbf{1}\{\mathscr{E}_L^c\} \prec 0$, the second term of (3.5.20) is bounded by $N^{-D'}$ for any fixed constant $D' > 0$ and

all large N. Thus,

$$\mathbb{E}\left[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}\right] \leq N^{-D'}. \tag{3.5.21}$$

In the second case, every index i_l that appears exactly once in $(i_1,\dots,i_L)$ appears as a lower index of $q^{(l')}$ for some $l' \neq l$. Call the number of such indices K. Then condition (b) of Lemma 44 implies that the total number of factors of the forms $Z^{(S)}_{ijk}[d]$, $B^{(S)}_{jk}$ for $j \neq k$, and $C^{(S)}_{jk}$ across all monomials $q^{(1)},\dots,q^{(L)}$ is at least K. Then, by the bounds of Lemma 42 and Proposition 36(d), we have

$$\mathbb{E}\left[\prod_{l=1}^{L} q^{(l)} \cdot \mathbf{1}\{\mathscr{E}_L\}\right] \prec (N^{-1}\Psi_N)^K \Phi_N^L. \tag{3.5.22}$$

Under the given condition $\Phi_N, \Psi_N \geq N^{-\upsilon}$, we have $N^{-D'} \leq (N^{-1}\Psi_N)^K \Phi_N^L$ for large enough D'. Then, combining the two cases (3.5.21) and (3.5.22) and applying this back to (3.5.19), we get

$$\mathbb{E}[m(i_1,\dots,i_L)] \prec (N^{-1}\Psi_N)^K \Phi_N^L \tag{3.5.23}$$

where K is the number of indices in $S = \{i_1,\dots,i_L\}$ that appear exactly once in $(i_1,\dots,i_L)$. Let J be the number of distinct indices in $S = \{i_1,\dots,i_L\}$ that appear at least twice in $(i_1,\dots,i_L)$. Then $2J + K \leq L$, and the number of index tuples $(i_1,\dots,i_L) \in [N]^L$ with these values of (J,K) is at most CN^{J+K}, for a constant $C = C(J,K) > 0$. Then, applying (3.5.23) back to (3.5.17) yields

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{N} Y_i^{(i)}[0]\right)^L \mathbf{1}\{\mathscr{E}_L\}\right] &\prec \max_{J,K\geq 0:\, 2J+K\leq L} N^{J+K} \cdot (N^{-1}\Psi_N)^K \Phi_N^L \\ &= \max_{J,K\geq 0:\, 2J+K\leq L} (\sqrt{N})^{2J} \Psi_N^K \Phi_N^L \leq \max(\Psi_N, \sqrt{N})^L \Phi_N^L.\end{aligned}$$

Finally, by Markov's inequality, the probability in (3.5.16) is at most

$$\max(\Psi_N, \sqrt{N})^{-L} \Phi_N^{-L} N^{-\varepsilon L} \cdot \mathbb{E}\left[\left(\sum_{i=1}^{N} Y_i^{(i)}[0]\right)^L \mathbf{1}\{\mathscr{E}_L\}\right] \prec N^{-\varepsilon L},$$

and (3.5.16) follows as desired under our initial choice $L = L(\varepsilon, D) > D/\varepsilon$. □

3.5.2 No Eigenvalues Outside the Support

We now prove Theorem 37. Let $m_N(z), \tilde{m}_N(z)$ be the Stieltjes transform of the N-dependent deterministic measures $\mu_N, \tilde{\mu}_N$. For each $z \in \mathbb{C}^+$, $\tilde{m}_N(z)$ is the unique root in $\mathbb{C}^+$ to the equation

$$z = -\frac{1}{\tilde{m}_N(z)} + \gamma_N \int \frac{\lambda}{1+\lambda \tilde{m}_N(z)} d\nu_N(\lambda), \tag{3.5.24}$$

and $m_N(z), \tilde{m}_N(z)$ are related by $\tilde{m}_N(z) = \gamma_N m_N(z) + (1-\gamma_N)(-1/z)$. Define the discrete set

$$\mathscr{T}_N = \{0\} \cup \{-1/\lambda : \lambda \in \mathrm{supp}(\nu_N)\}. \tag{3.5.25}$$

On the domain $\mathbb{C} \setminus \mathscr{T}_N$, we may define the formal inverse of (3.5.24),

$$z_N(\tilde{m}) = -\frac{1}{\tilde{m}} + \gamma_N \int \frac{\lambda}{1+\lambda \tilde{m}} d\nu_N(\lambda), \tag{3.5.26}$$

which is a finite-N analogue of (1.2.6). Let $\mathscr{S}_N$ be the deterministic support defined in (3.3.2), and let $U_N(\varepsilon)$ be the spectral domain (3.3.3). The following basic properties of $\mathscr{S}_N$ and $\tilde{m}_N(z)$ are known.

Proposition 45. Suppose Assumption 6(a) holds, and fix any $\varepsilon > 0$. Then there exist constants $C_0, c_0 > 0$, depending only on ε and the constants C, c of Assumption 6(a), such that for all $x \in \mathscr{S}_N$ we have $|x| \le C$, and for all $z = x + i\eta \in U_N(\varepsilon)$ we have

$$c < |\tilde{m}_N(z)| < C, \qquad c\eta \le |\operatorname{Im} \tilde{m}_N(z)| \le C\eta, \qquad \min_{\lambda \in \mathrm{supp}(\nu_N)} |1 + \lambda\, \tilde{m}_N(z)| \ge c$$

Proof. See [FJ22, Propositions A.3, B.1, B.2]. □

Let $m_{\widetilde{\boldsymbol{K}}}(z) = N^{-1}\operatorname{Tr}(\widetilde{\boldsymbol{K}} - z\boldsymbol{I})^{-1}$ be the Stieltjes transform of the empirical eigenvalue distribution of $\widetilde{\boldsymbol{K}} = \boldsymbol{G}\boldsymbol{G}^\top$. Since $\widetilde{\boldsymbol{K}}$ and $\boldsymbol{K} = \boldsymbol{G}^\top\boldsymbol{G}$ have the same eigenvalues up to $|N-n|$ 0's, we have

$$m_{\widetilde{\boldsymbol{K}}}(z) = \gamma_N m_{\boldsymbol{K}}(z) + (1-\gamma_N)(-1/z), \tag{3.5.27}$$

so in particular $m_{\widetilde{\boldsymbol{K}}}$ coincides with $\tilde{m}_{\boldsymbol{K}}^{(\emptyset)}$ from (3.5.1). We begin with a preliminary estimate for the Stieltjes transform $m_{\widetilde{\boldsymbol{K}}}(z)$ when $\operatorname{Im} z \geq N^{-1/11}$. Similar statements have been shown in [Sil95, BS98], and we provide an argument here following ideas of [BS98, Section 3] for later reference.

Lemma 46. *Fix any $\varepsilon > 0$, and suppose Assumption 6 holds. Then, uniformly over $z = x + i\eta \in U_N(\varepsilon)$ with $\operatorname{Im} z \geq N^{-1/11}$,*

$$m_{\widetilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z) \prec \frac{1}{\sqrt{N}\eta^4}.$$

Proof. Let $\boldsymbol{R}^{(i)}$ and $\tilde{m}_{\boldsymbol{K}}^{(i)}$ be as defined in (3.5.1) with $\boldsymbol{\Gamma} = z\boldsymbol{I}$. Applying the Sherman-Morrison-Woodbury formula

$$\boldsymbol{R} = \boldsymbol{R}^{(i)} - \frac{N^{-1}\boldsymbol{R}^{(i)}\boldsymbol{g}_i\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}}{1 + N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i}, \tag{3.5.28}$$

for any matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ we have

$$\begin{aligned}\operatorname{Tr}\boldsymbol{B} = \operatorname{Tr}(\boldsymbol{K} - z\boldsymbol{I})\boldsymbol{R}\boldsymbol{B} &= -z\operatorname{Tr}\boldsymbol{R}\boldsymbol{B} + \frac{1}{N}\sum_{i=1}^N \boldsymbol{g}_i^\top\boldsymbol{R}\boldsymbol{B}\boldsymbol{g}_i \\ &= -z\operatorname{Tr}\boldsymbol{R}\boldsymbol{B} + \frac{1}{N}\sum_{i=1}^N \frac{\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{B}\boldsymbol{g}_i}{1 + N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i}.\end{aligned} \tag{3.5.29}$$

Choosing $\boldsymbol{B} = \boldsymbol{I}$ in (3.5.29), applying $\operatorname{Tr}\boldsymbol{R} = n m_{\boldsymbol{K}} = N m_{\widetilde{\boldsymbol{K}}} + (n-N)(-1/z)$, and rearranging, we obtain the identity

$$m_{\widetilde{\boldsymbol{K}}} = -\frac{1}{Nz}\sum_{i=1}^N \frac{1}{1 + N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i}. \tag{3.5.30}$$

Now fix any deterministic matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, define

$$d_i = \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i - \frac{1}{N}\operatorname{Tr}\boldsymbol{R}\boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma},$$

and choose $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}$ in (3.5.29). Then, applying also the identity (3.5.30), we get

$$\begin{aligned}
&\operatorname{Tr}\boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1} \\
=\ & -z\operatorname{Tr}\boldsymbol{R}\boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1} - zm_{\tilde{\boldsymbol{K}}}\operatorname{Tr}\boldsymbol{R}\boldsymbol{A}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma} + \sum_{i=1}^{N}\frac{d_i}{1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i} \\
=\ & -z\operatorname{Tr}\boldsymbol{R}\boldsymbol{A} + \sum_{i=1}^{N}\frac{d_i}{1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i}. \qquad (3.5.31)
\end{aligned}$$

We proceed to bound d_i, where (for later purposes) we derive estimates in terms of the Frobenius norms of $\boldsymbol{R}, \boldsymbol{R}\boldsymbol{A}, \boldsymbol{R}^{(i)}, \boldsymbol{R}^{(i)}\boldsymbol{A}$ rather than their operator norms. Note that Assumption 6(c) implies, for any matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ independent of $\boldsymbol{g}_i$,

$$\|\boldsymbol{B}\boldsymbol{g}_i\|^2 = \boldsymbol{g}_i^\top \boldsymbol{B}^*\boldsymbol{B}\boldsymbol{g}_i \prec \operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{B}^*\boldsymbol{B} + \|\boldsymbol{B}^*\boldsymbol{B}\|_F \prec \|\boldsymbol{B}\|_F^2. \qquad (3.5.32)$$

We have also, by Assumption 6(c) and the Sherman-Morrison-Woodbury formula (3.5.28),

$$\begin{aligned}
N^{-1}|\operatorname{Tr}\boldsymbol{R}\boldsymbol{B} - \operatorname{Tr}\boldsymbol{R}^{(i)}\boldsymbol{B}| &= N^{-2}|1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i|^{-1}|\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{B}\boldsymbol{R}^{(i)}\boldsymbol{g}_i| \\
&\prec N^{-2}|1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i|^{-1}\Big(|\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(i)}\boldsymbol{B}\boldsymbol{R}^{(i)}| + \|\boldsymbol{R}^{(i)}\boldsymbol{B}\boldsymbol{R}^{(i)}\|_F\Big) \\
&\prec N^{-2}|1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i|^{-1}\|\boldsymbol{R}^{(i)}\boldsymbol{B}\|_F\|\boldsymbol{R}^{(i)}\|_F. \qquad (3.5.33)
\end{aligned}$$

Define $d_i = d_{i,1} + d_{i,2} + d_{i,3} + d_{i,4}$ where

$$\begin{aligned}
d_{i,1} &= N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}} \boldsymbol{\Sigma})^{-1} \boldsymbol{g}_i - N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1} \boldsymbol{g}_i, \\
d_{i,2} &= N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1} \boldsymbol{g}_i - N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}, \\
d_{i,3} &= N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1} - N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}, \\
d_{i,4} &= N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1} - N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R} \boldsymbol{A} (\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}} \boldsymbol{\Sigma})^{-1}.
\end{aligned} \tag{3.5.34}$$

Applying the identity (2.4.3) in Chapter 2, the definition of $\tilde{m}_{\boldsymbol{K}}^{(i)}$ in (3.5.1), and the bounds (3.5.32) and (3.5.33) (the latter with $\boldsymbol{B} = \boldsymbol{I}$),

$$\begin{aligned}
|d_{i,1}| &\le N^{-1} \|\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{A}\| \|(\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}} \boldsymbol{\Sigma})^{-1}\| \|(\tilde{m}_{\boldsymbol{K}}^{(i)} - m_{\tilde{\boldsymbol{K}}}) \boldsymbol{\Sigma}\| \|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\| \|\boldsymbol{g}_i\| \\
&\prec N^{-5/2} |1 + N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1} \|\boldsymbol{R}^{(i)} \boldsymbol{A}\|_F \|\boldsymbol{R}^{(i)}\|_F^2 \|(\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}} \boldsymbol{\Sigma})^{-1}\| \|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\|.
\end{aligned} \tag{3.5.35}$$

Applying Assumption 6(c),

$$|d_{i,2}| \prec N^{-1} \|\boldsymbol{R}^{(i)} \boldsymbol{A} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\|_F \le N^{-1} \|\boldsymbol{R}^{(i)} \boldsymbol{A}\|_F \|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\|. \tag{3.5.36}$$

Applying the Sherman-Morrison-Woodbury identity (3.5.28), $|\operatorname{Tr} \boldsymbol{u}\boldsymbol{v}^\top| \le \|\boldsymbol{u}\| \|\boldsymbol{v}\|$, and (3.5.32),

$$\begin{aligned}
|d_{i,3}| &\le N^{-2} |1 + N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1} \|\boldsymbol{\Sigma} \boldsymbol{R}^{(i)} \boldsymbol{g}_i\| \|\boldsymbol{g}_i^\top \boldsymbol{A} \boldsymbol{R}^{(i)} (\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\| \\
&\prec N^{-2} |1 + N^{-1} \boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1} \|\boldsymbol{R}^{(i)} \boldsymbol{A}\|_F \|\boldsymbol{R}^{(i)}\|_F \|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(i)} \boldsymbol{\Sigma})^{-1}\|.
\end{aligned} \tag{3.5.37}$$

Finally, applying (2.4.3) in Chapter 2, (3.5.33) (with $\boldsymbol{B} = \boldsymbol{I}$), and

$$|\operatorname{Tr} \boldsymbol{A}\boldsymbol{B}| \le \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F \le \sqrt{N} \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|,$$

$$|d_{i,4}| = N^{-1}\left|\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}\boldsymbol{A}(\boldsymbol{I}+\tilde{m}_{\boldsymbol{K}}^{(i)}\boldsymbol{\Sigma})^{-1}(\tilde{m}_{\boldsymbol{K}}^{(i)}-m_{\tilde{\boldsymbol{K}}})\boldsymbol{\Sigma}(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\right|$$
$$\prec N^{-5/2}|1+N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i|^{-1}\|\boldsymbol{R}\boldsymbol{A}\|_F\|\boldsymbol{R}\|_F^2\|(\boldsymbol{I}+\tilde{m}_{\boldsymbol{K}}^{(i)}\boldsymbol{\Sigma})^{-1}\|\|(\boldsymbol{I}+m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\|. \tag{3.5.38}$$

For the current proof, we apply (3.5.31) and the definitions (3.5.34) with $\boldsymbol{A}=\boldsymbol{I}$. Recalling $\operatorname{Tr}\boldsymbol{R} = n\,m_{\boldsymbol{K}} = N\,m_{\tilde{\boldsymbol{K}}} + (n-N)(-1/z)$ and rearranging (3.5.31) with $\boldsymbol{A}=\boldsymbol{I}$, we get the identity

$$z_N(m_{\tilde{\boldsymbol{K}}}) - z = -\frac{1}{m_{\tilde{\boldsymbol{K}}}}\cdot\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{1+N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i} \tag{3.5.39}$$

where $z_N(m) = -(1/m) + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}(\boldsymbol{I}+m\boldsymbol{\Sigma})^{-1}$ is the function defined in (3.5.26). For any $z = x+i\eta$ with $\eta > 0$, we have

$$|z(1+N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i)| \geq \operatorname{Im}[z(1+N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i)] \geq \operatorname{Im} z = \eta, \tag{3.5.40}$$
$$\max(\|\boldsymbol{R}\|_F, \|\boldsymbol{R}^{(i)}\|_F) \leq N^{1/2}\max(\|\boldsymbol{R}\|, \|\boldsymbol{R}^{(i)}\|) \leq N^{1/2}\eta^{-1}. \tag{3.5.41}$$

Here, the second inequalities of both (3.5.40) and (3.5.41) follow from the spectral representations of $\boldsymbol{R}, \boldsymbol{R}^{(i)}$, i.e. writing $(\lambda_j, \boldsymbol{v}_j)_{j=1}^n$ for the eigenvalues and unit eigenvectors of $\boldsymbol{K}^{(i)}$, we have

$$\begin{aligned}
\operatorname{Im}[z\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i] &= \operatorname{Im}\left[z\boldsymbol{g}_i^\top\left(\sum_{j=1}^{n}\frac{1}{\lambda_j - z}\boldsymbol{v}_j\boldsymbol{v}_j^\top\right)\boldsymbol{g}_i\right] = \sum_{j=1}^{n}\operatorname{Im}\frac{z}{\lambda_j - z}\cdot(\boldsymbol{g}_i^\top\boldsymbol{v}_j)^2 \\
&= \sum_{j=1}^{n}\frac{\lambda_j\operatorname{Im} z}{|\lambda_j - z|^2}\cdot(\boldsymbol{g}_i^\top\boldsymbol{v}_j)^2 \geq 0, \\
\|\boldsymbol{R}^{(i)}\| &= \left\|\sum_{j=1}^{n}\frac{1}{\lambda_j - z}\boldsymbol{v}_j\boldsymbol{v}_j^\top\right\| = \max_{1\leq j\leq n}|\lambda_j - z|^{-1} \leq \eta^{-1},
\end{aligned}$$

and similarly for $\|\boldsymbol{R}\|$. In particular, (3.5.40) and (3.5.41) imply

$$(1+N^{-1}\boldsymbol{g}_i^\top\boldsymbol{R}^{(i)}\boldsymbol{g}_i)^{-1} \prec \eta^{-1}, \qquad \|\boldsymbol{R}\|_F, \|\boldsymbol{R}^{(i)}\|_F \prec N^{1/2}\eta^{-1}. \tag{3.5.42}$$

Next, observe that if $m(z) = \int \frac{1}{\lambda - z} d\mu(\lambda)$ is the Stieltjes transform of any probability measure μ supported on $[-B, B]$, then for $z = x + i\eta$ with $\eta > 0$ and $|z| \le \varepsilon^{-1}$, we have

$$\operatorname{Im} m(z) = \int \frac{\eta}{|\lambda - z|^2} d\mu(\lambda) \ge c\eta, \quad |\operatorname{Re} m(z)| \le \int \frac{|\lambda - x|}{|\lambda - z|^2} d\mu(\lambda) \le (C/\eta) \operatorname{Im} m(z)$$

for some constants $C, c > 0$ depending on ε, B. Consequently, for any $\lambda \ge 0$, either $\lambda \cdot |\operatorname{Re} m(z)| < 1/2$ or $\lambda \cdot \operatorname{Im} m(z) \ge 2\eta/C$, so $|1 + \lambda m(z)| \ge \max(2, 2\eta/C)$. By Assumption 6(b) and Weyl's inequality, we have $\mathbf{1}\{\|\boldsymbol{K}\| > B\} \prec 0$ and $\mathbf{1}\{\|\boldsymbol{K}^{(i)}\| > B\} \prec 0$, and on the event where $\|\boldsymbol{K}\|, \|\boldsymbol{K}^{(i)}\| \le B$, we have that $m_{\tilde{\boldsymbol{K}}}, \tilde{m}^{(i)}_{\boldsymbol{K}}$ are Stieltjes transforms of probability measures supported on $[-B, B]$. Thus, this implies

$$|m_{\tilde{\boldsymbol{K}}}|^{-1} \le |\operatorname{Im} m_{\tilde{\boldsymbol{K}}}|^{-1} \prec \eta^{-1}, \quad \max(\|(\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}} \boldsymbol{\Sigma})^{-1}\|, \|(\boldsymbol{I} + \tilde{m}^{(i)}_{\boldsymbol{K}} \boldsymbol{\Sigma})^{-1}\|) \prec \eta^{-1}. \tag{3.5.43}$$

Applying these bounds (3.5.42) and (3.5.43) to (3.5.35)–(3.5.38), we get

$$d_i \prec N^{-1}\eta^{-6} + N^{-1/2}\eta^{-2} \le 2N^{-1/2}\eta^{-2}$$

for $\eta \ge N^{-1/11}$. Then, applying these bounds (3.5.42) and (3.5.43) also to (3.5.39), we get

$$z_N(m_{\tilde{\boldsymbol{K}}}) - z \prec \frac{1}{\sqrt{N}\eta^4}. \tag{3.5.44}$$

The proof is completed by the following stability argument: When $\eta \ge N^{-1/11}$, we have $1/(\sqrt{N}\eta^4) \ll \eta = \operatorname{Im} z$, so (3.5.44) implies in particular that

$$\mathbf{1}\{z_N(m_{\tilde{\boldsymbol{K}}}) \notin \mathbb{C}^+\} \prec 0. \tag{3.5.45}$$

On the event $z_N(m_{\tilde{\boldsymbol{K}}}) \in \mathbb{C}^+$, recalling the implicit definition of $\tilde{m}_N : \mathbb{C}^+ \to \mathbb{C}^+$ by (3.5.24), the

value $\tilde{m}_N(z_N(m_{\tilde{\boldsymbol{K}}}))$ must be the unique root $u \in \mathbb{C}^+$ to the equation

$$z_N(m_{\tilde{\boldsymbol{K}}}) = -\frac{1}{u} + \gamma_N \int \frac{\lambda}{1+\lambda u} d\nu_N(\lambda),$$

i.e. to the equation $z_N(m_{\tilde{\boldsymbol{K}}}) = z_N(u)$. This equation is satisfied by $u = m_{\tilde{\boldsymbol{K}}} \in \mathbb{C}^+$, so we deduce that $\tilde{m}_N(z_N(m_{\tilde{\boldsymbol{K}}})) = m_{\tilde{\boldsymbol{K}}}$. Then, applying that $z \in U_N(\varepsilon)$ and that $\tilde{m}_N : \mathbb{C}^+ \to \mathbb{C}^+$ is $(4/\varepsilon^2)$-Lipschitz over the domain $U_N(\varepsilon/2)$, we obtain from (3.5.44) that

$$\mathbf{1}\{z_N(m_{\tilde{\boldsymbol{K}}}) \in \mathbb{C}^+\}\Big(m_{\tilde{\boldsymbol{K}}} - \tilde{m}_N(z)\Big) = \mathbf{1}\{z_N(m_{\tilde{\boldsymbol{K}}}) \in \mathbb{C}^+\}\Big(\tilde{m}_N(z_N(m_{\tilde{\boldsymbol{K}}})) - \tilde{m}_N(z)\Big) \prec \frac{1}{\sqrt{N}\eta^4}.$$

Together with (3.5.45), this yields the lemma. □

Corollary 47. *Fix any $\varepsilon > 0$, and suppose Assumption 6 holds. Then there is a constant $C > 0$ such that uniformly over $z \in U_N(\varepsilon)$ with $\operatorname{Im} z \geq N^{-1/11}$,*

$$\mathbf{1}\{\|\boldsymbol{R}(z)\|_F > C\sqrt{N}\} \prec 0.$$

Proof. Since $m_{\tilde{\boldsymbol{K}}}(z) = \gamma_N m_{\boldsymbol{K}}(z) + (1-\gamma_N)(-1/z)$ and $\tilde{m}_N(z) = \gamma_N m_N(z) + (1-\gamma_N)(-1/z)$, Lemma 46 implies also

$$m_{\boldsymbol{K}}(z) - m_N(z) \prec \frac{1}{\sqrt{N}\eta^4} \ll \eta.$$

Observe that $\operatorname{Im} m_N(z) = \int \eta/|\lambda - z|^2 d\mu_N(\lambda) \leq \eta\varepsilon^{-2}$ for $z \in U_N(\varepsilon)$, so $\mathbf{1}\{\operatorname{Im} m_{\boldsymbol{K}}(z) > (1+\varepsilon^{-2})\eta\} \prec 0$. Then by the identity $\|\boldsymbol{R}(z)\|_F^2 = \sum_i 1/|z-\lambda_i(\boldsymbol{K})|^2 = (n/\eta)\operatorname{Im} m_{\boldsymbol{K}}(z)$, we get $\mathbf{1}\{\|\boldsymbol{R}(z)\|_F > C\sqrt{N}\} \prec 0$ for a constant $C = C(\varepsilon) > 0$, as desired. □

We may now apply Corollary 47 and the fluctuation averaging result of Lemma 41 to improve the estimate of Lemma 46 to the following result.

Lemma 48. *Fix any $\varepsilon > 0$, and suppose Assumption 6 holds. Then, uniformly over $z = x + i\eta \in$*

$U_N(\varepsilon)$ *with* $\operatorname{Im} z \geq N^{-1/11}$,

$$m_{\tilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z) \prec \frac{1}{N}.$$

Proof. We derive an improved estimate for (3.5.39). First, combining Lemma 46 with the bounds for $\tilde{m}_N(z)$ in Proposition 45, there are constants $C_0, c_0 > 0$ for which

$$\mathbf{1}\{|m_{\tilde{\boldsymbol{K}}}| > C_0\} \prec 0, \qquad \mathbf{1}\{|m_{\tilde{\boldsymbol{K}}}| < c_0\} \prec 0, \qquad \mathbf{1}\{\|(\boldsymbol{I} + m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\| > C_0\} \prec 0 \tag{3.5.46}$$

uniformly over $z \in U_N(\varepsilon)$ with $\operatorname{Im} z \geq N^{-1/11}$. Next, applying Assumption 6(c), we have also uniformly over $i \in [N]$,

$$\begin{aligned} N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i &= N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R}^{(i)} + O_\prec\Big(N^{-1}\|\boldsymbol{R}^{(i)}\|_F\Big) \\ &= N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R} + O_\prec\Big(N^{-2}|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1}\|\boldsymbol{R}^{(i)}\|_F^2\Big) + O_\prec\Big(N^{-1}\|\boldsymbol{R}^{(i)}\|_F\Big) \end{aligned}$$

where the second line follows from (3.5.33) applied with $\boldsymbol{B} = \boldsymbol{\Sigma}$. Applying $\|\boldsymbol{R}^{(i)}\|_F \prec N^{1/2}$ by Corollary 47 and the estimate $|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1} \prec \eta^{-1}$ from (3.5.42), this gives

$$1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i = 1 + N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R} + O_\prec\Big(N^{-1/2}\Big). \tag{3.5.47}$$

Then, applying this and $|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i|^{-1} \prec \eta^{-1}$ to (3.5.30),

$$m_{\tilde{\boldsymbol{K}}} = -\frac{1}{z} \cdot \frac{1}{1 + N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R}} + O_\prec\Big(N^{-1/2}\eta^{-2}\Big).$$

Together with the first bound of (3.5.46) and the bound $|z| \leq \varepsilon^{-1}$ for $z \in U_N(\varepsilon)$, this implies for a constant $c_0 > 0$ that $\mathbf{1}\{|1 + N^{-1} \operatorname{Tr} \boldsymbol{\Sigma} \boldsymbol{R}| < c_0\} \prec 0$, and thus $\mathbf{1}\{|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)} \boldsymbol{g}_i| < c_0\} \prec 0$.

Applying Corollary 47 and the above arguments now for $\boldsymbol{K}^{(S)}$ and $\boldsymbol{R}^{(S)}$ in place of $\boldsymbol{K}$ and $\boldsymbol{R}$, we obtain for any fixed $L \geq 1$ and some constants $C_0, c_0 > 0$, uniformly over $S \subset [N]$ with

$|S| \le L$, over $i \in S$, and over $z \in U_N(\varepsilon)$ with $\operatorname{Im} z \ge N^{-1/11}$,

$$\begin{gathered}\mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}| > C_0\} \prec 0, \quad \mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}| < c_0\} \prec 0, \quad \mathbf{1}\{\|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma})^{-1}\| > C_0\} \prec 0, \\ \mathbf{1}\{\|\boldsymbol{R}^{(S)}\|_F > C\sqrt{N}\} \prec 0, \quad \mathbf{1}\{|1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}| < c_0\} \prec 0, \\ \mathbf{1}\{|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(S)}\boldsymbol{g}_i| < c_0\} \prec 0.\end{gathered} \tag{3.5.48}$$

(We remark that a direct application of the above arguments for $\boldsymbol{K}^{(S)}$ yields the first three estimates of (3.5.48) for the quantity $\frac{N}{N-|S|}\tilde{m}_{\boldsymbol{K}}^{(S)} = \frac{1}{N-|S|}\operatorname{Tr}\boldsymbol{R}^{(S)} + \frac{n}{N-|S|}(-1/z)$ in place of $\tilde{m}_{\boldsymbol{K}}^{(S)}$, and the estimates for $\tilde{m}_{\boldsymbol{K}}^{(S)}$ then follow for slightly modified constants $C_0, c_0 > 0$ because $|S| \le L$.)

Finally, applying (3.5.47) and (3.5.48) back to (3.5.39) and (3.5.35)–(3.5.38) with $\boldsymbol{A} = \boldsymbol{I}$, we get $|d_{i,1}|, |d_{i,3}|, |d_{i,4}| \prec N^{-1}$, $|d_{i,2}| \prec N^{-1/2}$, and

$$\begin{aligned}|z_N(m_{\tilde{\boldsymbol{K}}}) - z| &\prec \left|\frac{1}{N}\sum_{i=1}^N \frac{d_{i,2}}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i}\right| + O_\prec\big(N^{-1}\big) \\ &= \frac{1}{N} \cdot \frac{1}{1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}} \cdot \left|\sum_{i=1}^N d_{i,2}\right| + O_\prec\big(N^{-1}\big).\end{aligned}$$

The statements of (3.5.48) verify the needed assumptions of Lemma 41 with $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{\Gamma} = z\boldsymbol{I}$, and $\Phi_N = \Psi_N = C\sqrt{N}$. Then Lemma 41 gives $\sum_{i=1}^N d_{i,2} \prec 1$, and hence

$$|z_N(m_{\tilde{\boldsymbol{K}}}) - z| \prec N^{-1}.$$

The proof is then completed by the same stability argument as in the conclusion of the proof of Lemma 46. □

Proof of Theorem 37. We apply the idea of [BS98, Section 6]. Let $z = x + i\eta$, where $\operatorname{dist}(x, \mathscr{S}_N) \ge \varepsilon$ and $\eta = N^{-1/11}$. Taking imaginary part in the estimate $m_{\tilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z) \prec N^{-1}$

of Lemma 48 and multiplying by η gives

$$\frac{1}{N}\sum_{j=1}^{N}\frac{\eta^2}{(\lambda_j(\widetilde{\boldsymbol{K}})-x)^2+\eta^2}-\int\frac{\eta^2}{(\lambda-x)^2+\eta^2}d\tilde{\mu}_N(\lambda)\prec\frac{\eta}{N}.$$

Fix any integer $P\geq 1$, and apply this instead at the point $z=x+i\sqrt{p}\,\eta$ for each $p=1,\ldots,P$. Then

$$\frac{1}{N}\sum_{j=1}^{N}\frac{\eta^2}{(\lambda_j(\widetilde{\boldsymbol{K}})-x)^2+p\eta^2}-\int\frac{\eta^2}{(\lambda-x)^2+p\eta^2}d\tilde{\mu}_N(\lambda)\prec\frac{\eta}{N}\text{ for all }p=1,\ldots,P.$$

Taking successive finite differences using

$$\frac{1}{r-q+1}\left(\frac{1}{\prod_{p=q}^{r}(\lambda-x)^2+p\eta^2}-\frac{1}{\prod_{p=q+1}^{r+1}(\lambda-x)^2+p\eta^2}\right)=\frac{\eta^2}{\prod_{p=q}^{r+1}(\lambda-x)^2+p\eta^2},$$

we then obtain

$$\frac{1}{N}\sum_{j=1}^{N}\frac{\eta^{2P}}{\prod_{p=1}^{P}[(\lambda_j(\widetilde{\boldsymbol{K}})-x)^2+p\eta^2]}-\int\frac{\eta^{2P}}{\prod_{p=1}^{P}[(\lambda-x)^2+p\eta^2]}d\tilde{\mu}_N(\lambda)\prec\frac{\eta}{N}.\tag{3.5.49}$$

Since $\operatorname{dist}(x,\mathscr{S}_N)\geq\varepsilon$, the second integral term of (3.5.49) is bounded by $C\eta^{2P}$ for a constant $C:=C(\varepsilon,P)>0$. Thus, we get

$$\frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\{\lambda_j(\widetilde{\boldsymbol{K}})\in(x-\eta,x+\eta)\}\leq\frac{C}{N}\sum_{j=1}^{N}\frac{\eta^{2P}}{\prod_{p=1}^{P}[(\lambda_j(\widetilde{\boldsymbol{K}})-x)^2+p\eta^2]}\prec\frac{\eta}{N}+\eta^{2P}$$

where the first inequality holds for a constant $C:=C(P)>0$. Finally, recalling $\eta=N^{-1/11}$ and taking any $P\geq 6$, we get $\eta/N+\eta^{2P}\ll 1/N$, hence

$$\mathbf{1}\big\{\text{there exists an eigenvalue of }\widetilde{\boldsymbol{K}}\text{ in }(x-\eta,x+\eta)\big\}\prec 0.$$

Recalling Assumption 6(b) and taking a union bound over x belonging to a η-net of $[-B,B]\setminus$

$(\mathscr{S}_N + (-\varepsilon, \varepsilon))$ (with cardinality at most $CN^{1/11}$), we obtain

$$\mathbf{1}\big\{\text{there exists an eigenvalue of } \widetilde{\boldsymbol{K}} \text{ in } \mathscr{S}_N + (-\varepsilon, \varepsilon)\big\} \prec 0.$$

The theorem follows from the observation that $\boldsymbol{K}$ has the same non-zero eigenvalues as $\widetilde{\boldsymbol{K}}$, and all 0 eigenvalues belong by definition to $\mathscr{S}_N$. □

3.5.3 Deterministic Equivalent for the Resolvent

In this section, we prove Theorem 38.

Lemma 49. *Suppose Assumption 6 holds. Let*

$$\gamma_N^{(S)} = \frac{n}{N - |S|}, \qquad \mu_N^{(S)} = \rho_{\gamma_N^{(S)}}^{\mathrm{MP}} \boxtimes \nu_N, \qquad \tilde{\mu}_N^{(S)} = \gamma_N^{(S)} \mu_N^{(S)} + (1 - \gamma_N^{(S)}) \delta_0$$

be the analogues of $\gamma_N, \mu_N, \tilde{\mu}_N$ defined with the dimension $N - |S|$ in place of N. Then for any fixed $\varepsilon > 0$ and $L \geq 1$, all large N, and all $S \subset [N]$ with $|S| \leq L$,

$$\mathrm{supp}(\tilde{\mu}_N^{(S)}) \subseteq \mathrm{supp}(\tilde{\mu}_N) + (-\varepsilon, \varepsilon)$$

Proof. Let $\mathscr{T}_N$ and $z_N : \mathbb{C} \setminus \mathscr{T}_N \to \mathbb{C}$ be as defined by (3.5.25) and (3.5.26). Define similarly

$$z_N^{(S)}(\tilde{m}) = -\frac{1}{\tilde{m}} + \gamma_N^{(S)} \int \frac{\lambda}{1 + \lambda \tilde{m}} d\nu_N(\lambda), \qquad z_N^{(S)} : \mathbb{C} \setminus \mathscr{T}_N \to \mathbb{C}.$$

We recall from Proposition 3 that $x \in \mathbb{R} \setminus \mathrm{supp}(\tilde{\mu}_N)$ if and only if there exists $\tilde{m} \in \mathbb{R} \setminus \mathscr{T}_N$ where $z_N(\tilde{m}) = x$ and $z_N'(\tilde{m}) > 0$; the analogous characterization holds for $\mathbb{R} \setminus \mathrm{supp}(\tilde{\mu}_N^{(S)})$ and $z_N^{(S)}(\tilde{m})$.

Now fix any $\varepsilon, L > 0$. By Proposition 45, there is a constant $C_0 > 0$ such that $\mathrm{supp}(\tilde{\mu}_N^{(S)}) \subseteq [-C_0, C_0]$ for all $|S| \leq L$ and all large N. Consider any $x \in [-C_0, C_0] \setminus (\mathrm{supp}(\tilde{\mu}_N) + (-\varepsilon, \varepsilon))$. Then $[x - \varepsilon/2, x + \varepsilon/2] \subset \mathbb{R} \setminus \mathrm{supp}(\tilde{\mu}_N)$, so $\tilde{m}_N$ is well-defined and increasing on $[x - \varepsilon/2, x + \varepsilon/2]$.

Define $[\tilde{m}_-, \tilde{m}_+] = [\tilde{m}_N(x-\varepsilon/2), \tilde{m}_N(x+\varepsilon/2)]$. Then Proposition 3 implies that z_N is increasing on $[\tilde{m}_-, \tilde{m}_+]$, and $z_N([\tilde{m}_-, \tilde{m}_+]) = [x-\varepsilon/2, x+\varepsilon/2]$. Again by Proposition 45, there is a constant $c > 0$ such that, for any such $x \in [-C_0, C_0] \setminus (\mathrm{supp}(\tilde{\mu}_N) + (-\varepsilon, \varepsilon))$, we have

$$\min_{y \in [x-\varepsilon/2, x+\varepsilon/2]} \min_{\lambda \in \mathrm{supp}(\nu_N)} |1 + \lambda \tilde{m}_N(y)| > c.$$

This then implies that there is a constant $C > 0$ for which

$$|z_N^{(S)}(\tilde{m}) - z_N(\tilde{m})| = |\gamma_N^{(S)} - \gamma_N| \cdot \left| \int \frac{\lambda}{1 + \lambda \tilde{m}} d\nu_N(\lambda) \right| \le \frac{C}{N} < \varepsilon/2$$

for all $\tilde{m} \in [\tilde{m}_-, \tilde{m}_+]$, $|S| \le L$, and large N. Then $z_N^{(S)}(\tilde{m}_-) < z_N(\tilde{m}_-) + \varepsilon/2 = x$ and $z_N^{(S)}(\tilde{m}_+) > z_N(\tilde{m}_+) - \varepsilon/2 = x$. [SC95, Theorem 4.3] shows that if $m_1, m_2 \in [\tilde{m}_-, \tilde{m}_+]$ satisfy $z_N^{(S)\prime}(m_1) \ge 0$ and $z_N^{(S)\prime}(m_2) \ge 0$, then $z_N^{(S)\prime}(m) > 0$ strictly for all $m \in [m_1, m_2]$. By this and the continuity and differentiability of $z_N^{(S)}$ on $[\tilde{m}_-, \tilde{m}_+]$, there must be a point $\tilde{m} \in (\tilde{m}_-, \tilde{m}_+)$ where $z_N^{(S)}(\tilde{m}) = x$ and $z_N^{(S)\prime}(\tilde{m}) > 0$ strictly. Then Proposition 3 implies that $x \notin \mathrm{supp}(\tilde{\mu}_N^{(S)})$. This holds for all $x \in [-C_0, C_0] \setminus (\mathrm{supp}(\tilde{\mu}_N) + (-\varepsilon, \varepsilon))$, implying $\mathrm{supp}(\tilde{\mu}_N^{(S)}) \subseteq \mathrm{supp}(\tilde{\mu}_N) + (-\varepsilon, \varepsilon)$ as desired. □

The following now applies Lemma 49 and Theorem 37 to extend the estimates (3.5.48) previously obtained over $\{z \in U_N(\varepsilon) : \mathrm{Im}\, z \ge N^{-1/11}\}$ to all of $U_N(\varepsilon)$.

Lemma 50. *Fix any $\varepsilon > 0$ and $L \ge 1$. Then for some constants $C_0, c_0 > 0$, uniformly over $z \in U_N(\varepsilon)$, $S \subset [N]$ with $|S| \le L$, and $i \in S$, we have*

$$\mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}(z)| > C_0\} \prec 0, \quad \mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}(z)| < c_0\} \prec 0, \quad \mathbf{1}\{\|(\boldsymbol{I} + \tilde{m}_{\boldsymbol{K}}^{(S)}(z)\boldsymbol{\Sigma})^{-1}\| > C_0\} \prec 0,$$
$$\mathbf{1}\{\|\boldsymbol{R}^{(S)}(z)\| > C_0\} \prec 0, \quad \mathbf{1}\{|1 + N^{-1}\,\mathrm{Tr}\,\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(z)| < c_0\} \prec 0,$$
$$\mathbf{1}\{|1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(S)}(z)\boldsymbol{g}_i| < c_0\} \prec 0.$$

Proof. By conjugation symmetry, it suffices to show the statements for $z \in U_N(\varepsilon)$ with $\mathrm{Im}\, z \ge$

0. Denote for simplicity $\boldsymbol{R}^{(S)} = \boldsymbol{R}^{(S)}(z)$ and $\tilde{m}_{\boldsymbol{K}}^{(S)} = \tilde{m}_{\boldsymbol{K}}^{(S)}(z)$. Let $\mathscr{S}_N^{(S)} = \mathrm{supp}(\mu_N^{(S)}) \cup \{0\} = \mathrm{supp}(\tilde{\mu}_N^{(S)}) \cup \{0\}$ where $\mu_N^{(S)}, \tilde{\mu}_N^{(S)}$ are as defined in Lemma 49. Then Theorem 37 applied to $\boldsymbol{K}^{(S)}$ guarantees that

$$\mathbf{1}\{\boldsymbol{K}^{(S)} \text{ has an eigenvalue outside } \mathscr{S}_N^{(S)} + (-\varepsilon/4, \varepsilon/4)\} \prec 0,$$

uniformly over all $S \subset [N]$ with $|S| \le L$. Note that $\mathscr{S}_N^{(S)} + (-\varepsilon/4, \varepsilon/4) \subseteq \mathscr{S}_N + (-\varepsilon/2, \varepsilon/2)$ by Lemma 49. Then, applying the bound $\|\boldsymbol{R}^{(S)}\| \le 1/\mathrm{dist}(z, \mathscr{S}_N^{(S)})$ and the condition $z \in U_N(\varepsilon)$, we get

$$\mathbf{1}\{\|\boldsymbol{R}^{(S)}\| > 2/\varepsilon\} \prec 0. \tag{3.5.50}$$

The remaining statements have already been shown for $z \in U_N(\varepsilon)$ with $\mathrm{Im}\, z \ge N^{-1/11}$ in (3.5.48). For $z = x + i\eta$ where $\eta \in [0, N^{-1/11}]$, define $z' = x + iN^{-1/11}$. On the event that $\boldsymbol{K}^{(S)}$ has no eigenvalues outside $\mathscr{S}_N + (-\varepsilon/2, \varepsilon/2)$, both $N^{-1}\,\mathrm{Tr}\,\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(z)$ and $\tilde{m}_{\boldsymbol{K}}^{(S)}(z) = N^{-1}\,\mathrm{Tr}\,\boldsymbol{R}^{(S)}(z) + \gamma_N(-1/z)$ are C-Lipschitz over $z \in U_N(\varepsilon)$ for a constant $C = C(\varepsilon) > 0$, and $N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(S)}(z)\boldsymbol{g}_i$ is $CN^{-1}\|\boldsymbol{g}_i\|^2$-Lipschitz where $N^{-1}\|\boldsymbol{g}_i\|^2 \prec 1$ by Assumption 6. Then

$$N^{-1}\,\mathrm{Tr}\,\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(z) - N^{-1}\,\mathrm{Tr}\,\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(z') \prec N^{-1/11}, \quad \tilde{m}_{\boldsymbol{K}}^{(S)}(z) - \tilde{m}_{\boldsymbol{K}}^{(S)}(z') \prec N^{-1/11},$$

$$N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(S)}(z)\boldsymbol{g}_i - N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(S)}(z')\boldsymbol{g}_i \prec N^{-1/11},$$

so the remaining statements of the lemma hold also for $z \in U_N(\varepsilon)$ with $\mathrm{Im}\, z \in [0, N^{-1/11}]$. □

Proof of Theorem 38. Again by conjugation symmetry, it suffices to show the result for $z \in U_N(\varepsilon)$ with $\mathrm{Im}\, z \ge 0$. Denote for simplicity $\boldsymbol{R}^{(S)} = \boldsymbol{R}^{(S)}(z)$ and $\tilde{m}_{\boldsymbol{K}}^{(S)} = \tilde{m}_{\boldsymbol{K}}^{(S)}(z)$. The first estimate of Lemma 50 implies

$$\mathbf{1}\{\|\boldsymbol{R}^{(S)}\|_F > C\sqrt{N}\} \prec 0, \qquad \mathbf{1}\{\|\boldsymbol{R}^{(S)}\boldsymbol{A}\|_F > C\|\boldsymbol{A}\|_F\} \prec 0 \tag{3.5.51}$$

uniformly over $z \in U_N(\varepsilon)$ and $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then also, by Assumption 6(c) and (3.5.33) applied with $\boldsymbol{B} = \boldsymbol{\Sigma}$,

$$\begin{aligned} 1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i &= 1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R} + O_\prec\Big(N^{-2}|1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i|^{-1}\|\boldsymbol{R}^{(i)}\|_F^2 + \|\boldsymbol{R}^{(i)}\|_F\Big) \\ &= 1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R} + O_\prec\Big(N^{-1/2}\Big). \end{aligned} \tag{3.5.52}$$

Let $d_i = d_{i,1} + d_{i,2} + d_{i,3} + d_{i,4}$ be as defined in (3.5.34) with $\boldsymbol{A} = \boldsymbol{I}$. Then, applying (3.5.51), (3.5.52), and the bounds of Lemma 50, we obtain exactly as in the proof of Lemma 48 (using again the fluctuation averaging result of Lemma 41) that, uniformly over $z \in U_N(\varepsilon)$, we have $|d_{i,1}|, |d_{i,3}|, |d_{i,4}| \prec N^{-1}$, $|d_{i,2}| \prec N^{-1/2}$, and

$$|z_N(m_{\tilde{\boldsymbol{K}}}) - z| \prec \frac{1}{N}\cdot\frac{1}{1+N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}}\cdot\left|\sum_{i=1}^N d_{i,2}\right| + O_\prec\big(N^{-1}\big) = O_\prec\big(N^{-1}\big).$$

Fix any $\iota > 0$. If $\operatorname{Im} z \geq N^{-1+\iota}$, then this implies $\mathbf{1}\{z_N(m_{\tilde{\boldsymbol{K}}}) \notin \mathbb{C}^+\} \prec 0$. By the same stability argument as in Lemma 46, we get $m_{\tilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z) \prec N^{-1}$ uniformly over $z \in U_N(\varepsilon)$ with $\operatorname{Im} z \geq N^{-1+\iota}$. For $\operatorname{Im} z \in [0, N^{-1+\iota}]$, on the event that all eigenvalues of $\boldsymbol{K}$ belong to $\mathscr{S}_N + (-\varepsilon/2, \varepsilon/2)$, we may apply that both $m_{\tilde{\boldsymbol{K}}}(z)$ and $\tilde{m}_N(z)$ are $C(\varepsilon)$-Lipschitz over $z \in U_N(\varepsilon)$ to compare values at $z = x + i\eta$ and $z' = x + iN^{-1+\iota}$. Applying $m_{\tilde{\boldsymbol{K}}}(z') - \tilde{m}_N(z') \prec N^{-1}$, we then get for any $D > 0$, all $z \in U_N(\varepsilon)$, some constant $C > 0$, and all large N,

$$\mathbb{P}[|m_{\tilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z)| > CN^{-1+\iota}] \leq N^{-D}.$$

Since $\iota > 0$ is arbitrary, this shows $m_{\tilde{\boldsymbol{K}}}(z) - \tilde{m}_N(z) \prec N^{-1}$ uniformly over $z \in U_N(\varepsilon)$. The bound $m_{\boldsymbol{K}}(z) - m_N(z) \prec N^{-1}$ then follows from $m_{\tilde{\boldsymbol{K}}}(z) = \gamma_N m_{\boldsymbol{K}}(z) + (1-\gamma_N)(-1/z)$ and $\tilde{m}_N(z) = \gamma_N m_N(z) + (1-\gamma_N)(-1/z)$.

For the estimate of $\operatorname{Tr}\boldsymbol{R}\boldsymbol{A}$, we apply the definition of $d_i = d_{i,1} + d_{i,2} + d_{i,3} + d_{i,4}$ from

(3.5.34) and the identity (3.5.31) now with this matrix $\boldsymbol{A}$. Then (3.5.31) gives

$$\operatorname{Tr}\left[\boldsymbol{R}\boldsymbol{A} - (-z\boldsymbol{I} - z m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\boldsymbol{A}\right] = \frac{1}{z}\sum_{i=1}^{N} \frac{d_i}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i}.$$

Applying (3.5.51), (3.5.52), and the bounds of Lemma 50 to (3.5.35)–(3.5.38), uniformly over $z \in U_N(\varepsilon)$ and $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, we have $|d_{i,1}|, |d_{i,3}|, |d_{i,4}| \prec N^{-3/2}\|\boldsymbol{A}\|_F$, $|d_{i,2}| \prec N^{-1}\|\boldsymbol{A}\|_F$, and hence

$$\left|\sum_{i=1}^{N} \frac{d_i}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}\boldsymbol{g}_i}\right| \prec \frac{1}{1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}}\left|\sum_{i=1}^{N} d_{i,2}\right| + O_\prec\left(N^{-1/2}\|\boldsymbol{A}\|_F\right).$$

Finally, applying Lemma 41 with $\boldsymbol{\Gamma} = z\boldsymbol{I}$, $\Psi_N(\boldsymbol{\Gamma}) = C\sqrt{N}$, and $\Phi_N(\boldsymbol{\Gamma},\boldsymbol{A}) = C\|\boldsymbol{A}\|_F$ (where we may assume without loss of generality $\|\boldsymbol{A}\|_F \in (N^{-\upsilon}, N^{\upsilon})$ by scale invariance of the desired estimate with respect to $\boldsymbol{A}$), we get $|\sum_i d_{i,2}| \prec N^{-1/2}\|\boldsymbol{A}\|_F$. Thus,

$$\operatorname{Tr}\left[\boldsymbol{R}\boldsymbol{A} - (-z\boldsymbol{I} - z m_{\tilde{\boldsymbol{K}}}\boldsymbol{\Sigma})^{-1}\boldsymbol{A}\right] \prec \frac{1}{\sqrt{N}}\|\boldsymbol{A}\|_F.$$

□

3.6 Analysis of Spiked Eigenstructure

We now consider the asymptotic setup of Section 3.3.2 and prove Corollary 39 and Theorem 40. As all the desired statements are invariant under conjugation of $\boldsymbol{\Sigma}$ by an orthogonal matrix, we may assume without loss of generality that $\boldsymbol{\Sigma}$ is diagonal and of the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_0 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_r = \operatorname{diag}(\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_r(\boldsymbol{\Sigma})), \qquad \boldsymbol{\Sigma}_0 = \operatorname{diag}(\lambda_{r+1}(\boldsymbol{\Sigma}), \ldots, \lambda_n(\boldsymbol{\Sigma})).$$

Denote the block decomposition of $\boldsymbol{G}$ corresponding to $\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_0$ as

$$\boldsymbol{G} = [\boldsymbol{G}_r, \boldsymbol{G}_0], \qquad \boldsymbol{G}_r \in \mathbb{R}^{N\times r}, \qquad \boldsymbol{G}_0 \in \mathbb{R}^{N\times(n-r)}.$$

We remind the reader that $\boldsymbol{G}_r$ and $\boldsymbol{G}_0$ need not be independent.

3.6.1 No Outliers Outside the Limit Support

We consider first the setting of $r = 0$, and prove Corollary 39 together with some uniform convergence properties of $\tilde{m}_N$ and z_N that will be used in the later analysis.

Recall the domain $\mathscr{T}_N$ and function $z_N : \mathbb{C} \setminus \mathscr{T}_N \to \mathbb{C}$ from (3.5.25) and (3.5.26), and their asymptotic analogues $\mathscr{T}$ and $z : \mathbb{C} \setminus \mathscr{T} \to \mathbb{C}$ from (1.2.5) and (1.2.6).

Lemma 51. *Suppose Assumption 6 holds, and Assumption 7 holds with $r = 0$. Then, as $N \to \infty$,*

(a) $z_N(\tilde{m})$ and its derivative $z_N'(\tilde{m})$ converge uniformly over compact subsets of $\mathbb{C} \setminus \mathscr{T}$ to $z(\tilde{m})$ and $z'(\tilde{m})$.

(b) For any $\varepsilon > 0$ and all large N,

$$\operatorname{supp}(\tilde{\mu}_N) \subseteq \operatorname{supp}(\tilde{\mu}) + (-\varepsilon, \varepsilon).$$

(c) $\tilde{m}_N(z)$ and its derivative $\tilde{m}_N'(z)$ converge uniformly over compact subsets of $\mathbb{C} \setminus \operatorname{supp}(\tilde{\mu})$ to $\tilde{m}(z)$ and $\tilde{m}'(z)$.

Proof. For part (a), let $K \subset \mathbb{C} \setminus \mathscr{T}$ be any fixed compact set. Then K does not intersect some sufficiently small open neighborhood of the compact domain $\mathscr{T}$. If Assumption 7 holds with $r = 0$, then $\mathscr{T}_N$ is contained in this open neighborhood of $\mathscr{T}$ for all large N, so $K \subset \mathbb{C} \setminus \mathscr{T}_N$, and both z_N and z are well-defined on K. The pointwise convergences $z_N(\tilde{m}) \to z(\tilde{m})$ and $z_N'(\tilde{m}) \to z'(\tilde{m})$ on K then follow from $\gamma_N \to \gamma$, the weak convergence $\nu_N \to \nu$, and the uniform boundedness of the functions $\lambda \mapsto \lambda/(1+\lambda\tilde{m})$ and $\lambda \mapsto \lambda^2/(1+\lambda\tilde{m})^2$ on an open neighborhood

of $\mathrm{supp}(\nu)$, for $\tilde{m} \in K$. This convergence is furthermore uniform because $\{z_N\}$ and $\{z_N'\}$ are both equicontinuous over K.

For part (b), consider any $x \notin \mathrm{supp}(\tilde{\mu}) + (-\varepsilon, \varepsilon)$. Then $[x-\varepsilon/2, x+\varepsilon/2] \subset \mathbb{R} \setminus \mathrm{supp}(\tilde{\mu})$, so $\tilde{m}$ is well-defined and increasing on $[x-\varepsilon/2, x+\varepsilon/2]$. Let $[\tilde{m}_-, \tilde{m}_+] = [\tilde{m}(x-\varepsilon/2), \tilde{m}(x+\varepsilon/2)]$. Then by Proposition 3, $z'(\tilde{m}) > 0$ for all $\tilde{m} \in [\tilde{m}_-, \tilde{m}_+]$, and $z([\tilde{m}_-, \tilde{m}_+]) = [x-\varepsilon/2, x+\varepsilon/2]$. The uniform convergence in part (a) implies for all large N that $z_N(\tilde{m}_-) < x$, $z_N(\tilde{m}_+) > x$, and $z_N'(\tilde{m}) > 0$ for all $\tilde{m} \in [\tilde{m}_-, \tilde{m}_+]$. Then there exists $\tilde{m} \in [\tilde{m}_-, \tilde{m}_+]$ where $z_N(\tilde{m}) = x$ and $z_N'(\tilde{m}) > 0$, implying by Proposition 3 that $x \notin \mathrm{supp}(\tilde{\mu}_N)$. So $\mathrm{supp}(\tilde{\mu}_N) \subseteq \mathrm{supp}(\tilde{\mu}) + (-\varepsilon, \varepsilon)$ as desired.

For part (c), let $K \subset \mathbb{C} \setminus \mathrm{supp}(\tilde{\mu})$ be any fixed compact set. Then K does not intersect some sufficiently small open neighborhood of the compact set $\mathrm{supp}(\tilde{\mu})$, so the inclusion of part (b) implies $K \subset \mathbb{C} \setminus \mathrm{supp}(\tilde{\mu}_N)$ for all large N, and both $\tilde{m}_N$ and $\tilde{m}$ are well-defined on K. The uniform convergence $\tilde{m}_N(z) \to \tilde{m}(z)$ and $\tilde{m}_N'(z) \to \tilde{m}'(z)$ on K then follow from the weak convergence $\tilde{\mu}_N \to \tilde{\mu}$, the uniform boundedness of the functions $\lambda \mapsto 1/(\lambda - z)$ and $\lambda \mapsto 1/(\lambda - z)^2$ on an open neighborhood of $\mathrm{supp}(\tilde{\mu})$ for $z \in K$, and the equicontinuity of $\{\tilde{m}_N\}$ and $\{\tilde{m}_N'\}$ on K. □

Proof of Corollary 39. By Lemma 51(b), for any fixed $\varepsilon > 0$, we have $\mathscr{S}_N + (-\varepsilon/2, \varepsilon/2) \subseteq \mathscr{S} + (-\varepsilon, \varepsilon)$ for all large N. Then by Theorem 37,

$$\begin{aligned} &\mathbf{1}\{\boldsymbol{K} \text{ has an eigenvalue in } \mathbb{R} \setminus (\mathscr{S} + (-\varepsilon, \varepsilon)) \\ &\leq \mathbf{1}\{\boldsymbol{K} \text{ has an eigenvalue in } \mathbb{R} \setminus (\mathscr{S}_N + (-\varepsilon/2, \varepsilon/2))\} \prec 0. \end{aligned}$$

□

3.6.2 Deterministic Equivalents for Generalized Resolvents

We next introduce two generalized resolvents for the matrix $\boldsymbol{K}$, and extend Theorem 38 to establish deterministic equivalents for these generalized resolvents.

Define the spectral domain

$$U(\varepsilon) = \left\{ z \in \mathbb{C} : |z| \leq \varepsilon^{-1}, \, \mathrm{dist}(z, \mathscr{S}) \geq \varepsilon \right\}$$

where $\mathscr{S}$ is the limit support set defined in (3.3.4). Given $z \in U(\varepsilon)$ and $\alpha \in \mathbb{C}$, define a diagonal matrix

$$\boldsymbol{\Gamma} := \boldsymbol{\Gamma}(z, \alpha) = z\boldsymbol{I}_n + \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top = \begin{pmatrix} (z+\alpha)\boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & z\boldsymbol{I}_{n-r} \end{pmatrix} \in \mathbb{C}^{n \times n}, \quad \boldsymbol{V}_r = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{n \times r}. \tag{3.6.1}$$

Define the first generalized resolvent

$$\boldsymbol{\mathscr{R}}(z, \alpha) = \begin{pmatrix} -\boldsymbol{\Gamma} & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I}_N \end{pmatrix}^{-1} \in \mathbb{C}^{(n+N) \times (n+N)}. \tag{3.6.2}$$

This matrix inverse exists if and only if the Schur complement $\boldsymbol{G}^\top \boldsymbol{G} - \boldsymbol{\Gamma} = \boldsymbol{K} - \boldsymbol{\Gamma}$ for its lower right block is invertible, in which case the upper-left block of $\boldsymbol{\mathscr{R}}(z, \alpha)$ is $\boldsymbol{R}(\boldsymbol{\Gamma}) = (\boldsymbol{K} - \boldsymbol{\Gamma})^{-1}$. The following provides a deterministic equivalent for this block of $\boldsymbol{\mathscr{R}}(z, \alpha)$.

Lemma 52. *Under the assumptions of Theorem 40, for any fixed $\varepsilon > 0$, there exist $C_0, \alpha_0 > 0$ (depending on ε) such that fixing any $\alpha \in \mathbb{C}$ with $|\alpha| > \alpha_0$, the following hold:*

(a) The event

$$\mathscr{E} = \left\{ \boldsymbol{\mathscr{R}}(z, \alpha) \text{ exists and } \|\boldsymbol{\mathscr{R}}(z, \alpha)\| \leq C_0 \text{ for all } z \in U(\varepsilon) \right\}$$

satisfies $\mathbf{1}\{\mathscr{E}^c\} \prec 0$.

(b) Uniformly over $z \in U(\varepsilon)$ and deterministic unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$,

$$\left\| \begin{pmatrix} \boldsymbol{v}_1^\top & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{v}_2 \\ \boldsymbol{0} \end{pmatrix} + \boldsymbol{v}_1^\top (\boldsymbol{\Gamma} + z \cdot \tilde{m}_{N,0}(z)\boldsymbol{\Sigma})^{-1} \boldsymbol{v}_2 \right\| \prec \frac{1}{\sqrt{N}}. \tag{3.6.3}$$

In the setting of Theorem 40(c), let $\boldsymbol{u} = \frac{1}{\sqrt{N}}(u_1, \ldots, u_N) \in \mathbb{R}^N$ be the additional given vector for which $\{(u_j, \boldsymbol{g}_j^\top)\}_{j=1}^N$ are independent vectors in $\mathbb{R}^{n+1}$. For $z \in U(\varepsilon)$ and $\alpha \in \mathbb{C}$, define

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \mathbb{E}[u^2] & \mathbb{E}[u\boldsymbol{g}]^\top \\ \mathbb{E}[u\boldsymbol{g}] & \boldsymbol{\Sigma} \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{3.6.4}$$

$$\widetilde{\boldsymbol{\Gamma}} = \widetilde{\boldsymbol{\Gamma}}(z,\alpha) = \begin{pmatrix} z+\alpha & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Gamma} \end{pmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}$$

where $\mathbb{E}[u^2]$ and $\mathbb{E}[u\boldsymbol{g}]$ denote the common values of $\mathbb{E}[u_j^2]$ and $\mathbb{E}[u_j \boldsymbol{g}_j]$ for $j = 1, \ldots, N$. Define the second generalized resolvent

$$\widetilde{\mathscr{R}}(z,\alpha) = \begin{pmatrix} -\widetilde{\boldsymbol{\Gamma}} & [\boldsymbol{u}, \boldsymbol{G}]^\top \\ [\boldsymbol{u}, \boldsymbol{G}] & -\boldsymbol{I} \end{pmatrix}^{-1} = \begin{pmatrix} -(z+\alpha) & \boldsymbol{0} & \boldsymbol{u}^\top \\ \boldsymbol{0} & -\boldsymbol{\Gamma} & \boldsymbol{G}^\top \\ \boldsymbol{u} & \boldsymbol{G} & -\boldsymbol{I}_N \end{pmatrix}^{-1} \in \mathbb{C}^{(n+1+N)\times(n+1+N)}. \tag{3.6.5}$$

We have the following deterministic equivalent for the upper-left block of $\widetilde{\mathscr{R}}(z,\alpha)$, which is analogous to Lemma 52.

Lemma 53. *Under the assumptions of Theorem 40(c), for any fixed $\varepsilon > 0$, there exist $C_0, \alpha_0 > 0$ (depending on ε) such that fixing any $\alpha \in \mathbb{C}$ with $|\alpha| > \alpha_0$, the following hold:*

(a) The event

$$\widetilde{\mathscr{E}} = \left\{ \widetilde{\mathscr{R}}(z,\alpha) \text{ exists and } \|\widetilde{\mathscr{R}}(z,\alpha)\| \le C_0 \text{ for all } z \in U(\varepsilon) \right\}$$

satisfies $\mathbf{1}\{\widetilde{\mathscr{E}}^c\} \prec 0$.

(b) Uniformly over $z \in U(\varepsilon)$ and deterministic unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^{n+1}$,

$$\left\| \begin{pmatrix} \boldsymbol{v}_1^\top & \mathbf{0} \end{pmatrix} \widetilde{\mathscr{R}}(z,\alpha) \begin{pmatrix} \boldsymbol{v}_2 \\ \mathbf{0} \end{pmatrix} + \boldsymbol{v}_1^\top \left(\widetilde{\boldsymbol{\Gamma}} + z \cdot \tilde{m}_{N,0}(z) \widetilde{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{v}_2 \right\| \prec \frac{1}{\sqrt{N}}. \tag{3.6.6}$$

In the remainder of this section, we prove Lemmas 52 and 53. Recall

$$\mu_{N,0} = \rho^{\mathrm{MP}}_{\gamma_{N,0}} \boxtimes \nu_{N,0}, \qquad \tilde{\mu}_{N,0} = \gamma_{N,0} \mu_{N,0} + (1 - \gamma_{N,0}) \delta_0.$$

Define the bulk components of the sample covariance and Gram matrices

$$\boldsymbol{K}_0 = \boldsymbol{G}_0^\top \boldsymbol{G}_0 \in \mathbb{R}^{(n-r)\times(n-r)}, \qquad \widetilde{\boldsymbol{K}}_0 = \boldsymbol{G}_0 \boldsymbol{G}_0^\top \in \mathbb{R}^{N \times N}. \tag{3.6.7}$$

Define also the N-dependent bulk spectral support and spectral domain

$$\begin{aligned} \mathscr{S}_{N,0} &= \mathrm{supp}(\mu_{N,0}) \cup \{0\} = \mathrm{supp}(\tilde{\mu}_{N,0}) \cup \{0\}, \\ U_{N,0}(\varepsilon) &= \{ z \in \mathbb{C} : |z| \le \varepsilon^{-1}, \mathrm{dist}(z, \mathscr{S}_{N,0}) \ge \varepsilon \}. \end{aligned} \tag{3.6.8}$$

Lemma 51(b) shows $\mathscr{S}_{N,0} \subseteq \mathscr{S} + (-\varepsilon/2, \varepsilon/2)$ for any fixed $\varepsilon > 0$ and all large N, so also $U(\varepsilon) \subseteq U_{N,0}(\varepsilon/2)$ for all large N. Thus, the results of Section 3.5 applied to $\boldsymbol{K}_0$, which hold uniformly over $z \in U_{N,0}(\varepsilon/2)$ for any fixed $\varepsilon > 0$, also hold uniformly over $z \in U(\varepsilon)$. In particular, the following is an immediate consequence of Corollary 39 and Theorem 38, which we record here for future reference.

Lemma 54. *Suppose Assumptions 6 and 7 hold. Then for any fixed $\varepsilon > 0$,*

$$\mathbf{1}\{\boldsymbol{K}_0 \text{ has an eigenvalue outside } \mathscr{S} + (-\varepsilon, \varepsilon)\} \prec 0.$$

Furthermore, uniformly over $z \in U(\varepsilon)$,

$$m_{\boldsymbol{K}_0} - m_{N,0}(z) \prec 1/N, \quad m_{\tilde{\boldsymbol{K}}_0} - \tilde{m}_{N,0}(z) \prec 1/N.$$

We now check that for sufficiently large $|\alpha|$, the generalized resolvent $\mathscr{R}(z,\alpha)$ exists and has bounded operator norm with high probability.

Proof of Lemma 52(a). Let

$$\mathscr{E}' = \Big\{ \text{all eigenvalues of } \boldsymbol{K}_0 \text{ belong to } \mathscr{S} + (-\varepsilon/2, \varepsilon/2), \text{ and } \|\boldsymbol{G}\| < \sqrt{B} \Big\}.$$

By Assumption 6(b) and Lemma 54, $\mathbf{1}\{\mathscr{E}'^c\} \prec 0$, so it suffices to show $\mathscr{E}' \subseteq \mathscr{E}$. On this event $\mathscr{E}'$, for any $z \in U(\varepsilon)$, we have that each eigenvalue of $\boldsymbol{K}_0$ is separated by at least $\varepsilon/2$ from z. Then

$$\mathscr{R}_0(z) := \begin{pmatrix} -z\boldsymbol{I}_{n-r} & \boldsymbol{G}_0^\top \\ \boldsymbol{G}_0 & -\boldsymbol{I}_N \end{pmatrix}^{-1} \in \mathbb{C}^{(n-r+N)\times(n-r+N)} \tag{3.6.9}$$

exists for all $z \in U(\varepsilon)$ because the Schur complement $\boldsymbol{K}_0 - z\boldsymbol{I}_{n-r}$ of its lower-right block is invertible. Furthermore, denoting $\boldsymbol{R}_0 = (\boldsymbol{K}_0 - z\boldsymbol{I}_{n-r})^{-1}$, we have $\|\boldsymbol{R}_0\| \le 2/\varepsilon$ and $\|\boldsymbol{G}_0\| \le \|\boldsymbol{G}\| < \sqrt{B}$, so

$$\|\mathscr{R}_0(z)\| = \left\| \begin{pmatrix} \boldsymbol{R}_0 & \boldsymbol{R}_0\boldsymbol{G}_0^\top \\ \boldsymbol{G}_0\boldsymbol{R}_0 & \boldsymbol{G}_0\boldsymbol{R}_0\boldsymbol{G}_0^\top - \boldsymbol{I}_N \end{pmatrix} \right\| \le C_1 \tag{3.6.10}$$

for some constant C_1 depending only on ε, B.

Now write $\mathscr{R}(z,\alpha)$ as defined in (3.6.2) in its block decomposition with blocks of sizes r

and $n-r+N$. Then the Schur complement of the upper left block of size $r\times r$ is given by

$$\boldsymbol{S} = -\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\mathscr{R}_0(z)\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix} - (\alpha+z)\boldsymbol{I}_r. \tag{3.6.11}$$

Notice that

$$\boldsymbol{S}\boldsymbol{S}^* = |\alpha+z|^2\boldsymbol{I}_r + \begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\mathscr{R}_0(z)\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix}\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\overline{\mathscr{R}_0(z)}\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix} \tag{3.6.12}$$

$$+(\bar\alpha+\bar z)\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\mathscr{R}_0(z)\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix} + (\alpha+z)\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\overline{\mathscr{R}_0(z)}\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix} \tag{3.6.13}$$

where the first two terms are positive semi-definite. Therefore, applying (3.6.10) and $\|\boldsymbol{G}_r\| \le \|\boldsymbol{G}\| < \sqrt{B}$ on the event $\mathscr{E}'$, there exist $\alpha_0, c_0 > 0$ depending only on ε, B, such that

$$\lambda_{\min}(\boldsymbol{S}\boldsymbol{S}^*) \ge |\alpha+z|^2 - 2(|\alpha|+|z|)\|\boldsymbol{G}_r\|^2\|\mathscr{R}_0(z)\| > c_0 \tag{3.6.14}$$

for any $z\in U(\varepsilon)$ and $|\alpha|>\alpha_0$. Consequently, under the event $\mathscr{E}'$, the Schur complement $\boldsymbol{S}$ in (3.6.11) is invertible with $\|\boldsymbol{S}^{-1}\| < c_0^{-1/2}$. Then $\mathscr{R}(z,\alpha)$ exists, and

$$\|\mathscr{R}(z,\alpha)\| = \left\|\begin{pmatrix}\boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\mathscr{R}_0(z)\\ -\mathscr{R}_0(z)\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix}\boldsymbol{S}^{-1} & \mathscr{R}_0(z)+\mathscr{R}_0(z)\begin{pmatrix}\mathbf{0}\\ \boldsymbol{G}_r\end{pmatrix}\boldsymbol{S}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{G}_r^\top\end{pmatrix}\mathscr{R}_0(z)\end{pmatrix}\right\| \le C_0 \tag{3.6.15}$$

for a constant $C_0>0$ depending only on ε, B. This shows $\mathscr{E}'\subseteq\mathscr{E}$ as desired. □

For the matrix $\boldsymbol{\Gamma}=\boldsymbol{\Gamma}(z,\alpha)$ in (3.6.1), recall the definitions of $\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})$ and $\tilde m_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})$ from

(3.5.1). The following provides an analogue of Lemma 50 for these quantities.

Lemma 55. *Fix any $\varepsilon > 0$ and $L \geq 1$. Then there exist $C_0, c_0, \alpha_0 > 0$ such that for any fixed $\alpha \in \mathbb{C}$ with $|\alpha| > \alpha_0$, uniformly over $S \subset [N]$ with $|S| \leq L$, over $j \in S$, and over $z \in U(\varepsilon)$,*

$$\mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})| > C_0\} \prec 0, \quad \mathbf{1}\{|\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})| < c_0\} \prec 0, \quad \mathbf{1}\{\|(z^{-1}\boldsymbol{\Gamma} + \tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\| > C_0\} \prec 0,$$
$$\mathbf{1}\{\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\| > C_0\} \prec 0, \quad \mathbf{1}\{|1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})| < c_0\} \prec 0,$$
$$\mathbf{1}\{|1 + N^{-1}\boldsymbol{g}_j^\top \boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{g}_j| < c_0\} \prec 0.$$

Proof. Suppose $|\alpha|$ is large enough so that Lemma 52(a) holds. Since $\boldsymbol{R}(\boldsymbol{\Gamma})$ is the upper-left block of $\mathscr{R}(z,\alpha)$, Lemma 52(a) applied with $\boldsymbol{G}^{(S)}$ in place of $\boldsymbol{G}$ shows that $\mathbf{1}\{\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\| > C_0\} \prec 0$ for a constant $C_0 > 0$, uniformly over $S \subset [N]$ with $|S| \leq L$ and over $z \in U(\varepsilon)$. For the remaining statements, let $\boldsymbol{G}_0^{(S)} \in \mathbb{R}^{(N-|S|)\times(n-r)}$ be the submatrix of $\boldsymbol{G}_0$ with the rows of S removed, and define

$$\boldsymbol{K}_0^{(S)} = \boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_0^{(S)}, \quad \widetilde{\boldsymbol{K}}_0^{(S)} = \boldsymbol{G}_0^{(S)}\boldsymbol{G}_0^{(S)\top},$$
$$\boldsymbol{R}_0^{(S)} = (\boldsymbol{K}_0^{(S)} - z\boldsymbol{I}_{n-r})^{-1}, \quad m_{\boldsymbol{K}_0}^{(S)} = \frac{1}{n-r}\operatorname{Tr}\boldsymbol{R}_0^{(S)}, \quad \tilde{m}_{\boldsymbol{K}_0}^{(S)} = \gamma_{N,0}m_{\boldsymbol{K}_0}^{(S)} + (1-\gamma_{N,0})\Big(-\frac{1}{z}\Big).$$

Then by Lemma 54 applied to $\widetilde{\boldsymbol{K}}_0^{(S)}$, also $\mathbf{1}\{\|\boldsymbol{R}_0^{(S)}\| > C_0\} \prec 0$ for a constant $C_0 > 0$.

Using these bounds, we first show the comparisons

$$|\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) - \tilde{m}_{\boldsymbol{K}_0}^{(S)}| \prec 1/N, \quad \Big|N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}) - N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}_0\boldsymbol{R}_0^{(S)}\Big| \prec 1/N. \tag{3.6.16}$$

For the first comparison, notice that in the decompositions into blocks of sizes r and $n-r$,

$$\frac{n-r}{n}m_{\boldsymbol{K}_0}^{(S)} = \frac{1}{n}\operatorname{Tr}\begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}_0^{(S)}\end{pmatrix}$$

and

$$
\begin{aligned}
m_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) &= \frac{1}{n}\operatorname{Tr}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma}) \\
&= \frac{1}{n}\operatorname{Tr}\begin{pmatrix} \boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_r^{(S)} - (\alpha+z)\boldsymbol{I}_r & \boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)} \\ \boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)} & \boldsymbol{K}_0^{(S)} - z\boldsymbol{I}_{N-|S|} \end{pmatrix}^{-1} \\
&= \frac{1}{n}\operatorname{Tr}\begin{pmatrix} (\boldsymbol{S}_r^{(S)})^{-1} & -(\boldsymbol{S}_r^{(S)})^{-1}\boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)} \\ -\boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)}(\boldsymbol{S}_r^{(S)})^{-1} & \boldsymbol{R}_0^{(S)} + \boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)}(\boldsymbol{S}_r^{(S)})^{-1}\boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)} \end{pmatrix},
\end{aligned}
$$

where

$$
\boldsymbol{S}_r^{(S)} := \boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_r^{(S)} - (\alpha+z)\boldsymbol{I}_r - \boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)} \tag{3.6.17}
$$

is the Schur complement of the lower-right block. We have $\|(\boldsymbol{S}_r^{(S)})^{-1}\| \le \|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\| \prec 1$, $\|\boldsymbol{R}_0^{(S)}\| \prec 1$, and by Assumption 6, $\|\boldsymbol{G}_0^{(S)}\| \prec 1$ and $\|\boldsymbol{G}_r^{(S)}\| \prec 1$. Combining these bounds and using $|\operatorname{Tr}\boldsymbol{A}| \le r\|\boldsymbol{A}\|$ when $\boldsymbol{A}$ has rank r (as follows from the von Neumann trace inequality),

$$
\begin{aligned}
&\left| m_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) - \frac{n-r}{n} m_{\boldsymbol{K}_0}^{(S)} \right| \\
&= \left| \frac{1}{n}\operatorname{Tr}\begin{pmatrix} (\boldsymbol{S}_r^{(S)})^{-1} & -(\boldsymbol{S}_r^{(S)})^{-1}\boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)} \\ -\boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)}(\boldsymbol{S}_r^{(S)})^{-1} & \boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)}(\boldsymbol{S}_r^{(S)})^{-1}\boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)} \end{pmatrix} \right| \\
&\le \frac{1}{n}|\operatorname{Tr}(\boldsymbol{S}_r^{(S)})^{-1}| + \frac{1}{n}|\operatorname{Tr}\boldsymbol{R}_0^{(S)}\boldsymbol{G}_0^{(S)\top}\boldsymbol{G}_r^{(S)}(\boldsymbol{S}_r^{(S)})^{-1}\boldsymbol{G}_r^{(S)\top}\boldsymbol{G}_0^{(S)}\boldsymbol{R}_0^{(S)}| \\
&\le \frac{r}{n}\|(\boldsymbol{S}_r^{(S)})^{-1}\| + \frac{r}{n}\|\boldsymbol{G}_r^{(S)}\|^2\|\boldsymbol{G}_0^{(S)}\|^2\|\boldsymbol{R}_0^{(S)}\|^2\|(\boldsymbol{S}_r^{(S)})^{-1}\| \prec 1/N.
\end{aligned}
$$

Then also $|m_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) - m_{\boldsymbol{K}_0}^{(S)}| \prec 1/N$ since $|m_{\boldsymbol{K}_0}^{(S)}| \le \|\boldsymbol{R}_0^{(S)}\| \prec 1$ and $(n-r)/n = 1 + O_\prec(1/N)$. Hence $|\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) - \tilde{m}_{\boldsymbol{K}_0}^{(S)}| \prec 1/N$ from the definitions $\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) = \gamma_N m_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma}) + (1-\gamma_N)(-1/z)$ and $\tilde{m}_{\boldsymbol{K}_0}^{(S)} = \gamma_{N,0} m_{\boldsymbol{K}_0}^{(S)} + (1-\gamma_{N,0})(-1/z)$, as $|1/z| \le \varepsilon$ for $z \in U(\varepsilon)$ and $\gamma_{N,0} = \gamma_N + O_\prec(1/N)$. The

proof of the second comparison of (3.6.16) is analogous, considering in addition

$$\frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}-\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_0\end{pmatrix}\right)\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})=\frac{1}{n}\operatorname{Tr}\begin{pmatrix}\boldsymbol{\Sigma}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\leq\frac{r}{n}\|\boldsymbol{\Sigma}_r\|\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\|\prec\frac{1}{N}. \quad (3.6.18)$$

Now, Lemma 50 applied with $\boldsymbol{K}_0$ shows, uniformly over $S\subset[N]$ with $|S|\leq L$ and over $z\in U(\varepsilon)$,

$$\mathbf{1}\{|\tilde{m}^{(S)}_{\boldsymbol{K}_0}|>C_0\}\prec 0,\quad \mathbf{1}\{|\tilde{m}^{(S)}_{\boldsymbol{K}_0}|<c_0\}\prec 0,\quad \mathbf{1}\{|1+N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}_0\boldsymbol{R}^{(S)}_0|<c_0\}\prec 0,$$

which together with (3.6.16) implies

$$\mathbf{1}\{|\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})|>C_0\}\prec 0,\quad \mathbf{1}\{|\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})|<c_0\}\prec 0,\quad \mathbf{1}\{|1+N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})|<c_0\}\prec 0$$

for adjusted constants $C_0,c_0>0$. Also by Assumption 6, uniformly over $j\in S$,

$$N^{-1}\boldsymbol{g}_j^\top\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{g}_j-N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\prec N^{-1}\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\|_F\leq N^{-1/2}\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\|\prec N^{-1/2}, \quad (3.6.19)$$

so $\mathbf{1}\{|1+N^{-1}\boldsymbol{g}_j^\top\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{g}_j|<c\}\prec 0$ for a constant $c>0$. Lastly, from the definition of $\boldsymbol{\Gamma}=\boldsymbol{\Gamma}(z,\alpha)$ in (3.6.1), we have

$$z^{-1}\boldsymbol{\Gamma}+\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}=\begin{pmatrix}\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}_r+(\frac{\alpha}{z}+1)\boldsymbol{I}_r & \mathbf{0}\\ \mathbf{0} & \tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}_0+\boldsymbol{I}_{n-r}\end{pmatrix}. \quad (3.6.20)$$

By (3.6.16) and Lemma 50, we have

$$\mathbf{1}\left\{\left\|\left(\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}_0+\boldsymbol{I}_{n-r}\right)^{-1}\right\|>C\right\}\prec 0 \quad (3.6.21)$$

for some constant $C>0$. We have already proved $\mathbf{1}\{|\tilde{m}^{(S)}_{\boldsymbol{K}}(\boldsymbol{\Gamma})|>C_0\}\prec 0$, and applying $\|\boldsymbol{\Sigma}_r\|\leq C$

under Assumption 6, we can deduce for the smallest singular value that

$$\sigma_{\min}\Big(\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}_r + (\alpha/z+1)\boldsymbol{I}_r\Big) \geq \frac{|\alpha|}{|z|} - 1 - |\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})|\|\boldsymbol{\Sigma}_r\| \geq c \tag{3.6.22}$$

on the event $\{|\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})| \leq C_0\}$, for any $z \in U(\varepsilon)$, $|\alpha| \geq \alpha_0$, and some $\alpha_0, c > 0$ depending on ε, C_0. Thus also

$$\mathbf{1}\{\|(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\| > C\} \prec 0 \tag{3.6.23}$$

for a constant $C > 0$, showing all statements of the lemma. □

Proof of Lemma 52(b). Recalling the form of $\mathscr{R}(z,\alpha)$ in (3.6.2), the quantity we wish to approximate is

$$\begin{pmatrix} \boldsymbol{v}_1^\top & \mathbf{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{v}_2 \\ \mathbf{0} \end{pmatrix} = \boldsymbol{v}_1^\top \boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{v}_2 = \boldsymbol{v}_1^\top (\boldsymbol{K} - \boldsymbol{\Gamma})^{-1}\boldsymbol{v}_2.$$

Analogous to (3.5.29) in the proof of Lemma 46, for any matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$, we have

$$\operatorname{Tr}\boldsymbol{B} = \operatorname{Tr}(\boldsymbol{K} - \boldsymbol{\Gamma})\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B} = -\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Gamma} + \frac{1}{N}\sum_{i=1}^{N} \frac{\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{g}_i}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i}. \tag{3.6.24}$$

Applying the definition $m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) = N^{-1}\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma}) + (1-\gamma_N)(-1/z)$ and the identity (3.6.24) with $\boldsymbol{B} = \boldsymbol{I}$, we obtain analogously to (3.5.30) that

$$\begin{aligned} m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) &= (1-\gamma_N)(-1/z) + \frac{1}{Nz}\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{\Gamma} - \frac{1}{N}\operatorname{Tr}(z^{-1}\boldsymbol{\Gamma} - \boldsymbol{I})\boldsymbol{R}(\boldsymbol{\Gamma}) \\ &= -\frac{1}{Nz}\sum_{i=1}^{N} \frac{1}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i} - \frac{1}{N}\operatorname{Tr}(z^{-1}\boldsymbol{\Gamma} - \boldsymbol{I})\boldsymbol{R}(\boldsymbol{\Gamma}). \end{aligned}$$

Then, noting that $z^{-1}\boldsymbol{\Gamma} - \boldsymbol{I}$ has rank r and hence $|N^{-1}\operatorname{Tr}(z^{-1}\boldsymbol{\Gamma} - \boldsymbol{I})\boldsymbol{R}(\boldsymbol{\Gamma})| \leq \frac{r}{N}\frac{|\alpha|}{|z|}\|\boldsymbol{R}(\boldsymbol{\Gamma})\| \prec N^{-1}$,

this gives

$$m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) = -\frac{1}{Nz}\sum_{i=1}^{N}\frac{1}{1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i} + O_\prec\big(N^{-1}\big). \tag{3.6.25}$$

Fixing the unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, let us now choose $\boldsymbol{A} = \boldsymbol{v}_2\boldsymbol{v}_1^\top$ and $\boldsymbol{B} = \boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\cdot\boldsymbol{\Sigma})^{-1}$ in (3.6.24), and define

$$\begin{aligned} d_i &= \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{g}_i - \frac{1}{N}\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Sigma} \\ &= \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}\big(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}\big)^{-1}\boldsymbol{g}_i - \frac{1}{N}\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{A}\big(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma}\big)^{-1}\boldsymbol{\Sigma}. \end{aligned}$$

Then, combining (3.6.24) and (3.6.25), we get

$$\begin{aligned} &\boldsymbol{v}_1^\top(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{v}_2 = \operatorname{Tr}\boldsymbol{B} \\ &= -\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Gamma} + \operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Sigma}\cdot\big(-z m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) + O_\prec\big(N^{-1}\big)\big) + \sum_{i=1}^{N}\frac{d_i}{1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i} \\ &= -z\cdot\boldsymbol{v}_1^\top\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{v}_2 + \sum_{i=1}^{N}\frac{d_i}{1+N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i} + O_\prec\big(N^{-1}\big), \end{aligned} \tag{3.6.26}$$

where the last equality applies the definition of $\boldsymbol{B}$ to combine the first two terms, and applies also $|\operatorname{Tr}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{B}\boldsymbol{\Sigma}| \le \|(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})\| \prec 1$ by Lemma 55 to obtain the $O_\prec\big(N^{-1}\big)$ remainder.

Considering a similar decomposition as in Lemma 46, we define $d_i = d_{i,1} + d_{i,2} + d_{i,3} + d_{i,4}$ where

$$\begin{aligned} d_{i,1} &= \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i - \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i, \\ d_{i,2} &= \frac{1}{N}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}\boldsymbol{g}_i - \frac{1}{N}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}, \\ d_{i,3} &= \frac{1}{N}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1} - \frac{1}{N}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}, \\ d_{i,4} &= \frac{1}{N}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m^{(i)}_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1} - \frac{1}{N}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{A}(z^{-1}\boldsymbol{\Gamma} + m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma})\boldsymbol{\Sigma})^{-1}. \end{aligned} \tag{3.6.27}$$

For $\boldsymbol{A} = \boldsymbol{v}_2\boldsymbol{v}_1^\top$, by the bound $\mathbf{1}\{\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\| > C_0\} \prec 0$ from Lemma 55, we have for a constant

$C > 0$ that

$$\mathbf{1}\Big\{\Big\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\Big\|_F > C\sqrt{N}\Big\} \prec 0, \quad \mathbf{1}\Big\{\Big\|\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})\boldsymbol{A}\Big\|_F > C\Big\} \prec 0 \tag{3.6.28}$$

uniformly over $z \in U(\varepsilon)$. Then, employing Lemma 55 and the same bounds as (3.5.35)–(3.5.38) from the proof of Lemma 50 (where here, the bounds for $\|\boldsymbol{R}^{(i)}\boldsymbol{A}\|_F, \|\boldsymbol{R}\boldsymbol{A}\|_F$ are improved by a factor of $N^{-1/2}$ because $\boldsymbol{A}$ is low-rank), we conclude that $|d_{i,1}|, |d_{i,3}|, |d_{i,4}| \prec N^{-3/2}$ and $|d_{i,2}| \prec N^{-1}$. Hence, applying also $1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i = 1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma}) + O_\prec\Big(N^{-1/2}\Big)$ as follows from (3.6.19) and the bound (3.5.33),

$$\sum_{i=1}^N \frac{d_i}{1 + N^{-1}\boldsymbol{g}_i^\top \boldsymbol{R}^{(i)}(\boldsymbol{\Gamma})\boldsymbol{g}_i} = \frac{1}{1 + N^{-1}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(\boldsymbol{\Gamma})} \cdot \sum_{i=1}^N d_{i,2} + O_\prec\Big(N^{-1/2}\Big).$$

By Lemma 41 applied with $\Psi_N(\boldsymbol{\Gamma}) = C\sqrt{N}$ and $\Phi_N(\boldsymbol{\Gamma}, \boldsymbol{A}) = C$ for a constant $C > 0$, we have $|\sum_i d_{i,2}| \prec N^{-1/2}$. Thus the above quantity is of size $O_\prec\Big(N^{-1/2}\Big)$, so applying this back to (3.6.26),

$$\boldsymbol{v}_1^\top (\boldsymbol{\Gamma} + z m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) \cdot \boldsymbol{\Sigma})^{-1} \boldsymbol{v}_2 + \boldsymbol{v}_1^\top \boldsymbol{R}(\boldsymbol{\Gamma})\boldsymbol{v}_2 \prec N^{-1/2}.$$

Finally, from (3.6.16) and Lemma 54 we have $m_{\tilde{\boldsymbol{K}}}(\boldsymbol{\Gamma}) = \tilde{m}_{N,0}(z) + O_\prec\big(N^{-1}\big)$, and applying this above completes the proof. □

Proof of Lemma 53. The proof is similar to Lemma 52, replacing r and n throughout by $r+1$ and $n+1$, $\boldsymbol{G}_r^{(S)}$ by $[\boldsymbol{u}^{(S)}, \boldsymbol{G}_r^{(S)}]$, $\boldsymbol{\Sigma}$ by $\widetilde{\boldsymbol{\Sigma}}$, and $\boldsymbol{R}^{(S)}(\boldsymbol{\Gamma})$ and $\tilde{m}_{\boldsymbol{K}}^{(S)}(\boldsymbol{\Gamma})$ by

$$\boldsymbol{R}^{(S)}(\widetilde{\boldsymbol{\Gamma}}) = \big([\boldsymbol{u}^{(S)}, \boldsymbol{G}^{(S)}]^\top [\boldsymbol{u}^{(S)}, \boldsymbol{G}^{(S)}] - \widetilde{\boldsymbol{\Gamma}}\big)^{-1}, \quad \tilde{m}_{\boldsymbol{K}}^{(S)}(\widetilde{\boldsymbol{\Gamma}}) = \frac{1}{N}\operatorname{Tr}\boldsymbol{R}^{(S)}(\widetilde{\boldsymbol{\Gamma}}) + \Big(1 - \frac{n+1}{N}\Big)\Big(-\frac{1}{z}\Big).$$

The only difference here is that $\widetilde{\boldsymbol{\Sigma}}$ is no longer diagonal, leading to the following minor modifica-

tions of the preceding proof: The bound

$$\frac{1}{n+1}\operatorname{Tr}\left(\widetilde{\boldsymbol{\Sigma}}-\begin{pmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\boldsymbol{\Sigma}_0\end{pmatrix}\right)\boldsymbol{R}^{(S)}(\widetilde{\boldsymbol{\Gamma}})\prec\frac{1}{N}$$

analogous to (3.6.18) follows upon noting that (with $\mathbb{E}[u\boldsymbol{g}]^\top=\begin{pmatrix}\mathbb{E}[u\boldsymbol{g}_r]^\top & \mathbb{E}[u\boldsymbol{g}_0]^\top\end{pmatrix}$)

$$\widetilde{\boldsymbol{\Sigma}}-\begin{pmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\boldsymbol{\Sigma}_0\end{pmatrix}=\begin{pmatrix}\mathbb{E}[u^2]&\mathbb{E}[u\boldsymbol{g}_r]^\top&\mathbb{E}[u\boldsymbol{g}_0]^\top\\ \mathbb{E}[u\boldsymbol{g}_r]&\boldsymbol{\Sigma}_r&\mathbf{0}\\ \mathbb{E}[u\boldsymbol{g}_0]&\mathbf{0}&\mathbf{0}\end{pmatrix}$$

still is of low rank, with rank at most $r+2$. Writing as shorthand $\tilde{m}_{\boldsymbol{K}}^{(S)}=\tilde{m}_{\boldsymbol{K}}^{(S)}(\widetilde{\boldsymbol{\Gamma}})$, the bound

$$\mathbf{1}\{\|(z^{-1}\widetilde{\boldsymbol{\Gamma}}+\tilde{m}_{\boldsymbol{K}}^{(S)}\widetilde{\boldsymbol{\Sigma}})^{-1}\|>C_0\}\prec 0$$

analogous to (3.6.23) follows from

$$(z^{-1}\widetilde{\boldsymbol{\Gamma}}+\tilde{m}_{\boldsymbol{K}}^{(S)}\widetilde{\boldsymbol{\Sigma}})^{-1}=\begin{pmatrix}\tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u^2]+\frac{\alpha}{z}+1&\tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_r]^\top&\tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_0]^\top\\ \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_r]&\tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma}_r+(\frac{\alpha}{z}+1)\boldsymbol{I}_r&\mathbf{0}\\ \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_0]&\mathbf{0}&\tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma}_0+\boldsymbol{I}\end{pmatrix}^{-1},$$

the bound $\mathbf{1}\{\|\tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma}_0+\boldsymbol{I}\|^{-1}>C\}\prec 0$ for the lower-right block as follows from (3.6.21), and

the bound for the inverse of its Schur-complement

$$\mathbf{1}\Bigg\{\Bigg\|\Bigg[\begin{pmatrix} \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u^2]+\frac{\alpha}{z}+1 & \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_r]^\top \\ \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_r] & \tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma}_r+(\frac{\alpha}{z}+1)\boldsymbol{I}_r \end{pmatrix} \\ - \begin{pmatrix} \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_0]^\top \\ \mathbf{0} \end{pmatrix} (\tilde{m}_{\boldsymbol{K}}^{(S)}\boldsymbol{\Sigma}_0+\boldsymbol{I})^{-1} \begin{pmatrix} \tilde{m}_{\boldsymbol{K}}^{(S)}\mathbb{E}[u\boldsymbol{g}_0] & \mathbf{0} \end{pmatrix}\Bigg]^{-1}\Bigg\| > C\Bigg\} \prec 0$$

which holds uniformly over $z \in U(\varepsilon)$ for any $|\alpha| > \alpha_0$ sufficiently large, by an argument analogous to (3.6.22). The remainder of the proof is identical to that of Lemma 52, and we omit the details. □

3.6.3 Analysis of Outliers

Let $\boldsymbol{V}_r, \boldsymbol{\Gamma}(z,\alpha), \boldsymbol{\mathscr{R}}(z,\alpha), \widetilde{\boldsymbol{\mathscr{R}}}(z,\alpha)$ be as defined in the preceding section. Consider the decomposition of $\widetilde{\boldsymbol{\mathscr{R}}}(z,\alpha)$ as in (3.6.5) into its blocks of dimensions 1, n, and N, and define

$$\widetilde{\mathscr{R}}_{11}(z,\alpha) := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^\top \widetilde{\boldsymbol{\mathscr{R}}}(z,\alpha) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{-z-\alpha+\boldsymbol{u}^\top\boldsymbol{u}-\boldsymbol{u}^\top\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G}-\boldsymbol{\Gamma}(z,\alpha))^{-1}\boldsymbol{G}^\top\boldsymbol{u}}, \tag{3.6.29}$$

$$\widetilde{\boldsymbol{\mathscr{R}}}_{1V}(z,\alpha) := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^\top \widetilde{\boldsymbol{\mathscr{R}}}(z,\alpha) \begin{pmatrix} 0 \\ \boldsymbol{V}_r \\ 0 \end{pmatrix} = -\widetilde{\mathscr{R}}_{11}(z,\alpha)\cdot\boldsymbol{u}^\top\boldsymbol{G}\Big(\boldsymbol{G}^\top\boldsymbol{G}-\boldsymbol{\Gamma}(z,\alpha)\Big)^{-1}\boldsymbol{V}_r, \tag{3.6.30}$$

where the second equality follow from the block matrix inversion of the lower 2×2 blocks of $\widetilde{\mathscr{R}}(z,\alpha)$, followed by block matrix inversion of the full matrix $\widetilde{\mathscr{R}}(z,\alpha)$. Set

$$\boldsymbol{M}_{\boldsymbol{K}}(z,\alpha) = \boldsymbol{I}_r + \alpha \begin{pmatrix} \boldsymbol{V}_r^\top & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix}. \tag{3.6.31}$$

Proposition 56. *Fix any $\varepsilon > 0$ and any $\alpha \in \mathbb{R}$ sufficiently large that satisfies Lemmas 52 and 53. Then on the event $\mathscr{E} \cap \widetilde{\mathscr{E}}$ of Lemmas 52 and 53, for all sufficiently large N,*

(a) $\widehat{\lambda} \in U(\varepsilon) \cap \mathbb{R}$ is an eigenvalue of $\boldsymbol{G}^\top \boldsymbol{G}$ if and only if $\det \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda},\alpha) = 0$, and its multiplicity as an eigenvalue of $\boldsymbol{G}^\top \boldsymbol{G}$ equals the dimension of $\ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda},\alpha)$.

(b) Let $\widehat{\boldsymbol{v}} \in \mathbb{R}^n$ be a unit eigenvector of $\boldsymbol{G}^\top \boldsymbol{G}$ (i.e. right singular vector of $\boldsymbol{G}$) corresponding to an eigenvalue $\widehat{\lambda} \in U(\varepsilon) \cap \mathbb{R}$. Then $\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}$ is a non-zero vector in $\ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda},\alpha)$, and

$$\frac{1}{\alpha^2} = \widehat{\boldsymbol{v}}^\top \boldsymbol{V}_r \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix}^\top \mathscr{R}(\widehat{\lambda},\alpha) \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \mathscr{R}(\widehat{\lambda},\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} \tag{3.6.32}$$

For any vector $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$\boldsymbol{v}^\top \widehat{\boldsymbol{v}} + \alpha \begin{pmatrix} \boldsymbol{v}^\top & \boldsymbol{0} \end{pmatrix} \mathscr{R}(\widehat{\lambda},\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = 0. \tag{3.6.33}$$

(c) Let $\boldsymbol{u}$ be as in Theorem 40(c), and let $\hat{\boldsymbol{u}} \in \mathbb{R}^N$ be a unit eigenvector of $\boldsymbol{G}\boldsymbol{G}^\top$ (i.e. left singular vector of $\boldsymbol{G}$) corresponding to the eigenvalue $\widehat{\lambda} \in U(\varepsilon) \cap \mathbb{R}$. Then

$$\boldsymbol{u}^\top \widehat{\boldsymbol{u}} = \frac{\alpha}{\widehat{\lambda}^{1/2} \widetilde{\mathscr{R}}_{11}(\widehat{\lambda},\alpha)} \widetilde{\mathscr{R}}_{1V}(\widehat{\lambda},\alpha) \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}. \tag{3.6.34}$$

Proof. For part (a), note that if $\widehat{\lambda}$ is an eigenvalue of $\boldsymbol{G}^\top \boldsymbol{G}$, i.e. $\widehat{\lambda}^{1/2}$ is a singular value of $\boldsymbol{G}$

with left and right unit singular vectors $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{v}}$, then

$$0 = \begin{pmatrix} -\widehat{\lambda}\boldsymbol{I}_n & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I}_N \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{v}} \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}} \end{pmatrix}$$

which implies, for any $\alpha \in \mathbb{R}$,

$$-\alpha \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = \begin{pmatrix} -\widehat{\lambda}\boldsymbol{I}_n - \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I}_N \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{v}} \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}} \end{pmatrix}.$$

Fixing $\alpha \in \mathbb{R}$ large enough, on the event $\mathscr{E}$ of Lemma 52, the generalized resolvent

$$\mathscr{R}(\widehat{\lambda}, \alpha) = \begin{pmatrix} -\widehat{\lambda}\boldsymbol{I}_n - \alpha \boldsymbol{V}_r \boldsymbol{V}_r^\top & \boldsymbol{G}^\top \\ \boldsymbol{G} & -\boldsymbol{I}_N \end{pmatrix}^{-1}$$

exists, and multiplying both sides by $\mathscr{R}(\widehat{\lambda}, \alpha)$ gives

$$\begin{pmatrix} \widehat{\boldsymbol{v}} \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}} \end{pmatrix} = -\alpha \mathscr{R}(\widehat{\lambda}, \alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}. \tag{3.6.35}$$

Then, multiplying by $(\boldsymbol{V}_r^\top \ \boldsymbol{0})$ on both sides and re-arranging, we get $\boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}, \alpha) \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = 0$.

We remark that if $\widehat{\lambda}$ is an eigenvalue of multiplicity k, and $\boldsymbol{G}$ has corresponding linearly independent left singular vectors $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_k$ and right singular vectors $\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_k$, then the vectors $\{(\widehat{\boldsymbol{v}}_j, \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}}_j)\}_{j=1}^k$ on the left side of (3.6.35) are linearly independent, implying that the vectors $\{\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_j\}_{j=1}^k$ on the right side must also be (non-zero and) linearly independent vectors in $\ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}, \alpha)$. Conversely, if $\{\boldsymbol{y}_j\}_{j=1}^k$ are linearly independent vectors in $\ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}, \alpha)$, then defining

$$\begin{pmatrix} \widehat{\boldsymbol{v}}_j \\ \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}}_j \end{pmatrix} = -\alpha \mathscr{R}(\widehat{\lambda}, \alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} \cdot \boldsymbol{y}_j$$

and multiplying by $(\boldsymbol{V}_r^\top\ \mathbf{0})$, we must have $\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_j = (-\boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda},\alpha)+\boldsymbol{I})\boldsymbol{y}_j = \boldsymbol{y}_j$. Thus the pairs $(\widehat{\boldsymbol{v}}_j, \widehat{\lambda}^{1/2}\widehat{\boldsymbol{u}}_j)$ are linearly independent vectors satisfying (3.6.35), and multiplying by $\mathscr{R}(\widehat{\lambda},\alpha)^{-1}$ and rearranging shows that $\widehat{\lambda}^{1/2}$ must be a singular value of $\boldsymbol{G}$ with multiplicity at least k, with corresponding singular vectors $\{(\widehat{\boldsymbol{v}}_j, \widehat{\boldsymbol{u}}_j)\}_{j=1}^k$. This establishes part (a).

For part (b), the above argument has shown $\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} \in \ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda},\alpha)$. Multiplying (3.6.35) on the left by

$$\begin{pmatrix} \boldsymbol{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and taking the squared norm (noting that $\widehat{\lambda}, \alpha$ and $\mathscr{R}(\widehat{\lambda},\alpha)$ here are real) shows (3.6.32). Multiplying (3.6.35) on the left by $(\boldsymbol{v}^\top\ \mathbf{0})$ shows (3.6.33). For part (c), multiplying (3.6.35) by $(\mathbf{0}\ \boldsymbol{u}^\top)$, we have

$$\widehat{\lambda}^{1/2}\boldsymbol{u}^\top\widehat{\boldsymbol{u}} = -\alpha \begin{pmatrix} \mathbf{0} \\ \boldsymbol{u} \end{pmatrix}^\top \mathscr{R}(\widehat{\lambda},\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \mathbf{0} \end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}} = -\alpha \boldsymbol{u}^\top \boldsymbol{G}\Big(\boldsymbol{G}^\top\boldsymbol{G} - \boldsymbol{\Gamma}(\widehat{\lambda},\alpha)\Big)^{-1}\boldsymbol{V}_r \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}$$

where the second equality follows from the block matrix inversion of $\mathscr{R}(\widehat{\lambda},\alpha)$. Then, recalling the forms of $\widetilde{\mathscr{R}}_{11}$ and $\widetilde{\mathscr{R}}_{1V}$ from (3.6.29) and (3.6.30), this gives

$$\boldsymbol{u}^\top\widehat{\boldsymbol{u}} = \frac{\alpha\widetilde{\mathscr{R}}_{1V}(\widehat{\lambda},\alpha)}{\widehat{\lambda}^{1/2}\widetilde{\mathscr{R}}_{11}(\widehat{\lambda},\alpha)} \cdot \boldsymbol{V}_r^\top\widehat{\boldsymbol{v}}$$

which is (3.6.34). □

For notational convenience, let us now introduce the shorthand

$$\psi_{N,0}(z) = z\tilde{m}_{N,0}(z), \qquad \psi(z) = z\tilde{m}(z).$$

By Lemma 52(b) applied with $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ being the columns of $\boldsymbol{V}_r$, we see that $\boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)$ is well-

approximated by the (deterministic, N-dependent) matrix

$$\boldsymbol{M}_N(z,\alpha) := \boldsymbol{I}_r - \alpha\Big((\alpha+z)\boldsymbol{I}_r + \psi_{N,0}(z)\operatorname{diag}(\lambda_1(\boldsymbol{\Sigma}),\dots,\lambda_r(\boldsymbol{\Sigma}))\Big)^{-1}. \tag{3.6.36}$$

To show Theorem 40(a), we translate this approximation into a comparison of the roots of $0 = \det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)$ and $0 = \det \boldsymbol{M}_N(z,\alpha)$, where the latter are explicitly given by $z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))$ for the function $z_{N,0}(\cdot)$ defined in (3.3.5).

Proof of Theorem 40(a). Let us fix any $\varepsilon > 0$ and $\alpha \in \mathbb{R}$ satisfying Lemmas 52 and 53, and denote

$$f_{N,i}(z,\alpha) = 1 - \frac{\alpha}{\alpha + z + \psi_{N,0}(z)\lambda_i(\boldsymbol{\Sigma})}$$

for each $i \in [r]$. Then $\det \boldsymbol{M}_N(z,\alpha) = \prod_{i=1}^r f_{N,i}(z,\alpha)$. Define also the limiting functions

$$f_i(z,\alpha) = 1 - \frac{\alpha}{\alpha + z + \psi(z)\lambda_i}, \qquad \boldsymbol{M}(z,\alpha) = \boldsymbol{I}_r - \alpha\Big((\alpha+z)\boldsymbol{I}_r + \psi(z)\operatorname{diag}(\lambda_1,\dots,\lambda_r)\Big)^{-1}$$

so $\det \boldsymbol{M}(z,\alpha) = \prod_{i=1}^r f_i(z,\alpha)$. We first analyze the roots of $0 = \det \boldsymbol{M}(z,\alpha)$: By the definition $\psi(z) = z\tilde{m}(z)$, observe that $z \in \mathbb{R} \setminus \operatorname{supp}(\tilde{\mu})$ satisfies $0 = \det \boldsymbol{M}(z,\alpha)$ if and only if either $z = 0$ or

$$\tilde{m}(z) = -1/\lambda_i \text{ for some } i \in [r].$$

(This condition is the same for any non-zero $\alpha \in \mathbb{R}$.) Let $\mathscr{T} = \{0\} \cup \{-1/\lambda : \lambda \in \operatorname{supp}(\nu)\}$ be as in (1.2.5) where ν is the limit spectral law of $\boldsymbol{\Sigma}_0$. Then $-1/\lambda_i \in \mathbb{R} \setminus \mathscr{T}$ for all $i \in [r]$ under Assumption 7, so Proposition 3 implies that $\tilde{m}(z) = -1/\lambda_i$ holds for some $z \in \mathbb{R} \setminus \operatorname{supp}(\tilde{\mu})$ if and only if $z'(-1/\lambda_i) > 0$, i.e. $i \in \mathscr{I}$. If $i \in \mathscr{I}$, then $\tilde{m}(z) = -1/\lambda_i$ holds for $z = z(-1/\lambda_i)$, and we must have $z(-1/\lambda_i) > 0$ strictly because for any $z \le 0$, we have $\tilde{m}(z) > 0$ (and hence $\tilde{m}(z) \neq -1/\lambda_i$) by the definition $\tilde{m}(z) = \int \frac{1}{x-z} d\tilde{\mu}(x)$. Thus the roots of $0 = \det \boldsymbol{M}(z,\alpha)$ in $\mathbb{R} \setminus \mathscr{S} = \mathbb{R} \setminus (\operatorname{supp}(\tilde{\mu}) \cup \{0\})$ — and hence in $U(\varepsilon) \cap \mathbb{R}$ for any sufficiently small $\varepsilon > 0$ — are

given precisely by

$$z_i := z(-1/\lambda_i) \text{ for } i \in \mathscr{I}.$$

Since $\tilde{m}'(z) = \int \frac{1}{(x-z)^2} d\tilde{\mu}(x) > 0$ for all $z \in \mathbb{R} \setminus \mathscr{S}$, and $\{\lambda_i : i \in \mathscr{I}\}$ are distinct by assumption, these values $\{z_i : i \in \mathscr{I}\}$ are simple roots of $0 = \det \boldsymbol{M}(z_i, \alpha)$. Then $(\det \boldsymbol{M})'(z_i, \alpha) \neq 0$ where $(\det \boldsymbol{M})'$ denotes the derivative in z.

Lemma 51(c) implies $\tilde{m}_{N,0}(z) \to \tilde{m}(z)$ and $\tilde{m}'_{N,0}(z) \to \tilde{m}'(z)$ uniformly over $z \in U(\varepsilon)$. Since also $\lambda_i(\boldsymbol{\Sigma}) \to \lambda_i$, we have $\det \boldsymbol{M}_N(z, \alpha) \to \det \boldsymbol{M}(z, \alpha)$ and $(\det \boldsymbol{M}_N)'(z, \alpha) \to (\det \boldsymbol{M})'(z, \alpha)$ uniformly over $z \in U(\varepsilon)$. This, together with the above condition $(\det \boldsymbol{M})'(z_i, \alpha) \neq 0$, imply that for all large N, the roots $z_{N,i} \in U(\varepsilon) \cap \mathbb{R}$ of $0 = \det \boldsymbol{M}_N(z, \alpha)$ are in 1-to-1 correspondence with, and converge to, the above roots $z_i \in U(\varepsilon) \cap \mathbb{R}$ of $0 = \det \boldsymbol{M}(z, \alpha)$. We note that $0 = \det \boldsymbol{M}_N(z, \alpha)$ if and only if either $z = 0$ or

$$\tilde{m}_{N,0}(z) = -1/\lambda_i(\boldsymbol{\Sigma}) \text{ for some } i \in [r]. \tag{3.6.37}$$

For each $i \in \mathscr{I}$, we have $\lambda_i(\boldsymbol{\Sigma}) \to \lambda_i$ where $z'(-1/\lambda_i) > 0$. Recall from Lemma 51(a) that $z_{N,0}(\tilde{m}) \to z(\tilde{m})$ and

$$z'_{N,0}(\tilde{m}) \to z'(\tilde{m})$$

uniformly over compact subsets of $\mathbb{R} \setminus \mathscr{T}$. Then $z'_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) \to z'(-1/\lambda_i)$, so also

$$z'_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) > 0$$

for all large N. Then Proposition 3 implies that (3.6.37) holds for $z_{N,i} := z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))$. We have $z_{N,i} \to z_i = z(-1/\lambda_i)$, so these must be the roots of $\det \boldsymbol{M}_N(z, \alpha)$ in $U(\varepsilon) \cap \mathbb{R}$. Thus we have shown that for any sufficiently small $\varepsilon > 0$ and all large N, the roots $z \in U(\varepsilon) \cap \mathbb{R}$ of

$0 = \det \boldsymbol{M}_N(z,\alpha)$ are precisely the values

$$z_{N,i} := z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) \text{ for } i \in \mathscr{I},$$

and $z_{N,i} \to z_i > 0$ for each $i \in \mathscr{I}$.

Finally, we apply Lemma 52(b) with $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ being the columns of $\boldsymbol{V}_r$. On the event $\mathscr{E}$ of Lemma 52(a), we have

$$\|\boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)\| \le C, \qquad \left\|\boldsymbol{M}_{\boldsymbol{K}}(z,\alpha) - \boldsymbol{M}_{\boldsymbol{K}}(z',\alpha)\right\| \le C|z-z'| \tag{3.6.38}$$

for some $C > 0$ and all $z, z' \in U(\varepsilon/2)$. Also $|\tilde{m}_{N,0}(z)|, |\tilde{m}'_{N,0}(z)| < C$ for a constant $C > 0$, all $z \in U(\varepsilon)$, and all large N, and thus

$$\|\boldsymbol{M}_N(z,\alpha)\| \le C, \qquad \left\|\boldsymbol{M}_N(z,\alpha) - \boldsymbol{M}_N(z',\alpha)\right\| \le C|z-z'| \tag{3.6.39}$$

for some $C > 0$ and all $z, z' \in U(\varepsilon/2)$. Then, applying Lemma 52(b) and the Lipschitz bounds of (3.6.38) and (3.6.39) to take a union bound over a sufficiently fine covering net of $U(\varepsilon/2)$, we get

$$\sup_{z \in U(\varepsilon/2)} \|\boldsymbol{M}_N(z,\alpha) - \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)\| \prec 1/\sqrt{N}. \tag{3.6.40}$$

Applying also the first bounds of (3.6.38) and (3.6.39), this gives

$$\sup_{z \in U(\varepsilon/2)} |\det \boldsymbol{M}_N(z,\alpha) - \det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)| \prec 1/\sqrt{N}. \tag{3.6.41}$$

Since $\det \boldsymbol{M}_N(z,\alpha)$ and $\det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)$ are both holomorphic over $z \in U(\varepsilon/2)$ on this event $\mathscr{E}$, the Cauchy integral formula then implies

$$\sup_{z \in U(\varepsilon)} |(\det \boldsymbol{M}_N)'(z,\alpha) - (\det \boldsymbol{M}_{\boldsymbol{K}})'(z,\alpha)| \prec 1/\sqrt{N}.$$

In particular, combining with the uniform convergence statements $\det \boldsymbol{M}_N(z,\alpha) \to \det \boldsymbol{M}(z,\alpha)$ and $(\det \boldsymbol{M}_N)'(z,\alpha) \to (\det \boldsymbol{M})'(z,\alpha)$ over $z \in U(\varepsilon)$ as argued above, this shows that on an event $\mathscr{E}$ satisfying $\mathbf{1}\{\mathscr{E}^c\} \prec 0$ and for some $\delta_N \to 0$, we have

$$\sup_{z \in U(\varepsilon) \cap \mathbb{R}} |\det \boldsymbol{M}(z,\alpha) - \det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)|, |(\det \boldsymbol{M})'(z,\alpha) - (\det \boldsymbol{M}_{\boldsymbol{K}})'(z,\alpha)| < \delta_N.$$

Thus, on this event $\mathscr{E}$ and as $N \to \infty$, the roots $\widehat{\lambda}_i \in U(\varepsilon) \cap \mathbb{R}$ of $0 = \det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)$ are also in 1-to-1 correspondence with, and converge to, the roots $z_i \in U(\varepsilon) \cap \mathbb{R}$ of $0 = \det \boldsymbol{M}(z,\alpha)$. Furthermore, the condition $(\det \boldsymbol{M})'(z_i,\alpha) \neq 0$ implies that $|(\det \boldsymbol{M}_N)'(z,\alpha)|$ and $|(\det \boldsymbol{M}_{\boldsymbol{K}})'(z,\alpha)|$ are bounded away from 0 in a neighborhood of each such root z_i, so (3.6.41) then implies that the corresponding roots $\widehat{\lambda}_i$ and $z_{N,i}$ of $0 = \det \boldsymbol{M}_{\boldsymbol{K}}(z,\alpha)$ and $0 = \det \boldsymbol{M}_N(z,\alpha)$ satisfy

$$|\widehat{\lambda}_i - z_{N,i}| \prec 1/\sqrt{N}.$$

Proposition 56 shows that on this event $\mathscr{E}$, these roots $\{\widehat{\lambda}_i : i \in \mathscr{I}\}$ are precisely the eigenvalues of $\boldsymbol{G}^\top \boldsymbol{G}$ in $U(\varepsilon) \cap \mathbb{R}$. By the definition of $\boldsymbol{M}(z_i,\alpha)$, each root z_i of $\det \boldsymbol{M}(z_i,\alpha)$ is such that $\ker \boldsymbol{M}(z_i,\alpha)$ has dimension 1. Since $\mathbf{1}\{\mathscr{E}\}(\widehat{\lambda}_i - z_i) \to 0$, we have $\mathbf{1}\{\mathscr{E}\}\|\boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i,\alpha) - \boldsymbol{M}(z_i,\alpha)\| \to 0$, so $\ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i,\alpha)$ also has dimension 1 on this event $\mathscr{E}$ for all large N. Then Proposition 56 implies that the eigenvalues $\{\widehat{\lambda}_i : i \in \mathscr{I}\}$ of $\boldsymbol{G}^\top \boldsymbol{G}$ are simple, and thus in 1-to-1 correspondence with $\{\lambda_i : i \in \mathscr{I}\}$. This proves part (a) of the theorem. □

Lemma 57. *Under the assumptions of Theorem 40, for any fixed $\varepsilon > 0$, there exists $\alpha_0 > 0$ such*

that fixing any $\alpha \in \mathbb{C}$ with $|\alpha| > \alpha_0$, uniformly over $z \in U(\varepsilon)$,

$$\left\| \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix}^\top \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix} - \left((\alpha+z)\boldsymbol{I}_r + \psi_{N,0}(z)\boldsymbol{\Sigma}_r\right)^{-2}\left(\boldsymbol{I}_r + \psi'_{N,0}(z)\boldsymbol{\Sigma}_r\right) \right\| \prec \frac{1}{\sqrt{N}}.$$

Proof. Fix any $\alpha \in \mathbb{C}$ satisfying Lemma 52, and denote

$$f_N(z,\alpha) := \begin{pmatrix} \boldsymbol{V}_r^\top & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0} \end{pmatrix}, \qquad g_N(z,\alpha) := -\left((\alpha+z)\boldsymbol{I}_r + \psi_{N,0}(z)\boldsymbol{\Sigma}_r\right)^{-1}.$$

Applying Lemma 52(b) and the Lipschitz continuity statements of (3.6.38) and (3.6.39) to take a union bound over a sufficiently fine covering net of $U(\varepsilon/2)$, we have

$$\sup_{z \in U(\varepsilon/2)} \|f_N(z,\alpha) - g_N(z,\alpha)\| \prec 1/\sqrt{N}.$$

Then by the Cauchy integral formula, $\sup_{z\in U(\varepsilon)} \|f'_N(z,\alpha) - g'_N(z,\alpha)\| \prec 1/\sqrt{N}$ where f'_N and g'_N denote the entrywise derivatives in z. The lemma follows, since differentiating $\mathscr{R}(z,\alpha)$ in (3.6.2) shows

$$f'_N(z,\alpha) = \begin{pmatrix} \boldsymbol{V}_r^\top & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \mathscr{R}(z,\alpha) \begin{pmatrix} \boldsymbol{V}_r \\ \boldsymbol{0}, \end{pmatrix}$$

while $g'_N(z,\alpha) = ((\alpha+z)\boldsymbol{I}_r + \psi_{N,0}(z)\boldsymbol{\Sigma}_r)^{-2}(\boldsymbol{I}_r + \psi'_{N,0}(z)\boldsymbol{\Sigma}_r)$. □

Proof of Theorem 40(b). Let $\widehat{\boldsymbol{v}}_i$ be the given unit-norm eigenvector of $\boldsymbol{K}$ with eigenvalue $\widehat{\lambda}_i$. Let $z_{N,i} = z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))$ and $z_i = z(-1/\lambda_i)$. Then, fixing any $\alpha \in \mathbb{R}$ large enough to satisfy Lemmas 52 and 53, Proposition 56(b) shows that $\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i \in \ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i, \alpha)$. By (3.6.40), (3.6.39),

and the bound $|\widehat{\lambda}_i - z_{N,i}| \prec N^{-1/2}$ of part (a) of the theorem already proven, we have

$$\begin{aligned}\left\|\boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i,\alpha) - \boldsymbol{M}_N(z_{N,i},\alpha)\right\| &\leq \left\|\boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i,\alpha) - \boldsymbol{M}_N(\widehat{\lambda}_i,\alpha)\right\| + \left\|\boldsymbol{M}_N(\widehat{\lambda}_i,\alpha) - \boldsymbol{M}_N(z_{N,i},\alpha)\right\| \\ &\prec N^{-1/2}. \end{aligned} \tag{3.6.42}$$

Let $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r$ denote the columns of $\boldsymbol{V}_r$, which are the unit eigenvectors of $\boldsymbol{\Sigma}$. Then, applying $\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i \in \ker \boldsymbol{M}_{\boldsymbol{K}}(\widehat{\lambda}_i,\alpha)$, (3.6.42), and the definition of $\boldsymbol{M}_N(z,\alpha)$, and noting that $\psi_{N,0}(z_{N,i}) = z_{N,i}\tilde{m}_{N,0}(z_{N,i}) = -z_{N,i}/\lambda_i(\boldsymbol{\Sigma})$, we have

$$\left\|\boldsymbol{M}_N(z_{N,i},\alpha)\cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i\right\|^2 = \sum_{j=1}^r \left(1 - \frac{\alpha}{\alpha + z_{N,i}(1-\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma}))}\right)^2 (\boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i)^2 \prec 1/N.$$

For each $j \in [r]\setminus\{i\}$, we have that $z_{N,i}(1-\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma}))$ is bounded away from 0 as $N\to\infty$ because $z_{N,i}\to z_i > 0$ and $\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma}) \to \lambda_j/\lambda_i \neq 1$. So this implies

$$|\boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i|^2 \prec 1/N \quad \text{ for all } \quad j \in [r]\setminus\{i\}. \tag{3.6.43}$$

At the same time, applying Lemma 57 and $|\widehat{\lambda}_i - z_{N,i}| \prec N^{-1/2}$ to bound (3.6.32) in Proposition 56(b), we have

$$\begin{aligned}\frac{1}{\alpha^2} &= \widehat{\boldsymbol{v}}_i^\top \boldsymbol{V}_r \begin{pmatrix}\boldsymbol{V}_r \\ \boldsymbol{0}\end{pmatrix}^\top \boldsymbol{\mathscr{R}}(z_{N,i},\alpha)\begin{pmatrix}\boldsymbol{I}_n & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{\mathscr{R}}(z_{N,i},\alpha)\begin{pmatrix}\boldsymbol{V}_r \\ \boldsymbol{0}\end{pmatrix}\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i + O_\prec\Big(N^{-1/2}\Big) \\ &= \widehat{\boldsymbol{v}}_i^\top \boldsymbol{V}_r \Big((\alpha+z_{N,i})\boldsymbol{I}_r + \psi_{N,0}(z_{N,i})\boldsymbol{\Sigma}_r\Big)^{-2}\Big(\boldsymbol{I}_r + \psi'_{N,0}(z_{N,i})\boldsymbol{\Sigma}_r\Big)\boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i + O_\prec\Big(N^{-1/2}\Big) \\ &= |\boldsymbol{v}_i^\top \widehat{\boldsymbol{v}}_i|^2 \cdot \frac{1+\psi'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma})}{\alpha^2} \\ &\qquad + \sum_{j\neq i} |\boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i|^2 \cdot \frac{1+\psi'_{N,0}(z_{N,i})\lambda_j(\boldsymbol{\Sigma})}{(\alpha + z_{N,i}(1-\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma})))^2} + O_\prec\Big(N^{-1/2}\Big) \\ &= |\boldsymbol{v}_i^\top \widehat{\boldsymbol{v}}_i|^2 \cdot \frac{1+\psi'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma})}{\alpha^2} + O_\prec\Big(N^{-1/2}\Big), \end{aligned} \tag{3.6.44}$$

the last equality applying (3.6.43). Observe that

$$\begin{aligned}1+\psi'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma}) &= 1+z_{N,i}\tilde{m}'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma})+\tilde{m}_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma})\\ &= z_{N,i}\tilde{m}'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma}) = z_{N,i}\lambda_i(\boldsymbol{\Sigma})/z'_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})),\end{aligned}$$

where the last two equalities use $z_{N,i} = z_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))$ and $\tilde{m}_{N,0}(\cdot)$ is the inverse function of $z_{N,0}(\cdot)$. Then, multiplying by $\alpha^2/(1+\psi'_{N,0}(z_{N,i})\lambda_i(\boldsymbol{\Sigma}))$ we obtain

$$|\boldsymbol{v}_i^\top \widehat{\boldsymbol{v}}_i|^2 = \frac{z'_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))}{z_{N,i}\lambda_i(\boldsymbol{\Sigma})} + O_\prec\Big(N^{-1/2}\Big) = \varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) + O_\prec\Big(N^{-1/2}\Big),$$

where we recall $\varphi_{N,0}$ from (3.3.5). We have $\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma})) \to \varphi(-1/\lambda_i) = z'(-1/\lambda_i)/(\lambda_i z_i) > 0$, so taking a square root gives

$$|\boldsymbol{v}_i^\top \widehat{\boldsymbol{v}}_i| = \sqrt{\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))} + O_\prec\Big(N^{-1/2}\Big). \tag{3.6.45}$$

Finally, for any unit vector $\boldsymbol{v} \in \mathbb{R}^n$, by (3.6.33) in Proposition 56(b), Lemma 52(b), and the bound $|\widehat{\lambda}_i - z_{N,i}| \prec N^{-1/2}$ in part (a) of the theorem already shown, we know that

$$\begin{aligned}\boldsymbol{v}^\top \widehat{\boldsymbol{v}}_i &= -\alpha \cdot \begin{pmatrix}\boldsymbol{v}^\top & \mathbf{0}\end{pmatrix} \mathscr{R}(z_{N,i},\alpha) \begin{pmatrix}\boldsymbol{V}_r \\ \mathbf{0}\end{pmatrix} \cdot \boldsymbol{V}_r^\top \widehat{\boldsymbol{v}}_i + O_\prec\Big(N^{-1/2}\Big)\\ &= -\alpha \sum_{j=1}^r \frac{\boldsymbol{v}^\top \boldsymbol{v}_j \cdot \boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i}{\alpha + z_{N,i} + \psi_{N,0}(z_{N,i})\lambda_j(\boldsymbol{\Sigma})} + O_\prec\Big(N^{-1/2}\Big)\\ &= -\alpha \sum_{j=1}^r \frac{\boldsymbol{v}^\top \boldsymbol{v}_j \cdot \boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i}{\alpha + z_{N,i}\cdot(1-\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma}))} + O_\prec\Big(N^{-1/2}\Big).\end{aligned}$$

Applying (3.6.43) and (3.6.45), only the summand with $j = i$ contributes, and we obtain as desired

$$|\boldsymbol{v}^\top \widehat{\boldsymbol{v}}_i| = \sqrt{\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))} \cdot |\boldsymbol{v}^\top \boldsymbol{v}_i| + O_\prec\Big(N^{-1/2}\Big).$$

□

Proof of Theorem 40(c). Applying Lemma 53(b) and block matrix inversion of $\widetilde{\boldsymbol{\Gamma}}+\psi_{N,0}(z)\widetilde{\boldsymbol{\Sigma}}$ to the definitions of $\widetilde{\mathscr{R}}_{11}$ and $\widetilde{\boldsymbol{\mathscr{R}}}_{1V}$ in (3.6.29) and (3.6.30), we have

$$\left|\widetilde{\mathscr{R}}_{11}(z,\alpha)+\Big(z+\alpha+\psi_{N,0}(z)\cdot\mathbb{E}[u^2]-\psi_{N,0}(z)^2\cdot\mathbb{E}[u\boldsymbol{g}]^\top(\boldsymbol{\Gamma}+\psi_{N,0}(z)\boldsymbol{\Sigma})^{-1}\mathbb{E}[u\boldsymbol{g}]\Big)^{-1}\right|\prec\frac{1}{\sqrt{N}},$$
$$\left\|\widetilde{\boldsymbol{\mathscr{R}}}_{1V}(z,\alpha)-\frac{\psi_{N,0}(z)\cdot\mathbb{E}[u\boldsymbol{g}]^\top(\boldsymbol{\Gamma}+\psi_{N,0}(z)\boldsymbol{\Sigma})^{-1}\boldsymbol{V}_r}{z+\alpha+\psi_{N,0}(z)\cdot\mathbb{E}[u^2]-\psi_{N,0}(z)^2\cdot\mathbb{E}[u\boldsymbol{g}]^\top(\boldsymbol{\Gamma}+\psi_{N,0}(z)\boldsymbol{\Sigma})^{-1}\mathbb{E}[u\boldsymbol{g}]}\right\|\prec\frac{1}{\sqrt{N}}.$$

Hence,

$$\left\|\frac{\widetilde{\boldsymbol{\mathscr{R}}}_{1V}(z,\alpha)}{\widetilde{\mathscr{R}}_{11}(z,\alpha)}+\psi_{N,0}(z)\cdot\mathbb{E}[u\boldsymbol{g}]^\top\boldsymbol{V}_r\cdot((\alpha+z)\boldsymbol{I}_r+\psi_{N,0}(z)\boldsymbol{\Sigma}_r)^{-1}\right\|\prec\frac{1}{\sqrt{N}}.$$

Applying this and the bound $|\widehat{\lambda}_i-z_{N,i}|\prec N^{-1/2}$ to Proposition 56(c),

$$\begin{aligned}\boldsymbol{u}^\top\widehat{\boldsymbol{u}}_i=\frac{\alpha}{\widehat{\lambda}_i^{1/2}}\frac{\widetilde{\boldsymbol{\mathscr{R}}}_{1V}(\widehat{\lambda}_i,\alpha)}{\widetilde{\mathscr{R}}_{11}(\widehat{\lambda}_i,\alpha)}\cdot\boldsymbol{V}_r^\top\widehat{\boldsymbol{v}}_i&=-\frac{\alpha}{\sqrt{z_{N,i}}}\sum_{j=1}^r\frac{\psi_{N,0}(z_{N,i})\cdot\mathbb{E}[u\boldsymbol{g}]^\top\boldsymbol{v}_j\cdot\boldsymbol{v}_j^\top\widehat{\boldsymbol{v}}_i}{\alpha+z_{N,i}+\psi_{N,0}(z_{N,i})\lambda_j(\boldsymbol{\Sigma})}+O_\prec\Big(N^{-1/2}\Big)\\&=-\frac{\alpha}{\sqrt{z_{N,i}}}\sum_{j=1}^r\frac{\psi_{N,0}(z_{N,i})\cdot\mathbb{E}[u\boldsymbol{g}]^\top\boldsymbol{v}_j\cdot\boldsymbol{v}_j^\top\widehat{\boldsymbol{v}}_i}{\alpha+z_{N,i}(1-\lambda_j(\boldsymbol{\Sigma})/\lambda_i(\boldsymbol{\Sigma}))}+O_\prec\Big(N^{-1/2}\Big).\end{aligned}$$

Then, applying again (3.6.43) and (3.6.45), only the summand with $j=i$ contributes, and this gives

$$|\boldsymbol{u}^\top\widehat{\boldsymbol{u}}_i|=\frac{|\mathbb{E}[u\boldsymbol{g}]^\top\boldsymbol{v}_i|\cdot|\psi_{N,0}(z_{N,i})|\sqrt{\varphi_{N,0}(-1/\lambda_i(\boldsymbol{\Sigma}))}}{\sqrt{z_{N,i}}}+O_\prec\Big(N^{-1/2}\Big).$$

Recalling $\psi_{N,0}(z_{N,i})=-z_{N,i}/\lambda_i(\boldsymbol{\Sigma})$, this yields part (c) of the theorem. □

3.7 Proofs for Propagation of Spiked Eigenstructure in Deep NNs

We next prove Theorems 33 and 34. Section 3.7.1 first establishes these results for a one-hidden-layer NN, $L = 1$. We then apply this result for $L = 1$ inductively in Section 3.7.2 to obtain these results for general L. Section 3.7.3 proves Corollary 35.

3.7.1 Spike Analysis for One-hidden-layer CK

Consider the setup in Section 3.2 with a single hidden layer $L = 1$. In this setting, let us simplify notation and denote

$$\boldsymbol{X} = \boldsymbol{X}_0, \qquad \boldsymbol{W} = \boldsymbol{W}_0, \qquad d = d_0, \qquad N = d_1,$$

$$\boldsymbol{Y} = \boldsymbol{X}_1 = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{X}), \qquad \boldsymbol{K} = \boldsymbol{K}_1 = \boldsymbol{Y}^\top \boldsymbol{Y}.$$

We denote the rows of $\boldsymbol{W}$ and columns of $\boldsymbol{X}$ respectively by

$$\boldsymbol{w}_i^\top \in \mathbb{R}^d \text{ for } i \in [N], \qquad \boldsymbol{x}_\alpha \in \mathbb{R}^d \text{ for } \alpha \in [n].$$

We write $\mathbb{E}_{\boldsymbol{w}}$ for the expectation over a standard Gaussian vector $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I})$ in $\mathbb{R}^d$.

Note that for a sufficiently large constant $B > 0$ (depending on $\mathrm{supp}(\nu)$ and $\lambda_1, \ldots, \lambda_r$), Assumption 4 implies that the event

$$\mathscr{E}(\boldsymbol{X}) = \Big\{ \|\boldsymbol{X}\| < B,\ |\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta| < \tau_n \text{ and } |\|\boldsymbol{x}_\alpha\| - 1| < \tau_n \text{ for all } \alpha \neq \beta \in [n] \Big\} \tag{3.7.1}$$

holds almost surely for all large n. Throughout this section, we use the following argument: Since $\boldsymbol{W} \equiv \boldsymbol{W}^{(n)}$ is independent of $\boldsymbol{X} \equiv \boldsymbol{X}^{(n)}$, and $\mathscr{E}(\boldsymbol{X}^{(n)})$ holds for all large n with probability one over $\{\boldsymbol{X}^{(n)}\}_{n=1}^\infty$, to prove any almost-sure statement, it suffices to show that the statement holds with probability one over $\{\boldsymbol{W}^{(n)}\}_{n=1}^\infty$, for any deterministic matrices $\{\boldsymbol{X}^{(n)}\}_{n=1}^\infty$ satisfying

$\mathscr{E}(\boldsymbol{X}^{(n)})$. Therefore, we assume in the remainder of this section that $\boldsymbol{X}$ is deterministic and satisfies $\mathscr{E}(\boldsymbol{X})$ for all large n, and write $\mathbb{E}$ and $\mathbb{P}$ for the expectation and probability over *only* the random weight matrix $\boldsymbol{W}$, respectively.

We will apply Theorem 40 to a centered version of $\boldsymbol{Y}$,

$$\boldsymbol{G} := \boldsymbol{Y} - \mathbb{E}\boldsymbol{Y} = \frac{1}{\sqrt{N}}[\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N]^\top, \qquad \boldsymbol{g}_i^\top := \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}) - \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})].$$

Note that these rows $\boldsymbol{g}_i^\top$ are i.i.d. with mean $\boldsymbol{0}$ and covariance

$$\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})] - \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})]^\top \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})] \in \mathbb{R}^{n \times n}. \tag{3.7.2}$$

Lemma 58. *Suppose Assumptions 3, 4, and 5 hold, with $L = 1$ and deterministic $\boldsymbol{X}$. Then*

$$\|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})]\| \to 0, \qquad \|\mathbb{E}\boldsymbol{Y}\| \to 0.$$

Proof. Denote $\xi \sim \mathcal{N}(0,1)$. Applying $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma'(\xi)\xi] = \mathbb{E}[\sigma''(\xi)] = 0$, and a Taylor approximation of σ, for any $\alpha \in [n]$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_\alpha)] &= \mathbb{E}[\sigma(\|\boldsymbol{x}_\alpha\|\xi)] - \mathbb{E}[\sigma(\xi)] \\ &= \mathbb{E}[\sigma'(\xi)\xi(\|\boldsymbol{x}_\alpha\| - 1)] + \mathbb{E}[\sigma''(\eta)\xi^2(\|\boldsymbol{x}_\alpha\| - 1)^2] = \mathbb{E}[\sigma''(\eta)\xi^2(\|\boldsymbol{x}_\alpha\| - 1)^2]\end{aligned}$$

for some η between ξ and $\|\boldsymbol{x}_i\|\xi$. Then, applying $|\sigma''(x)| \le \lambda_\sigma$ and the τ_n-orthonormality of $\boldsymbol{X}$ under $\mathscr{E}(\boldsymbol{X})$,

$$|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_\alpha)]| \le \lambda_\sigma \tau_n^2.$$

This gives $\|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})]\| \le \lambda_\sigma \tau_n^2 \sqrt{n} \to 0$, so also $\|\mathbb{E}\boldsymbol{Y}\| = \left\| \frac{1}{\sqrt{N}} \boldsymbol{1}_N \cdot \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})] \right\| \to 0$. □

Next, we utilize Lemma 84 in Chapter 4 to derive an approximation of $\boldsymbol{\Sigma}$ by the linearized

matrix

$$\boldsymbol{\Sigma}_{\text{lin}} := b_\sigma^2 \boldsymbol{X}^\top \boldsymbol{X} + (1 - b_\sigma^2) \boldsymbol{I}_n \tag{3.7.3}$$

in the operator norm.

Lemma 59. *Suppose Assumptions 3, 4, and 5 hold, with $L = 1$ and deterministic $\boldsymbol{X}$.*

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{lin}}\| \to 0.$$

Consequently, ordering $\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_n(\boldsymbol{\Sigma})$ in the same order as $\lambda_1(\boldsymbol{X}^\top \boldsymbol{X}), \ldots, \lambda_n(\boldsymbol{X}^\top \boldsymbol{X})$,

$$\sup_{i \in [n]} \left| b_\sigma^2 \lambda_i(\boldsymbol{X}^\top \boldsymbol{X}) + (1 - b_\sigma^2) - \lambda_i(\boldsymbol{\Sigma}) \right| \to 0. \tag{3.7.4}$$

Proof. Denote $\xi \sim \mathcal{N}(0,1)$. Let $\zeta_k(\sigma) = \mathbb{E}[\sigma(\xi) h_k(\xi)]$ be the k-th Hermite coefficient of σ, where $h_k(x)$ is the k-th Hermite polynomial normalized so that $\mathbb{E}[h_k(\xi)^2] = 1$. Note that by Gaussian integration by parts and the assumption $\mathbb{E}[\sigma''(\xi)] = 0$,

$$\zeta_1(\sigma) = \mathbb{E}[\xi \sigma(\xi)] = \mathbb{E}[\sigma'(\xi)] = b_\sigma, \tag{3.7.5}$$

$$\sqrt{2}\, \zeta_2(\sigma) = \mathbb{E}[(\xi^2 - 1)\sigma(\xi)] = \mathbb{E}[\xi \sigma'(\xi)] = \mathbb{E}[\sigma''(\xi)] = 0. \tag{3.7.6}$$

Then by Lemma 84 in Chapter 4 and the first statement of Lemma 58, we have

$$\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}\| \le \|\boldsymbol{\Sigma}_0 - \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})]\| + \|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})] - \boldsymbol{\Sigma}\| \to 0$$

where

$$\boldsymbol{\Sigma}_0 = \zeta_1(\sigma)^2 \boldsymbol{X}^\top \boldsymbol{X} + \zeta_3(\sigma)^2 (\boldsymbol{X}^\top \boldsymbol{X})^{\odot 3} + (1 - \zeta_1(\sigma)^2 - \zeta_3(\sigma)^2) \boldsymbol{I}_n.$$

(Here, examination of the proof of Lemma 84 shows that the condition $\sum_\alpha (\|\boldsymbol{x}_\alpha\| - 1)^2 \le B^2$ for (ε, B)-orthonormality is not used when $\zeta_2(\sigma) = 0$, and the remaining conditions of (ε, B)-

orthonormality hold under $\mathscr{E}(\boldsymbol{X})$.) The lemma then follows upon observing that under $\mathscr{E}(\boldsymbol{X})$,

$$\begin{aligned}\left\|(\boldsymbol{X}^\top\boldsymbol{X})^{\odot 3}-\boldsymbol{I}_n\right\| &\le \left\|\mathrm{diag}((\boldsymbol{X}^\top\boldsymbol{X})^{\odot 3}-\boldsymbol{I}_n)\right\|+\left\|\mathrm{offdiag}(\boldsymbol{X}^\top\boldsymbol{X})^{\odot 3}\right\|_F \\ &\le \max_{\alpha\in[n]}\left|\|\boldsymbol{x}_\alpha\|^6-1\right|+n\cdot\max_{\alpha\neq\beta\in[n]}|\boldsymbol{x}_\alpha^\top\boldsymbol{x}_\beta|^3 \le C(\tau_n+n\tau_n^3),\end{aligned}$$

so that $\|\boldsymbol{\Sigma}_{\mathrm{lin}}-\boldsymbol{\Sigma}_0\|=\zeta_3(\sigma)^2\left\|(\boldsymbol{X}^\top\boldsymbol{X})^{\odot 3}-\boldsymbol{I}_n\right\|\to 0$ when $\lim_{n\to\infty}\tau_n\cdot n^{1/3}=0$. □

Theorem 40 will provide a characterization of outlier eigenvalues of $\boldsymbol{K}$ that are separated from $\mathscr{S}_1=\mathrm{supp}(\mu_1)\cup\{0\}$, which is different from $\mathrm{supp}(\mu_1)$ when $\gamma_1<1$. For $\gamma_1<1$, we augment this statement with a small-ball argument to bound the smallest eigenvalue of $\boldsymbol{K}$, using the following result of [Yas16, Theorem 2.1].

Lemma 60 ([Yas16])**.** *Let $\boldsymbol{G}=\frac{1}{\sqrt{N}}[\boldsymbol{g}_1,\dots,\boldsymbol{g}_N]^\top\in\mathbb{R}^{N\times n}$ where the rows $\boldsymbol{g}_i\in\mathbb{R}^n$ are i.i.d. and equal in law to $\boldsymbol{g}\in\mathbb{R}^n$. Define*

$$\boldsymbol{\Sigma}=\mathbb{E}\boldsymbol{g}\boldsymbol{g}^\top,\qquad c_{\boldsymbol{g}}=\inf_{\boldsymbol{v}\in\mathbb{R}^n:\|\boldsymbol{v}\|=1}\mathbb{E}|\boldsymbol{g}^\top\boldsymbol{v}|,\qquad L_{\boldsymbol{g}}(\delta,\iota)=\sup_{\boldsymbol{\Pi}:\mathrm{rank}(\boldsymbol{\Pi})\ge\iota n}\mathbb{P}\left[|\boldsymbol{\Pi}\boldsymbol{g}|^2\le\delta\,\mathrm{rank}(\boldsymbol{\Pi})\right]$$

where the latter supremum is taken over all orthogonal projections $\boldsymbol{\Pi}\in\mathbb{R}^{n\times n}$ with rank at least $\iota\cdot n$.

Suppose $\lambda_{\max}(\boldsymbol{\Sigma})\le 1$, $c(\boldsymbol{g})\ge c$, and $n/N\le y$ for some constants $c>0$ and $y\in(0,1)$. Then there exist constants $s_0,\iota>0$ depending only on (c,y) such that for any $\delta\in(0,1)$ and $s>0$,

$$\mathbb{P}[\lambda_{\min}(\boldsymbol{G}^\top\boldsymbol{G})\ge(s_0-L(\delta,\iota;\boldsymbol{g})-s)\delta]\ge 1-2e^{-yns^2/2}.$$

Lemma 61. *Suppose Assumptions 3, 4, and 5 hold, with $L=1$ and deterministic $\boldsymbol{X}$. Let $\boldsymbol{G}=\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y}$.*

(a) If $\gamma_1\ge 1$, then $0\in\mathrm{supp}(\mu_1)$.

(b) If $\gamma_1<1$, then there is a constant $c>0$ such that $\lambda_{\min}(\boldsymbol{G}^\top\boldsymbol{G})>c$ almost surely for all

large n.

Proof. If $\gamma_1 > 1$ strictly, then by definition

$$\mu_1 = \frac{1}{\gamma_1}\tilde{\mu}_1 + \frac{\gamma_1 - 1}{\gamma_1}\delta_0$$

is a mixture of $\tilde{\mu}_1$ and a point mass at 0, so $0 \in \text{supp}(\mu_1)$. If $\gamma = 1$, then $\mu_1 = \tilde{\mu}_1$. In this case, recall from Proposition 3 that $\text{supp}(\tilde{\mu}_1)$ is characterized by the function

$$z(\tilde{m}) = -\frac{1}{\tilde{m}} + \gamma_1 \int \frac{\lambda}{1+\lambda\tilde{m}} d\nu_0(\lambda).$$

When $\gamma_1 = 1$, we have for all $\tilde{m} \in (0,\infty)$ that

$$z(\tilde{m}) < 0, \qquad z'(\tilde{m}) = \frac{1}{\tilde{m}^2} - \int \frac{\lambda^2}{(1+\lambda\tilde{m})^2} d\nu(\lambda) > 0,$$

so $z(\tilde{m})$ increases from $-\infty$ to 0 over the positive line $\tilde{m} \in (0,\infty)$. Suppose by contradiction that $0 \notin \text{supp}(\tilde{\mu})$. Then by Proposition 3, there must be a point $\tilde{m} \in \mathbb{R} \setminus \mathscr{T}$ where $z(\tilde{m}) = 0$ and $z'(\tilde{m}) > 0$ strictly, implying that there is an open interval $(\tilde{m}_-, \tilde{m}_+) \ni \tilde{m}$ on which $z(\cdot)$ increases from $z(\tilde{m}_-) < 0$ to $z(\tilde{m}_+) > 0$. We must have $\tilde{m} < 0$ by the above behavior of $z(\cdot)$ on $(0,\infty)$, and the range $[z(\tilde{m}_-), z(\tilde{m}_+)]$ must overlap with $[z(a), z(b)]$ for some sufficiently large $a, b \in (0,\infty)$. But this contradicts the non-intersecting property shown in [SC95, Theorem 4.4]. So also in this case $0 \in \text{supp}(\tilde{\mu}_1) = \text{supp}(\mu_1)$, showing part (a).

For part (b), we apply Lemma 60. Under $\mathscr{E}(\boldsymbol{X})$, the condition $b_\sigma \neq 0$ implies $c_0 < \lambda_{\min}(\boldsymbol{\Sigma}_{\text{lin}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\text{lin}}) < C_0$ for some constants $C_0, c_0 > 0$. Hence, also,

$$c_0 < \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) < C_0 \tag{3.7.7}$$

for all large n, by Lemma 59. We assume without loss of generality that $\lambda_{\max}(\boldsymbol{\Sigma}) \leq 1$ as needed in Lemma 60; otherwise, the following argument may be applied to a rescaling of $\boldsymbol{\Sigma}$ and $\boldsymbol{G}$.

To lower bound $c_{\boldsymbol{g}}$ in Lemma 60, observe that for any unit vector $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$\mathbb{E}[(\boldsymbol{g}^\top \boldsymbol{v})^2] = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} > c_0.$$

Viewing $F(\boldsymbol{w}) = \boldsymbol{g}^\top \boldsymbol{v} = \sigma(\boldsymbol{w}^\top \boldsymbol{X})\boldsymbol{v} = \sum_{\alpha=1}^n v_\alpha \sigma(\boldsymbol{w}^\top \boldsymbol{x}_\alpha)$ as a function of $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I})$, we have $\nabla F(\boldsymbol{w}) = \sum_{\alpha=1}^n v_\alpha \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_\alpha)\boldsymbol{x}_\alpha = \boldsymbol{X}(\boldsymbol{v} \odot \sigma'(\boldsymbol{w}^\top \boldsymbol{X}))$ where $\sigma'(\cdot)$ is applied coordinatewise and $\odot$ is the coordinatewise product. Then, applying $|\sigma'(x)| \le \lambda_\sigma$, observe that $\|\nabla F(\boldsymbol{w})\| \le \|\boldsymbol{X}\| \cdot \|\boldsymbol{v} \odot \sigma'(\boldsymbol{w}^\top \boldsymbol{X})\| \le \lambda_\sigma \|\boldsymbol{X}\|$, so $F(\boldsymbol{w})$ is C-Lipschitz in $\boldsymbol{w}$ for a constant $C > 0$ (not depending on $\boldsymbol{v}$) on the event $\mathscr{E}(\boldsymbol{X})$. This implies by Gaussian concentration-of-measure that $\boldsymbol{g}^\top \boldsymbol{v}$ is sub-Gaussian, i.e. for some constants $C, c > 0$ and any $t > 0$, $\mathbb{P}[|\boldsymbol{g}^\top \boldsymbol{v}| \ge t] \le Ce^{-ct^2}$. Integrating this tail bound, for some constant $t > 0$ sufficiently large, we have $\mathbb{E}[(\boldsymbol{g}^\top \boldsymbol{v})^2 \mathbf{1}_{\{|\boldsymbol{g}^\top \boldsymbol{v}| > t\}}] \le c_0/2$, and hence

$$c_0 < \mathbb{E}[(\boldsymbol{g}^\top \boldsymbol{v})^2] \le \mathbb{E}[(\boldsymbol{g}^\top \boldsymbol{v})^2 \mathbf{1}_{\{|\boldsymbol{g}^\top \boldsymbol{v}| \le t\}}] + \frac{c_0}{2} \le t \cdot \mathbb{E}|\boldsymbol{g}^\top \boldsymbol{v}| + \frac{c_0}{2}.$$

So $\mathbb{E}|\boldsymbol{g}^\top \boldsymbol{v}| \ge c_0/(2t)$, and hence $c_{\boldsymbol{g}} \ge c_0/(2t) > c$ for a constant $c > 0$.

Now let $s_0, \iota > 0$ be the constants depending on (c, γ_1) in the statement of Lemma 60. By the nonlinear Hanson-Wright inequality of (4.4.4), for any orthogonal projection $\boldsymbol{\Pi} \in \mathbb{R}^{n \times n}$, any $t > 0$, and some constant $c > 0$, we have

$$\mathbb{P}\big[|\boldsymbol{g}^\top \boldsymbol{\Pi} \boldsymbol{g} - \mathbb{E}\boldsymbol{g}^\top \boldsymbol{\Pi} \boldsymbol{g}| > t\big] \le 2e^{-c\min(t^2/\|\boldsymbol{\Pi}\|_F^2, t/\|\boldsymbol{\Pi}\|)}.$$

Here $\mathbb{E}\boldsymbol{g}^\top \boldsymbol{\Pi} \boldsymbol{g} = \mathrm{Tr}\,\boldsymbol{\Pi}\boldsymbol{\Sigma} > c_0 \mathrm{rank}(\boldsymbol{\Pi})$, $\|\boldsymbol{\Pi}\|_F^2 = \mathrm{rank}(\boldsymbol{\Pi})$, and $\|\boldsymbol{\Pi}\| = 1$, so applying this with $t = (c_0/2)\mathrm{rank}(\boldsymbol{\Pi})$ yields

$$\mathbb{P}[|\boldsymbol{\Pi}\boldsymbol{g}|^2 \le (c_0/2)\mathrm{rank}(\boldsymbol{\Pi})] \le 2e^{-c'\mathrm{rank}(\boldsymbol{\Pi})}.$$

Then, choosing $\delta = c_0/2$, we get $L_{\boldsymbol{g}}(\delta, \iota) \to 0$ as $n \to \infty$. Then Lemma 60 implies $\lambda_{\min}(\boldsymbol{G}^\top \boldsymbol{G}) > s_0\delta/2$ almost surely for all large n, as desired. □

The following is the main result of this section, showing that Theorems 33 and 34 hold in this setting of $L = 1$.

Lemma 62. *Theorems 33 and 34 hold for a single layer $L = 1$. Furthermore, $\boldsymbol{Y}$ is $C\tau_n$-orthonormal for some constant $C > 0$, almost surely for all large n.*

Proof. We condition on $\boldsymbol{X}$ as discussed at the start of this section, and apply Theorem 40 to the centered matrix $\boldsymbol{G} = \boldsymbol{Y} - \mathbb{E}\boldsymbol{Y} = \frac{1}{\sqrt{N}}[\boldsymbol{g}_1, \dots, \boldsymbol{g}_N]^\top$. Let us verify Assumption 6 for $\boldsymbol{G}$: We have shown Assumption 6(a) in (3.7.7). The rows $\boldsymbol{g}_i$ are sub-Gaussian as shown in the above proof of Lemma 61(b), so Assumption 6(b) holds by [Ver10, Eq. (5.26)], and Assumption 6(d) holds by [JNG+19, Lemma 2]. The nonlinear Hanson-Wright inequality of (4.4.4) implies

$$|\boldsymbol{g}_i^\top \boldsymbol{A}\boldsymbol{g}_i - \mathrm{Tr}\boldsymbol{A}\boldsymbol{\Sigma}| \prec \|\boldsymbol{A}\|_F$$

uniformly over $i \in [N]$ and deterministic matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Furthermore, it is clear from the argument preceding (4.4.4) that for any $i \neq j \in [N]$, the joint vector $(\boldsymbol{g}_i, \boldsymbol{g}_j) \in \mathbb{R}^{2n}$ also satisfies Lipschitz concentration, hence

$$\left| \begin{pmatrix} \boldsymbol{g}_i^\top & \boldsymbol{g}_j^\top \end{pmatrix} \boldsymbol{B} \begin{pmatrix} \boldsymbol{g}_i \\ \boldsymbol{g}_j \end{pmatrix} - \mathrm{Tr}\boldsymbol{B} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{pmatrix} \right| \prec \|\boldsymbol{B}\|_F$$

uniformly over $i \neq j \in [N]$ and deterministic matrices $\boldsymbol{B} \in \mathbb{C}^{2n\times 2n}$. Applying this with

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A} & \boldsymbol{0} \end{pmatrix}$$

verifies both statements of Assumption 6(c).

Next, we check Assumption 7 for the population covariance matrix $\boldsymbol{\Sigma}$. Combining

Assumption 4 and (3.7.4) from Lemma 59, we have

$$\frac{1}{n-r}\sum_{i=r+1}^{n}\delta_{\lambda_i(\mathbf{\Sigma})} \to \nu_0 := b_\sigma^2 \otimes \mu_0 \oplus (1-b_\sigma^2) \text{ weakly,} \tag{3.7.8}$$

$$\lambda_i(\mathbf{\Sigma}) \to -\frac{1}{s_{i,0}} := b_\sigma^2 \lambda_i + (1-b_\sigma^2) \notin \operatorname{supp}(\nu_0) \text{ for } i = 1,\dots,r. \tag{3.7.9}$$

Here, the statement $-1/s_{i,0} \notin \operatorname{supp}(\nu_0)$ in (3.7.9) follows from the assumptions $\lambda_i \notin \operatorname{supp}(\mu_0)$ and $b_\sigma \neq 0$. This then implies by the definition of $\mathscr{T}_1$ that $s_{i,0} \in \mathbb{R} \setminus \mathscr{T}_1$, as claimed in Theorem 34(a). Furthermore, for any fixed $\varepsilon > 0$ and all large n, Assumption 4 and (3.7.4) imply also that

$$\lambda_i(\mathbf{\Sigma}) \in \operatorname{supp}(\nu_0) + (-\varepsilon, \varepsilon) \text{ for all } i \geq r+1.$$

Thus Assumption 7 holds for $\mathbf{\Sigma}$ as $n \to \infty$.

Then we can apply Theorems 38 and 40 for $\bar{\boldsymbol{K}} := \boldsymbol{G}^\top \boldsymbol{G}$. The Stieltjes transform approximation in Theorem 38 and Lemma 51(c) together imply $m_{\bar{\boldsymbol{K}}}(z) \to m_1(z)$ almost surely for each fixed $z \in \mathbb{C}^+$, where $m_1(z)$ is the Stieltjes transform of the measure $\mu_1 = \rho_{\gamma_1}^{\mathrm{MP}} \boxtimes \nu_0$. This implies the weak convergence

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(\bar{\boldsymbol{K}})} \to \mu_1 \text{ a.s.} \tag{3.7.10}$$

Theorem 40(a,b) further justifies:

- Let $z_1(\cdot)$ and $\mathscr{I}_1$ be defined by (3.2.6) and (3.2.7) with $\ell = 1$. Then for any sufficiently small constant $\varepsilon > 0$, almost surely for all large n, there is a 1-to-1 correspondence between the eigenvalues of $\bar{\boldsymbol{K}}$ outside $\mathscr{S}_1 + (-\varepsilon, \varepsilon)$ and $\{i : i \in \mathscr{I}_1\}$. Furthermore, for each $i \in \mathscr{I}_1$,

  $$\lambda_i(\bar{\boldsymbol{K}}) \to z_1(s_{i,0}) > 0. \tag{3.7.11}$$

  almost surely as $n \to \infty$.

- Let $\varphi_1(\cdot)$ be defined by (3.2.6) with $\ell = 1$. For each $i \in \mathscr{I}_1$, let $\boldsymbol{v}_i(\bar{\boldsymbol{K}}) \in \mathbb{R}^n$ be a unit-norm

eigenvector of $\bar{\boldsymbol{K}}$ corresponding to $\lambda_i(\bar{\boldsymbol{K}})$, and for each $j \in [r]$, let $\boldsymbol{v}_j(\boldsymbol{\Sigma})$ be a unit-norm eigenvector of $\boldsymbol{\Sigma}$ corresponding to $\lambda_j(\boldsymbol{\Sigma})$. Then almost surely as $n \to \infty$, for each $i \in \mathscr{I}_1$ and $j \in [r]$,

$$|\boldsymbol{v}_j(\boldsymbol{\Sigma})^\top \boldsymbol{v}_i(\bar{\boldsymbol{K}})| \to \sqrt{\varphi_1(s_{i,0})} \cdot \mathbf{1}\{i = j\} \tag{3.7.12}$$

where $\varphi_1(s_{i,0}) > 0$. Moreover, letting $\boldsymbol{v} \in \mathbb{R}^n$ be any unit vector independent of $\boldsymbol{W}$, almost surely

$$|\boldsymbol{v}^\top \boldsymbol{v}_i(\bar{\boldsymbol{K}})| - \sqrt{\varphi_1(s_{i,0})} \cdot |\boldsymbol{v}^\top \boldsymbol{v}_i(\boldsymbol{\Sigma})| \to 0. \tag{3.7.13}$$

If $\gamma_1 \geq 1$, then Lemma 61(a) shows that $\operatorname{supp}(\mu_1) = \operatorname{supp}(\mu_1) \cup \{0\} = \mathscr{S}_1$. If $\gamma_1 < 1$, then Lemma 61(b) shows that for any sufficiently small constant $\varepsilon > 0$, $\bar{\boldsymbol{K}}$ has no eigenvalues in $[0, \varepsilon)$ almost surely for all large n. Thus, in both cases, the first statement above in fact establishes a 1-to-1 correspondence between $\{i : i \in \mathscr{I}_1\}$ and all eigenvalues of $\bar{\boldsymbol{K}}$ outside $\operatorname{supp}(\mu_1) + (-\varepsilon, \varepsilon)$, almost surely for all large n.

To translate these statements to the non-centered matrix $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$, recall from Lemma 58 that $\|\boldsymbol{Y} - \boldsymbol{G}\| \to 0$ almost surely, and from Assumption 6(b) verified above that $\mathbf{1}\{\|\boldsymbol{G}^\top \boldsymbol{G}\| > B'\} \prec 0$ for a constant $B' > 0$. Then, almost surely as $n \to \infty$,

$$\left\|\boldsymbol{K} - \bar{\boldsymbol{K}}\right\| \to 0.$$

Therefore, by Weyl's inequality and (3.7.10), the empirical eigenvalue distribution $\widehat{\mu}_1$ of $\boldsymbol{K}$ converges also to μ_1 weakly a.s., as claimed in Theorem 33. And, by (3.7.11), almost surely for all large n, the eigenvalues $\widehat{\lambda}_{i,1}$ of $\boldsymbol{K}$ outside $\operatorname{supp}(\mu_1) + (-\varepsilon, \varepsilon)$ are also in 1-to-1 correspondence with $\{i : i \in \mathscr{I}_1\}$, where $\widehat{\lambda}_{i,1} \to z_1(s_{i,0})$ for each $i \in \mathscr{I}_1$. In particular, if $r = 0$, then also $|\mathscr{I}_1| = 0$, so $\boldsymbol{K}$ has no eigenvalues outside $\operatorname{supp}(\mu_1) + (-\varepsilon, \varepsilon)$. This proves Theorem 33.

For each $i \in \mathscr{I}_1$, let $\widehat{\boldsymbol{v}}_{i,1} \in \mathbb{R}^n$ be a unit-norm eigenvector of $\boldsymbol{K}$ corresponding to $\widehat{\lambda}_{i,1}$.

Then by the Davis-Kahan Theorem [DK70], we may choose a sign for $\widehat{\boldsymbol{v}}_{i,1}$ such that

$$\left\|\widehat{\boldsymbol{v}}_{i,1} - \boldsymbol{v}_i(\bar{\boldsymbol{K}})\right\| \le \frac{\sqrt{2}\|\boldsymbol{K} - \bar{\boldsymbol{K}}\|}{\operatorname{dist}(\widehat{\lambda}_{i,1}, \operatorname{spec}(\bar{\boldsymbol{K}}) \setminus \{\lambda_i(\bar{\boldsymbol{K}})\})}.$$

We note that $\widehat{\lambda}_{i,1} \to z_1(s_{i,0})$ a.s., which is distinct from the limit values $\{z_1(s_{j,0}) : j \in \mathscr{I}_1 \setminus \{i\}\}$ of $\{\lambda_j(\bar{\boldsymbol{K}}) : \mathscr{I}_1 \setminus \{i\}\}$ by bijectivity of the map $z_1(\cdot)$ in Proposition 3. Furthermore $z_1(s_{i,0})$ falls outside $\operatorname{supp}(\mu_1) + (-\varepsilon, \varepsilon)$ for sufficiently small $\varepsilon > 0$, which contains all other eigenvalues of $\bar{\boldsymbol{K}}$. Thus $\operatorname{dist}(\widehat{\lambda}_{i,1}, \operatorname{spec}(\bar{\boldsymbol{K}}) \setminus \{\lambda_i(\bar{\boldsymbol{K}})\}) \ge c$ for a constant $c > 0$ almost surely for all large n, so

$$\left\|\widehat{\boldsymbol{v}}_{i,1} - \boldsymbol{v}_i(\bar{\boldsymbol{K}})\right\| \to 0 \text{ a.s.}$$

Similarly, by the convergence $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{lin}}\| \to 0$ and the assumption $b_\sigma \neq 0$, we have

$$\left\|\boldsymbol{v}_j(\boldsymbol{\Sigma}) - \boldsymbol{v}_j\right\| \to 0$$

for each $j \in [r]$, where $\boldsymbol{v}_j$ is the unit-norm eigenvector of $\boldsymbol{\Sigma}_{\text{lin}}$ corresponding to its eigenvalue $b_\sigma^2 \lambda_j(\boldsymbol{X}^\top \boldsymbol{X}) + (1 - b_\sigma^2)$, i.e. the eigenvector of $\boldsymbol{X}^\top \boldsymbol{X}$ corresponding to $\lambda_j(\boldsymbol{X}^\top \boldsymbol{X})$. Then (3.7.12) and (3.7.13) imply also

$$|\boldsymbol{v}_j^\top \widehat{\boldsymbol{v}}_i|^2 \to \varphi_1(s_{i,0}) \cdot \mathbf{1}\{i = j\}, \qquad |\boldsymbol{v}^\top \widehat{\boldsymbol{v}}_i|^2 - \varphi_1(s_{i,0}) \cdot |\boldsymbol{v}^\top \boldsymbol{v}_i|^2 \to 0.$$

This shows all claims of Theorem 34 for $L = 1$.

Finally, on $\mathscr{E}(\boldsymbol{X})$, the matrix $\boldsymbol{Y}$ is $C\tau_n$-orthonormal for a constant $C > 0$ by Lemma 16(b). Notice that the proof of Lemma 16(b) again does not use the condition $\sum_\alpha (\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \le B^2$ of (ε, B)-orthonormality therein, and the remaining conditions of (ε, B)-orthonormality hold under $\mathscr{E}(\boldsymbol{X})$. This shows the last claim of the lemma. □

3.7.2 Spike Analysis for Multiple Layers

We now prove Theorem 34 by inductively applying the result for $L = 1$ through multiple layers. We follow the notations of Section 3.2.

Proof of Theorem 34. Suppose inductively that Assumption 4 holds with $\boldsymbol{X}_{\ell-1}$ in place of $\boldsymbol{X}_0$, and all conclusions of Theorem 34 hold for $\boldsymbol{K}_\ell$. The base case of $\ell = 1$ follows from Lemma 62.

Then the last statement of Lemma 62 implies that $\boldsymbol{X}_\ell$ is τ_n'-orthonormal almost surely for all large n, for some τ_n' satisfying $\tau_n' \cdot n^{1/3} \to 0$. Furthermore, the conclusions of Theorem 34(b,c) for $\boldsymbol{K}_\ell$ imply that statements (a) and (b) of Assumption 4 also hold for $\boldsymbol{X}_\ell$, in the following sense: Let $r_\ell = |\mathscr{I}_\ell|$. Then

$$\frac{1}{n-|r_\ell|}\sum_{i\notin\mathscr{I}_\ell}\delta_{\lambda_i(\boldsymbol{X}_\ell^\top\boldsymbol{X}_\ell)} \to \mu_\ell \text{ weakly a.s.}$$

For any fixed $\varepsilon > 0$, almost surely for all large n, $\widehat{\lambda}_{i,\ell} := \lambda_i(\boldsymbol{X}_\ell^\top\boldsymbol{X}_\ell) \in \operatorname{supp}(\mu_\ell) + (-\varepsilon,\varepsilon)$ for all $i \notin \mathscr{I}_\ell$. Furthermore, for each $i \in \mathscr{I}_\ell$, $\widehat{\lambda}_{i,\ell} \to z_\ell(s_{i,\ell-1}) \notin \operatorname{supp}(\mu_\ell)$.

Then we may apply Lemma 62 with input data $\boldsymbol{X} = \boldsymbol{X}_\ell$ in place of $\boldsymbol{X}_0$. This shows that for any fixed $\varepsilon > 0$ and all large n, there is a 1-to-1 correspondence between the eigenvalues $\widehat{\lambda}_{i,\ell+1}$ of $\boldsymbol{K}_{\ell+1}$ outside $\operatorname{supp}(\mu_{\ell+1}) + (-\varepsilon,\varepsilon)$ and $\{i : i \in \mathscr{I}_{\ell+1}\}$, where $\widehat{\lambda}_{i,\ell+1} \to z_{\ell+1}\big(s_{i,\ell}\big) > 0$ a.s., and $s_{i,\ell} \in \mathbb{R} \setminus \mathscr{T}_{\ell+1}$. Moreover, for any unit vector $\boldsymbol{v} \in \mathbb{R}^n$ independent of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{\ell+1}$,

$$|\widehat{\boldsymbol{v}}_{i,\ell+1}^\top\boldsymbol{v}|^2 - \varphi_{\ell+1}(s_{i,\ell}) \cdot |\widehat{\boldsymbol{v}}_{i,\ell}^\top\boldsymbol{v}|^2 \to 0,$$

where also $\varphi_{\ell+1}(s_{i,\ell}) > 0$. Then by the induction hypothesis for $|\widehat{\boldsymbol{v}}_{i,\ell}^\top\boldsymbol{v}|^2$,

$$|\widehat{\boldsymbol{v}}_{i,\ell+1}^\top\boldsymbol{v}|^2 \to \prod_{k=1}^{\ell+1}\varphi_k(s_{i,k-1}) \cdot |\boldsymbol{v}_i^\top\boldsymbol{v}|^2,$$

and specializing to $\boldsymbol{v} = \boldsymbol{v}_j$ for $j \in [r]$ gives

$$|\widehat{\boldsymbol{v}}_{i,\ell+1}^\top \boldsymbol{v}_j|^2 \to \prod_{k=1}^{\ell+1} \varphi_k(s_{i,k-1}) \cdot \mathbf{1}\{i = j\}.$$

This verifies all conclusions of Theorem 34 for $\boldsymbol{K}_{\ell+1}$, completing the induction. □

3.7.3 Corollary for Signal-Plus-Noise Input Data

Proof of Corollary 35. It is shown in [BGN12, Section 3.1] that asymptotically as $d, n \to \infty$ with $n/d \to \gamma_0$, the data matrix $\boldsymbol{X}$ has a spike singular value corresponding to θ_i if and only if $\theta_i > \gamma_0^{1/4}$, in which case

$$\lambda_i(\boldsymbol{K}_0) \to \lambda_i := \frac{(1+\theta_i^2)(\gamma_0+\theta_i^2)}{\theta_i^2}, \qquad |\boldsymbol{b}_i^\top \boldsymbol{v}_i|^2 \to 1 - \frac{\gamma_0(1+\theta_i^2)}{\theta_i^2(\theta_i^2+\gamma_0)}$$

where $\boldsymbol{v}_i$ is the unit eigenvector of the input Gram matrix $\boldsymbol{K}_0 = \boldsymbol{X}^\top \boldsymbol{X}$. Thus claims (a) and (b) of Assumption 4 hold with $r = |\{i : \theta_i > \gamma_0^{1/4}\}|$, $\mu_0 = \rho_{\gamma_0}^{\mathrm{MP}}$ being the standard Marčenko-Pastur law, and $\lambda_i = (1+\theta_i^2)(\gamma_0+\theta_i^2)/\theta_i^2$ being the above values.

We note that $\boldsymbol{X}$ is $n^{-1/2+\varepsilon}$-orthonormal for any $\varepsilon > 0$ almost surely for all large n, by the given condition $\max_{1 \le i \le r} \|\boldsymbol{b}_i\|_\infty < n^{-1/2+\varepsilon}$ and the bounds, for any $\alpha, \beta \in [n]$,

$$\begin{aligned}
\|\boldsymbol{x}_\alpha\| &= \|\boldsymbol{z}_\alpha\| + \sum_{i=1}^r O_\prec(\|\boldsymbol{a}_i\||\theta_i||b_{i,\alpha}|) = \|\boldsymbol{z}_\alpha\| + O_\prec\Big(n^{-1/2+\varepsilon}\Big) = 1 + O_\prec\Big(n^{-1/2+\varepsilon}\Big), \\
\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta &= \boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta + \sum_{i=1}^r O_\prec\Big(|\theta_i|\Big(|\boldsymbol{a}_i^\top \boldsymbol{z}_\alpha||b_{i,\alpha}| + |\boldsymbol{a}_i^\top \boldsymbol{z}_\beta||b_{i,\beta}|\Big) + \theta_i^2 \|\boldsymbol{a}_i\|^2 |b_{i,\alpha} b_{i,\beta}|\Big) \\
&= O_\prec\Big(n^{-1/2+\varepsilon}\Big).
\end{aligned}$$

Hence Theorem 34 applies, showing that $\boldsymbol{K}_\ell$ has an outlier eigenvalue corresponding to each input signal θ_i if and only if $\theta_i > \gamma_0^{1/4}$ and $i \in \mathscr{I}_\ell$. The statement (3.2.9) follows from Theorem 34(c) applied with $\boldsymbol{v} = \boldsymbol{b}_i$. □

3.8 Acknowledgment

Chapter 3 is extracted from the preprint “Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. *arXiv preprint* arXiv:2402.10127 (2024)”. The thesis author is the co-author of this paper.

Chapter 4

Deformed Semicircle Law for Ultra-Wide NNs

In this Chapter, we study the random CK and NTK matrices of a two-layer fully connected neural network with input data $X \in \mathbb{R}^{d_0 \times n}$, given by $f : \mathbb{R}^{d_0 \times n} \to \mathbb{R}^n$ such that

$$f(X) := \frac{1}{\sqrt{d_1}} \boldsymbol{a}^\top \sigma(WX), \tag{4.0.1}$$

where $W \in \mathbb{R}^{d_1 \times d_0}$ is the weight matrix for the first layer, $\boldsymbol{a} \in \mathbb{R}^{d_1}$ are the second layer weights, and σ is a nonlinear activation function applied to the matrix WX element-wisely. We assume that all entries of $\boldsymbol{a}$ and W are independently identically distributed by the standard Gaussian $\mathcal{N}(0,1)$. We will always view the input data X as a deterministic matrix (independent of the random weights in $\boldsymbol{a}$ and W) with certain assumptions.

In terms of random matrix theory, we study the difference between these two kernel matrices (CK and NTK) and their expectations with respect to random weights, showing both asymptotic and non-asymptotic behaviors of these differences as the width of the first hidden layer d_1 is growing faster than the number of samples n. As an extension of Chapter 2, we prove that when $n/d_1 \to 0$, the centered CK and NTK with appropriate normalization have the limiting eigenvalue distribution given by a deformed semicircle law, determined by the training data spectrum and the nonlinear activation function. To prove this global law, we further set up a limiting law theorem for centered sample covariance matrices with dependent structures

and a nonlinear version of the Hanson-Wright inequality. These two results are very general, which makes them potentially applicable to different scenarios beyond our neural network model. For the non-asymptotic analysis, we establish concentration inequalities between the random kernel matrices and their expectations. As a byproduct, we provide lower bounds of the smallest eigenvalues of CK and NTK, which are essential for the global convergence of gradient-based optimization methods when training a wide neural network [OS20, NM20, Ngu21]. Because of the non-asymptotic results for kernel matrices, we can also describe how close the performances of the random feature regression and the limiting kernel regression are with a general dataset, which allows us to compute the limiting training error and generalization error for the random feature regression via its corresponding kernel regression in the ultra-wide regime.

4.1 Related Work

General sample covariance matrices

We observe that the random matrix $Y \in \mathbb{R}^{d_1 \times n}$ defined above has independent and identically distributed rows. Hence, $Y^\top Y$ is a generalized sample covariance matrix. We first inspect a more general sample covariance matrix Y whose rows are independent copies of some random vector $\mathbf{y} \in \mathbb{R}^n$. Assuming n and d_1 both go to infinity but $n/d_1 \to 0$, we aim to study the limiting empirical eigenvalue distribution of centered Wishart matrices in the form of

$$\frac{1}{\sqrt{nd_1}}\Big(Y^\top Y - \mathbb{E}[Y^\top Y]\Big), \tag{4.1.1}$$

with certain conditions on $\mathbf{y}$. This regime is also related to the ultra-high dimensional setting in statistics [QLY23].

This regime has been studied for decades starting in [BY88], where Y has i.i.d. entries and $\mathbb{E}[Y^\top Y] = d_1\,\mathrm{Id}$. In this setting, by the moment method, one can obtain the semicircle law. This normalized model also arises in quantum theory with respect to random induced states (see [Aub12, AS17, CYZ18]). The largest eigenvalue of such a normalized sample covariance

matrix has been considered in [CP12]. Subsequently, [CP15, LY16, YXZ22, QLY23] analyzed the fluctuations for the linear spectral statistics of this model and applied this result to hypothesis testing for the covariance matrix. A spiked model for sample covariance matrices in this regime was recently studied in [Fel23b]. This kind of semicircle law also appears in many other random matrix models. For instance, [Jia04] showed this limiting law for normalized sample correlation matrices. Also, the semicircle law for centered sample covariance matrices has already been applied in machine learning: [GKZ19] controlled the generalization error of shallow neural networks with quadratic activation functions by the moments of this limiting semicircle law; [GZR22] derived a semicircle law of the fluctuation matrix between stochastic batch Hessian and the deterministic empirical Hessian of deep neural networks.

For general sample covariance, [WP14] considered the form $Y = BXA^{1/2}$ with deterministic A and B, where X consists of i.i.d. entries with mean zero and variance one. The same result has been proved in [Bao12] by generalized Stein's method. Unlike previous results, [Xie13] tackled the general case, only assuming Y has independent rows with some deterministic covariance Φ_n. Though this is similar to our model in Section 4.5, we will consider more general assumptions on each row of Y, which can be directly verified in our neural network models.

Infinite-width kernels and the smallest eigenvalues of empirical kernels

Besides the above asymptotic spectral fluctuation, we provide non-asymptotic concentrations of centered CK and NTK in spectral norm and a corresponding result for the NTK. In the infinite-width networks, where $d_1 \to \infty$ and n are fixed, both CK and NTK will converge to their expected kernels. This has been investigated in [DFS16, SGGSD17, LBN$^+$18, MHR$^+$18] for the CK and [JGH18, DZPS19a, AZLS19, ADH$^+$19b, LRZ20] for the NTK. Such kernels are also called infinite-width kernels in literature. In this current work, we present the precise probability bounds for concentrations of CK and NTK around their infinite-width kernels, where the difference is of order $\sqrt{n/d_1}$. Our results permit more general activation functions and input data X only with pairwise approximate orthogonality, albeit similar concentrations have been

applied in [AKM$^+$17, SY19, AP20, MZ20, HXAP20].

A corollary of our concentration is the explicit lower bounds of the smallest eigenvalues of the CK and the NTK. Such extreme eigenvalues of the NTK have been utilized to prove the global convergence of gradient descent algorithms of wide neural networks since the NTK governs the gradient flow in the training process, see Section 1.1 in Chapter 1. The smallest eigenvalue of NTK is also crucial for proving generalization bounds and memorization capacity in [ADH$^+$19a, MZ20]. Analogous to Theorem 3.1 in [MZ20], our lower bounds are given by the Hermite coefficients of the activation function σ. Besides, the lower bound of NTK for multi-layer ReLU networks is analyzed in [NMM21].

Random feature regression and limiting kernel regression

Another byproduct of our concentration results is to measure the difference of performance between random feature regression with respect to $\frac{1}{\sqrt{d_1}}Y$ and corresponding kernel regression when $d_1/n \to \infty$. Random feature regression can be viewed as the linear regression of the last hidden layer, and its performance has been studied in, for instance, [PW17, LLC18, MM22, LCM20, GLK$^+$20, HL20, LD21, MMM21, LGC$^+$21b] under the linear-width regime. This linear-width regime is also known as the high-dimensional regime, while our ultra-wide regime is also called a highly overparameterized regime in literature, see [MM22]. In this regime, the CK matrix $\frac{1}{d_1}Y^\top Y$ is not concentrated around its expectation

$$\Phi := \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top X)^\top \sigma(\boldsymbol{w}^\top X)] \tag{4.1.2}$$

under the spectral norm, where $\boldsymbol{w}$ is the standard normal random vector in $\mathbb{R}^{d_0}$. But the limiting spectrum of CK is exploited to characterize the asymptotic performance and double descent phenomenon of random feature regression when $n, d_0, d_1 \to \infty$ proportionally. Several works have also utilized this regime to depict the performance of the ultra-wide random network by letting $d_1/n \to \psi \in (0,\infty)$ first, getting the asymptotic performance and then taking $\psi \to \infty$ (see [MM22, YBM21]). However, there is still a difference between this sequential limit and

the ultra-wide regime. Before these results, random feature regression has already attracted significant attention in that it is a random approximation of the RKHS defined by population kernel function $K : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ such that

$$K(\mathbf{x}, \mathbf{z}) := \mathbb{E}_{\boldsymbol{w}}[\sigma(\langle \boldsymbol{w}, \mathbf{x} \rangle)\sigma(\langle \boldsymbol{w}, \mathbf{z} \rangle)], \tag{4.1.3}$$

when width d_1 is sufficiently large [RR07, Bac13, RR17, Bac17]. We point out that Theorem 9 of [AKM$^+$17] has the same order $\sqrt{n/d_1}$ of the approximation as ours, despite only for random Fourier features.

In this Chapter, the concentration between empirical kernel induced by $\frac{1}{d_1} Y^\top Y$ and the population kernel matrix K defined in (4.1.3) for X leads to the control of the differences of training/test errors between random feature regression and kernel regression, which were previously concerned by [AKM$^+$17, JSS$^+$20, MZ20, MMM21] in different cases. Specifically, [JSS$^+$20] obtained the same kind of estimation but considered random features sampled from Gaussian Processes. Our results explicitly show how large width d_1 should be so that the random feature regression gets the same asymptotic performance as kernel regression [MMM21]. With these estimations, we can take the limiting test error of the kernel regression to predict the limiting test error of random feature regression as $n/d_1 \to 0$ and $d_0, n \to \infty$. We refer [LR20, LRZ20, LLS21, MMM21], [BMR21, Section 4.3] and references therein for more details in high-dimensional kernel ridge/ridgeless regressions. We emphasize that the optimal prediction error of random feature regression in linear-width regime is actually achieved in the ultra-wide regime, which boils down to the limiting kernel regression, see [MM22, MMM21, YBM21, LGC$^+$21b]. This is one of the motivations for studying the ultra-wide regime and the limiting kernel regression.

In the end, we would like to mention the idea of spectral-norm approximation for the expected kernel Φ, which helps us describe the asymptotic behavior of limiting kernel regression. For specific activation σ, kernel Φ has an explicit formula, see [LLC18, LC18b, LCM20],

whereas generally, it can be expanded in terms of the Hermite expansion of σ [PW17, MM22, FW20]. Thanks to pairwise approximate orthogonality introduced in Definition 4, we can approximate Φ in the spectral norm for general deterministic data X. This pairwise approximate orthogonality defines how orthogonal is within different input vectors of X. With certain i.i.d. assumption on X, [LRZ20] and [BMR21, Section 4.3], where the scaling $d_0 \propto n^{\alpha}$, for $\alpha \in (0,1]$, determined which degree of the polynomial kernel is sufficient to approximate Φ. Instead, our theory leverages the approximate orthogonality among general datasets X to obtain a similar approximation. Our analysis presumably indicates that the weaker orthogonality X has, the higher degree of the polynomial kernel we need to approximate the kernel Φ.

4.2 Preliminaries

Before stating our main results, we describe our model with assumptions. We first consider the output of the first hidden layer and empirical *Conjugate Kernel* (CK):

$$Y := \sigma(WX) \quad \text{and} \quad \frac{1}{d_1} Y^\top Y. \tag{4.2.1}$$

Observe that the rows of matrix Y are independent and identically distributed since only W is random and X is deterministic. Let the i-th row of Y be $\mathbf{y}_i^\top$, for $1 \le i \le d_1$. Then, we obtain a sample covariance matrix,

$$Y^\top Y = \sum_{i=1}^{d_1} \mathbf{y}_i \mathbf{y}_i^\top, \tag{4.2.2}$$

which is the sum of d_1 independent rank-one random matrices in $\mathbb{R}^{n\times n}$. Let the second moment of any row $\mathbf{y}_i$ be (4.1.2). Later on, we will approximate Φ based on the assumptions of input data X.

Next, we define the empirical *Neural Tangent Kernel* (NTK) for (4.0.1), denoted by

$H \in \mathbb{R}^{n\times n}$. From Section 2.2.3, the (i,j)-th entry of H can be explicitly written as

$$H_{ij} := \frac{1}{d_1}\sum_{r=1}^{d_1}\Big(\sigma(\boldsymbol{w}_r^\top \mathbf{x}_i)\sigma(\boldsymbol{w}_r^\top \mathbf{x}_j) + a_r^2\sigma'(\boldsymbol{w}_r^\top \mathbf{x}_i)\sigma'(\boldsymbol{w}_r^\top \mathbf{x}_j)\mathbf{x}_i^\top \mathbf{x}_j\Big), \quad 1 \le i,j \le n, \tag{4.2.3}$$

where $\boldsymbol{w}_r$ is the r-th row of weight matrix W, $\mathbf{x}_i$ is the i-th column of matrix X, and a_r is r-th entry of the output layer $\boldsymbol{a}$. In the matrix form, H can be written by

$$H := \frac{1}{d_1}\Big(Y^\top Y + (S^\top S)\odot(X^\top X)\Big), \tag{4.2.4}$$

where the α-th column of S is given by

$$\operatorname{diag}(\sigma'(W\mathbf{x}_\alpha))\boldsymbol{a}, \quad \forall 1 \le \alpha \le n. \tag{4.2.5}$$

We introduce the following assumptions for the random weights, nonlinear activation function σ, and input data. These assumptions are basically carried on from Chapter 2.

Assumption 8. The entries of W and $\boldsymbol{a}$ are i.i.d. and distributed by $\mathcal{N}(0,1)$.

Assumption 9. Activation function $\sigma(x)$ is a Lipschitz function with the Lipschitz constant $\lambda_\sigma \in (0,\infty)$. Assume that σ is centered and normalized with respect to $\xi \sim \mathcal{N}(0,1)$ such that

$$\mathbb{E}[\sigma(\xi)] = 0, \qquad \mathbb{E}[\sigma^2(\xi)] = 1. \tag{4.2.6}$$

Define constants a_σ and $b_\sigma \in \mathbb{R}$ by

$$b_\sigma := \mathbb{E}[\sigma'(\xi)], \qquad a_\sigma := \mathbb{E}[\sigma'(\xi)^2]. \tag{4.2.7}$$

Furthermore, σ satisfies *either* of the following:

1. $\sigma(x)$ is twice differentiable with $\sup_{x\in\mathbb{R}}|\sigma''(x)| \le \lambda_\sigma$, or

2. $\sigma(x)$ is a piece-wise linear function defined by

$$\sigma(x) = \begin{cases} ax+b, & x > 0, \\ cx+b, & x \leq 0, \end{cases}$$

for some constants $a, b, c \in \mathbb{R}$ such that (4.2.6) holds.

Analogously to [HXAP20], our Assumption 9 permits σ to be the commonly used activation functions, including ReLU, Sigmoid, and Tanh, although we have to center and normalize the activation functions to guarantee (4.2.6). Such normalized activation functions exclude some trivial spike in the limiting spectra of CK and NTK [BP21, FW20]. The foregoing assumptions ensure our nonlinear Hanson-Wright inequality in the proof. As a future direction, going beyond Gaussian weights and Lipschitz activation functions may involve different types of concentration inequalities.

Next, we present the conditions of the deterministic input data X and the asymptotic regime for this Chapter. Recall the definition of (ε, B)-orthonormal property for our data matrix X in Definition 4.

Assumption 10. Let $n, d_0, d_1 \to \infty$ such that

(a) $\gamma := n/d_1 \to 0$;

(b) X is (ε_n, B)-orthonormal such that $n\varepsilon_n^4 \to 0$ as $n \to \infty$;

(c) The empirical spectral distribution $\hat{\mu}_0$ of $X^\top X$ converges weakly to a fixed and non-degenerate probability distribution $\mu_0 \neq \delta_0$ on $[0, \infty)$.

In above (b), the (ε_n, B)-orthonormal property with $n\varepsilon_n^4 = o(1)$ is a quantitative version of *pairwise approximate orthogonality* for the column vectors of the data matrix $X \in \mathbb{R}^{d_0 \times n}$. When $d_0 \asymp n$, it holds, with high probability, for many random X with independent columns $\mathbf{x}_\alpha$, including the anisotropic Gaussian vectors $\mathbf{x}_\alpha \sim \mathcal{N}(0, \Sigma)$ with $\mathrm{tr}(\Sigma) = 1$ and $\|\Sigma\| \lesssim 1/n$, vectors

generated by Gaussian mixture models, and vectors satisfying the log-Sobolev inequality or convex Lipschitz concentration property. Specifically, when $\mathbf{x}_\alpha$'s are independently sampled from the unit sphere $\mathbb{S}^{d_0-1}$, X is (ε_n, B)-orthonormal with high probability where $\varepsilon_n = O\Big(\sqrt{\frac{\log(n)}{n}}\Big)$ and $B = O(1)$ as $n \asymp d_0$. In this case, for any $\ell > 2$, we have $n\varepsilon_n^\ell \to 0$. In our theory, we always treat X as a deterministic matrix. However, our results also work for random input X independent of weights W and $\boldsymbol{a}$ by conditioning on the high probability event that X satisfies (ε_n, B)-orthonormal property. Unlike data vectors with independent entries, our assumption is promising to analyze real-world datasets [LGC+21b] and establish some n-dependent deterministic equivalents like [LCM20].

The following Hermite polynomials are crucial to the approximation of Φ in our analysis.

Definition 63 (Normalized Hermite polynomials)**.** The r-th normalized Hermite polynomial is given by

$$h_r(x) = \frac{1}{\sqrt{r!}}(-1)^r e^{x^2/2}\frac{d^r}{dx^r}e^{-x^2/2}.$$

Here $\{h_r\}_{r=0}^\infty$ form an orthonormal basis of $L^2(\mathbb{R}, \Gamma)$, where Γ denotes the standard Gaussian distribution. For $\sigma_1, \sigma_2 \in L^2(\mathbb{R}, \Gamma)$, the inner product is defined by

$$\langle \sigma_1, \sigma_2 \rangle = \int_{-\infty}^{\infty} \sigma_1(x)\sigma_2(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx.$$

Every function $\sigma \in L^2(\mathbb{R}, \Gamma)$ can be expanded as a Hermite polynomial expansion

$$\sigma(x) = \sum_{r=0}^{\infty} \zeta_r(\sigma) h_r(x),$$

where $\zeta_r(\sigma)$ is the r-th Hermite coefficient defined by

$$\zeta_r(\sigma) := \int_{-\infty}^{\infty} \sigma(x) h_r(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx.$$

In the following statements and proofs, we denote $\xi \sim \mathcal{N}(0,1)$. Based on Defini-

tion 63, let us further denote that, for any $k \in \mathbb{N}$, $\zeta_k(\sigma) = \mathbb{E}[\sigma(\xi)h_k(\xi)]$. Specifically, $b_\sigma := \mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi \cdot \sigma(\xi)] = \zeta_1(\sigma)$. Let $f_k(x) = x^k$. We define the inner-product kernel random matrix $f_k(X^\top X) \in \mathbb{R}^{n\times n}$ by applying f_k entrywise to $X^\top X$. Define a deterministic matrix

$$\Phi_0 := \boldsymbol{\mu}\boldsymbol{\mu}^\top + \sum_{k=1}^{3} \zeta_k(\sigma)^2 f_k(X^\top X) + (1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2)\,\mathrm{Id}, \tag{4.2.8}$$

where the α-th entry of $\boldsymbol{\mu} \in \mathbb{R}^n$ is $\sqrt{2}\zeta_2(\sigma)\cdot(\|\mathbf{x}_\alpha\| - 1)$ for $1 \le \alpha \le n$. We will employ Φ_0 as an approximation of the population covariance Φ in (4.1.2) in the spectral norm when $n\varepsilon_n^4 \to 0$.

4.3 Main Results

4.3.1 Spectra of the Centered CK and NTK

Our first result is a deformed semicircle law for the CK matrix. Denote $\tilde{\mu}_0 = (1-b_\sigma)^2 \oplus b_\sigma^2 \otimes \mu_0$. The limiting law of our centered and normalized CK matrix is depicted by $\mu_s \boxtimes \tilde{\mu}_0$, where μ_s is the standard semicircle law and the notation $\boxtimes$ is the *free multiplicative convolution* in free probability theory.

Theorem 64 (Limiting spectral distribution for the conjugate kernel)**.** *Suppose Assumptions 8, 9 and 10 of the input matrix X hold, the empirical eigenvalue distribution of*

$$\frac{1}{\sqrt{d_1 n}}\Big(Y^\top Y - \mathbb{E}[Y^\top Y]\Big) \tag{4.3.1}$$

converges weakly to

$$\mu := \mu_s \boxtimes \Big((1-b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_0\Big) = \mu_s \boxtimes \tilde{\mu}_0 \tag{4.3.2}$$

almost surely as $n, d_0, d_1 \to \infty$. Furthermore, if $d_1\varepsilon_n^4 \to 0$ as $n, d_0, d_1 \to \infty$, then the empirical eigenvalue distribution of

$$\sqrt{\frac{d_1}{n}}\left(\frac{1}{d_1}Y^\top Y - \Phi_0\right) \tag{4.3.3}$$

also converges weakly to the probability measure μ almost surely, whose Stieltjes transform $m(z)$ is defined by

$$m(z) + \int_{\mathbb{R}} \frac{d\tilde{\mu}_0(x)}{z + \beta(z)x} = 0 \tag{4.3.4}$$

for each $z \in \mathbb{C}^+$, where $\beta(z) \in \mathbb{C}^+$ is the unique solution to

$$\beta(z) + \int_{\mathbb{R}} \frac{x d\tilde{\mu}_0(x)}{z + \beta(z)x} = 0. \tag{4.3.5}$$

Suppose that we additionally have $b_\sigma = 0$, i.e. $\mathbb{E}[\sigma'(\xi)] = 0$. In this case, our Theorem 64 shows that the limiting spectral distribution of (4.1.1) is the semicircle law, and from (4.3.2), the deterministic data matrix X does not have an effect on the limiting spectrum. See Figure 4.1 for a cosine-type σ with $b_\sigma = 0$. The only effect of the nonlinearity in μ is the coefficient b_σ in the deformation $\tilde{\mu}_0$.

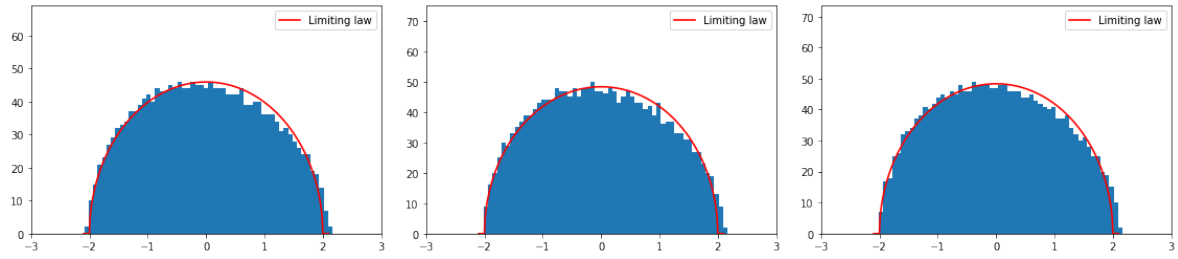


Figure 4.1. Simulations for ESDs of (4.3.3) and theoretical predication (red curves) of the limiting law μ where activation $\sigma(x) \propto \cos(x)$ satisfies Assumption 9 with $b_\sigma = 0$, and X is a standard Gaussian random matrix. Dimension parameters are given by $n = 1.9 \times 10^3$, $d_0 = 2 \times 10^3$ and $d_1 = 2 \times 10^5$ (left); $n = 2 \times 10^3$, $d_0 = 1.9 \times 10^3$ and $d_1 = 2 \times 10^5$ (middle); $n = 2 \times 10^3$, $d_0 = 2 \times 10^3$ and $d_1 = 2 \times 10^5$ (right).

Figure 4.2 is a simulation of the limiting spectral distribution of CK with activation function $\sigma(x)$ given by $\arctan(x)$ after proper shifting and scaling. The red curves are implemented by the self-consistent equations (4.3.4) and (4.3.5) in Theorem 64. In Section 4.5, we present general random matrix models with similar limiting eigenvalue distribution as μ whose Stieltjes transform is also determined by (4.3.4) and (4.3.5).

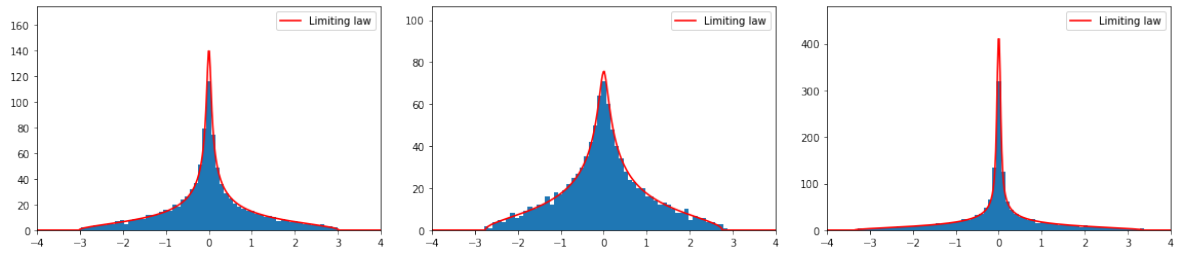


Figure 4.2. Simulations for ESDs of (4.3.3) and theoretical predication (red curves) of the limiting law μ where activation $\sigma(x) \propto \arctan(x)$ satisfies Assumption 9 and X is a standard Gaussian random matrix: $n = 10^3$, $d_0 = 10^3$ and $d_1 = 10^5$ (left); $n = 10^3$, $d_0 = 1.5 \times 10^3$ and $d_1 = 10^5$ (middle); $n = 1.5 \times 10^3$, $d_0 = 10^3$ and $d_1 = 10^5$ (right).

Theorem 64 can be extended to the NTK model as well. Denote by

$$\Psi := \frac{1}{d_1}\mathbb{E}[S^\top S] \odot (X^\top X) \in \mathbb{R}^{n\times n}. \tag{4.3.6}$$

As an approximation of Ψ in the spectral norm, we define

$$\Psi_0 := \left(a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma)\right) \mathrm{Id} + \sum_{k=0}^{2} \eta_k^2(\sigma) f_{k+1}(X^\top X), \tag{4.3.7}$$

where f_k's are defined in (4.2.8), a_σ is defined in (4.2.7), and the k-th Hermite coefficient of σ' is

$$\eta_k(\sigma) := \mathbb{E}[\sigma'(\xi) h_k(\xi)]. \tag{4.3.8}$$

Then, a similar deformed semicircle law can be obtained for the empirical NTK matrix H.

Theorem 65 (Limiting spectral distribution of the NTK)**.** *Under Assumptions 8, 9 and 10 of the input matrix X, the empirical eigenvalue distribution of*

$$\sqrt{\frac{d_1}{n}}(H - \mathbb{E}[H]) \tag{4.3.9}$$

weakly converges to $\mu = \mu_s \boxtimes \left((1-b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_0\right)$ almost surely as $n, d_0, d_1 \to \infty$ and $n/d_1 \to 0$.

Furthermore, suppose that $\varepsilon_n^4 d_1 \to 0$, then the empirical eigenvalue distribution of

$$\sqrt{\frac{d_1}{n}}(H - \Phi_0 - \Psi_0) \tag{4.3.10}$$

weakly converges to μ almost surely, where Φ_0 and Ψ_0 are defined in (4.2.8) *and* (4.3.7)*, respectively.*

4.3.2 Non-asymptotic Estimations for Kernels

With our nonlinear Hanson-Wright inequality (Corollary 77), we attain the following concentration bound on the CK matrix in the spectral norm.

Theorem 66. *With Assumption 8, assume X satisfies $\sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 1)^2 \le B^2$ for a constant $B \ge 0$, and σ is λ_σ-Lipschitz with $\mathbb{E}[\sigma(\xi)] = 0$. Then with probability at least $1 - 4e^{-2n}$,*

$$\left\| \frac{1}{d_1} Y^\top Y - \Phi \right\| \le C\left(\sqrt{\frac{n}{d_1}} + \frac{n}{d_1} \right) \lambda_\sigma^2 \|X\|^2 + 32B\lambda_\sigma^2 \|X\| \sqrt{\frac{n}{d_1}}, \tag{4.3.11}$$

where $C > 0$ is a universal constant.

Remark. *Theorem 66 ensures that the empirical spectral measure μ_n of the centered random matrix $\sqrt{\frac{d_1}{n}}\left(\frac{1}{d_1} Y^\top Y - \Phi \right)$ has a bounded support for all sufficiently large n. Together with the global law in Theorem 64, our concentration inequality* (4.3.11) *is* tight *up to a constant factor. Additionally, by the weak convergence of μ_n to μ proved in Theorem 64, we can take a test function x^2 to obtain that*

$$\int_{\mathbb{R}} x^2 d\mu_n(x) \to \int_{\mathbb{R}} x^2 d\mu(x), \quad \textit{i.e.,} \quad \frac{\sqrt{d_1}}{n} \left\| \frac{1}{d_1} Y^\top Y - \Phi \right\|_F \to \left(\int_{\mathbb{R}} x^2 d\mu(x) \right)^{\frac{1}{2}}$$

almost surely, as $n, d_1 \to \infty$ and $d_1/n \to \infty$. Therefore, the fluctuation of $\frac{1}{d_1} Y^\top Y$ around Φ under the Frobenius norm is exactly of order $n/\sqrt{d_1}$.

Based on the foregoing estimation, we have the following lower bound on the smallest eigenvalue of the empirical conjugate kernel, denoted by $\lambda_{\min}\left(\frac{1}{d_1} Y^\top Y \right)$.

Theorem 67. *Suppose Assumptions 8 and 9 hold and σ is not a linear function, X is (ε_n, B)-orthonormal. Then with probability at least $1-4e^{-2n}$,*

$$\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right) \geq 1-\sum_{i=1}^{3}\zeta_i(\sigma)^2 - C_B\varepsilon_n^2\sqrt{n} - C\left(\sqrt{\frac{n}{d_1}}+\frac{n}{d_1}\right)\lambda_\sigma^2 B^2,$$

where C_B is a constant depending on B. In particular, if $\varepsilon_n^4 n = o(1), B = O(1), d_1 = \omega(n)$, then with high probability,

$$\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right) \geq 1-\sum_{i=1}^{3}\zeta_i(\sigma)^2 - o(1).$$

A related result in [OS20, Theorem 5] assumed $\|\mathbf{x}_j\| = 1$ for all $j \in [n]$, $\lambda_\sigma \leq B, |\sigma(0)| \leq B$, $d_1 \geq C\log^2(n)\frac{n}{\lambda_{\min}(\Phi)}$ and obtained $\frac{1}{d_1}\lambda_{\min}(Y^\top Y) \geq \lambda_{\min}(\Phi) - o(1)$. We relax the assumption on the column vectors of X, and extend the range of d_1 down to $d_1 = \Omega(n)$, to guarantee that $\frac{1}{d_1}\lambda_{\min}(Y^\top Y)$ is lower bounded by an absolute constant, with an extra assumption that $\mathbb{E}[\sigma(\xi)] = 0$. This assumption can always be satisfied by shifting the activation function with a proper constant. Our bound for $\lambda_{\min}(\Phi)$ is derived via Hermite polynomial expansion, similar to [OS20, Lemma 15]. However, we apply an ε-net argument for concentration bound for $\frac{1}{d_1}Y^\top Y$ around Φ, while a matrix Chernoff concentration bound with truncation was used in [OS20, Theorem 5].

Additionally, the concentration for the NTK matrix H can be obtained in the next theorem. Recall that H is defined by (4.2.4) and the columns of S are defined by (4.2.5) with Assumption 8.

Theorem 68. *Suppose $d_1 \geq \log n$, and σ is λ_σ-Lipschitz. Then with probability at least $1-n^{-7/3}$,*

$$\left\|\frac{1}{d_1}(S^\top S - \mathbb{E}[S^\top S]) \odot (X^\top X)\right\| \leq 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}. \tag{4.3.12}$$

Moreover, if the assumptions in Theorem 66 hold, then with probability at least $1-n^{-7/3}-4e^{-2n}$,

$$\|H - \mathbb{E}H\| \leq C\left(\sqrt{\frac{n}{d_1}}+\frac{n}{d_1}\right)\lambda_\sigma^2\|X\|^2 + 32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{d_1}} + 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}. \tag{4.3.13}$$

Compared to Proposition D.3 in [HXAP20], we assume $\boldsymbol{a}$ is a Gaussian vector instead of a Rademacher random vector and attain a better bound. If $a_i \in \{+1,-1\}$, then one can apply matrix Bernstein inequality for the sum of bounded random matrices. In our case, the boundedness condition is not satisfied. Section S1.1 in [AP20] applied matrix Bernstein inequality for the sum of bounded random matrices when $\boldsymbol{a}$ is a Gaussian vector, but the boundedness condition does not hold in Equation (S7) of [AP20].

Based on Theorem 68, we get a lower bound for the smallest eigenvalue of the NTK.

Theorem 69. *Under Assumptions 8 and 9, suppose that X is (ε_n, B)-orthonormal, σ is not a linear function, and $d_1 \geq \log n$. Then with probability at least $1 - n^{-7/3}$,*

$$\lambda_{\min}(H) \geq a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma) - C_B \varepsilon_n^4 n - 10\lambda_\sigma^4 B^4 \sqrt{\frac{\log n}{d_1}},$$

where C_B is a constant depending only on B, and $\eta_k(\sigma)$ is defined in (4.3.8)*. In particular, if $\varepsilon_n^4 n = o(1)$, $B = O(1)$, and $d_1 = \omega(\log n)$, then with high probability,*

$$\lambda_{\min}(H) \geq \left(a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma) \right)(1 - o(1)).$$

We relax the assumption in [NMM21] to $d_1 = \omega(\log n)$ for the 2-layer case and our result is applicable beyond the ReLU activation function and to more general assumptions on X. Our proof strategy is different from [NMM21]. In [NMM21], the authors used the inequality $\lambda_{\min}((S^\top S) \odot (X^\top X)) \geq \min_i \|S_i\|^2 \cdot \lambda_{\min}(X^\top X)$ where S_i is the i-th column of S. Then, getting the lower bound is reduced to show the concentration of the 2-norm of the column vectors of S. Here we apply a matrix concentration inequality to $(S^\top S) \odot (X^\top X)$ and gain a weaker assumption on d_1 to ensure the lower bound on $\lambda_{\min}(H)$.

Remark. *In Theorems 67 and 69, we exclude the linear activation function. When $\sigma(x) = x$, it is easy to check both $\frac{1}{d_1}\lambda_{\min}(Y^\top Y)$ and $\lambda_{\min}(H)$ will trivially determined by $\lambda_{\min}(X^\top X)$, which can be vanishing. In this case, the lower bounds of the smallest eigenvalues of CK and NTK rely*

on the assumption of μ_0 or the distribution of X. For instance, when the entries of X are i.i.d. Gaussian random variables, $\lambda_{\min}(X^\top X)$ has been analyzed in [Sil85].

4.3.3 Training and Test Errors for Random Feature Regression

We apply the results of the preceding sections to a two-layer neural network at random initialization defined in (4.0.1), to estimate the training errors and test errors with mean-square losses for random feature regression under the ultra-wide regime where $d_1/n \to \infty$ and $n \to \infty$. In this model, we take the random feature $\frac{1}{\sqrt{d_1}}\sigma(WX)$ and consider the regression with respect to $\theta \in \mathbb{R}^{d_1}$ based on

$$f_\theta(X) := \frac{1}{\sqrt{d_1}}\theta^\top \sigma(WX),$$

with training data $X \in \mathbb{R}^{d_0 \times n}$ and training labels $y \in \mathbb{R}^n$. Considering the ridge regression with ridge parameter $\lambda \geq 0$ and squared loss defined by

$$L(\theta) := \|f_\theta(X)^\top - y\|^2 + \lambda\|\theta\|^2, \tag{4.3.14}$$

we can conclude that the minimization $\hat{\theta} := \arg\min_\theta L(\theta)$ has an explicit solution

$$\hat{\theta} = \frac{1}{\sqrt{d_1}} Y \left(\frac{1}{d_1} Y^\top Y + \lambda \,\mathrm{Id}\right)^{-1} y, \tag{4.3.15}$$

where $Y = \sigma(WX)$ is defined in (4.2.1). When σ is nonlinear, by Theorem 67, it is feasible to take inverse in (4.3.15) for any $\lambda \geq 0$. Hence, in the following results, we will focus on *nonlinear* activation functions[1]. In general, the optimal predictor for this random feature with respect to (4.3.14) is

$$\hat{f}_\lambda^{(RF)}(\mathbf{x}) := \frac{1}{\sqrt{d_1}}\hat{\theta}^\top \sigma(W\mathbf{x}) = K_n(\mathbf{x}, X)(K_n(X,X) + \lambda \,\mathrm{Id})^{-1} y, \tag{4.3.16}$$

[1]As Remark 4.3.2 stated, when $\sigma(x) = x$, $\lambda_{\min}$ of CK may be possibly vanishing. To include the linear activation function, we can alternatively assume that the ridge parameter λ is *strictly* positive and focus on random feature *ridge* regressions.

where we define an empirical kernel $K_n(\cdot,\cdot) : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ as

$$K_n(\mathbf{x},\mathbf{z}) := \frac{1}{d_1}\sigma(W\mathbf{x})^\top \sigma(W\mathbf{z}) = \frac{1}{d_1}\sum_{i=1}^{d_1} \sigma(\langle \boldsymbol{w}_i, \mathbf{x}\rangle)\sigma(\langle \boldsymbol{w}_i, \mathbf{z}\rangle). \tag{4.3.17}$$

The n-dimension row vector is given by

$$K_n(\mathbf{x}, X) = [K_n(\mathbf{x},\mathbf{x}_1), \ldots, K_n(\mathbf{x},\mathbf{x}_n)], \tag{4.3.18}$$

and the (i,j) entry of $K_n(X,X)$ is defined by $K_n(\mathbf{x}_i,\mathbf{x}_j)$, for $1 \le i,j \le n$.

Analogously, consider any kernel function $K(\cdot,\cdot) : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$. The optimal kernel predictor with a ridge parameter $\lambda \ge 0$ for the kernel ridge regression is given by (see [RR07, AKM$^+$17, LR20, JSS$^+$20, LLS21, BMR21] for more details)

$$\hat{f}_\lambda^{(K)}(\mathbf{x}) := K(\mathbf{x}, X)(K(X,X) + \lambda\,\mathrm{Id})^{-1} y, \tag{4.3.19}$$

where $K(X,X)$ is an $n \times n$ matrix such that its (i,j) entry is $K(\mathbf{x}_i,\mathbf{x}_j)$, and $K(\mathbf{x},X)$ is a row vector in $\mathbb{R}^n$ similarly with (4.3.18). We compare the characteristics of the two different predictors $\hat{f}_\lambda^{(RF)}(\mathbf{x})$ and $\hat{f}_\lambda^{(K)}(\mathbf{x})$ when the kernel function K is defined in (4.1.3). Denote the optimal predictors for random features and kernel K on training data X by

$$\hat{f}_\lambda^{(RF)}(X) = \left(\hat{f}_\lambda^{(RF)}(\mathbf{x}_1), \ldots, \hat{f}_\lambda^{(RF)}(\mathbf{x}_n)\right)^\top,$$
$$\hat{f}_\lambda^{(K)}(X) = \left(\hat{f}_\lambda^{(K)}(\mathbf{x}_1), \ldots, \hat{f}_\lambda^{(K)}(\mathbf{x}_n)\right)^\top,$$

respectively. Notice that, in this case, $K(X,X) \equiv \Phi$ defined in (4.1.2) and $K_n(X,X)$ is the random empirical CK matrix $\frac{1}{d_1}Y^\top Y$ defined in (4.2.1).

We aim to compare the training and test errors for these two predictors in ultra-wide

random neural networks, respectively. Let *training errors* of these two predictors be

$$E_{\text{train}}^{(K,\lambda)} := \frac{1}{n}\|\hat{f}_{\lambda}^{(K)}(X) - y\|^2 = \frac{\lambda^2}{n}\|(K(X,X) + \lambda\,\mathrm{Id})^{-1}y\|^2, \tag{4.3.20}$$

$$E_{\text{train}}^{(RF,\lambda)} := \frac{1}{n}\|\hat{f}_{\lambda}^{(RF)}(X) - y\|^2 = \frac{\lambda^2}{n}\|(K_n(X,X) + \lambda\,\mathrm{Id})^{-1}y\|^2. \tag{4.3.21}$$

In the following theorem, we show that, with high probability, the training error of the random feature regression model can be approximated by the corresponding kernel regression model with the same ridge parameter $\lambda \geq 0$ for ultra-wide neural networks.

Theorem 70 (Training error approximation)**.** *Suppose Assumptions 8, 9 and 10 hold, and σ is not a linear function. Then, for all large n, with probability at least $1 - 4e^{-2n}$,*

$$\left|E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)}\right| \leq \frac{C_1}{\sqrt{nd_1}}\left(\sqrt{\frac{n}{d_1}} + C_2\right)\|y\|^2, \tag{4.3.22}$$

where constants C_1 and C_2 only depend on λ, B and σ.

Next, to investigate the test errors (or generalization errors), we introduce further assumptions on the data and the target function that we want to learn from training data. Denote the true regression function by $f^* : \mathbb{R}^{d_0} \to \mathbb{R}$. Then, the training labels are defined by

$$y = f^*(X) + \boldsymbol{\varepsilon} \quad \text{and} \quad f^*(X) = (f^*(\mathbf{x}_1), \ldots, f^*(\mathbf{x}_n))^\top, \tag{4.3.23}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the training label noise. For simplicity, we further impose the following assumptions, analogously to [LD21].

Assumption 11. Assume that the target function is a linear function $f^*(\mathbf{x}) = \langle \boldsymbol{\beta}^*, \mathbf{x}\rangle$, where random vector satisfies $\boldsymbol{\beta}^* \sim \mathcal{N}(0, \sigma_{\boldsymbol{\beta}}^2\,\mathrm{Id})$. Then, in this case, the training label vector is given by $y = X^\top\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\varepsilon}^2\,\mathrm{Id})$ independent with $\boldsymbol{\beta}^* \in \mathbb{R}^{d_0}$.

Assumption 12. Suppose that training dataset

$$X = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d_0 \times n}$$

satisfies (ε_n, B)-orthonormal condition with $n\varepsilon_n^4 = o(1)$, and a test data $\mathbf{x} \in \mathbb{R}^{d_0}$ is independent with X and y such that $\tilde{X} := [\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}] \in \mathbb{R}^{d_0 \times (n+1)}$ is also (ε_n, B)-orthonormal. For convenience, we further assume the population covariance of the test data is $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] = \frac{1}{d_0}\,\mathrm{Id}$.

Remark. *Our Assumption 12 of the test data* $\mathbf{x}$ *ensures the same statistical behavior as training data in* X*, but we do not have any explicit assumption of the distribution of* $\mathbf{x}$*. It is promising to adopt such assumptions to handle statistical models with real-world data [LC18b, LCM20]. Besides, it is possible to extend our analysis to general population covariance for* $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top]$.

For any predictor $\hat{f}$, define the *test error* (generalization error) by

$$\mathscr{L}(\hat{f}) := \mathbb{E}_{\mathbf{x}}[|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})|^2]. \tag{4.3.24}$$

We first present the following approximation of the test error of a random feature predictor via its corresponding kernel predictor.

Theorem 71 (Test error approximation)**.** *Suppose that Assumptions 8, 9, 11 and 12 hold, and* σ *is not a linear function. Then, for any* $\varepsilon \in (0, 1/2)$*, the difference of test errors satisfies*

$$\left| \mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x})) - \mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x})) \right| = o\Big((n/d_1)^{\frac{1}{2}-\varepsilon} \Big), \tag{4.3.25}$$

with probability $1 - o(1)$*, when* $n/d_1 \to 0$ *and* $n \to \infty$.

Theorems 70 and 71 verify that the random feature regression achieves the same asymptotic errors as the kernel regression, as long as $n/d_1 \to \infty$. This is closely related to [MMM21, Theorem 1] with different settings. Based on that, we can compute the asymptotic training and

test errors for the random feature model by calculating the corresponding quantities for the kernel regression in the ultra-wide regime where $n/d_1 \to 0$.

Theorem 72 (Asymptotic training and test errors)**.** *Suppose Assumptions 8 and 9 hold, and σ is not a linear function. Suppose the target function f^*, training data X and test data $\mathbf{x} \in \mathbb{R}^{d_0}$ satisfy Assumptions 11 and 12. For any $\lambda \geq 0$, let the effective ridge parameter be*

$$\lambda_{\text{eff}}(\lambda, \sigma) := \frac{1 + \lambda - b_\sigma^2}{b_\sigma^2}. \tag{4.3.26}$$

If the training data has some limiting eigenvalue distribution $\mu_0 = \lim \text{spec} X^\top X$ as $n \to \infty$ and $n/d_0 \to \gamma \in (0, \infty)$, then when $n/d_1 \to 0$ and $n \to \infty$, the training error satisfies

$$E_{\text{train}}^{(RF,\lambda)} \xrightarrow{\mathbb{P}} \frac{\sigma_{\boldsymbol{\beta}}^2 \lambda^2}{\gamma b_\sigma^4} \mathscr{V}_K(\lambda_{\text{eff}}(\lambda, \sigma)) + \frac{\sigma_\varepsilon^2 \lambda^2}{\gamma(1 + \lambda - b_\sigma^2)^2} (\mathscr{B}_K(\lambda_{\text{eff}}(\lambda, \sigma)) - 1 + \gamma), \tag{4.3.27}$$

and the test error satisfies

$$\mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x})) \xrightarrow{\mathbb{P}} \sigma_{\boldsymbol{\beta}}^2 \mathscr{B}_K(\lambda_{\text{eff}}(\lambda, \sigma)) + \sigma_\varepsilon^2 \mathscr{V}_K(\lambda_{\text{eff}}(\lambda, \sigma)), \tag{4.3.28}$$

where the bias and variance functions are defined by

$$\mathscr{B}_K(\nu) := (1 - \gamma) + \gamma \nu^2 \int_{\mathbb{R}} \frac{1}{(x + \nu)^2} d\mu_0(x), \tag{4.3.29}$$

$$\mathscr{V}_K(\nu) := \gamma \int_{\mathbb{R}} \frac{x}{(x + \nu)^2} d\mu_0(x). \tag{4.3.30}$$

We emphasize that in the proof of Theorem 72, we also get n-dependent deterministic equivalents for training/test errors of the kernel regression to approximate the performance of random feature regression. This is akin to [LCM20, Theorem 3] and [BMR21, Theorem 4.13], but in different regimes. In the following Figure 4.3, we present implementations of test errors for random feature regressions on standard Gaussian random data as input X and their limits

(4.3.28). In Figure 4.3, regularization parameters are $\lambda = 10^{-3}$ (left) and $\lambda = 10^{-6}$ (right). Here, the activation function σ is a re-scaled Sigmoid function, $\sigma_\varepsilon = 1$ and $\sigma_\beta = 2$. We fix $d_0 = 500$, varying values of sample sizes n and widths d_1. In other words, we fix n, d_0, only let $d_1 \to \infty$, and use empirical spectral distribution of $X^\top X$ to approximate μ_0 in $\mathscr{B}_K(\lambda_{\text{eff}}(\lambda,\sigma))$ and $\mathscr{V}_K(\lambda_{\text{eff}}(\lambda,\sigma))$, which is actually the n-dependent deterministic equivalent. However, for Gaussian random matrix X, μ_0 is actually a Marčenko-Pastur law with ratio γ, so $\mathscr{B}_K(\lambda_{\text{eff}}(\lambda,\sigma))$ and $\mathscr{V}_K(\lambda_{\text{eff}}(\lambda,\sigma))$ can be computed explicitly according to [LD21, Definition 1]. In Figure 4.3, test errors in solid lines with error bars are computed using an independent test set of size 5000. We average our results over 50 repetitions. Limiting test errors in black dash lines are computed by (4.3.28), and we take μ_0 to be the corresponding Marčenko–Pastur distributions.

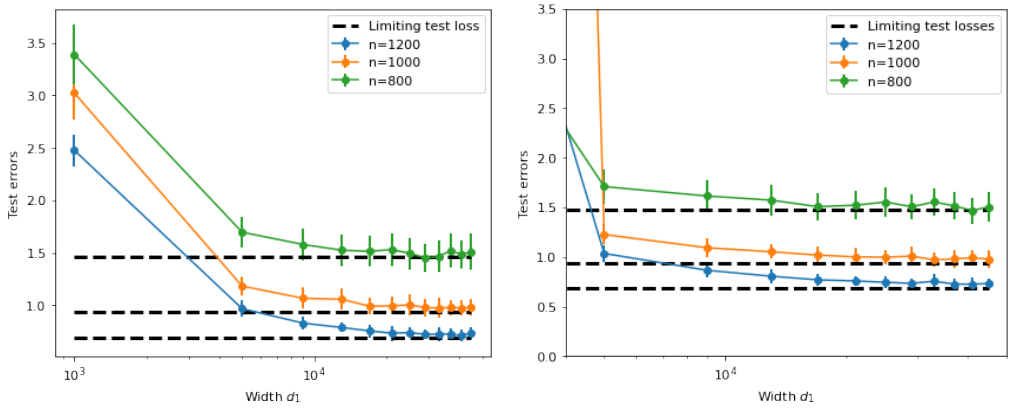


Figure 4.3. Simulations for the test errors of random feature regressions with centered Gaussian random matrix as input X and regularization parameter $\lambda = 10^{-3}$ (left) and $\lambda = 10^{-6}$ (right). Limiting test errors in black dash lines are computed by (4.3.28), and we take μ_0 to be the corresponding Marčenko–Pastur distributions.

Remark (Implicit regularization)**.** *For nonlinear σ, the effective ridge parameter* (4.3.26) *can be viewed as an inflated ridge parameter since $b_\sigma^2 \in [0,1)$ and $\lambda_{\text{eff}} > \lambda \geq 0$. This λ_{eff} leads to* implicit regularization *for our random feature and kernel ridge regressions even for the ridgeless regression with $\lambda = 0$ [LR20, MZ20, JSS+20, BMR21]. This effective ridge parameter λ_{eff} also shows the effect of the nonlinearity in the random feature and kernel regressions induced by ultra-wide neural networks.*

For convenience, we only consider the linear target function f^*, but in general, the above theorems can also be obtained for nonlinear target functions, for instance, f^* is a nonlinear single-index model. Under (ε_n, B)-orthonormal assumption with $n\varepsilon_n^4 \to 0$, our expected kernel $K(X,X) \equiv \Phi$ is approximated in terms of

$$\lim \operatorname{spec} K(X,X) = \lim \operatorname{spec} \left(b_\sigma^2 X^\top X + (1 - b_\sigma^2) \operatorname{Id} \right), \tag{4.3.31}$$

whence, this kernel regression can only learn linear functions. So if f^* is nonlinear, the limiting test error should be decomposed into the linear part as (4.3.28) and the nonlinear component as a noise [BMR21, Theorem 4.13]. For more conclusions of this kernel machine, we refer to [LR20, LRZ20, LLS21, MMM21].

4.3.4 Neural Tangent Kernel Regression

In parallel to the above results, we can obtain a similar analysis of the limiting training and test errors for random feature regression in (4.3.16) with empirical NTK given by either $K_n(X,X) = \frac{1}{d_1}(S^\top S) \odot (X^\top X)$ or $K_n(X,X) = H$. This random feature regression also refers to *neural tangent regression* [MZ20]. With the help of our concentration results in Theorem 68 and the lower bound of the smallest eigenvalues in Theorem 69, we can directly extend the above Theorems 70, 71 and 72 to this neural tangent regression. We omit the proofs in these cases and only state the results as follows.

If $K_n(X,X) = \frac{1}{d_1}(S^\top S) \odot (X^\top X)$ with expected kernel $K(X,X) = \Psi$ defined by (4.3.6), the limiting training and test errors of this neural tangent regression can be approximated by the kernel regression with respect to Ψ, as long as $d_1 = \omega(\log n)$. Analogously to (4.3.31), we have an additional approximation

$$\lim \operatorname{spec} \Psi = \lim \operatorname{spec} \left(b_\sigma^2 X^\top X + (a_\sigma - b_\sigma^2) \operatorname{Id} \right). \tag{4.3.32}$$

Under the same assumptions of Theorem 72 and replacing $n/d_1 \to 0$ with $d_1 = \omega(\log n)$, we can conclude that the test error of this neural tangent regression has the same limit as (4.3.28) but changing the effective ridge parameter (4.3.26) into $\lambda_{\text{eff}}(\lambda,\sigma) = \frac{a_\sigma+\lambda-b_\sigma^2}{b_\sigma^2}$. This result is akin to [MZ20, Corollary 3.2] but permits more general assumptions on X. The limiting training error of this neural tangent regression can be obtained by slightly modifying the coefficient in (4.3.27).

Similarly, if $K_n(X,X) = H$ defined by (4.2.4) possesses an expected kernel $K(X,X) = \Phi + \Psi$, this neural tangent regression in (4.3.16) is close to kernel regression (4.3.19) with kernel

$$K(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \mathbf{x})\sigma(\boldsymbol{w}^\top \mathbf{x})] + \mathbb{E}_{\boldsymbol{w}}[\sigma'(\boldsymbol{w}^\top \mathbf{x})\sigma'(\boldsymbol{w}^\top \mathbf{x})]\mathbf{x}^\top \mathbf{z},$$

under the ultra-wide regime, $n/d_1 \to 0$. Combining (4.3.31) and (4.3.32), Theorem 72 can directly be extended to this neural tangent regression but replacing (4.3.26) with $\lambda_{\text{eff}}(\lambda,\sigma) = \frac{a_\sigma+1+\lambda-2b_\sigma^2}{2b_\sigma^2}$. Section 6.1 of [AP20] also calculated this limiting test error when data X is isotropic Gaussian.

4.4 A Non-linear Hanson-Wright Inequality

We give an improved version of Lemma 1 in [LLC18] with a simple proof based on a Hanson-Wright inequality for random vectors with dependence [Ada15]. This serves as the concentration tool for us to prove the deformed semicircle law in Section 4.6 and provide bounds on extreme eigenvalues in Section 4.7. First, we define some concentration properties for random vectors.

Definition 73 (Concentration property)**.** Let X be a random vector in $\mathbb{R}^n$. We say X has the *K-concentration property* with constant K if for any 1-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}|f(X)| < \infty$ and for any $t > 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp(-t^2/K^2). \tag{4.4.1}$$

There are many distributions of random vectors satisfying K-concentration property, including uniform random vectors on the sphere, unit ball, hamming or continuous cube, uniform random permutation, etc. See [Ver18, Chapter 5] for more details.

Definition 74 (Convex concentration property)**.** Let X be a random vector in $\mathbb{R}^n$. We say X has the *K-convex concentration property* with the constant K if for any 1-Lipschitz convex function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\mathbb{E}|f(X)| < \infty$ and for any $t > 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp(-t^2/K^2).$$

We will apply the following result from [Ada15] to the nonlinear setting.

Lemma 75 (Theorem 2.5 in [Ada15])**.** *Let X be a mean zero random vector in $\mathbb{R}^n$. If X has the K-convex concentration property, then for any $n \times n$ matrix A and any $t > 0$,*

$$\mathbb{P}(|X^\top AX - \mathbb{E}(X^\top AX)| \geq t) \leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2}{2K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\right\}\right)$$

for some universal constant $C > 1$.

Theorem 76. *Let $\boldsymbol{w} \in \mathbb{R}^{d_0}$ be a random vector with K-concentration property, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be a deterministic matrix. Define $\mathbf{y} = \sigma(\boldsymbol{w}^\top X)^\top$, where σ is λ_σ-Lipschitz, and $\Phi = \mathbb{E}\mathbf{y}\mathbf{y}^\top$. Let A be an $n \times n$ deterministic matrix.*

1. *If $\mathbb{E}[\mathbf{y}] = 0$, for any $t > 0$,*

$$\mathbb{P}\Big(|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi| \geq t\Big) \tag{4.4.2}$$

$$\leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2}{2K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\right\}\right), \tag{4.4.3}$$

where $C > 0$ is an absolute constant.

2. *If* $\mathbb{E}[\mathbf{y}] \neq 0$, *for any* $t > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi| > t\Big) \\
&\leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{4K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg) \\
&\quad + 2\exp\bigg(-\frac{t^2}{16K^2\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\bigg).
\end{aligned}$$

for some constant $C > 0$.

Proof. Let f be any 1-Lipschitz convex function. Since $\mathbf{y} = \sigma(\boldsymbol{w}^\top X)^\top$, $f(\mathbf{y}) = f(\sigma(\boldsymbol{w}^\top X)^\top)$ is a $\lambda_\sigma\|X\|$-Lipschitz function of $\boldsymbol{w}$. Then by the Lipschitz concentration property of $\boldsymbol{w}$ in (4.4.1), we obtain

$$\mathbb{P}(|f(\mathbf{y}) - \mathbb{E}f(\mathbf{y})| \geq t) \leq 2\exp\bigg(-\frac{t^2}{K^2\lambda_\sigma^2\|X\|^2}\bigg).$$

Therefore, $\mathbf{y}$ satisfies the $K\lambda_\sigma\|X\|$-convex concentration property. Define $\tilde{f}(\mathbf{x}) = f(\mathbf{x} - \mathbb{E}\mathbf{y})$, then $\tilde{f}$ is also a convex 1-Lipschitz function and $\tilde{f}(\mathbf{y}) = f(\mathbf{y} - \mathbb{E}\mathbf{y})$. Hence $\tilde{\mathbf{y}} := \mathbf{y} - \mathbb{E}\mathbf{y}$ also satisfies the $K\lambda_\sigma\|X\|$-convex concentration property. Applying Theorem 75 to $\tilde{\mathbf{y}}$, we have for any $t > 0$,

$$\mathbb{P}(|\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})| \geq t) \leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{2K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg). \quad (4.4.4)$$

Since $\mathbb{E}\tilde{\mathbf{y}} = 0$, the inequality above implies (4.4.2). Note that

$$\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}}) = (\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi) - \tilde{\mathbf{y}}^\top A\mathbb{E}\mathbf{y} - \mathbb{E}\mathbf{y}^\top A\tilde{\mathbf{y}},$$

Hence,

$$\begin{aligned}\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi &= (\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})) + (\mathbf{y}-\mathbb{E}\mathbf{y})^\top (A+A^\top)\mathbb{E}\mathbf{y} \\ &= (\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})) + (\mathbf{y}^\top (A+A^\top)\mathbb{E}\mathbf{y} - \mathbb{E}\mathbf{y}^\top (A+A^\top)\mathbb{E}\mathbf{y}). \end{aligned} \tag{4.4.5}$$

Since $\mathbf{y}^\top (A+A^\top)\mathbb{E}\mathbf{y}$ is a $(2\|A\|\|\mathbb{E}\mathbf{y}\|\|X\|\lambda_\sigma)$-Lipschitz function of $\boldsymbol{w}$, by the Lipschitz concentration property of $\boldsymbol{w}$, we have

$$\mathbb{P}(|(\mathbf{y}-\mathbb{E}\mathbf{y})^\top (A+A^\top)\mathbb{E}\mathbf{y}| \geq t) \leq 2\exp\left(-\frac{t^2}{4K^2(\|A\|\|\mathbb{E}\mathbf{y}\|\|X\|\lambda_\sigma)^2}\right). \tag{4.4.6}$$

Then combining (4.4.4), (4.4.5), and (4.4.6), we have

$$\begin{aligned}&\mathbb{P}(|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi| \geq t) \\ &\leq \mathbb{P}(|\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}} - \mathbb{E}(\tilde{\mathbf{y}}^\top A\tilde{\mathbf{y}})| \geq t/2) + \mathbb{P}(|(\mathbf{y}-\mathbb{E}\mathbf{y})^\top (A+A^\top)\mathbb{E}\mathbf{y}| \geq t/2) \\ &\leq 2\exp\left(-\frac{1}{2C}\min\left\{\frac{t^2}{4K^4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{K^2\lambda_\sigma^2\|X\|^2\|A\|}\right\}\right) \\ &\quad + 2\exp\left(-\frac{t^2}{16K^2\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\right).\end{aligned} \tag{4.4.7}$$

This finishes the proof. □

Since the Gaussian random vector $\boldsymbol{w} \sim \mathcal{N}(0, I_{d_0})$ satisfies the K-concentration inequality with $K = \sqrt{2}$ (see for example [BLM13]), we have the following corollary.

Corollary 77. *Let $\boldsymbol{w} \sim \mathcal{N}(0, I_{d_0})$, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be a deterministic matrix. Define $\mathbf{y} = \sigma(\boldsymbol{w}^\top X)^\top$, where σ is λ_σ-Lipschitz, and $\Phi = \mathbb{E}\mathbf{y}\mathbf{y}^\top$. Let A be an $n \times n$ deterministic matrix.*

1. *If $\mathbb{E}[\mathbf{y}] = 0$, for any $t > 0$,*

$$\mathbb{P}\Big(|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi| \geq t\Big) \leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2}{4\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\right\}\right) \tag{4.4.8}$$

for some absolute constant $C > 0$.

2. *If $\mathbb{E}[\mathbf{y}] \neq 0$, for any $t > 0$,*

$$\mathbb{P}\Big(|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi| > t\Big) \leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{8\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg) \tag{4.4.9}$$

$$+2\exp\bigg(-\frac{t^2}{32\lambda_\sigma^2\|X\|^2\|A\|^2\|\mathbb{E}\mathbf{y}\|^2}\bigg)$$

$$\leq 2\exp\bigg(-\frac{1}{C}\min\bigg\{\frac{t^2}{8\lambda_\sigma^4\|X\|^4\|A\|_F^2}, \frac{t}{\lambda_\sigma^2\|X\|^2\|A\|}\bigg\}\bigg) \tag{4.4.10}$$

$$+2\exp\bigg(-\frac{t^2}{32\lambda_\sigma^2\|X\|^2\|A\|^2 t_0}\bigg), \tag{4.4.11}$$

where

$$t_0 := 2\lambda_\sigma^2\sum_{i=1}^n(\|\mathbf{x}_i\| - 1)^2 + 2n(\mathbb{E}\sigma(\xi))^2, \quad \xi \sim \mathscr{N}(0,1). \tag{4.4.12}$$

Remark. *Compared to [LLC18, Lemma 1], we identify the dependence on $\|A\|_F$ and $\mathbb{E}\mathbf{y}$ in the probability estimate. By* (1.4.1)*, we obtain a similar inequality to the one in [LLC18] with a better dependence on n. Moreover, our bound in t_0 is independent of d_0, while the corresponding term t_0 in [LLC18, Lemma 1] depends on $\|X\|$ and d_0. In particular, when $\mathbb{E}\sigma(\xi) = 0$ and X is (ε_n, B)-orthonormal, t_0 is of order 1. Hence,* (4.4.9) *with the special choice of t_0 is the key ingredient in the proof of Theorem 66 to get a concentration of the spectral norm for CK.*

Proof of Corollary 77. We only need to prove (4.4.9), since other statements follow immediately by taking $K = \sqrt{2}$. Let $\mathbf{x}_i$ be the i-th column of X. Then

$$\|\mathbb{E}\mathbf{y}\|^2 = \|\mathbb{E}\sigma(\boldsymbol{w}^\top X)\|^2 = \sum_{i=1}^n[\mathbb{E}\sigma(\boldsymbol{w}^\top\mathbf{x}_i)]^2.$$

Let $\xi \sim \mathcal{N}(0,1)$. We have

$$|\mathbb{E}\sigma(\boldsymbol{w}^\top \mathbf{x}_i)| = |\mathbb{E}\sigma(\xi\|\mathbf{x}_i\|)| \le \mathbb{E}|(\sigma(\xi\|\mathbf{x}_i\|) - \sigma(\xi))| + |\mathbb{E}\sigma(\xi)|$$
$$\le \lambda_\sigma \mathbb{E}|\xi(\|\mathbf{x}_i\| - 1)| + |\mathbb{E}\sigma(\xi)| \le \lambda_\sigma |\|\mathbf{x}_i\| - 1| + |\mathbb{E}\sigma(\xi)|. \tag{4.4.13}$$

Therefore

$$\|\mathbb{E}\mathbf{y}\|^2 \le \sum_{i=1}^n (\lambda_\sigma(\|\mathbf{x}_i\| - 1) + |\mathbb{E}\sigma(\xi)|)^2 \le \sum_{i=1}^n 2\lambda_\sigma^2(\|\mathbf{x}_i\| - 1)^2 + 2(\mathbb{E}\sigma(\xi))^2 \tag{4.4.14}$$
$$= 2\lambda_\sigma^2 \sum_{i=1}^n (\|\mathbf{x}_i\| - 1)^2 + 2n(\mathbb{E}\sigma(\xi))^2 = t_0,$$

and (4.4.9) holds. □

We include the following corollary about the variance of $\mathbf{y}^\top A\mathbf{y}$, which will be used in Section 4.6 to study the spectrum of the CK and NTK.

Corollary 78. *Under the same assumptions of Corollary 77, we further assume that $t_0 \le C_1 n$, and $\|A\|, \|X\| \le C_2$. Then as $n \to \infty$,*

$$\frac{1}{n^2}\mathbb{E}\left[\left|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi\right|^2\right] \to 0.$$

Proof. Notice that $\|A\|_F \le \sqrt{n}\|A\|$. Thanks to Theorem 77 (2), we have that for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\mathbf{y}^\top A\mathbf{y} - \mathrm{Tr}A\Phi\right| > t\right) \le 4\exp\left(-Cn\min\{t^2, t\}\right), \tag{4.4.15}$$

where constant $C > 0$ only relies on C_1, C_2, λ_σ, and K. Therefore, we can compute the variance

in the following way:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^2}\left|\mathbf{y}^\top A\mathbf{y}-\operatorname{Tr}A\Phi\right|^2\right] &= \int_0^\infty \mathbb{P}\left(\frac{1}{n^2}\left|\mathbf{y}^\top A\mathbf{y}-\operatorname{Tr}A\Phi\right|^2 > s\right)ds \\ &\le 4\int_0^\infty \exp\bigl(-Cn\min\{s,\sqrt{s}\}\bigr)ds \\ &= 4\int_0^1 \exp\bigl(-Cn\sqrt{s}\bigr)ds + 4\int_1^{+\infty}\exp(-Cns)ds \to 0,\end{aligned}$$

as $n\to\infty$. Here, we use the dominant convergence theorem for the first integral in the last line. □

4.5 Limiting Law for General Centered Sample Covariance Matrices

Independent of the subsequent sections, this section focuses on the generalized sample covariance matrix where the dimension of the feature is much smaller than the sample size. We will later interpret such sample covariance matrix specifically for our neural network applications. Under certain weak assumptions, we prove the limiting eigenvalue distribution of the normalized sample covariance matrix satisfies two self-consistent equations, which are subsumed into a deformed semicircle law. Our findings in this section demonstrate some degree of universality, indicating that they hold across various random matrix models and may have implications for other related fields.

Theorem 79. *Suppose $\mathbf{y}_1,\dots,\mathbf{y}_d\in\mathbb{R}^n$ are independent random vectors with the same distribution of a random vector $\mathbf{y}\in\mathbb{R}^n$. Assume that $\mathbb{E}[\mathbf{y}]=\boldsymbol{0}$, $\mathbb{E}[\mathbf{y}\mathbf{y}^\top]=\Phi_n\in\mathbb{R}^{n\times n}$, where Φ_n is a deterministic matrix whose limiting eigenvalue distribution is $\mu_\Phi\neq\delta_0$. Assume $\|\Phi_n\|\le C$ for some constant C. Define $A_n:=\sqrt{\frac{d}{n}}\bigl(\frac{1}{d}\sum_{i=1}^d\mathbf{y}_i\mathbf{y}_i^\top-\Phi_n\bigr)$ and $R(z):=(A_n-z\,\mathrm{Id})^{-1}$. For any*

$z \in \mathbb{C}^+$ and any deterministic matrices D_n with $\|D_n\| \leq C$, suppose that as $n, d \to \infty$ and $n/d \to 0$,

$$\operatorname{tr} R(z) D_n - \mathbb{E}[\operatorname{tr} R(z) D_n] \xrightarrow{\text{a.s.}} 0, \tag{4.5.1}$$

and

$$\frac{1}{n^2} \mathbb{E}\left[\left| \mathbf{y}^\top D_n \mathbf{y} - \operatorname{Tr} D_n \Phi_n \right|^2 \right] \to 0. \tag{4.5.2}$$

Then the empirical eigenvalue distribution of matrix A_n weakly converges to μ almost surely, whose Stieltjes transform $m(z)$ is defined by (1.2.7) *in Chapter 1 for each $z \in \mathbb{C}^+$, where $\beta(z) \in \mathbb{C}^+$ in* (1.2.7) *is the unique solution to* (1.2.7) *in Chapter 1. In particular, $\mu = \mu_s \boxtimes \mu_\Phi$.*

Remark. *In [Xie13], it was assumed that $\frac{d}{n^3} \mathbb{E}\left| \mathbf{y}^\top D_n \mathbf{y} - \operatorname{Tr} D_n \Phi_n \right|^2 \to 0$, where $n^3/d \to \infty$ and $n/d \to 0$ as $n \to \infty$. By martingale difference, this condition implies* (4.5.1)*. However, we are not able to verify a certain step in the proof of [Xie13]. Hence, we will not directly adopt the result of [Xie13] but consider a more general situation without assuming $n^3/d \to \infty$. The weakest conditions we found are conditions* (4.5.1) *and* (4.5.2)*, which can be verified in our nonlinear random model.*

The self-consistent equations we derived are consistent with the results in [Bao12, Xie13], where they studied the empirical spectral distribution of separable sample covariance matrices in the regime $n/d \to 0$ under different assumptions. When $n \to \infty$ and $n/d \to 0$, our goal is to prove that the Stieltjes transform $m_n(z)$ of the empirical eigenvalue distribution of A_n and $\beta_n(z) := \operatorname{tr}[R(z)\Phi_n]$ point-wisely converges to $m(z)$ and $\beta(z)$, respectively.

For the rest of this section, we first prove a series of lemmas to get n-dependent deterministic equivalents related to (1.2.7) and (1.2.8) and then deduce the proof of Theorem 79 at the end of this section. Recall $A_n = \sqrt{\frac{d}{n}}\left(\frac{1}{d}\sum_{i=1}^d \mathbf{y}_i \mathbf{y}_i^\top - \Phi_n\right)$, $R(z) = (A_n - z\,\mathrm{Id})^{-1}$, and $\mathbf{y}$ is a random vector independent of A_n with the same distribution of $\mathbf{y}_i$.

Lemma 80. *Under the assumptions of Theorem 79, for any $z \in \mathbb{C}^+$, as $d, n \to \infty$,*

$$\operatorname{tr} D + z\mathbb{E}[\operatorname{tr} R(z)D] + \mathbb{E}\left[\frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y} \cdot \frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1+\sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}\right] = o(1), \tag{4.5.3}$$

where $D \in \mathbb{R}^{n\times n}$ is any deterministic matrix such that $\|D\| \le C$, for some constant C.

Proof. Let $z = u + iv \in \mathbb{C}^+$ where $u \in \mathbb{R}$ and $v > 0$. Let

$$\hat{R} := \left(\frac{1}{\sqrt{dn}}\sum_{j=1}^{d+1}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\right)^{-1},$$

where $\mathbf{y}_j$'s are independent copies of $\mathbf{y}$ defined in Theorem 79. Notice that, for any deterministic matrix $D \in \mathbb{R}^{n\times n}$,

$$D = \hat{R}\left(\frac{1}{\sqrt{dn}}\sum_{j=1}^{d+1}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\right)D \tag{4.5.4}$$

$$= \frac{1}{\sqrt{dn}}\hat{R}\left(\sum_{i=1}^{d+1}\mathbf{y}_i\mathbf{y}_i^\top\right)D - \sqrt{\frac{d}{n}}\hat{R}\Phi_n D - z\hat{R}D. \tag{4.5.5}$$

Without loss of generality, we assume $\|D\| \le 1$. Taking normalized trace, we have

$$\operatorname{tr} D + z\operatorname{tr}[\hat{R}D] = \frac{1}{\sqrt{dn}}\frac{1}{n}\sum_{i=1}^{d+1}\mathbf{y}_i^\top D\hat{R}\mathbf{y}_i - \sqrt{\frac{d}{n}}\operatorname{tr}[\hat{R}\Phi_n D]. \tag{4.5.6}$$

For each $1 \le i \le d+1$, Sherman–Morrison formula (Lemma 96) implies

$$\hat{R} = R^{(i)} - \frac{R^{(i)}\mathbf{y}_i\mathbf{y}_i^\top R^{(i)}}{\sqrt{dn} + \mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}, \tag{4.5.7}$$

where the leave-one-out resolvent $R^{(i)}$ is defined as

$$R^{(i)} := \left(\frac{1}{\sqrt{dn}}\sum_{1\le j\le d+1, j\ne i}\mathbf{y}_j\mathbf{y}_j^\top - \sqrt{\frac{d}{n}}\Phi_n - z\,\mathrm{Id}\right)^{-1}.$$

Hence, by (4.5.7), we obtain

$$\frac{1}{\sqrt{dn}}\frac{1}{n}\sum_{i=1}^{d+1}\mathbf{y}_i^\top D\hat{R}\mathbf{y}_i = \frac{1}{n}\sum_{i=1}^{d+1}\frac{\mathbf{y}_i^\top DR^{(i)}\mathbf{y}_i}{\sqrt{dn}+\mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}. \tag{4.5.8}$$

Combining equations (4.5.6) and (4.5.8), and applying expectation at both sides implies

$$\begin{aligned}\operatorname{tr} D + z\mathbb{E}[\operatorname{tr}\hat{R}D] &= \frac{1}{n}\sum_{i=1}^{d+1}\mathbb{E}\left[\frac{\mathbf{y}_i^\top DR^{(i)}\mathbf{y}_i}{\sqrt{dn}+\mathbf{y}_i^\top R^{(i)}\mathbf{y}_i}\right] - \sqrt{\frac{d}{n}}\mathbb{E}\operatorname{tr}\hat{R}\Phi_n D \\ &= \frac{d+1}{n}\mathbb{E}\left[\frac{\mathbf{y}^\top DR(z)\mathbf{y}}{\sqrt{dn}+\mathbf{y}^\top R(z)\mathbf{y}}\right] - \sqrt{\frac{d}{n}}\mathbb{E}\operatorname{tr}\hat{R}\Phi_n D, \end{aligned}\tag{4.5.9}$$

because of the assumption that all $\mathbf{y}_i$'s have the same distribution as vector $\mathbf{y}$ for all $i \in [d+1]$. With (4.5.9), to prove (4.5.3), we will first show that when $n, d \to \infty$,

$$\sqrt{\frac{d}{n}}\big(\mathbb{E}[\operatorname{tr}\hat{R}\Phi_n D] - \mathbb{E}[\operatorname{tr}R(z)\Phi_n D]\big) = o(1), \tag{4.5.10}$$

$$\mathbb{E}[\operatorname{tr}\hat{R}D] - \mathbb{E}[\operatorname{tr}R(z)D] = o(1), \tag{4.5.11}$$

$$\frac{1}{n}\mathbb{E}\left[\frac{\mathbf{y}^\top DR(z)\mathbf{y}}{\sqrt{dn}+\mathbf{y}^\top R(z)\mathbf{y}}\right] = o(1). \tag{4.5.12}$$

Recall that

$$\hat{R} - R(z) = \frac{1}{\sqrt{dn}}R(z)\left(\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top\right)\hat{R},$$

and spectral norms $\|\hat{R}\|, \|R(z)\| \le 1/v$ due to Proposition 12. Notice that $\|\Phi_n\| \le C$. Hence, we can deduce that

$$\begin{aligned}\sqrt{\frac{d}{n}}\big|\mathbb{E}[\operatorname{tr}\hat{R}\Phi_n D] - \mathbb{E}[\operatorname{tr}R(z)\Phi_n D]\big| &\le \frac{1}{n}\mathbb{E}[|\operatorname{tr}R(z)\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top\hat{R}\Phi_n D|] \\ &\le \frac{1}{n^2}\mathbb{E}[\|\hat{R}\Phi_n DR(z)\|\cdot\|\mathbf{y}_{d+1}\|^2] \\ &= \frac{C}{v^2n^2}\mathbb{E}[\operatorname{Tr}\mathbf{y}_{d+1}\mathbf{y}_{d+1}^\top] = \frac{C\operatorname{Tr}\Phi_n}{v^2n^2} \le \frac{C^2}{v^2n} \to 0,\end{aligned}$$

as $n \to \infty$. The same argument can be applied to the error of $\mathbb{E}[\operatorname{tr}\hat{R}D] - \mathbb{E}[\operatorname{tr}R(z)D]$. Therefore

(4.5.10) and (4.5.11) hold. For (4.5.12), we denote $\tilde{\mathbf{y}} := \mathbf{y}/(nd)^{1/4}$ and observe that

$$\frac{1}{n}\mathbb{E}\left[\frac{\mathbf{y}^\top DR(z)\mathbf{y}}{\sqrt{dn}+\mathbf{y}^\top R(z)\mathbf{y}}\right] = \frac{1}{n}\mathbb{E}\left[\frac{\tilde{\mathbf{y}}^\top DR(z)\tilde{\mathbf{y}}}{1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\right]. \tag{4.5.13}$$

Let $R(z) = \sum_{i=1}^n \frac{1}{\lambda_i - z}\mathbf{u}_i\mathbf{u}_i^\top$ be the eigen-decomposition of $R(z)$. Then

$$\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}/\|\tilde{\mathbf{y}}\|^2 = \sum_{i=1}^n \frac{1}{\lambda_i - z}\frac{(\langle \mathbf{u}_i, \tilde{\mathbf{y}}\rangle)^2}{\|\tilde{\mathbf{y}}\|^2} := \int \frac{1}{x-z} d\mu_{\tilde{\mathbf{y}}} \tag{4.5.14}$$

is the Stieltjes transform of a discrete measure $\mu_{\tilde{\mathbf{y}}} = \sum_{i=1}^n \frac{(\langle \mathbf{u}_i, \tilde{\mathbf{y}}\rangle)^2}{\|\tilde{\mathbf{y}}\|^2}\delta_{\lambda_i}$. Then, we can control the real part of $\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}$ by Lemma 98:

$$\left|\mathrm{Re}(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})\right| \le v^{-1/2}\|\tilde{\mathbf{y}}\|\left(\mathrm{Im}(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})\right)^{1/2}. \tag{4.5.15}$$

We now separately consider two cases in the following:

- If the right-hand side of the above inequality (4.5.15) is at most $1/2$, then

$$\left|1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}\right| \ge \left|1+\mathrm{Re}(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})\right| \ge \frac{1}{2},$$

  which results in

$$\left|\frac{\tilde{\mathbf{y}}^\top DR(z)\tilde{\mathbf{y}}}{1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\right| \le \frac{C}{\sqrt{dn}}\|\mathbf{y}\|^2. \tag{4.5.16}$$

- When $v^{-1/2}\|\tilde{\mathbf{y}}\|\left(\mathrm{Im}(\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})\right)^{1/2} > 1/2$, we know that

$$\begin{aligned}\left|\frac{\tilde{\mathbf{y}}^\top DR(z)\tilde{\mathbf{y}}}{1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}}\right| &\le \frac{\|\tilde{\mathbf{y}}^\top D\|\|R(z)\tilde{\mathbf{y}}\|}{|\mathrm{Im}(1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}})|} = \frac{\|\tilde{\mathbf{y}}^\top D\|\|R(z)\tilde{\mathbf{y}}\|}{\tilde{\mathbf{y}}^\top \mathrm{Im}(R(z))\tilde{\mathbf{y}}} \\ &\le \frac{\|\tilde{\mathbf{y}}^\top D\|}{\left(v\tilde{\mathbf{y}}^\top \mathrm{Im}(R(z))\tilde{\mathbf{y}}\right)^{1/2}} \le \frac{2\|\tilde{\mathbf{y}}^\top D\|\|\tilde{\mathbf{y}}\|}{v} \le \frac{C\|\mathbf{y}\|^2}{v\sqrt{nd}},\end{aligned} \tag{4.5.17}$$

where we exploit the fact that (see also Equation (A.1.11) in [BS10])

$$\|R(z)\tilde{\mathbf{y}}\| = (\tilde{\mathbf{y}}^\top R(\bar{z})R(z)\tilde{\mathbf{y}})^{1/2} = \left(\frac{1}{v}\tilde{\mathbf{y}}^\top \operatorname{Im}(R(z))\tilde{\mathbf{y}}\right)^{1/2}.$$

Finally, combining (4.5.16) and (4.5.17) in the above two cases, we can conclude the asymptotic result (4.5.12) because $\mathbb{E}\|\mathbf{y}\|^2 = \operatorname{Tr}\Phi_n \leq Cn$ in terms of the assumptions of Theorem 79.

Then with (4.5.10), (4.5.11), and (4.5.12), we get

$$\operatorname{tr} D + z\mathbb{E}[\operatorname{tr} R(z)D] = \mathbb{E}\left[\frac{\sqrt{\frac{d}{n}}\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}}{1+\frac{1}{\sqrt{dn}}\mathbf{y}^\top R(z)\mathbf{y}} - \sqrt{\frac{d}{n}}\operatorname{tr} R(z)\Phi_n D\right] + o(1), \tag{4.5.18}$$

as $n \to \infty$. We utilize the notion $\mathbb{E}_{\mathbf{y}}$ to clarify the expectation only with respect to random vector $\mathbf{y}$, conditioning on other independent random variables. So the conditional expectation is $\mathbb{E}_{\mathbf{y}}\left[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\right] = \operatorname{tr} DR(z)\Phi_n$ and

$$\mathbb{E}\left[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\right] = \mathbb{E}\left[\mathbb{E}_{\mathbf{y}}\left[\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\right]\right] = \mathbb{E}[\operatorname{tr} R(z)\Phi_n D].$$

Therefore, based on (4.5.18), the conclusion (4.5.3) holds. □

Next, we apply the quadratic concentration condition (4.5.2) to simplify (4.5.3).

Lemma 81. *Under the assumptions of Theorem 79, condition* (4.5.2) *of Theorem 79 implies that*

$$\mathbb{E}\left[\frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\cdot\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1+\sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}\right] = \mathbb{E}\left[\frac{\operatorname{tr} DR(z)\Phi_n \operatorname{tr} R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr} R(z)\Phi_n}\right] + o(1), \tag{4.5.19}$$

for each $z \in \mathbb{C}^+$ and any deterministic matrix D with $\|D\| \leq C$.

Proof. Let us denote

$$\delta_n := \frac{\frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}\cdot\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}}{1+\sqrt{\frac{n}{d}}\frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y}} - \frac{\operatorname{tr} DR(z)\Phi_n \operatorname{tr} R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr} R(z)\Phi_n},$$

$$Q_1 := \frac{1}{n}\mathbf{y}^\top DR(z)\mathbf{y}, \quad Q_2 := \frac{1}{n}\mathbf{y}^\top R(z)\mathbf{y},$$

$\bar{Q}_1 := \mathbb{E}_{\mathbf{y}}[Q_1] = \operatorname{tr} DR(z)\Phi_n$, and $\bar{Q}_2 := \mathbb{E}_{\mathbf{y}}[Q_1] = \operatorname{tr} R(z)\Phi_n$. In other words, δ_n can be expressed by

$$\begin{aligned}
\delta_n &= \frac{Q_1 Q_2}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\bar{Q}_1\bar{Q}_2}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} \\
&= \frac{Q_1\left(Q_2+\sqrt{\frac{d}{n}}\right)}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\sqrt{\frac{d}{n}}Q_1}{1+\sqrt{\frac{n}{d}}Q_2} - \frac{\bar{Q}_1\left(\bar{Q}_2+\sqrt{\frac{d}{n}}\right)}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} + \frac{\sqrt{\frac{d}{n}}\bar{Q}_1}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} \\
&= \sqrt{\frac{d}{n}}(Q_1-\bar{Q}_1) + \frac{\sqrt{\frac{d}{n}}(\bar{Q}_1-Q_1)}{1+\sqrt{\frac{n}{d}}\bar{Q}_2} + \frac{\sqrt{\frac{n}{d}}Q_1\sqrt{\frac{d}{n}}(\bar{Q}_2-Q_2)}{\left(1+\sqrt{\frac{n}{d}}\bar{Q}_2\right)\left(1+\sqrt{\frac{n}{d}}Q_2\right)}.
\end{aligned}$$

Observe that $\mathbb{E}[\bar{Q}_i] = \mathbb{E}[Q_i]$ for $i = 1,2$. Thus, δ_n has the same expectation as the last term

$$\Delta_n := \frac{Q_1(\bar{Q}_2-Q_2)}{\left(1+\sqrt{\frac{n}{d}}\bar{Q}_2\right)\left(1+\sqrt{\frac{n}{d}}Q_2\right)},$$

since we can first take the expectation for $\mathbf{y}$ conditioning on the resolvent $R(z)$ and then take the expectation for $R(z)$. Besides, notice that $|\bar{Q}_1|, |\bar{Q}_2| \le \frac{C}{v}$ uniformly. Hence, $\sqrt{\frac{n}{d}}\bar{Q}_2$ converges to zero uniformly and there exists some constant $C > 0$ such that

$$\left|\frac{1}{1+\sqrt{\frac{n}{d}}\bar{Q}_2}\right| \le C, \tag{4.5.20}$$

for all large d and n. In addition, observe that

$$\frac{\sqrt{\frac{n}{d}}Q_1}{1+\sqrt{\frac{n}{d}}Q_2} = \frac{\tilde{\mathbf{y}}^\top DR(z)\tilde{\mathbf{y}}}{1+\tilde{\mathbf{y}}^\top R(z)\tilde{\mathbf{y}}},$$

where $\tilde{\mathbf{y}}$ is defined in the proof of Lemma 80. In terms of (4.5.16) and (4.5.17), we verify that

$$\left| \frac{Q_1}{1+\sqrt{\frac{n}{d}}Q_2} \right| \leq \frac{C\|\mathbf{y}\|^2}{n}, \tag{4.5.21}$$

where $C > 0$ is some constant depending on v. Next, recall that condition (4.5.2) exposes that

$$\mathbb{E}(Q_2 - \bar{Q}_2)^2 \to 0 \quad \text{and} \quad \mathbb{E}(\|\mathbf{y}\|^2/n - \operatorname{tr}\Phi_n)^2 \to 0 \tag{4.5.22}$$

as $n \to \infty$. The first convergence is derived by viewing $D_n = R(z)$ and taking expectation conditional on $R(z)$. To sum up, we can bound $|\Delta_n|$ based on (4.5.20) and (4.5.21) in the subsequent way:

$$|\Delta_n| \leq \frac{C\|\mathbf{y}\|^2}{n}|\bar{Q}_2 - Q_2| \leq C\big|\|\mathbf{y}\|^2/n - \operatorname{tr}\Phi_n\big| \cdot |\bar{Q}_2 - Q_2| + C|\operatorname{tr}\Phi_n| \cdot |\bar{Q}_2 - Q_2|.$$

Here, $|\operatorname{tr}\Phi_n| \leq \|\Phi_n\|$ and $\|\Phi_n\|$ is uniformly bounded by some constant. Then, by Hölder's inequality, (4.5.22) implies that $\mathbb{E}[|\Delta_n|] \to 0$, as n approaching to infinity. This concludes $\mathbb{E}[\delta_n] = \mathbb{E}[\Delta_n]$ converges to zero. □

Lemma 82. *Under assumptions of Theorem 79, we can conclude that*

$$\lim_{n,d\to\infty} (\operatorname{tr} D + z\mathbb{E}[\operatorname{tr} R(z)D] + \mathbb{E}[\operatorname{tr} DR(z)\Phi_n]\mathbb{E}[\operatorname{tr} R(z)\Phi_n]) = 0$$

holds for each $z \in \mathbb{C}^+$ and deterministic matrix D with uniformly bounded spectral norm.

Proof. Based on Lemma 80 and Lemma 81, (4.5.19) and (4.5.3) yield

$$\operatorname{tr} D + z\mathbb{E}[\operatorname{tr} R(z)D] + \mathbb{E}\left[\frac{\operatorname{tr} DR(z)\Phi_n \operatorname{tr} R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr} R(z)\Phi_n}\right] = o(1).$$

As $|\operatorname{tr} R(z)D|$ and $|\operatorname{tr} R(z)D\Phi_n|$ are bounded by some constants uniformly and almost surely, for

sufficiently large d and n, $|\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n| < 1/2$ and

$$\left|\mathbb{E}\left[\frac{\operatorname{tr}DR(z)\Phi_n \operatorname{tr}R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}\right] - \mathbb{E}[\operatorname{tr}DR(z)\Phi_n \operatorname{tr}R(z)\Phi_n]\right|$$
$$\leq \mathbb{E}\left[|\operatorname{tr}R(z)D| \cdot |\operatorname{tr}R(z)D\Phi_n| \cdot \left|\frac{\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}{1+\sqrt{\frac{n}{d}}\operatorname{tr}R(z)\Phi_n}\right|\right] \leq 2C\sqrt{\frac{n}{d}} \to 0,$$

as $n/d \to 0$. Hence,

$$\operatorname{tr}D + z\mathbb{E}[\operatorname{tr}R(z)D] + \mathbb{E}[\operatorname{tr}DR(z)\Phi_n \operatorname{tr}R(z)\Phi_n] = o(1). \tag{4.5.23}$$

Considering $D_n = \Phi_n$ in (4.5.1), we can get almost sure convergence for $\operatorname{tr}DR(z)\Phi_n \cdot (\operatorname{tr}R(z)\Phi_n - \mathbb{E}[\operatorname{tr}R(z)\Phi_n])$ to zero. Thus by dominated convergence theorem,

$$\lim_{n\to\infty} \mathbb{E}[\operatorname{tr}DR(z)\Phi_n \cdot (\operatorname{tr}R(z)\Phi_n - \mathbb{E}[\operatorname{tr}R(z)\Phi_n])] \to 0.$$

So we can replace the third term at the right-hand side of (4.5.23) with

$$\mathbb{E}[\operatorname{tr}DR(z)\Phi_n]\mathbb{E}[\operatorname{tr}R(z)\Phi_n]$$

to obtain the conclusion. □

Proof of Theorem 79. Fix any $z \in \mathbb{C}^+$. Denote the Stieltjes transform of empirical spectrum of A_n and its expectation by $m_n(z) := \operatorname{tr}R(z)$ and $\bar{m}_n(z) := \mathbb{E}[m_n(z)]$ respectively. Let $\beta_n(z) := \operatorname{tr}R(z)\Phi_n$ and $\bar{\beta}_n(z) := \mathbb{E}[\beta_n(z)]$. Notice that $m_n(z), \bar{m}_n(z), \beta_n$ and $\bar{\beta}_n(z)$ are all in $\mathbb{C}^+$ and uniformly and almost surely bounded by some constant. By choosing $D = \mathrm{Id}$ in Lemma 82, we conclude

$$\lim_{n,d\to\infty} \left(1 + z\bar{m}_n(z) + \bar{\beta}_n(z)^2\right) = 0. \tag{4.5.24}$$

Likewise, in Lemma 82, we consider $D = \left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1}\Phi_n$. Let

$$U = \left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1}.$$

Because $\|\Phi_n\|$ is uniformly bounded, $\|D\| \leq C\|U\|$. In terms of Proposition 12, we only need to provide a lower bound for the imaginary part of U. Observe that $\mathrm{Im}\,U = \mathrm{Im}\,\bar{\beta}_n(z)\Phi_n + v\,\mathrm{Id} \succeq v\,\mathrm{Id}$ since $\lambda_{\min}(\Phi_n) \geq 0$ and $\mathrm{Im}\,\bar{\beta}_n(z) > 0$. Thus, $\|D\| \leq Cv^{-1}$ for all n. Meanwhile, we have the equation $\bar{\beta}_n(z)\Phi_n D = \Phi_n - zD$ and hence,

$$\bar{\beta}_n(z)\mathbb{E}[\mathrm{tr}\,R(z)\Phi_n D] = \mathbb{E}[\mathrm{tr}\,R(z)\Phi_n D]\mathbb{E}[\mathrm{tr}\,R(z)\Phi_n] = \bar{\beta}_n(z) - z\mathbb{E}[\mathrm{tr}\,R(z)D].$$

So applying Lemma 82 again, we have another limiting equation $\mathrm{tr}\,D + \bar{\beta}_n(z) \to 0$. In other words,

$$\lim_{n,d\to\infty}\left(\mathrm{tr}\left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1}\Phi_n + \bar{\beta}_n(z)\right) = 0. \tag{4.5.25}$$

Thanks to the identity

$$\bar{\beta}_n(z)\,\mathrm{tr}\left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1}\Phi_n - 1 = -z\,\mathrm{tr}\left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1},$$

we can modify (4.5.24) and (4.5.25) to get

$$\lim_{n,d\to\infty}\left(\bar{m}_n(z) + \mathrm{tr}\left(\bar{\beta}_n(z)\Phi_n + z\,\mathrm{Id}\right)^{-1}\right) = 0. \tag{4.5.26}$$

Since $\bar{\beta}_n(z)$ and $\bar{m}_n(z)$ are uniformly bounded, for any subsequence in n, there is a further convergent sub-subsequence. We denote the limit of such sub-subsequence by $\beta(z)$ and $m(z) \in \mathbb{C}^+$ respectively. Hence, by (4.5.25) and (4.5.26), one can conclude

$$\lim_{n,d\to\infty}\left(\beta(z) + \mathrm{tr}(\beta(z)\Phi_n + z\,\mathrm{Id})^{-1}\Phi_n\right) = 0.$$

Because of the convergence of the empirical eigenvalue distribution of Φ_n, we obtain the fixed point equation (1.2.8) for $\beta(z)$. Analogously, we can also obtain (1.2.7) for $m(z)$ and $\beta(z)$. The existence and the uniqueness of the solutions to (1.2.7) and (1.2.8) are proved in [BZ10, Theorem 2.1] and [WP14, Section 3.4], which implies the convergence of $\bar{m}_n(z)$ and $\bar{\beta}_n(z)$ to $m(z)$ and $\beta(z)$ governed by the self-consistent equations (1.2.7) and (1.2.8) as $n \to \infty$, respectively.

Then, by virtue of condition (4.5.1) in Theorem 79, we know $m_n(z) - \bar{m}_n(z) \xrightarrow{\text{a.s.}} 0$ and $\beta_n(z) - \bar{\beta}_n(z) \xrightarrow{\text{a.s.}} 0$. Therefore, the empirical Stieltjes transform $m_n(z)$ converges to $m(z)$ almost surely for each $z \in \mathbb{C}^+$. Recall that the Stieltjes transform of μ is $m(z)$. By the standard Stieltjes continuity theorem (see, for example, [BS10, Theorem B.9]), this finally concludes the weak convergence of empirical eigenvalue distribution of A_n to μ.

Now we show $\mu = \mu_s \boxtimes \mu_\Phi$. The fixed point equations (1.2.7) and (1.2.8) induce

$$\beta^2(z) + 1 + zm(z) = 0, \tag{4.5.27}$$

since $\beta(z) \in \mathbb{C}^+$ for any $z \in \mathbb{C}^+$. Together with (1.2.7), we attain the same self-consistent equations for the convergence of the empirical spectral distribution of the Wigner-type matrix studied in [BZ10, Theorem 1.1].

Define W_n, the n-by-n Wigner matrix, as a Hermitian matrix with independent entries

$$\{W_n[i,j] : \mathbb{E}[W_n[i,j]] = 0,\ \mathbb{E}[W_n[i,j]^2] = 1,\ 1 \le i \le j \le n\}.$$

The Wigner-type matrix studied in [BZ10, Definition 1.2] is indeed $\frac{1}{\sqrt{n}}\Phi_n^{1/2} W_n \Phi_n^{1/2}$. Hence, such Wigner-type matrix $\frac{1}{\sqrt{n}}\Phi_n^{1/2} W_n \Phi_n^{1/2}$ has the same limiting spectral distribution as A_n defined in Theorem 79. Both limits are determined by self-consistent equations (1.2.7) and (4.5.27).

On the other hand, based on [AGZ10, Theorem 5.4.5], $\frac{1}{\sqrt{n}} W_n$ and Φ_n are almost surely asymptotically free, i.e., the empirical distribution of $\{\frac{1}{\sqrt{n}} W_n, \Phi_n\}$ converges almost surely to the law of $\{\mathbf{s}, \mathbf{d}\}$, where $\mathbf{s}$ and $\mathbf{d}$ are two free non-commutative random variables ($\mathbf{s}$ is a semicircle

element and $\mathbf{d}$ has the law μ_Φ). Thus, the limiting spectral distribution μ of $\frac{1}{\sqrt{n}}\Phi_n^{1/2}W_n\Phi_n^{1/2}$ is the free multiplicative convolution between μ_s and μ_Φ. This implies $\mu = \mu_s \boxtimes \mu_\Phi$ in our setting. $\square$

4.6 Proofs of Theorem 64 and Theorem 65

To prove Theorem 64, we first establish the following proposition to analyze the difference between Stieltjes transform of (4.3.1) and its expectation. This will assist us to verify condition (4.5.1) in Theorem 79. The proof is based on Lemma 23 in Chapter 2.

Proposition 83. *Let $D \in \mathbb{R}^{n\times n}$ be any deterministic symmetric matrix with a uniformly bounded spectral norm. Following the notions in Theorem 64, assume $\|X\| \le C$ for some constant C and Assumption 9 holds. Let $R(z)$ be the resolvent*

$$\left(\frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - \mathbb{E}[Y^\top Y]\right) - z\,\mathrm{Id}\right)^{-1},$$

for any fixed $z \in \mathbb{C}^+$. Then, there exist some constants $s, n_0 > 0$ such that for all $n > n_0$ and any $t > 0$,

$$\mathbb{P}(|\mathrm{tr}R(z)D - \mathbb{E}[\mathrm{tr}R(z)D]| > t) \le 2e^{-cnt^2}.$$

Proof. Define function $F : \mathbb{R}^{d_1\times d_0} \to \mathbb{R}$ by $F(W) := \mathrm{tr}R(z)D$. Fix any $W, \Delta \in \mathbb{R}^{d_1\times d_0}$ where $\|\Delta\|_F = 1$, and let $W_t = W + t\Delta$. We want to verify $F(W)$ is a Lipschitz function in W with respect to the Frobenius norm. First, recall

$$R(z)^{-1} = \frac{1}{\sqrt{d_1 n}}\sigma(WX)^\top\sigma(WX) - \sqrt{\frac{d_1}{n}}\Phi - z\,\mathrm{Id},$$

where the last two terms are deterministic with respect to W. Hence,

$$\begin{aligned}
\operatorname{vec}(\Delta)^\top(\nabla F(W)) &= \frac{d}{dt}\Big|_{t=0} F(W_t) \\
&= -\operatorname{tr} R(z)\left(\frac{d}{dt}\Big|_{t=0} R(z)^{-1}\right)R(z)D \\
&= -\frac{1}{\sqrt{d_1 n}}\operatorname{tr} R(z)\left(\frac{d}{dt}\Big|_{t=0}\sigma(W_t X)^\top \sigma(W_t X)\right)R(z)D \\
&= -\frac{2}{\sqrt{d_1 n}}\operatorname{tr} R(z)\left(\sigma(WX)^\top \cdot \frac{d}{dt}\Big|_{t=0}\sigma(W_t X)\right)R(z)D \\
&= -\frac{2}{\sqrt{d_1 n}}\operatorname{tr} R(z)\Big(\sigma(WX)^\top \cdot \big(\sigma'(WX)\odot(\Delta X)\big)\Big)R(z)D,
\end{aligned}$$

where $\odot$ is the Hadamard product, and σ' is applied entrywise. Here we utilize the formula

$$\partial R(z) = -R(z)(\partial(R(z)^{-1}))R(z)$$

and $R(z) = R(z)^\top$. Proposition 12 in Chapter 2 implies that $\|R(z)\| \le \frac{1}{|\operatorname{Im} z|}$. Therefore, based on the assumption of D, we have

$$\Big|\operatorname{vec}(\Delta)^\top(\nabla F(W))\Big| \le \frac{C}{\sqrt{d_1 n}}\|R(z)\sigma(WX)^\top\| \cdot \|\sigma'(WX)\odot(\Delta X)\|,$$

for some constant $C > 0$. For the first term in the product on the right-hand side,

$$\begin{aligned}
&\left(\frac{1}{\sqrt{d_1 n}}\|R(z)\sigma(WX)^\top\|\right)^2 \\
=&\frac{1}{\sqrt{d_1 n}}\left\|R(z)\left(\frac{1}{\sqrt{d_1 n}}\sigma(WX)^\top\sigma(WX)\right)R(z)^*\right\| \\
\le&\frac{1}{\sqrt{d_1 n}}\left(\|R(z)R(z)^{-1}R(z)^*\| + \left\|R(z)\left(\sqrt{\frac{d_1}{n}}\Phi + z\,\mathrm{Id}\right)R(z)^*\right\|\right) \\
\le&\frac{1}{\sqrt{d_1 n}}\left(\|R(z)\| + \|R(z)\|^2\left(\sqrt{\frac{d_1}{n}}\|\Phi\| + |z|\right)\right) \le \frac{C}{n}.
\end{aligned}$$

For the second term,

$$\|\sigma'(WX)\odot(\Delta X)\| \le \|\sigma'(WX)\odot(\Delta X)\|_F \le \lambda_\sigma\|\Delta X\|_F \le \lambda_\sigma\|\Delta\|_F\cdot\|X\| \le C.$$

Thus, $|\operatorname{vec}(\Delta)^\top(\nabla F(W))| \le C/\sqrt{n}$. This holds for every Δ such that $\|\Delta\|_F = 1$, so $F(W)$ is $C/\sqrt{n}$-Lipschitz in W with respect to the Frobenius norm. Then the result follows from the Gaussian concentration inequality for Lipschitz functions. □

Next, we investigate the approximation of $\Phi = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top X)^\top\sigma(\boldsymbol{w}^\top X)]$ via the Hermite polynomials $\{h_k\}_{k\ge 0}$. The orthogonality of Hermite polynomials allows us to write Φ as a series of kernel matrices. Then we only need to estimate each kernel matrix in this series. The proof is directly based on [GMMM19, Lemma 2]. The only difference is that we consider the deterministic input data X with the (ε_n, B)-orthonormal property, while in Lemma 2 of [GMMM19], the matrix X is formed by independent Gaussian vectors.

Lemma 84. *Recall the definition of Φ_0 in* (4.2.8)*. If X is (ε_n, B)-orthonormal and Assumption 9 holds, then we have the spectral norm bound*

$$\|\Phi - \Phi_0\| \le C_B\varepsilon_n^2\sqrt{n},$$

where C_B is a constant depending on B. Suppose that $\varepsilon_n^2\sqrt{n} \to 0$ as $n\to\infty$, then $\|\Phi\| \le C$ uniformly for some constant C independent of n.

Proof. By Assumption 9, we know that

$$\xi_0(\sigma) = 0, \quad \sum_{k=1}^{\infty}\zeta_k^2(\sigma) = \mathbb{E}[\sigma(\xi)^2] = 1.$$

For any fixed t, $\sigma(tx) \in L^2(\mathbb{R},\Gamma)$. This is because $\sigma(x) \in L^2(\mathbb{R},\Gamma)$ is a Lipschitz function and

by triangle inequality $|\sigma(tx)-\sigma(x)| \le \lambda_\sigma |tx-x|$, we have, for $\xi \sim \mathcal{N}(0,1)$,

$$\mathbb{E}(\sigma(t\xi)^2) \le \mathbb{E}(|\sigma(\xi)| + \lambda_\sigma |t\xi - \xi|)^2 < \infty. \tag{4.6.1}$$

For $1 \le \alpha \le n$, let $\sigma_\alpha(x) := \sigma(\|\mathbf{x}_\alpha\| x)$ and the Hermite expansion of σ_a can be written as

$$\sigma_\alpha(x) = \sum_{k=0}^{\infty} \zeta_k(\sigma_\alpha) h_k(x),$$

where the coefficient $\zeta_k(\sigma_\alpha) = \mathbb{E}[\sigma_\alpha(\xi) h_k(\xi)]$. Let unit vectors be $\mathbf{u}_\alpha = \mathbf{x}_\alpha / \|\mathbf{x}_\alpha\|$, for $1 \le \alpha \le n$. So for $1 \le \alpha, \beta \le n$, the (α, β) entry of Φ is

$$\Phi_{\alpha\beta} = \mathbb{E}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)\sigma(\boldsymbol{w}^\top \mathbf{x}_\beta)] = \mathbb{E}[\sigma_\alpha(\xi_\alpha)\sigma_\beta(\xi_\beta)],$$

where $(\xi_\alpha, \xi_\beta) = (\boldsymbol{w}^\top \mathbf{u}_\alpha, \boldsymbol{w}^\top \mathbf{u}_\beta)$ is a Gaussian random vector with mean zero and covariance

$$\begin{pmatrix} 1 & \mathbf{u}_\alpha^\top \mathbf{u}_\beta \\ \mathbf{u}_\alpha^\top \mathbf{u}_\beta & 1 \end{pmatrix}. \tag{4.6.2}$$

By the orthogonality of Hermite polynomials with respect to Γ and Lemma 99, we can obtain

$$\mathbb{E}[h_j(\xi_\alpha) h_k(\xi_\beta)] = \mathbb{E}[h_j(\boldsymbol{w}^\top \mathbf{u}_\alpha) h_k(\boldsymbol{w}^\top \mathbf{u}_\beta)] = \delta_{j,k} (\mathbf{u}_\alpha^\top \mathbf{u}_\beta)^k,$$

which leads to

$$\Phi_{\alpha\beta} = \sum_{k=0}^{\infty} \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)(\mathbf{u}_\alpha^\top \mathbf{u}_\beta)^k. \tag{4.6.3}$$

For any $k \in \mathbb{N}$, let T_k be an n-by-n matrix with (α, β)-th entry

$$(T_k)_{\alpha\beta} := \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)(\mathbf{u}_\alpha^\top \mathbf{u}_\beta)^k. \tag{4.6.4}$$

Specifically, for any $k \in \mathbb{N}$, we have

$$T_k = D_k f_k(X^\top X) D_k,$$

where D_k is the diagonal matrix $\mathrm{diag}(\zeta_k(\sigma_\alpha)/\|\mathbf{x}_\alpha\|^k)_{\alpha\in[n]}$.

At first, we consider twice differentiable σ in Assumption 9. Similar with [GMMM19, Equation (26)], for any $\varepsilon > 0$ and $|t-1| \le \varepsilon$, we take the Taylor approximation of $\sigma(tx)$ at point x, then there exists η between tx and x such that

$$\sigma(tx) - \sigma(x) = \sigma'(x)x(t-1) + \frac{1}{2}\sigma''(\eta)x^2(t-1)^2.$$

Replacing x by ξ and taking expectation, since σ'' is uniformly bounded, we can get

$$\left|\mathbb{E}[\sigma(t\xi) - \sigma(\xi)] - \mathbb{E}[\sigma'(\xi)\xi](t-1)\right| \le C|t-1|^2 \le C\varepsilon_n^2, \tag{4.6.5}$$

For $k \ge 1$, the Lipschitz condition for σ yields

$$|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)| \le C|\|\mathbf{x}_\alpha\| - 1| \cdot \mathbb{E}[|\xi| \cdot |h_k(\xi)|] \le C\varepsilon_n, \tag{4.6.6}$$

where constant C does not depend on k. As for piece-wise linear σ, it is not hard to see

$$\mathbb{E}[\sigma(t\xi) - \sigma(\xi)] = \mathbb{E}[\sigma'(\xi)\xi](t-1). \tag{4.6.7}$$

Now, we begin to approximate T_k separately based on (4.6.5), (4.6.6) and (4.6.7). Denote $\mathrm{diag}(A)$ the diagonal submatrix of a matrix A.

(1) Approximation for $\sum_{k\ge4}(T_k - \mathrm{diag}(T_k))$. At first, we estimate the L^2 norm with respect to Γ of the function σ_α. Recall that $\|\sigma_\alpha\|_{L^2} = \mathbb{E}[\sigma_\alpha(\xi)^2]^{1/2}$. Because $\|\sigma\|_{L^2} = 1$ and σ

is a Lipschitz function, we have

$$\sup_{1\le\alpha\le n}\|\sigma-\sigma_\alpha\|_{L^2}=\mathbb{E}[(\sigma(\xi)-\sigma_\alpha(\xi))^2]^{1/2}\le C|\|\mathbf{x}_\alpha\|-1|, \tag{4.6.8}$$

$$\sup_{1\le\alpha\le n}\|\sigma_\alpha\|_{L^2}\le 1+C\varepsilon_n. \tag{4.6.9}$$

Hence, $\|\sigma_\alpha\|_{L^2}$ is uniformly bounded with some constant for all large n. Next, we estimate the off-diagonal entries of T_k when $k\ge 4$. From (4.6.4), we obtain that

$$\begin{aligned}\left\|\sum_{k\ge4}(T_k-\mathrm{diag}(T_k))\right\| &\le \left\|\sum_{k\ge4}(T_k-\mathrm{diag}(T_k))\right\|_F \le \sum_{k\ge4}\|T_k-\mathrm{diag}(T_k)\|_F\\ &\le \sum_{k\ge4}\left(\sup_{\alpha\neq\beta}|\mathbf{u}_\alpha^\top\mathbf{u}_\beta|^k\right)\left[\sum_{\alpha,\beta=1}^n\zeta_k(\sigma_\alpha)^2\zeta_k(\sigma_\beta)^2\right]^{\frac12}\\ &\le \left(\sup_{\alpha\neq\beta}|\mathbf{u}_\alpha^\top\mathbf{u}_\beta|^4\right)\sum_{\alpha=1}^n\sum_{k=0}^\infty\zeta_k(\sigma_\alpha)^2\\ &\le n\cdot\left(\sup_{\alpha\neq\beta}\frac{|\mathbf{x}_\alpha^\top\mathbf{x}_\beta|^4}{\|\mathbf{x}_\alpha\|^4\|\mathbf{x}_\beta\|^4}\right)\sup_{1\le\alpha\le n}\|\sigma_\alpha\|_{L^2}^2\le Cn\cdot\varepsilon_n^4,\end{aligned} \tag{4.6.10}$$

when n is sufficiently large.

(2) Approximation for T_0. Recall $\mathbb{E}[\sigma(\xi)]=0$ and by Gaussian integration by part,

$$\mathbb{E}[\sigma'(\xi)\xi]=\mathbb{E}[\xi\int_0^\xi\sigma'(x)xdx]=\mathbb{E}[\xi^2\sigma(\xi)]-\mathbb{E}[\xi\int_0^\xi\sigma(x)dx]=\mathbb{E}[\xi^2\sigma(\xi)]-\mathbb{E}[\sigma(\xi)].$$

Then, we have

$$\mathbb{E}[\sigma'(\xi)\xi]=\mathbb{E}[(\xi^2-1)\sigma(\xi)]=\mathbb{E}[\sqrt{2}h_2(\xi)\sigma(\xi)]=\sqrt{2}\zeta_2(\sigma).$$

If σ is twice differentiable, then $\mathbb{E}[\sigma''(\xi)]=\sqrt{2}\zeta_2(\sigma)$ as well.

Thus, taking $t = \|\mathbf{x}_\alpha\|$ in (4.6.5) and (4.6.7) implies that for any $1 \le \alpha \le n$,

$$\left|\zeta_0(\sigma_\alpha) - \sqrt{2}\zeta_2(\sigma)(\|\mathbf{x}_\alpha\| - 1)\right| \le C\varepsilon_n^2. \tag{4.6.11}$$

Define $\boldsymbol{\nu}^\top := (\zeta_0(\sigma_1), \ldots, \zeta_0(\sigma_n))$, then $T_0 = \boldsymbol{\nu}\boldsymbol{\nu}^\top$. Recall the definition of *mu* in (4.2.8). Then, (4.6.11) ensures that

$$\|\boldsymbol{\mu} - \boldsymbol{\nu}\| \le C\sqrt{n}\varepsilon_n^2.$$

Applying the (ε_n, B)-orthonormal property of $\mathbf{x}_\alpha$ yields

$$\|\boldsymbol{\mu}\|^2 = 2\zeta_2(\sigma)^2 \sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\| - 1)^2 \le 2\zeta_2(\sigma)^2 \sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\|^2 - 1)^2 \le 2B^2\zeta_2(\sigma)^2. \tag{4.6.12}$$

Hence the difference between T_0 and $\boldsymbol{\mu}\boldsymbol{\mu}^\top$ is controlled by

$$\|T_0 - \boldsymbol{\mu}\boldsymbol{\mu}^\top\| \le \|\boldsymbol{\mu} - \boldsymbol{\nu}\|(2\|\boldsymbol{\mu}\| + \|\boldsymbol{\nu} - \boldsymbol{\mu}\|) \le C\sqrt{n}\varepsilon_n^2. \tag{4.6.13}$$

(3) Approximation for T_k for $k = 1,2,3$. For $0 \le k \le 3$, Assumption 10 and (4.6.6) show that

$$\begin{aligned}\left|\zeta_k(\sigma_\alpha)/\|\mathbf{x}_\alpha\|^k - \zeta_k(\sigma)\right| &\le \frac{1}{\|\mathbf{x}_\alpha\|^k}\left[|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)| + |\zeta_k(\sigma)| \cdot |\|\mathbf{x}_\alpha\|^k - 1|\right] \\ &\le \frac{C\varepsilon_n + C_1|\|\mathbf{x}_\alpha\| - 1|}{(1-\varepsilon_n)^k} \le C_2\varepsilon_n,\end{aligned} \tag{4.6.14}$$

when n is sufficiently large. Notice that $T_k = D_k f_k(X^\top X)D_k$, where D_k is the diagonal matrix. Hence, by (4.6.14),

$$\|D_k - \zeta_k(\sigma)\,\mathrm{Id}\| \le C_2\varepsilon_n.$$

And for $k = 1, 2, 3$, by the triangle inequality,

$$\|T_k - \zeta_k(\sigma)^2 f_k(X^\top X)\| = \|D_k f_k(X^\top X) D_k - \zeta_k(\sigma)^2 f_k(X^\top X)\|$$
$$\leq \|D_k - \zeta_k(\sigma)\,\mathrm{Id}\| \cdot \|f_k(X^\top X)\|(|\zeta_k(\sigma)| + \|D_k - \zeta_k(\sigma)\,\mathrm{Id}\|) \leq C\varepsilon_n \|f_k(X^\top X)\|.$$

When $k = 1$, $f_1(X^\top X) = X^\top X$ and $\|X^\top X\| \leq \|X\|^2 \leq B^2$. When $k = 2$,

$$f_2(X^\top X) = (X^\top X) \odot (X^\top X).$$

From Lemma 95 in Appendix 4.9, we have that

$$\|f_2(X^\top X)\| \leq \max_{1 \leq \alpha, \beta \leq n} |\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \cdot \|X\|^2 \leq B^2(1 + \varepsilon_n). \tag{4.6.15}$$

So the left-hand side of (4.6.15) is bounded. Analogously, we can verify $\|f_3(X^\top X)\|$ is also bounded. Therefore, we have

$$\|T_k - \zeta_k(\sigma)^2 f_k(X^\top X)\| \leq C\varepsilon_n, \tag{4.6.16}$$

for some constant C and $k = 1, 2, 3$ when n is sufficiently large.

(4) Approximation for $\sum_{k \geq 4} \mathrm{diag}(T_k)$. Since $\mathbf{u}_\alpha^\top \mathbf{u}_\alpha = 1$, we know

$$\sum_{k \geq 4} \mathrm{diag}(T_k) = \mathrm{diag}\left(\sum_{k \geq 4} \zeta_k(\sigma_\alpha)^2 \right)_{\alpha \in [n]} = \mathrm{diag}\left(\|\sigma_\alpha\|_{L^2}^2 - \sum_{k=0}^{4} \zeta_k(\sigma_\alpha)^2 \right)_{\alpha \in [n]}.$$

First, by (4.6.8) and (4.6.9), we can claim that

$$|\|\sigma_\alpha\|_{L^2}^2 - 1| = |\|\sigma_\alpha\|_{L^2}^2 - \|\sigma\|_{L^2}^2| \leq C\|\sigma_\alpha - \sigma\|_{L^2} \leq C\varepsilon_n.$$

Second, in terms of (4.6.14), we obtain

$$|\zeta_k(\sigma_\alpha)^2 - \zeta_k(\sigma)^2| \le C|\zeta_k(\sigma_\alpha) - \zeta_k(\sigma)| \le C\varepsilon_n,$$

for $k = 1, 2$ and 3. Combining these together, we conclude that

$$\begin{aligned} & \left\| \sum_{k\ge 4} \operatorname{diag}(T_k) - (1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2)\,\mathrm{Id} \right\| \\ \le \quad & \max_{1\le\alpha\le n} \left| (\|\sigma_\alpha\|_{L^2}^2 - 1) - \sum_{k=0}^{4} (\zeta_k(\sigma_\alpha)^2 - \zeta_k(\sigma)^2) \right| \le C\varepsilon_n. \end{aligned} \tag{4.6.17}$$

Recall

$$\Phi_0 = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \sum_{k=1}^{3} \zeta_k(\sigma)^2 f_k(X^\top X) + (1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2)\,\mathrm{Id}\,.$$

In terms of approximations (4.6.10), (4.6.13), (4.6.16) and (4.6.17), we can finally manifest

$$\|\Phi - \Phi_0\| \le C\big(\varepsilon_n + \sqrt{n}\varepsilon_n^2 + n\varepsilon_n^4\big) \le C\sqrt{n}\varepsilon_n^2, \tag{4.6.18}$$

for some constant $C > 0$ as $\sqrt{n}\varepsilon_n^2 \to 0$. The spectral norm bound of Φ is directly deduced by the spectral norm bound of Φ_0 based on (4.6.12) and (4.6.15), together with (4.6.18). □

Remark (Optimality of ε_n)**.** *For general deterministic data X, our pairwise orthogonality assumption with rate $n\varepsilon_n^4 = o(1)$ is optimal for the approximation of Φ by Φ_0 in the spectral norm. If we relax the decay rate of ε_n in Assumption 10, the above approximation may require including terms of higher-degree $f_k(X^\top X)$ for $k \ge 4$ in Φ_0, which will lead to the invalidation of some of our following results and simplifications. This weaker regime has been considered in our follow-up work [WZ23].*

Next, we continue to provide an additional estimate for Φ, but in the Frobenius norm to further simplify the limiting spectral distribution of Φ.

Lemma 85. *If Assumptions 9 and 10 hold, then* Φ *has the same limiting spectrum as* $b_\sigma^2 X^\top X + (1-b_\sigma^2)\,\mathrm{Id}$ *when* $n \to \infty$, *i.e.*

$$\lim \operatorname{spec} \Phi = \lim \operatorname{spec} \left(b_\sigma^2 X^\top X + (1-b_\sigma^2)\,\mathrm{Id} \right) = b_\sigma^2 \mu_0 + (1-b_\sigma^2).$$

Proof. By the definition of b_σ, we know that $b_\sigma = \zeta_1(\sigma)$. As a direct deduction of Lemma 84, the limiting spectrum of Φ is identical to the limiting spectrum of Φ_0. To prove this lemma, it suffices to check the Frobenius norm of the difference between Φ_0 and $\zeta_1(\sigma)^2 X^\top X + (1-\zeta_1(\sigma)^2)\,\mathrm{Id}$. Notice that

$$\begin{aligned} &\Phi_0 - \zeta_1(\sigma)^2 X^\top X - (1-\zeta_1(\sigma)^2)\,\mathrm{Id} \\ =\ &\boldsymbol{\mu}\boldsymbol{\mu}^\top + \zeta_2(\sigma)^2 f_2(X^\top X) + \zeta_3(\sigma)^2 f_3(X^\top X) - (\zeta_2(\sigma)^2 + \zeta_3(\sigma)^2)\,\mathrm{Id}\,. \end{aligned}$$

By the definition of vector *mu* and the assumption of X, we have

$$\|\boldsymbol{\mu}\boldsymbol{\mu}^\top\|_F = \|\boldsymbol{\mu}\|^2 = 2\zeta_2^2(\sigma) \sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\| - 1)^2 \le 2\zeta_2^2(\sigma) B^2. \tag{4.6.19}$$

For $k = 2, 3$, the Frobenius norm can be controlled by

$$\begin{aligned} \|f_k(X^\top X) - \mathrm{Id}\,\|_F^2 &= \sum_{\alpha,\beta=1}^{n} \left((\mathbf{x}_\alpha^\top \mathbf{x}_\beta)^k - \delta_{\alpha\beta} \right)^2 \\ &\le n(n-1)\varepsilon_n^{2k} + \sum_{\alpha=1}^{n} (\|\mathbf{x}_\alpha\|^{2k} - 1)^2 \le n^2 \varepsilon_n^{2k} + Cn\varepsilon_n^2. \end{aligned}$$

Hence, as $n \to \infty$, we have

$$\frac{1}{n}\|\boldsymbol{\mu}\boldsymbol{\mu}^\top\|_F^2, \quad \frac{1}{n}\|f_k(X^\top X) - \mathrm{Id}\,\|_F^2 \to 0, \ \text{ for } \ k = 2, 3,$$

as $n\varepsilon_n^4 \to 0$. Then we conclude that

$$\frac{1}{n}\|\Phi_0 - \zeta_1(\sigma)^2 X^\top X - (1-\zeta_1(\sigma)^2)\,\mathrm{Id}\,\|_F^2 \le C(n\varepsilon_n^4 + \varepsilon_n^2) \to 0.$$

Hence, $\lim \operatorname{spec} \Phi$ is the same as $\lim \operatorname{spec}\left(\zeta_1(\sigma)^2 X^\top X + (1-\zeta_1(\sigma)^2)\,\mathrm{Id}\right)$ when $n \to \infty$, due to Proposition 13. □

Moreover, the proof of Lemma 85 can be modified to prove (4.3.32), so we omit its proof. Now, based on Corollary 78, Proposition 83, Lemma 84, and Lemma 85, applying Theorem 79 for general sample covariance matrices, we can finish the proof of Theorem 64.

Proof of Theorem 64. Based on Corollary 78 and Proposition 83, we can verify the conditions (4.5.1) and (4.5.2) in Theorem 79. By Lemma 84 and Lemma 85, we know that the limiting eigenvalue distributions of Φ and $(1-b_\sigma^2)\,\mathrm{Id} + b_\sigma^2 X^\top X$ are identical and $\|\Phi\|$ is uniformly bounded. So the limiting eigenvalue distribution of Φ denoted by μ_Φ is just $(1-b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_0$. Hence, the first conclusion of Theorem 64 follows from Theorem 79.

For the second part of this theorem, we consider the difference

$$\begin{aligned}
&\frac{1}{n}\left\| \frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - \mathbb{E}[Y^\top Y]\right) - \frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - d_1\Phi_0\right)\right\|_F^2 \\
\le\ & \frac{d_1}{n^2}\|\Phi - \Phi_0\|_F^2 \le \frac{d_1}{n}\|\Phi - \Phi_0\|^2 \le d_1\varepsilon_n^4 \to 0,
\end{aligned}$$

where we employ Lemma 84 and the assumption $d_1\varepsilon_n^4 = o(1)$. Thus, because of Proposition 13, $\frac{1}{\sqrt{d_1 n}}\left(Y^\top Y - d_1\Phi_0\right)$ has the same limiting eigenvalue distribution as (4.3.1), $\mu_s \boxtimes ((1-b_\sigma^2) \oplus b_\sigma^2 \otimes \mu_0)$. This finishes the proof of Theorem 64. □

Next, we move to study the empirical NTK and its corresponding limiting eigenvalue distribution. Similarly, we first verify that such NTK concentrates around its expectation and

then simplify this expectation by some deterministic matrix only depending on the input data matrix X and nonlinear activation σ. The following lemma can be obtained from (4.3.12) in Theorem 68.

Lemma 86. *Suppose that Assumption 8 holds,* $\sup_{x\in\mathbb{R}}|\sigma'(x)| \le \lambda_\sigma$ *and* $\|X\| \le B$. *Then if* $d_1 = \omega(\log n)$, *we have*

$$\frac{1}{d_1}\left\|(S^\top S)\odot(X^\top X) - \mathbb{E}[(S^\top S)\odot(X^\top X)]\right\| \to 0, \tag{4.6.20}$$

almost surely as $n, d_0, d_1 \to \infty$. *Moreover, if* $d_1/n \to \infty$ *as* $n \to \infty$, *then almost surely*

$$\frac{1}{\sqrt{nd_1}}\left\|(S^\top S)\odot(X^\top X) - \mathbb{E}[(S^\top S)\odot(X^\top X)]\right\| \to 0. \tag{4.6.21}$$

Lemma 87. *Suppose* X *is* (ε_n, B)*-orthonormal. Under Assumption 9, we have*

$$\|\Psi - \Psi_0\| \le C_B \varepsilon_n^4 n, \tag{4.6.22}$$

where Ψ *and* Ψ_0 *are defined in* (4.3.6) *and* (4.3.7)*, respectively, and* C_B *is a constant depending on* B.

Proof. We can directly apply methods in the proof of Lemma 84. Notice that (4.2.3) and (4.2.5) imply

$$\mathbb{E}[S^\top S] = d_1 \mathbb{E}[\sigma'(\boldsymbol{w}^\top X)^\top \sigma'(\boldsymbol{w}^\top X)],$$

for any standard Gaussian random vector $\boldsymbol{w} \sim \mathcal{N}(0, \mathrm{Id})$. Recall that (4.3.8) defines the k-th coefficient of Hermite expansion of $\sigma'(x)$ by $\eta_k(\sigma)$ for any $k \in \mathbb{N}$. Then, Assumption 9 indicates $b_\sigma = \eta_0(\sigma)$ and $a_\sigma = \sum_{k=0}^\infty \eta_k^2(\sigma)$. For $1 \le \alpha \le n$, we introduce $\phi_\alpha(x) := \sigma'(\|\mathbf{x}_\alpha\| x)$ and the Hermite expansion of this function as

$$\phi_\alpha(x) = \sum_{k=0}^\infty \zeta_k(\phi_\alpha) h_k(x),$$

where the coefficient $\zeta_k(\sigma_\alpha) = \mathbb{E}[\phi_\alpha(\xi)h_k(\xi)]$. Let $\mathbf{u}_\alpha = \mathbf{x}_\alpha/\|\mathbf{x}_\alpha\|$, for $1 \le \alpha \le n$. So for $1 \le \alpha, \beta \le n$, the (α,β)-entry of Ψ is

$$\Psi_{\alpha\beta} = \mathbb{E}[\phi_\alpha(\xi_\alpha)\phi_\beta(\xi_\beta)] \cdot (\mathbf{x}_a^\top \mathbf{x}_\beta),$$

where $(\xi_\alpha, \xi_\beta) = (\boldsymbol{w}^\top \mathbf{u}_\alpha, \boldsymbol{w}^\top \mathbf{u}_\beta)$ is a Gaussian random vector with mean zero and covariance (4.6.2). Following the derivation of formula (4.6.3), we obtain

$$\Psi_{\alpha\beta} = \sum_{k=0}^{\infty} \frac{\zeta_k(\phi_\alpha)\zeta_k(\phi_\beta)}{\|\mathbf{x}_\alpha\|^k \|\mathbf{x}_\beta\|^k} (\mathbf{x}_\alpha^\top \mathbf{x}_\beta)^{k+1}. \tag{4.6.23}$$

For any $k \in \mathbb{N}$, let $T_k \in \mathbb{R}^{n\times n}$ be an n-by-n matrix with (α,β) entry

$$(T_k)_{\alpha\beta} := \frac{\zeta_k(\phi_\alpha)\zeta_k(\phi_\beta)}{\|\mathbf{x}_\alpha\|^k \|\mathbf{x}_\beta\|^k} (\mathbf{x}_\alpha^\top \mathbf{x}_\beta)^{k+1}.$$

We can write $T_k = D_k f_{k+1}(X^\top X) D_k$ for any $k \in \mathbb{N}$, where D_k is $\mathrm{diag}(\zeta_k(\phi_\alpha)/\|\mathbf{x}_\alpha\|^k)$. Then, adopting the proof of (4.6.16), we can similarly conclude that

$$\|T_k - \eta_k^2(\sigma) f_{k+1}(X^\top X)\| \le C\varepsilon_n,$$

for some constant C and $k = 0, 1, 2$, when n is sufficiently large. Likewise, (4.6.10) indicates

$$\left\| \sum_{k\ge 3} (T_k - \mathrm{diag}(T_k)) \right\| \le C\varepsilon_n^4 n,$$

and a similar proof of (4.6.17) implies that

$$\left\| \sum_{k\ge 3} \mathrm{diag}(T_k) - \left(a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma) \right) \mathrm{Id} \right\| \le C\varepsilon_n.$$

Based on these approximations, we can conclude the result of this lemma. □

Proof of Theorem Theorem 65. The first part of the statement is a straight consequence of (4.6.21) and Theorem 64. Denote by $A := \sqrt{\frac{d_1}{n}}(H - \mathbb{E}[H])$ and $B := \sqrt{\frac{d_1}{n}}\Big(\frac{1}{d_1}Y^\top Y - \Phi\Big)$. Observe that

$$B - A = \frac{1}{\sqrt{nd_1}}\Big[(S^\top S) \odot (X^\top X) - \mathbb{E}[(S^\top S) \odot (X^\top X)]\Big].$$

Hence, (4.6.21) indicates $\|B - A\| \to 0$ as $n \to \infty$. This convergence implies that limiting laws of A and B are identical because of Lemma 97.

The second part is because of Lemma 84 and Lemma 87. From (4.2.4) and (4.3.6), $\mathbb{E}[H] = \Phi + \Psi$. Then almost surely,

$$\begin{aligned}
&\left\| \sqrt{\frac{d_1}{n}}(H - \mathbb{E}[H]) - \sqrt{\frac{d_1}{n}}(H - \Phi_0 - \Psi_0) \right\| = \sqrt{\frac{d_1}{n}} \|\Phi_0 + \Psi_0 - \mathbb{E}[H]\| \\
\leq &\sqrt{\frac{d_1}{n}} (\|\Phi - \Phi_0\| + \|\Psi - \Psi_0\|) \leq \sqrt{\frac{d_1}{n}} \big(\sqrt{n}\varepsilon_n^2 + n\varepsilon_n^4\big) \to 0,
\end{aligned}$$

as $\varepsilon_n^4 d_1 \to 0$ by the assumption of Theorem 65. Therefore, the limiting eigenvalue distribution of (4.3.10) is the same as (4.3.9). □

4.7 Proof of the Concentration for Extreme Eigenvalues

In this section, we obtain the estimates of the extreme eigenvalues for the CK and NTK we studied in Section 4.6. The limiting spectral distribution of $\frac{1}{\sqrt{d_1 n}}(Y^\top Y - \mathbb{E}[Y^\top Y])$ tells us the bulk behavior of the spectrum. An estimation of the extreme eigenvalues will show that the eigenvalues are confined in a finite interval with high probability. We first provide a non-asymptotic bound on the concentration of $\frac{1}{d_1}Y^\top Y$ under the spectral norm. The proof is based on the Hanson-Wright inequality we proved in Section 4.4 and an ε-net argument.

Proof of Theorem 66. Recall notations in Section 4.2. Define

$$M := \frac{1}{\sqrt{d_1 n}} Y^\top Y = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} \mathbf{y}_i \mathbf{y}_i^\top,$$
$$M - \mathbb{E}M = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} (\mathbf{y}_i \mathbf{y}_i^\top - \mathbb{E}[\mathbf{y}_i \mathbf{y}_i^\top]) = \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} (\mathbf{y}_i \mathbf{y}_i^\top - \Phi),$$

where $\mathbf{y}_i^\top = \sigma(\boldsymbol{w}_i^\top X)$.

For any fixed $\mathbf{z} \in \mathbb{S}^{n-1}$, we have

$$\begin{aligned}
\mathbf{z}^\top (M - \mathbb{E}M)\mathbf{z} &= \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} [\langle \mathbf{z}, \mathbf{y}_i \rangle^2 - \mathbf{z}^\top \Phi \mathbf{z}] \\
&= \frac{1}{\sqrt{d_1 n}} \sum_{i=1}^{d_1} [\mathbf{y}_i^\top (\mathbf{z}\mathbf{z}^\top) \mathbf{y}_i - \mathrm{Tr}(\Phi \mathbf{z}\mathbf{z}^\top)] \\
&= (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})^\top A_{\mathbf{z}} (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}) - \mathrm{Tr}(A_{\mathbf{z}} \tilde{\Phi}), \qquad (4.7.1)
\end{aligned}$$

where

$$A_{\mathbf{z}} = \frac{1}{\sqrt{d_1 n}} \begin{bmatrix} \mathbf{z}\mathbf{z}^\top & & \\ & \ddots & \\ & & \mathbf{z}\mathbf{z}^\top \end{bmatrix} \in \mathbb{R}^{nd_1 \times nd_1}, \quad \tilde{\Phi} = \begin{bmatrix} \Phi & & \\ & \ddots & \\ & & \Phi \end{bmatrix} \in \mathbb{R}^{nd_1 \times nd_1},$$

and column vector $(\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}) \in \mathbb{R}^{nd_1}$ is the concatenation of column vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{d_1}$. Then

$$(\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})^\top = \sigma((\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{d_1})^\top \tilde{X})$$

with block matrix

$$\tilde{X} = \begin{bmatrix} X & & \\ & \ddots & \\ & & X \end{bmatrix}.$$

Notice that

$$\|A_{\mathbf{z}}\| = \frac{1}{\sqrt{d_1 n}}, \quad \|A_{\mathbf{z}}\|_F = \frac{1}{\sqrt{n}}, \quad \|\tilde{X}\| = \|X\|.$$

Denote $\tilde{\mathbf{y}} = (\mathbf{y}_1, \ldots, \mathbf{y}_{d_1})$. With (4.4.14), we obtain

$$\begin{aligned}\|\mathbb{E}\tilde{\mathbf{y}}\|^2 = d_1\|\mathbb{E}\mathbf{y}\|^2 &\leq d_1\left(2\lambda_\sigma^2\sum_{i=1}^n(\|\mathbf{x}_i\|^2-1)^2 + 2n(\mathbb{E}\sigma(\xi))^2\right)\\ &= d_1\left(2\lambda_\sigma^2\sum_{i=1}^n(\|\mathbf{x}_i\|^2-1)^2\right) \leq 2d_1\lambda_\sigma^2 B^2,\end{aligned}$$

where the last line is from the assumptions on X and σ. When $B \neq 0$, applying (4.4.9) to (4.7.1) implies

$$\begin{aligned}&\mathbb{P}\Big(|(\mathbf{y}_1,\ldots,\mathbf{y}_{d_1})^\top A_{\mathbf{z}}(\mathbf{y}_1,\ldots,\mathbf{y}_{d_1}) - \mathrm{Tr}(A_{\mathbf{z}}\tilde{\Phi})| \geq t\Big)\\ &\leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2 n}{8\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{\lambda_\sigma^2\|X\|^2}\right\}\right) + 2\exp\left(-\frac{t^2 d_1 n}{32\lambda_\sigma^2\|X\|^2\|\mathbb{E}\tilde{\mathbf{y}}\|^2}\right)\\ &\leq 2\exp\left(-\frac{1}{C}\min\left\{\frac{t^2 n}{8\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{\lambda_\sigma^2\|X\|^2}\right\}\right) + 2\exp\left(-\frac{t^2 n}{64\lambda_\sigma^4 B^2\|X\|^2}\right).\end{aligned}$$

Let subset $\mathscr{N}$ be a $1/2$-net on $\mathbb{S}^{n-1}$ with $|\mathscr{N}| \leq 5^n$ (see e.g. [Ver18, Corollary 4.2.13]), then

$$\|M - \mathbb{E}M\| \leq 2\sup_{\mathbf{z}\in\mathscr{N}}|\mathbf{z}^\top(M-\mathbb{E}M)\mathbf{z}|.$$

Taking a union bound over $\mathscr{N}$ yields

$$\begin{aligned}\mathbb{P}(\|M-\mathbb{E}M\| \geq 2t) \leq 2\exp&\left(n\log 5 - \frac{1}{C}\min\left\{\frac{t^2 n}{16\lambda_\sigma^4\|X\|^4}, \frac{t\sqrt{d_1 n}}{2\lambda_\sigma^2\|X\|^2}\right\}\right)\\ &+ 2\exp\left(n\log 5 - \frac{t^2 n}{64\lambda_\sigma^4 B^2\|X\|^2}\right).\end{aligned}$$

We then can set

$$t = \left(8\sqrt{C} + 8C\sqrt{\frac{n}{d_1}}\right)\lambda_\sigma^2\|X\|^2 + 16B\lambda_\sigma^2\|X\|,$$

to conclude

$$\mathbb{P}\left(\|M - \mathbb{E}M\| \geq \left(16\sqrt{C} + 16C\sqrt{\frac{n}{d_1}}\right)\lambda_\sigma^2\|X\|^2 + 32B\lambda_\sigma^2\|X\|\right) \leq 4e^{-2n}.$$

Since

$$\left\|\frac{1}{d_1}Y^\top Y - \Phi\right\| = \sqrt{\frac{n}{d_1}}\|M - \mathbb{E}M\|,$$

the upper bound in (4.3.11) is then verified. When $B = 0$, we can apply (4.4.8) and follow the same steps to get the desired bound. □

By the concentration inequality in Theorem 66, we can get a lower bound on the smallest eigenvalue of the conjugate kernel $\frac{1}{d_1}Y^\top Y$ as follows.

Lemma 88. *Assume X satisfies $\sum_{i=1}^n (\|\mathbf{x}_i\|^2 - 1)^2 \leq B^2$ for a constant $B > 0$, and σ is λ_σ-Lipschitz with $\mathbb{E}\sigma(\xi) = 0$. Then with probability at least $1 - 4e^{-2n}$,*

$$\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right) \geq \lambda_{\min}(\Phi) - C\left(\sqrt{\frac{n}{d_1}} + \frac{n}{d_1}\right)\lambda_\sigma^2\|X\|^2 - 32B\lambda_\sigma^2\|X\|\sqrt{\frac{n}{d_1}}. \tag{4.7.2}$$

Proof. By Weyl's inequality [AGZ10, Corollary A.6], we have

$$\left|\lambda_{\min}\left(\frac{1}{d_1}Y^\top Y\right) - \lambda_{\min}(\Phi)\right| \leq \left\|\frac{1}{d_1}Y^\top Y - d_1\Phi\right\|.$$

Then (4.7.2) follows from (4.3.11). □

The lower bound in (4.7.2) relies on $\lambda_{\min}(\Phi)$. Under certain assumptions on X and σ,

we can guarantee that $\lambda_{\min}(\Phi)$ is bounded below by an absolute constant.

Lemma 89. *Assume σ is not a linear function and $\sigma(x)$ is Lipschitz. Then*

$$\sup\{k \in \mathbb{N} : \zeta_k(\sigma)^2 > 0\} = \infty. \tag{4.7.3}$$

Proof. Suppose that $\sup\{k \in \mathbb{N} : \zeta_k(\sigma)^2 > 0\}$ is finite. Then σ is a polynomial of degree at least 2 from our assumption, which is a contradiction to the fact that σ is Lipschitz. Hence, (4.7.3) holds. □

Lemma 90. *Assume Assumption 9 holds, σ is not a linear function, and X satisfies (ε_n, B)-orthonormal property. Then,*

$$\lambda_{\min}(\Phi) \geq 1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2 - C_B\varepsilon_n^2\sqrt{n}. \tag{4.7.4}$$

Remark. *This bound will not hold when σ is a linear function. Suppose σ is a linear function, under Assumption 9, we must have $\sigma(x) = x$ and $\Phi = X^\top X$. Then we will not have a lower bound on $\lambda_{\min}(\Phi)$ based on the Hermite coefficients of σ.*

Proof of Lemma 90. From Lemma 84, under our assumptions, we know that

$$\|\Phi - \Phi_0\| \leq C_B\varepsilon_n^2\sqrt{n}.$$

where Φ_0 is given by (4.2.8). Thus, $\lambda_{\min}(\Phi) \geq \lambda_{\min}(\Phi_0) - C_B\varepsilon_n^2\sqrt{n}$,

and, from Weyl's inequality [AGZ10, Theorem A.5], we have

$$\lambda_{\min}(\Phi_0) \geq \sum_{k=1}^{3} \zeta_k(\sigma)^2\lambda_{\min}(f_k(X^\top X)) + (1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2).$$

Note that $f_k(X^\top X) = K_k^\top K_k$, where $K_k \in \mathbb{R}^{d_0^k \times n}$, and each column of K_k is given by the k-th

Kronecker product $\mathbf{x}_i \otimes \cdots \otimes \mathbf{x}_i$. Hence, $f_k(X^\top X)$ is positive semi-definite. Therefore,

$$\lambda_{\min}(\Phi_0) \geq 1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2.$$

Since σ is nonlinear and Lipschitz, (4.7.3) holds for σ. Therefore,

$$1 - \zeta_1(\sigma)^2 - \zeta_2(\sigma)^2 - \zeta_3(\sigma)^2 = \sum_{k=4}^{\infty} \zeta_k(\sigma)^2 > 0,$$

and (4.7.4) holds. □

Theorem 67 then follows directly from Lemma 88 and Lemma 90.

Next, we move on to non-asymptotic estimations for NTK. Recall that the empirical NTK matrix H is given by (4.2.4) and the α-th column of S is defined by $\mathrm{diag}(\sigma'(W\mathbf{x}_\alpha))\boldsymbol{a}$, for $1 \leq \alpha \leq n$, in (4.2.5).

The i-th row of S is given by $\mathbf{z}_i^\top := \sigma'(\boldsymbol{w}_i^\top X)a_i$, and $\mathbb{E}[\mathbf{z}_i] = 0$, where a_i is the i-th entry of $\boldsymbol{a}$. Define $D_\alpha = \mathrm{diag}(\sigma'(\boldsymbol{w}_\alpha^\top X)a_\alpha)$, for $1 \leq \alpha \leq d_1$. We can rewrite $(S^\top S) \odot (X^\top X)$ as

$$(S^\top S) \odot (X^\top X) = \sum_{\alpha=1}^{d_1} a_\alpha^2 D_\alpha X^\top X D_\alpha.$$

Let us define L and further expand it as follows:

$$\begin{aligned} L &:= \frac{1}{d_1}(S^\top S - \mathbb{E}[S^\top S]) \odot (X^\top X) && (4.7.5) \\ &= \frac{1}{d_1}\sum_{i=1}^{d_1}(\mathbf{z}_i\mathbf{z}_i^\top - \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^\top]) \odot (X^\top X) \\ &= \frac{1}{d_1}\sum_{i=1}^{d_1}\Big(D_i(X^\top X)D_i - \mathbb{E}[D_i(X^\top X)D_i]\Big) = \frac{1}{d_1}\sum_{i=1}^{d_1} Z_i. && (4.7.6) \end{aligned}$$

Here Z_i is a centered random matrix, and we can apply matrix Bernstein's inequality to show the concentration of L. Since Z_i does not have an almost sure bound on the spectral norm, we will

use the following sub-exponential version of the matrix Bernstein inequality from [Tro12].

Lemma 91 ([Tro12], Theorem 6.2)**.** *Let Z_k be independent Hermitian matrices of size $n \times n$. Assume*

$$\mathbb{E}Z_i = 0, \quad \left\|\mathbb{E}[Z_i^p]\right\| \leq \frac{1}{2}p!R^{p-2}a^2,$$

for any integer $p \geq 2$. Then for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{d_1} Z_i\right\| \geq t\right) \leq n\exp\left(-\frac{t^2}{2d_1a^2 + 2Rt}\right). \tag{4.7.7}$$

Proof of Theorem 68. From (4.7.6), $\mathbb{E}Z_i = 0$, and

$$\|Z_i\| \leq \|D_i\|^2\|XX^\top\| + \mathbb{E}\|D_i\|^2\|XX^\top\| \leq C_1(a_i^2 + 1),$$

where $C_1 = \lambda_\sigma^2\|X\|^2$ and where $a_i \sim \mathcal{N}(0,1)$ is the i-th entry of the second layer weight $\boldsymbol{a}$. Then

$$\begin{aligned}\|\mathbb{E}[Z_i^p]\| \leq \mathbb{E}\|Z_i\|^p \leq C_1^{2p}\mathbb{E}(a_i^2+1)^p &\leq C_1^{2p}\sum_{k=1}^{p}\binom{p}{k}(2k-1)!! \\ &= C_1^{2p}p!\sum_{k=1}^{p}\frac{(2k-1)!!}{k!(p-k)!} \leq C_1^{2p}p!\sum_{k=1}^{p}2^k \leq 2(2C_1^2)^p p!.\end{aligned}$$

So we can take $R = 2C_1^2, a^2 = 8C_1^4$ in (4.7.7) and obtain

$$\mathbb{P}\left(\left\|\sum_{i=1}^{d_1} Z_i\right\| \geq t\right) \leq n\exp\left(-\frac{t^2}{16d_1C_1^4 + 4C_1^2t}\right).$$

Hence, L defined in (4.7.5) has a probability bound:

$$\mathbb{P}(\|L\| \geq t) = \mathbb{P}\left(\frac{1}{d_1}\left\|\sum_{i=1}^{d_1} Z_i\right\| \geq t\right) \leq n\exp\left(-\frac{t^2d_1}{16C_1^4 + 4C_1^2t}\right).$$

Take $t = 10C_1^2\sqrt{\log n/d_1}$. Under the assumption that $d_1 \geq \log n$, we conclude that, with high

probability at least $1-n^{-7/3}$,

$$\|L\| \leq 10C_1^2\sqrt{\frac{\log n}{d_1}}. \tag{4.7.8}$$

Thus, as a corollary, the two statements in Lemma 86 follow from (4.7.8). Meanwhile, since

$$\|H-\mathbb{E}H\| \leq \left\|\frac{1}{d_1}Y^\top Y-\Phi\right\| + \|L\|,$$

the bound in (4.3.13) follows from Theorem 66 and (4.7.8). □

We now proceed to provide a lower bound of $\lambda_{\min}(H)$ from Theorem 68.

Proof of Theorem 69. Note that from (4.2.4), (4.3.6) and (4.7.5), we have

$$\begin{aligned}\lambda_{\min}(H) &\geq \frac{1}{d_1}\lambda_{\min}((S^\top S)\odot(X^\top X))\\ &\geq \frac{1}{d_1}\lambda_{\min}((\mathbb{E}S^\top S)\odot(X^\top X)) - \|L\| = \lambda_{\min}(\Psi)-\|L\|.\end{aligned}$$

Then with Lemma 87, we can get

$$\lambda_{\min}(H) \geq \lambda_{\min}(\Psi_0) - C\varepsilon_n^4 n - \|L\| \geq \left(a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma)\right) - C\varepsilon_n^4 n - \|L\|.$$

Therefore, from Theorem 68, with probability at least $1-n^{-7/3}$,

$$\begin{aligned}\lambda_{\min}(H) &\geq a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) - C\varepsilon_n^4 n - 10\lambda_\sigma^4\|X\|^4\sqrt{\frac{\log n}{d_1}}\\ &\geq a_\sigma - \sum_{k=0}^{2}\eta_k^2(\sigma) - C\varepsilon_n^4 n - 10\lambda_\sigma^4 B^4\sqrt{\frac{\log n}{d_1}}.\end{aligned}$$

Since σ is Lipschitz and non-linear, we know $\sigma'(x)$ is not a linear function (including the constant function) and $|\sigma'(x)|$ is bounded. Suppose that $\sigma'(x)$ has finite many non-zero Hermite

coefficients, $\sigma(x)$ is a polynomial, then we get a contradiction. Hence, the Hermite coefficients of σ' satisfy

$$\sup\{k \in \mathbb{N} : \eta_k^2(\sigma) > 0\} = \infty \text{ and } a_\sigma - \sum_{k=0}^{2} \eta_k^2(\sigma) = \sum_{k=3}^{\infty} \eta_k^2(\sigma) > 0. \tag{4.7.9}$$

This finishes the proof. □

4.8 Proofs of Theorem 70 and Theorem 72

By definitions, the random matrix $K_n(X,X)$ is $\frac{1}{d_1} Y^\top Y$ and the kernel matrix $K(X,X) = \Phi$ is defined in (4.1.2). These two matrices have already been analyzed in Theorem 66 and Theorem 67, so we will apply these results to estimate how great the difference between training errors of random feature regression and its corresponding kernel regression.

Proof of Theorem 70. Denote $K_\lambda := (K + \lambda \operatorname{Id})$. From the definitions of training errors in (4.3.20) and (4.3.21), we have

$$\begin{aligned}
&\left| E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)} \right| = \frac{1}{n} \left| \|\hat{f}_\lambda^{(RF)}(X) - y\|^2 - \|\hat{f}_\lambda^{(K)}(X) - y\|^2 \right| \\
&= \frac{\lambda^2}{n} \left| \operatorname{Tr}[(K(X,X) + \lambda \operatorname{Id})^{-2} y y^\top] - \operatorname{Tr}[(K_n(X,X) + \lambda \operatorname{Id})^{-2} y y^\top] \right| \\
&= \frac{\lambda^2}{n} \left| y^\top \left[(K(X,X) + \lambda \operatorname{Id})^{-2} - (K_n(X,X) + \lambda \operatorname{Id})^{-2} \right] y \right| \\
&\le \frac{\lambda^2}{n} \|(K(X,X) + \lambda \operatorname{Id})^{-2} - (K_n(X,X) + \lambda \operatorname{Id})^{-2}\| \cdot \|y\|^2 \\
&\le \frac{\lambda^2 \|y\|^2}{n \lambda_{\min}^2(K(X,X)) \lambda_{\min}^2(K_n(X,X))} \|(K_\lambda^2 - (K_n(X,X) + \lambda \operatorname{Id})^2\|.
\end{aligned} \tag{4.8.1}$$

Here, in (4.8.1), we employ the identity (2.4.3) in Chapter 2 for $A = (K(X,X) + \lambda \operatorname{Id})^{-2}$ and $B = (K_n(X,X) + \lambda \operatorname{Id})^{-2}$, and the fact that

$$\|(K(X,X) + \lambda \operatorname{Id})^{-1}\| \le \lambda_{\min}^{-1}(K(X,X))$$

and $(K_n(X,X)+\lambda\,\mathrm{Id})^{-1}\| \le \lambda_{\min}^{-1}(K_n(X,X))$. Next, before providing uniform upper bounds for $\lambda_{\min}^{-2}(K(X,X))$ and $\lambda_{\min}^{-2}(K_n(X,X))$ in (4.8.1), we can first get a bound for the last term of (4.8.1) as follows:

$$\begin{aligned}
&\|(K(X,X)+\lambda\,\mathrm{Id})^2-(K_n(X,X)+\lambda\,\mathrm{Id})^2\| \\
&=\|K^2(X,X)-K_n^2(X,X)+2\lambda(K(X,X)-K_n(X,X))\| \\
&\le \|K^2(X,X)-K_n^2(X,X)\|+2\lambda\|(K(X,X)-K_n(X,X))\| \\
&\le \Big(\|K_n(X,X)-K(X,X)\|+2\|K(X,X)\|+2\lambda\Big)\cdot\|K(X,X)-K_n(X,X)\| \\
&\le C\left(\sqrt{\frac{n}{d_1}}+C\right)\sqrt{\frac{n}{d_1}}. \qquad (4.8.2)
\end{aligned}$$

for some constant $C>0$, with probability at least $1-4e^{-2n}$, where the last bound in (4.8.2) is due to Theorem 66 and Lemma 101 in Appendix 4.9. Additionally, combining Theorem 66 and Theorem 67, we can easily get

$$\|(K_n(X,X)+\lambda\,\mathrm{Id})^{-1}\| \le \lambda_{\min}^{-1}(K_n(X,X)) \le C \qquad (4.8.3)$$

for all large n and some universal constant C, under the same event that (4.8.2) holds. Theorem 90 also shows $\lambda_{\min}^{-1}(K(X,X)) \le C$ for all large n. Hence, with the upper bounds for $\lambda_{\min}^{-2}(K(X,X))$ and $\lambda_{\min}^{-2}(K_n(X,X))$, (4.3.22) follows from the bounds of (4.8.1) and (4.8.2). □

For ease of notation, we denote $K := K(X,X)$ and $K_n := K_n(X,X)$. Hence, from (4.3.24), we can further decompose the test errors for K and K_n into

$$\begin{aligned}
\mathscr{L}(\hat{f}_\lambda^{(K)}) = \mathbb{E}_{\mathbf{x}}[|f^*(\mathbf{x})|^2] \\
&+\mathrm{Tr}\Big[(K+\lambda\,\mathrm{Id})^{-1}yy^\top(K+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)]\Big] \qquad (4.8.4)\\
&-2\,\mathrm{Tr}\big[(K+\lambda\,\mathrm{Id})^{-1}y\mathbb{E}_{\mathbf{x}}[f^*(\mathbf{x})K(\mathbf{x},X)]\big],
\end{aligned}$$

and

$$\begin{aligned}\mathscr{L}(\hat{f}_\lambda^{(RF)}) = \mathbb{E}_{\mathbf{x}}[|f^*(\mathbf{x})|^2] \\ + \operatorname{Tr}\Big[(K_n+\lambda\,\mathrm{Id})^{-1}yy^\top(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}[K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)]\Big] \\ - 2\operatorname{Tr}\big[(K_n+\lambda\,\mathrm{Id})^{-1}y\mathbb{E}_{\mathbf{x}}[f^*(\mathbf{x})K_n(\mathbf{x},X)]\big].\end{aligned} \tag{4.8.5}$$

Let us denote

$$\begin{aligned}E_1 &:= \operatorname{Tr}\Big[(K_n+\lambda\,\mathrm{Id})^{-1}yy^\top(K_n+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}[K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)]\Big],\\ \bar{E}_1 &:= \operatorname{Tr}\Big[(K+\lambda\,\mathrm{Id})^{-1}yy^\top(K+\lambda\,\mathrm{Id})^{-1}\mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)]\Big],\\ E_2 &:= \operatorname{Tr}\Big[(K_n+\lambda\,\mathrm{Id})^{-1}y\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}K_n(\mathbf{x},X)]\Big],\\ \bar{E}_2 &:= \operatorname{Tr}\Big[(K+\lambda\,\mathrm{Id})^{-1}y\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]\Big].\end{aligned}$$

As we can see, to compare the test errors between random feature and kernel regression models, we need to control $|E_1-\bar{E}_1|$ and $|E_2-\bar{E}_2|$. Firstly, it is necessary to study the concentrations of

$$\mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)-K_n(\mathbf{x},X)^\top K_n(\mathbf{x},X)]$$

and

$$\mathbb{E}_{\mathbf{x}}[f^*(\mathbf{x})(K(\mathbf{x},X)-K_n(\mathbf{x},X))].$$

Lemma 92. *Under Assumption 9 for σ and Assumption 12 for $\mathbf{x}$ and X, with probability at least $1-4e^{-2n}$, we have*

$$\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\| \le C\sqrt{\frac{n}{d_1}}, \tag{4.8.6}$$

where $C>0$ is a universal constant. Here, we only consider the randomness of the weight matrix in $K_n(\mathbf{x},X)$ defined by (4.3.17) *and* (4.3.18).

Proof. We consider $\tilde{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}]$, its corresponding kernels $K_n(\tilde{X}, \tilde{X})$ and $K(\tilde{X}, \tilde{X}) \in \mathbb{R}^{(n+1)\times(n+1)}$. Under Assumption 12, we can directly apply Theorem 66 to get the concentration of $K_n(\tilde{X}, \tilde{X})$ around $K(\tilde{X}, \tilde{X})$, namely,

$$\left\| K_n(\tilde{X}, \tilde{X}) - K(\tilde{X}, \tilde{X}) \right\| \leq C\sqrt{\frac{n}{d_1}}, \tag{4.8.7}$$

with probability at least $1 - 4e^{-2n}$. Meanwhile, we can write $K_n(\tilde{X}, \tilde{X})$ and $K(\tilde{X}, \tilde{X})$ as block matrices:

$$K_n(\tilde{X}, \tilde{X}) = \begin{pmatrix} K_n(X, X) & K_n(X, \mathbf{x}) \\ K_n(\mathbf{x}, X) & K_n(\mathbf{x}, \mathbf{x}) \end{pmatrix} \text{ and } K(\tilde{X}, \tilde{X}) = \begin{pmatrix} K(X, X) & K(X, \mathbf{x}) \\ K(\mathbf{x}, X) & K(\mathbf{x}, \mathbf{x}) \end{pmatrix}.$$

Since the ℓ_2-norm of any row is bounded above by the spectral norm of its entire matrix, we complete the proof of (4.8.6). □

Lemma 93. *Assume that training labels satisfy Assumption 11 and $\|X\| \leq B$, then for any deterministic $A \in \mathbb{R}^{n\times n}$, we have*

$$\mathrm{Var}\left(y^\top A y\right), \mathrm{Var}\left(\boldsymbol{\beta}^{*\top} A y\right) \leq c\|A\|_F^2,$$

where constant c only depends on $\sigma_{\boldsymbol{\beta}}$, σ_ε and B. Moreover,

$$\mathbb{E}[y^\top A y] = \sigma_{\boldsymbol{\beta}}^2 \mathrm{Tr} A X^\top X + \sigma_\varepsilon^2 \mathrm{Tr} A, \quad \mathbb{E}[\boldsymbol{\beta}^{*\top} A y] = \sigma_{\boldsymbol{\beta}}^2 \mathrm{Tr} A X^\top.$$

Proof. We follow the idea in Lemma C.8 of [MM22] to investigate the variance of the quadratic form for the Gaussian random vector by

$$\mathrm{Var}(g^\top A g) = \|A\|_F^2 + \mathrm{Tr}(A^2) \leq 2\|A\|_F^2, \tag{4.8.8}$$

for any deterministic square matrix A and standard normal random vector g. Notice that the quadratic form

$$y^\top Ay = g^\top \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 XAX^\top & \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} XA \\ \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} AX^\top & \sigma_\varepsilon^2 A \end{pmatrix} g, \tag{4.8.9}$$

where g is a standard Gaussian random vector in $\mathbb{R}^{d_0+n}$. Similarly, the second quadratic form can be written as

$$\boldsymbol{\beta}^{*\top} Ay = g^\top \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 AX^\top & \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} A \\ \mathbf{0} & \mathbf{0} \end{pmatrix} g.$$

Let

$$\tilde{A}_1 := \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 XAX^\top & \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} XA \\ \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} AX^\top & \sigma_\varepsilon^2 A \end{pmatrix}, \quad \tilde{A}_2 := \begin{pmatrix} \sigma_{\boldsymbol{\beta}}^2 AX^\top & \sigma_\varepsilon \sigma_{\boldsymbol{\beta}} A \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

By (4.8.8), we know $\mathrm{Var}\big(y^\top Ay\big) \le 2\|\tilde{A}_1\|_F^2$ and $\mathrm{Var}\big(\boldsymbol{\beta}^{*\top} Ay\big) \le 2\|\tilde{A}_2\|_F^2$. Since

$$\|\tilde{A}_1\|_F^2 = \sigma_{\boldsymbol{\beta}}^4 \|XAX^\top\|_F^2 + \sigma_\varepsilon^2 \sigma_{\boldsymbol{\beta}}^2 \|XA\|_F^2 + \sigma_\varepsilon^2 \sigma_{\boldsymbol{\beta}}^2 \|AX^\top\|_F^2 + \sigma_\varepsilon^4 \|A\|_F^2 \le c\|A\|_F^2$$

and similarly $\|\tilde{A}_2\|_F \le c\|A\|_F^2$ for a constant c, we can complete the proof. □

As a remark, in Lemma 93, for simplicity, we only provide a variance control for the quadratic forms to obtain convergence in probability in the following proofs of Theorems 71 and 72. However, we can apply Hanson-Wright inequalities in Section 4.4 to get more precise probability bounds and consider non-Gaussian distributions for $\boldsymbol{\beta}^*$ and ε.

Proof of Theorem 71. Based on the preceding expansions of $\mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x}))$ and $\mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ in (4.8.4) and (4.8.5), we need to control the right-hand side of

$$\left|\mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x})) - \mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))\right| \le |E_1 - \bar{E}_1| + 2|\bar{E}_2 - E_2|.$$

In the subsequent procedure, we first take the concentrations of E_1 and E_2 with respect to normal random vectors $\boldsymbol{\beta}^*$ and ε, respectively. Then, we apply Theorem 66 and Lemma 92 to complete

the proof of (4.3.25). For simplicity, we start with the second term

$$
\begin{aligned}
|\bar{E}_2 - E_2| &\le \left|\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))](K_n+\lambda\,\mathrm{Id})^{-1}y\right| \\
&\quad + \left|\boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]\big((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1}\big)y\right| \\
&\le |I_1-\bar{I}_1|+|I_2-\bar{I}_2|+|\bar{I}_1|+|\bar{I}_2|, \qquad (4.8.10)
\end{aligned}
$$

where I_1 and I_2 are quadratic forms defined below

$$
\begin{aligned}
I_1 &:= \boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))](K_n+\lambda\,\mathrm{Id})^{-1}y, \\
I_2 &:= \boldsymbol{\beta}^{*\top}\mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]\big((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1}\big)y,
\end{aligned}
$$

and their expectations with respect to random vectors $\boldsymbol{\beta}^*$ and ε are denoted by

$$
\begin{aligned}
\bar{I}_1 &:= \mathbb{E}_{\varepsilon,\boldsymbol{\beta}^*}[I_1] = \sigma_{\boldsymbol{\beta}}^2\,\mathrm{Tr}\Big(\mathbb{E}_{\mathbf{x}}[\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))](K_n+\lambda\,\mathrm{Id})^{-1}X^\top\Big), \\
\bar{I}_2 &:= \mathbb{E}_{\varepsilon,\boldsymbol{\beta}^*}[I_2] = \sigma_{\boldsymbol{\beta}}^2\,\mathrm{Tr}\Big(\big((K_n+\lambda\,\mathrm{Id})^{-1}-(K+\lambda\,\mathrm{Id})^{-1}\big)X^\top\mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]\Big).
\end{aligned}
$$

We first consider the randomness of the weight matrix in K_n and define the event $\mathscr{E}$ where both (4.8.3) and (4.8.7) hold. Then, Theorem 67 and the proof of Lemma 92 indicate that event $\mathscr{E}$ occurs with probability at least $1-4e^{-2n}$ for all large n. Notice that $\mathscr{E}$ does not rely on the randomness of test data $\mathbf{x}$.

We now consider $A = \mathbb{E}_{\mathbf{x}}[\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))](K_n+\lambda\,\mathrm{Id})^{-1}$ in Lemma 93. Conditioning on event $\mathscr{E}$, we have

$$
\begin{aligned}
\|A\|_F^2 &\le \mathbb{E}_{\mathbf{x}}\left[\left\|\mathbf{x}(K_n(\mathbf{x},X)-K(\mathbf{x},X))^\top\right\|_F^2\right]\cdot\left\|(K_n+\lambda\,\mathrm{Id})^{-1}X^\top\right\|^2 \\
&\le \|X\|^2\left\|(K_n+\lambda\,\mathrm{Id})^{-1}\right\|^2\cdot\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\|^2\|K_n(\mathbf{x},X)-K(\mathbf{x},X)\|^2\big] \le C\frac{n}{d_1}, \qquad (4.8.11)
\end{aligned}
$$

for some constant C, where we utilize the assumption $\mathbb{E}[\|\mathbf{x}\|^2]=1$. Hence, based on Lemma 93,

we know $\mathrm{Var}_{\varepsilon,\boldsymbol{\beta}^*}(I_1) \le cn/d_1$, for some constant c. By Chebyshev's inequality and event $\mathscr{E}$,

$$\mathbb{P}\Big(|I_1 - \bar{I}_1| \ge (n/d_1)^{\frac{1-\varepsilon}{2}}\Big) \le c\left(\frac{n}{d_1}\right)^{\varepsilon} + 4e^{-2n}, \tag{4.8.12}$$

for any $\varepsilon \in (0,1/2)$. Hence, $(d_1/n)^{\frac{1}{2}-\varepsilon} \cdot |I_1 - \bar{I}_1| = o(1)$ with probability $1-o(1)$, when $n/d_1 \to 0$ and $n \to \infty$.

Likewise, when $A = \mathbb{E}_{\mathbf{x}}[\mathbf{x}K(\mathbf{x},X)]\big((K_n + \lambda\,\mathrm{Id})^{-1} - (K+\lambda\,\mathrm{Id})^{-1}\big)$, we can apply (2.4.3) and

$$\|K(\mathbf{x},X)\| \le \|K(\tilde{X},\tilde{X})\| \le C\lambda_\sigma^2 B^2, \tag{4.8.13}$$

due to Lemma 101 in Appendix 4.9, to obtain $\|A\|_F^2 \le Cn/d_1$ conditionally on event $\mathscr{E}$. Then, similarly, Lemma 93 shows $\mathrm{Var}_{\varepsilon,\boldsymbol{\beta}^*}(I_2) \le cn/d_1$. Therefore, (4.8.12) also holds for $|I_2 - \bar{I}_2|$.

Moreover, conditioning on the event $\mathscr{E}$,

$$\begin{aligned}
|\bar{I}_1| &= \sigma_{\boldsymbol{\beta}}^2 \Big| \mathbb{E}_{\mathbf{x}}\Big[(K_n(\mathbf{x},X) - K(\mathbf{x},X))(K_n + \lambda\,\mathrm{Id})^{-1} X^\top \mathbf{x}\Big]\Big| \\
&\le \sigma_{\boldsymbol{\beta}}^2 \mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\| \cdot \|K_n(\mathbf{x},X) - K(\mathbf{x},X)\| \cdot \|X\| \cdot \big\|(K_n + \lambda\,\mathrm{Id})^{-1}\big\|\big], \\
&\le \sigma_{\boldsymbol{\beta}}^2 \mathbb{E}_{\mathbf{x}}\big[\|\mathbf{x}\|^2\big]^{\frac{1}{2}} \mathbb{E}_{\mathbf{x}}\Big[\|K_n(\mathbf{x},X) - K(\mathbf{x},X)\|^2\Big]^{\frac{1}{2}} \|X\| \big\|(K_n + \lambda\,\mathrm{Id})^{-1}\big\| \le C\sqrt{\frac{n}{d_1}},
\end{aligned} \tag{4.8.14}$$

for some constant C. In the same way, with (4.8.13), $|\bar{I}_2| \le C\sqrt{\frac{n}{d_1}}$ on the event $\mathscr{E}$. Therefore, from (4.8.10), we can conclude $|\bar{E}_2 - E_2| = o\Big((n/d_1)^{1/2-\varepsilon}\Big)$ for any $\varepsilon \in (0,1/2)$, with probability $1-o(1)$, when $n/d_1 \to 0$ and $n \to \infty$.

Analogously, the first term $|\bar{E}_1 - E_1|$ is controlled by the following four quadratic forms

$$|\bar{E}_1 - E_1| \le \sum_{i=1}^{4} \Big| y^\top A_i y \Big|,$$

where we define by $J_i := y^\top A_i y$ for $1 \le i \le 4$ and

$$
\begin{aligned}
A_1 &:= (K_n + \lambda \,\mathrm{Id})^{-1} \mathbb{E}_{\mathbf{x}}[K_n(\mathbf{x}, X)^\top (K_n(\mathbf{x}, X) - K(\mathbf{x}, X))](K_n + \lambda \,\mathrm{Id})^{-1}, \\
A_2 &:= (K_n + \lambda \,\mathrm{Id})^{-1} \mathbb{E}_{\mathbf{x}}[(K_n(\mathbf{x}, X) - K(\mathbf{x}, X))^\top K(\mathbf{x}, X)](K_n + \lambda \,\mathrm{Id})^{-1}, \\
A_3 &:= \big((K_n + \lambda \,\mathrm{Id})^{-1} - (K + \lambda \,\mathrm{Id})^{-1}\big) \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)](K_n + \lambda \,\mathrm{Id})^{-1}, \\
A_4 &:= (K + \lambda \,\mathrm{Id})^{-1} \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)] \big((K_n + \lambda \,\mathrm{Id})^{-1} - (K + \lambda \,\mathrm{Id})^{-1}\big).
\end{aligned}
$$

Similarly with (4.8.11) and (4.8.14), it is not hard to verify $\|A_i\|_F \le C\sqrt{n/d_1}$ and $|\mathbb{E}_{\varepsilon, \boldsymbol{\beta}^*}[J_i]| \le C\sqrt{n/d_1}$ conditioning on the event $\mathscr{E}$. Then, like (4.8.12), we can invoke Lemma 93 for each A_i to apply Chebyshev's inequality and conclude $|\bar{E}_1 - E_1| = o\Big((n/d_1)^{1/2-\varepsilon}\Big)$ with probability $1 - o(1)$ when $d_1/n \to \infty$, for any $\varepsilon \in (0, 1/2)$. □

Lemma 94. *With Assumptions 9 and 12, for (ε_n, B)-orthonormal X, we have that*

$$
\begin{aligned}
\left\| \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)] - \frac{b_\sigma^4}{d_0} X^\top X \right\| &\le \left\| \mathbb{E}_{\mathbf{x}}[K(\mathbf{x}, X)^\top K(\mathbf{x}, X)] - \frac{b_\sigma^4}{d_0} X^\top X \right\|_F \\
&\le C\sqrt{n}\varepsilon_n^2, \qquad (4.8.15) \\
\left\| \mathbb{E}_{\mathbf{x}}[\mathbf{x} K(\mathbf{x}, X)] - \frac{b_\sigma^2}{d_0} X \right\| &\le \left\| \mathbb{E}_{\mathbf{x}}[\mathbf{x} K(\mathbf{x}, X)] - \frac{b_\sigma^2}{d_0} X \right\|_F \le C\sqrt{n}\varepsilon_n^2, \qquad (4.8.16)
\end{aligned}
$$

for some constant $C > 0$.

Proof. By Lemma 100, we have an entrywise approximation

$$
|K(\mathbf{x}, \mathbf{x}_i) - b_\sigma^2 \mathbf{x}^\top \mathbf{x}_i| \le C\lambda_\sigma \varepsilon_n^2,
$$

for any $1 \le i \le n$. Hence, $\|K(\mathbf{x}, X) - b_\sigma^2 \mathbf{x}^\top X\| \le C\lambda_\sigma \sqrt{n}\varepsilon_n^2$. Assumption 12 of $\mathbf{x}$ implies that

$\frac{b_\sigma^4}{d_0} X^\top X = b_\sigma^4 \mathbb{E}_{\mathbf{x}}[X^\top \mathbf{x}\mathbf{x}^\top X]$. Then, we can verify (4.8.15) based on the following approximation

$$\begin{aligned}
&\left\| \mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)] - \frac{b_\sigma^4}{d_0} X^\top X \right\|_F \le \mathbb{E}_{\mathbf{x}}\Big[\Big\| K(\mathbf{x},X)^\top K(\mathbf{x},X) - b_\sigma^4 X^\top \mathbf{x}\mathbf{x}^\top X \Big\|_F \Big] \\
&\le \mathbb{E}_{\mathbf{x}}\Big[\Big\| K(\mathbf{x},X)^\top \Big(K(\mathbf{x},X) - b_\sigma^2 \mathbf{x}^\top X \Big) \Big\|_F + b_\sigma^2 \Big\| \Big(K(\mathbf{x},X)^\top - b_\sigma^2 X^\top \mathbf{x} \Big) \mathbf{x}^\top X \Big\|_F \Big] \\
&\le \mathbb{E}_{\mathbf{x}}\Big[\| K(\mathbf{x},X) - b_\sigma^2 \mathbf{x}^\top X \| \Big(\| K(\mathbf{x},X) \| + \| b_\sigma^2 \mathbf{x}^\top X \| \Big) \Big] \le C\sqrt{n}\varepsilon_n^2,
\end{aligned}$$

for some universal constant C. The same argument can also be employed to prove (4.8.16), so details will be omitted here. □

Proof of Theorem 72. From (4.3.22) and (4.3.25), we can easily conclude that

$$E_{\text{train}}^{(RF,\lambda)} - E_{\text{train}}^{(K,\lambda)} \xrightarrow{\mathbb{P}} 0, \tag{4.8.17}$$

$$\mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x})) - \mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x})) \xrightarrow{\mathbb{P}} 0, \tag{4.8.18}$$

as $n \to \infty$ and $n/d_1 \to 0$. Therefore, to study the training error $E_{\text{train}}^{(RF,\lambda)}$ and the test error $\mathscr{L}(\hat{f}_\lambda^{(RF)}(\mathbf{x}))$ of random feature regression, it suffices to analyze the asymptotic behaviors of $E_{\text{train}}^{(K,\lambda)}$ and $\mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ for the kernel regression, respectively. In the rest of the proof, we will first analyze the test error $\mathscr{L}(\hat{f}_\lambda^{(K)}(\mathbf{x}))$ and then compute the training error $E_{\text{train}}^{(K,\lambda)}$ under the ultra-wide regime.

Recall that $K_\lambda = (K + \lambda\,\mathrm{Id})$ and the test error is given by

$$\mathscr{L}(\hat{f}_\lambda^{(K)}) = \frac{1}{d_0}\|\boldsymbol{\beta}^*\|^2 + L_1 - 2L_2, \tag{4.8.19}$$

where $L_1 := y^\top K_\lambda^{-1} \mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)] K_\lambda^{-1} y$, $L_2 := \boldsymbol{\beta}^{*\top} \mathbb{E}_{\mathbf{x}}[\mathbf{x} K(\mathbf{x},X)] K_\lambda^{-1} y$. The spectral norm of K_λ is bounded from above and the smallest eigenvalue is bounded from below by some positive constants.

We first focus on the last two terms L_1 and L_2 in the test error. Let us define

$$\widetilde{L}_1 := \frac{b_\sigma^4}{d_0} y^\top K_\lambda^{-1} X^\top X K_\lambda^{-1} y \quad \text{and} \quad \widetilde{L}_2 := \frac{b_\sigma^2}{d_0} \boldsymbol{\beta}^{*\top} X K_\lambda^{-1} y.$$

Then, we obtain two quadratic forms

$$\begin{aligned} L_1 - \widetilde{L}_1 &= y^\top K_\lambda^{-1} \left(\mathbb{E}_{\mathbf{x}}[K(\mathbf{x},X)^\top K(\mathbf{x},X)] - \frac{b_\sigma^4}{d_0} X^\top X \right) K_\lambda^{-1} y =: y^\top A_1 y, \\ L_2 - \widetilde{L}_2 &= \boldsymbol{\beta}^{*\top} \left(\mathbb{E}_{\mathbf{x}}[\mathbf{x} K(\mathbf{x},X)] - \frac{b_\sigma^2}{d_0} X \right) K_\lambda^{-1} y =: \boldsymbol{\beta}^{*\top} A_2 y, \end{aligned}$$

where $\|A_1\|_F$ and $\|A_2\|_F$ are at most $C\sqrt{n}\varepsilon_n^2$ for some constant $C > 0$, due to Lemma 94. Hence, applying Lemma 93 for these two quadratic forms, we have $\mathrm{Var}(L_i - \widetilde{L}_i) \le cn\varepsilon_n^4 \to 0$ as $n \to \infty$. Additionally, Lemma 93 and the proof of Lemma 94 verify that $\mathbb{E}[y^\top A_1 y]$ and $\mathbb{E}[\boldsymbol{\beta}^{*\top} A_2 y]$ are vanishing as $n \to \infty$. Therefore, $L_i - \widetilde{L}_i$ converges to zero in probability for $i = 1, 2$. So we can move to analyze $\widetilde{L}_1$ and $\widetilde{L}_2$ instead. Copying the above procedure, we can separately compute the variances of $\widetilde{L}_1$ and $\widetilde{L}_2$ with respect to $\boldsymbol{\beta}^*$ and ε, and then apply Lemma 93. Then, $|\widetilde{L}_1 - \bar{L}_1|$ and $|\widetilde{L}_2 - \bar{L}_2|$ will converge to zero in probability as $n, d_0 \to \infty$, where

$$\begin{aligned} \bar{L}_1 &:= \mathbb{E}_{\varepsilon, \boldsymbol{\beta}^*}[\widetilde{L}_1] = \frac{b_\sigma^4 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X K_\lambda^{-1} X^\top X + \frac{b_\sigma^4 \sigma_\varepsilon^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X K_\lambda^{-1}, \\ \bar{L}_2 &:= \mathbb{E}_{\varepsilon, \boldsymbol{\beta}^*}[\widetilde{L}_2] = \frac{b_\sigma^2 \sigma_{\boldsymbol{\beta}}^2 n}{d_0} \operatorname{tr} K_\lambda^{-1} X^\top X. \end{aligned}$$

To obtain the last approximation, we define $\bar{K}(X,X) := b_\sigma^2 X^\top X + (1 - b_\sigma^2) \operatorname{Id}$ and

$$\bar{K}_\lambda := b_\sigma^2 X^\top X + (1 + \lambda - b_\sigma^2) \operatorname{Id}. \tag{4.8.20}$$

We aim to replace K_λ by $\bar{K}_\lambda$ in $\bar{L}_1$ and $\bar{L}_2$. Recalling the identity (2.4.3), we have

$$K_\lambda^{-1} - \bar{K}_\lambda^{-1} = \bar{K}_\lambda^{-1} (K(X,X) - \bar{K}(X,X)) K_\lambda^{-1}.$$

Since σ is not a linear function, $1-b_\sigma^2>0$. Then, with (4.8.3), the proof of Lemma 85 indicates

$$\left\|K_\lambda^{-1}-\bar{K}_\lambda^{-1}\right\|_F \le C\sqrt{n^2\varepsilon_n^4+n\varepsilon_n^2}, \tag{4.8.21}$$

where we apply the fact that $\lambda_{\min}(\bar{K}(X,X)) \ge 1-b_\sigma^2>0$. Let us denote

$$L_1^0 := \frac{b_\sigma^4\sigma_{\boldsymbol{\beta}}^2 n}{d_0}\operatorname{tr}\bar{K}_\lambda^{-1}X^\top X\bar{K}_\lambda^{-1}X^\top X + \frac{b_\sigma^4\sigma_\varepsilon^2 n}{d_0}\operatorname{tr}\bar{K}_\lambda^{-1}X^\top X\bar{K}_\lambda^{-1}, \tag{4.8.22}$$

$$L_2^0 := \frac{b_\sigma^2\sigma_{\boldsymbol{\beta}}^2 n}{d_0}\operatorname{tr}\bar{K}_\lambda^{-1}X^\top X. \tag{4.8.23}$$

Notice that for any matrices $A,B\in\mathbb{R}^{n\times n}$, $\|AB\|_F \le \|A\|\|B\|_F, |\operatorname{Tr}(AB)| \le \|A\|_F\|B\|_F$. Then, with the help of (4.8.21) and uniform bounds of the spectral norms of $X^\top X$, K_λ^{-1} and $\bar{K}_\lambda^{-1}$, we obtain that

$$\begin{aligned}
&|\bar{L}_1 - L_1^0| \\
\le\ & \frac{b_\sigma^4\sigma_{\boldsymbol{\beta}}^2}{d_0}\left|\operatorname{Tr}K_\lambda^{-1}X^\top X(K_\lambda^{-1}-\bar{K}_\lambda^{-1})X^\top X\right| + \frac{b_\sigma^4\sigma_{\boldsymbol{\beta}}^2}{d_0}\left|\operatorname{Tr}(K_\lambda^{-1}-\bar{K}_\lambda^{-1})X^\top X\bar{K}_\lambda^{-1}X^\top X\right| \\
&+ \frac{b_\sigma^4\sigma_\varepsilon^2}{d_0}\left|\operatorname{Tr}(K_\lambda^{-1}-\bar{K}_\lambda^{-1})X^\top X\bar{K}_\lambda^{-1}\right| + \frac{b_\sigma^4\sigma_\varepsilon^2}{d_0}\left|\operatorname{Tr}K_\lambda^{-1}X^\top X(K_\lambda^{-1}-\bar{K}_\lambda^{-1})\right| \\
\le\ & \frac{C\sqrt{n}}{d_0}\left\|K_\lambda^{-1}-\bar{K}_\lambda^{-1}\right\|_F \le C\frac{n}{d_0}\sqrt{n\varepsilon_n^4+\varepsilon_n^2}\to 0,
\end{aligned}$$

as $n\to\infty$, $n/d_0\to\gamma$ and $n\varepsilon_n^4\to 0$. Combining all the approximations, we conclude that L_i and L_i^0 have identical limits in probability for $i=1,2$. On the other hand, based on the assumption of X and definitions in (4.8.20), (4.8.22) and (4.8.23), it is not hard to check that

$$\begin{aligned}
\lim_{n\to\infty} L_1^0 &= b_\sigma^4\sigma_{\boldsymbol{\beta}}^2\gamma\int_{\mathbb{R}}\frac{x^2}{(b_\sigma^2 x+1+\lambda-b_\sigma^2)^2}d\mu_0(x) + b_\sigma^4\sigma_\varepsilon^2\gamma\int_{\mathbb{R}}\frac{x}{(b_\sigma^2 x+1+\lambda-b_\sigma^2)^2}d\mu_0(x), \\
\lim_{n\to\infty} L_2^0 &= b_\sigma^2\sigma_{\boldsymbol{\beta}}^2\gamma\int_{\mathbb{R}}\frac{x}{b_\sigma^2 x+1+\lambda-b_\sigma^2}d\mu_0(x).
\end{aligned}$$

Therefore, L_1 and L_2 converge in probability to the above limits, respectively, as $n\to\infty$. In the

end, we apply the concentration of the quadratic form $\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^*$ in (4.8.19) to get $\frac{1}{d_0}\|\boldsymbol{\beta}^*\|^2 \xrightarrow{\mathbb{P}} \sigma_{\boldsymbol{\beta}}^2$. Then, by (4.8.18), we can get the limit in (4.3.28) for the test error $\mathscr{L}(\hat{f}_\lambda^{(RF)})$. As a byproduct, we can even use L_1^0 and L_2^0 to form an n-dependent deterministic equivalent of $\mathscr{L}(\hat{f}_\lambda^{(RF)})$ as well.

Thanks to Lemma 93, the training error, $E_{\text{train}}^{(K,\lambda)} = \frac{\lambda^2}{n} y^\top K_\lambda^{-2} y$, analogously, concentrates around its expectation with respect to $\boldsymbol{\beta}^*$ and ε, which is $\sigma_{\boldsymbol{\beta}}^2\lambda^2 \operatorname{tr} K_\lambda^{-2} X^\top X + \sigma_\varepsilon^2 \lambda^2 \operatorname{tr} K_\lambda^{-2}$. Moreover, because of (4.8.21), we can further substitute K_λ^{-2} by $\bar{K}_\lambda^{-2}$ defined in (4.8.20). Hence, we know that, asymptotically,

$$\left| E_{\text{train}}^{(K,\lambda)} - \sigma_{\boldsymbol{\beta}}^2\lambda^2 \operatorname{tr} \bar{K}_\lambda^{-2} X^\top X - \sigma_\varepsilon^2\lambda^2 \operatorname{tr} \bar{K}_\lambda^{-2} \right| \xrightarrow{\mathbb{P}} 0,$$

where as $n, d_0 \to \infty$,

$$\lim_{n\to\infty} \sigma_{\boldsymbol{\beta}}^2\lambda^2 \operatorname{tr} \bar{K}_\lambda^{-2} X^\top X = \sigma_{\boldsymbol{\beta}}^2\lambda^2 \int_{\mathbb{R}} \frac{x}{(b_\sigma^2 x + 1 + \lambda - b_\sigma^2)^2} d\mu_0(x), \tag{4.8.24}$$

$$\lim_{n\to\infty} \sigma_\varepsilon^2\lambda^2 \operatorname{tr} \bar{K}_\lambda^{-2} = \sigma_\varepsilon^2\lambda^2 \int_{\mathbb{R}} \frac{1}{(b_\sigma^2 x + 1 + \lambda - b_\sigma^2)^2} d\mu_0(x). \tag{4.8.25}$$

The last two limits are due to $\mu_0 = \lim \operatorname{spec} X^\top X$ as $n, d_0 \to \infty$. Therefore, by (4.8.17), we obtain our final result (4.3.27) in Theorem 72. □

4.9 Auxiliary Lemmas

Lemma 95 (Equation (3.7.9) in [Joh90])**.** *Let A, B be two $n \times n$ matrices, A be positive semidefinite, and $A \odot B$ be the Hadamard product between A and B. Then,*

$$\|A \odot B\| \le \max_{i,j} |A_{ij}| \cdot \|B\|. \tag{4.9.1}$$

Lemma 96 (Sherman–Morrison formula, [Bar51])**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square*

matrix and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ *are column vectors. Then*

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}}. \tag{4.9.2}$$

Lemma 97 (Theorem A.45 in [BS10])**.** *Let* A, B *be two* $n \times n$ *Hermitian matrices. Then* A *and* B *have the same limiting spectral distribution if* $\|A - B\| \to 0$ *as* $n \to \infty$*.*

Lemma 98 (Theorem B.11 in [BS10])**.** *Let* $z = x + iv \in \mathbb{C}, v > 0$ *and* $s(z)$ *be the Stieltjes transform of a probability measure. Then* $|\operatorname{Re} s(z)| \le v^{-1/2}\sqrt{\operatorname{Im} s(z)}$*.*

Lemma 99 (Lemma D.2 in [NM20])**.** *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *such that* $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ *and* $\boldsymbol{w} \sim \mathcal{N}(0, I_d)$*. Let* h_j *be the* j*-th normalized Hermite polynomial given in Definition 63. Then*

$$\mathbb{E}_{\boldsymbol{w}}[h_j(\langle \boldsymbol{w}, \mathbf{x}\rangle) h_k(\langle \boldsymbol{w}, \mathbf{y}\rangle)] = \delta_{jk}\langle \mathbf{x}, \mathbf{y}\rangle^k.$$

Lemma 100. *Recall the definition of* Φ *in* (4.1.2)*. Under Assumption 9, if* X *is* (ε, B)*-orthonormal with sufficiently small* ε*, then for a universal constant* $C > 0$ *and any* $\alpha \neq \beta \in [n]$*, we have*

$$|\Phi_{\alpha\beta} - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta| \le C\varepsilon^2,$$
$$|\mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \mathbf{x}_\alpha)]| \le C\varepsilon.$$

Proof. When σ is twice differentiable in Assumption 9, this result follows from Lemma 16. When σ is a piece-wise linear function defined in case 2 of Assumption 9, the second inequality follows from (4.6.7) with $t = \|\mathbf{x}_\alpha\|$. For the first inequality, the Hermite expansion of $\Phi_{\alpha\beta}$ is given by (4.6.3) with coefficients $\zeta_k(\sigma_\alpha) = \mathbb{E}[\sigma(\|\mathbf{x}_\alpha\|\xi) h_k(\xi)]$ for $k \in \mathbb{N}$. Observe that the piece-wise linear function in case 2 of Assumption 9 satisfies

$$\zeta_k(\sigma_\alpha) = \|\mathbf{x}_\alpha\| \zeta_k(\sigma), \text{ for } k \ge 1,$$
$$\zeta_0(\sigma_\alpha) = b(1 - \|\mathbf{x}_\alpha\|),$$

because of condition (4.2.6) for σ. Recall $\mathbf{u}_\alpha = \mathbf{x}_\alpha/\|\mathbf{x}_\alpha\|$ and $\zeta_1(\sigma) = b_\sigma$. Then, analogously to the derivation of (4.6.10), there exists some constant $C > 0$ such that

$$|\Phi_{\alpha\beta} - b_\sigma^2 \mathbf{x}_\alpha^\top \mathbf{x}_\beta| = \left| \sum_{k\neq 1} \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)(\mathbf{u}_\alpha^\top \mathbf{u}_\beta)^k \right| \\ \leq b^2(1-\|\mathbf{x}_\alpha\|)(1-\|\mathbf{x}_\beta\|) + \frac{|\mathbf{x}_\alpha^\top \mathbf{x}_\beta|^2}{\|\mathbf{x}_\alpha\|\|\mathbf{x}_\beta\|}\|\sigma\|_{L^2}^2 \leq C\varepsilon^2,$$

for $\varepsilon \in (0,1)$ and (ε, B)-orthonormal X. This completes the proof of this lemma. □

With the above lemma, the proof of Lemma 17 directly yields the following lemma.

Lemma 101. *Under the same assumptions as Lemma 100, there exists a constant $C > 0$ such that $\|K(X,X)\| \leq CB^2$. Additionally, with Assumption 12, we have $\|K(\tilde{X},\tilde{X})\| \leq CB^2$.*

4.10 Acknowledgment

Chapter 4 is extracted from "Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability* 34, no. 2 (2024): 1896-1947". The thesis author is the co-author of this paper.

Chapter 5

Spectral Analysis in Trained Neural Networks

In many theoretical results, the conjugate kernel (CK) and the neural tangent kernel (NTK) remain fixed throughout training, which leads to a kernel gradient descent with the initial kernel [JGH18, BM19], whereas in practice the spectra of the weight matrix, CK, and NTK of the NN change tremendously while learning the features from the training data [MM19, MM21, FW20, CHS20, OJMDF21]. In this Chapter, under the linear-width regime, we experimentally and theoretically explore the following question:

How do the spectra of weight and kernel matrices of the NN evolve during the training process?

This question is crucial to extend our understanding beyond the kernel regime. It will help us analyze the generalization of the NN in instances when it performs better than the kernel machine. For this case, the spectral properties of the trained NN could be entirely different from the initial kernel [Lon21, BGL$^+$21, SB21]. Also, various spectral properties of weight and kernel matrices can reveal different features learned by different training procedures [WHS22]. Understanding the dynamics of the spectral properties may aid in finding better approaches to training and tuning hyper-parameters for NNs. From a theoretical perspective, random matrix theory (RMT) can be further exploited to study and elucidate the NN training under the proportional limit in high dimensions [LCKS91, PB17, LLC18, PW18, HN20, MM22].

5.1 Related Work

Global convergence of GD for ultra wide NNs.

A recent line of work has shown the global convergences of the learning dynamics of gradient-based methods in a certain overparameterized regime, e.g. [DZPS19b, DLL$^+$19a, OFLS19, OS19, Ngu21, LZB22, PC22, Cha22]. We refer to Table 1 of [PC22] as a summary of these recent results. Most of the theorems in the literature require $h \gg n$, which implies that the NTK is almost static during training, while [OS19, Ngu21] can consider LWR under some specific assumptions. Recently, [Cha22] established a new criterion for the convergence of GD which results in the global convergence of general NNs with finite width h and $d \geq n$.

Beyond NTK regime.

Under the proportional limit, the initial kernel regression can only learn a linear component of the target [GMMM21]. Thus, it is reasonable to consider the cases beyond the NTK regime. To this end, [DGA20, HY20] considered the dynamics of NTK throughout training while [AZLL19, BL20] have shown a second-order approximation of NTK, outperforming the initial kernel. In addition, there are many theoretical works analyzing when a NN outperforms the initial kernels in some specific settings: [LMZ20] proved a two-layer ReLU NN that is shown to beat any kernel method; [KWLS21] verified a two-layer CNN with some simple dataset can outperform the initial NTK for image classifications; [BES$^+$22] showed a NN can escape the kernel regime by only taking one specific large gradient step; [DLS22] showed a specific gradient-based training can even learn polynomials with low-dimensional latent representation.

Evolution of NTK and alignment in NNs.

The feature learning can be characterized by the evolution of the kernel during training [FDP$^+$20, OJMDF21, Lon21, ABP22, LHAR22]. Specifically, [Lon21] studied the hard-margin SVM for "after kernels" which are the CK and NTK matrices of trained NNs. One of the effective ways of depicting how the kernels evolve during training is to capture the evolution of kernel alignment [BGL$^+$21, SB21, ABP22, LHAR22]. Kernel alignments between kernels and

training labels essentially reveal how the NN accelerates training [SB21]. Also, several papers showed that the top eigenfunctions of the kernel align with the target function learned by the NN [KI20, OJMMF20, OJMDF21]. This becomes an efficient way of analyzing how NNs learn features through a particular gradient-based optimization.

Large learning rate regime.

As mentioned earlier, the large learning rate may contribute to feature learning. The benefits of large-learning-rate training have been studied from different aspects [LWM19, Nak20, BMR22, AVPVF23]. Specifically, [LM20] observed that training dynamics with large learning rates differ from the small learning rate regime, where the latter regime exhibits monotone and fast convergence of training loss but may not generalize well on test data. At the early phase of training, [JSF$^+$20] showed using lower learning rates may result in finding a region of the loss surface with worse conditioning of kernel and Hessian matrices. In [Lon21], the after kernels of NNs trained with larger learning rates generalize better and stay more stable. [LBD$^+$20] raised a "catapult mechanism", where gradient descent dynamics converge to flatter minima for extremely large learning rates. There is a transition as a function of the learning rate, from lazy training to the catapult regime. Section 5.2.2 illustrates a similar transition in our situations.

Heavy-tailed phenomenon.

The heavy-tailed phenomenon has appeared in many places in deep learning theory. [MM19] and [MM21] observed that many state-of-the-art pre-trained models obtain heavy-tailed weight spectra. More precisely, these spectra have a "5+1" phase transition which relates to different degrees of regularization of the NN. With this heavy-tailed self-regularization theory, [MPM21] further showed how to distinguish well-trained and poorly trained models by a power-law-based approximation. [MY23] classified trained weight spectra into three types: Marčenko–Pastur law, bulk with (few) outliers, and heavy-tailed spectra. We extend this classification to both weight and kernel matrices in Figure 5.1. Additionally, similarly to the discussion in 5.2.3, [MY23] showed that the difficulty of the classification problem is related to the emer-

gence of heavy-tailed spectra in weight matrices. This heavy-tailed phenomenon can be used to construct metrics for evaluating the generalization of NNs [MPM21, YTH+22], and early stopping of NNs to avoid over-fitting [MY23].

5.2 Empirical Study for the Spectra in Trained NNs

In this section, we empirically investigate a two-layer NN with synthetic data. This setting is promising for future theoretical studies by virtue of RMT. We will showcase the evolution of its spectral properties through the training process of a two-layer NN in (1.1.1) with $L = 2$ defined by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{1}{\sqrt{h}} \sum_{i=1}^{h} v_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}/\sqrt{d}). \tag{5.2.1}$$

At initialization, we assume that the first hidden-layer $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_h]^\top \in \mathbb{R}^{h \times d}$ is composed of independent standard normal random vectors and all entries of $\boldsymbol{v} := [v_1, \ldots, v_h]^\top \in \mathbb{R}^h$ are independently distributed either by $\mathcal{N}(0,1)$ or by $\mathrm{Unif}(-1,1)$. We consider the dataset size n to be proportional to width h and feature dimension d, i.e. linear-width regime (LWR).

Assumption 13 (Synthetic dataset and teacher model)**.** Training data is $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, where $\boldsymbol{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. The training labels $\boldsymbol{y} = [y_1, \ldots, y_n]$ are defined by $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$, for $i \in [n]$, where $f^* : \mathbb{R}^d \to \mathbb{R}$ is the teacher model, and ε_i is centered sub-Gaussian noise with variance σ_ε^2.

One of the simplest nonlinear teacher models we can generate is the single-index model, namely $f^*(\boldsymbol{x}) = \sigma^*(\boldsymbol{x}^\top \boldsymbol{\beta})$ for a fixed vector $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$ and nonlinear function σ^*; the hidden feature is simply $\boldsymbol{\beta} \in \mathbb{R}^d$. In general, we can consider a multiple-index model

$$f^*(\boldsymbol{x}) = \frac{1}{k} \sum_{i=1}^{k} \sigma^*(\boldsymbol{x}^\top \boldsymbol{\beta}_i) \tag{5.2.2}$$

where $\boldsymbol{\beta}_i$ are some orthogonal unit vectors. We will specifically consider a mixture of single-

Table 5.1. Four models with the same architecture ($n = 2000$, $h = 1500$, $d = 1000$, and σ is normalized tanh), but different choices of initial learning rates and optimization tools. The last column summarizes the spectral behavior of weight and kernel matrices after training.

	Optimization	Learning rate η	R^2 score	Test error	Spectra
Case 1	GD	5.0	0.63582	0.36381	Invariant Bulk
Case 2	SGD	0.1	0.60605	0.36879	Invariant Bulk
Case 3	SGD	22.0	0.76081	0.23791	Bulk+spike
Case 4	Adam	0.092	**0.78829**	**0.21071**	Heavy tail
	Lazy regime		0.68092	0.3185	

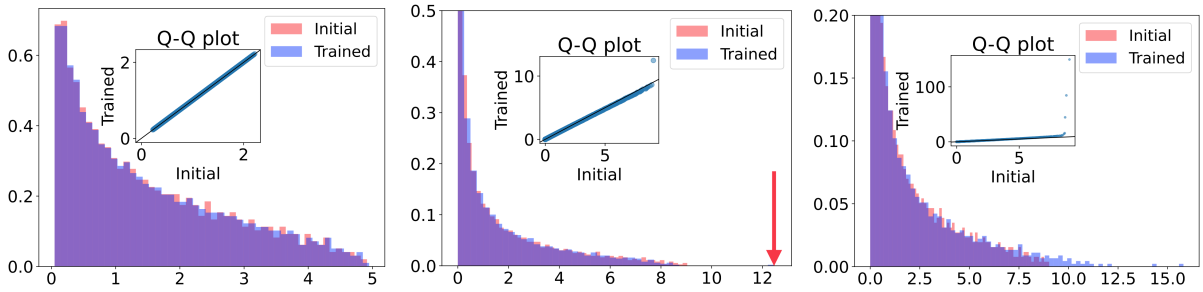
index and quadratic models as our teacher model in this section:

$$f^*(\boldsymbol{x}) = \sigma^*(\boldsymbol{x}^\top \boldsymbol{\beta}) + \frac{\tau}{d}\|\boldsymbol{x}\|^2, \tag{5.2.3}$$

for some nonlinear target σ^*, signal $\boldsymbol{\beta}$ and constant τ. Here, the norm term of $\boldsymbol{x}$ in (5.2.3) is designed to make the teacher model more complicated to be learned. All our empirical results still hold when $\tau = 0$. The advantage of this toy model is that we can easily extract the spectral behaviors over training and then compare them with the kernel machine. We use lazy training defined in (1.1.9) as our *benchmark* to assist us in determining whether a neural network outperforms the associated kernel machine (Table 5.1).

Following the above assumptions and constructions, we show different spectral properties (Figure 5.1) for this two-layer NN using different training procedures (Table 5.1). Figure 5.1 exhibits three types of spectra after training: unchanged bulk distribution, bulk with one spike, and heavy tail in spectra. Putting things together, Table 5.1 exhibits close relationships between the spectra and the generalization of the NN. These different spectral properties actually reveal disparate features learned via different training strategies.

Table 5.1 compares the test errors and R^2 scores for different optimization cases and the lazy training. By tuning the hyper-parameters, we can find specific situations where NN outperforms the lazy training. Here in Table 5.1, $n = 2000$, $h = 1500$, $d = 1000$, and σ is



(a) Weight spectra in Case 1. (b) CK spectra in Case 2. (c) CK spectra in Case 3.

Figure 5.1. Different spectral behaviors in Table 5.1: (a) The initial and trained spectra of $\boldsymbol{W}$ in Case 1. The spectrum is invariant based on the Q-Q subplot. (b) The initial and trained spectra of $\boldsymbol{K}^{\mathrm{CK}}$ in Case 3. There is an outlier (red arrow) after training. (c) The initial and trained spectra of $\boldsymbol{K}^{\mathrm{CK}}$ in Case 4.

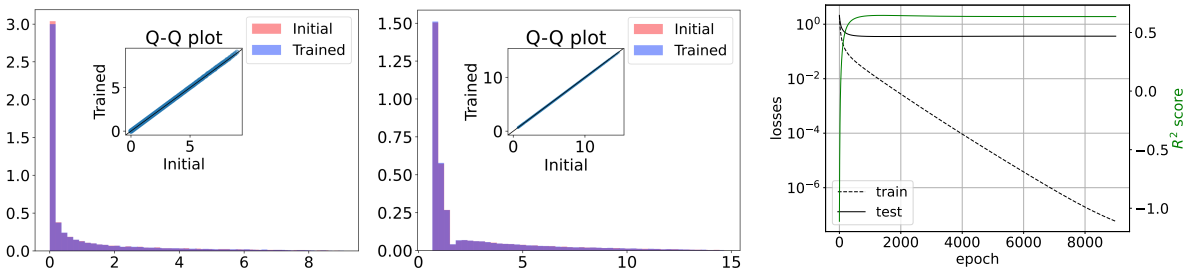
normalized tanh in (5.2.1). The training label noise $\sigma_\varepsilon = 0.3$ and the teacher model is defined by (5.2.3) with σ^* a normalized *softplus* and $\tau = 0.2$. We observe that simply choosing an optimizer and learning rate can affect the shapes of the final spectra and the performance of the NN, as measured by R^2 scores and test errors. Figure 5.1 presents the spectra of weight/kernel matrices of the initial/trained NN in different cases of Table 5.1. Notice that the subfigures in Figure 5.1 present the Q-Q plots to compare the initial and trained spectra for different cases.

We now further explore the spectral behaviors in different cases of Table 5.1 by clarifying how the spectra evolve through different training processes and how this evolution may affect the NN. Following Figure 5.1, we study the training processes case-by-case: invariant bulk, spikes outside the bulk, and heavy-tailed distribution. Figure 5.1 summarizes these three situations for both weights and empirical kernels.

5.2.1 Invariant Spectra Throughout Training

In Figure 5.1(a), we observe the bulk distributions of the weight matrix and CK in Cases 1&2 remain globally unchanged (invariant) over certain training processes, respectively. We now further explore the invariant spectra in Cases 1&2 when training NNs.

In the setting of Case 1, Figures 5.1(a) and 5.2 present results of the weight matrix and NTK on GD training and indicate evidence of kernel regime in Case 1. This shows that, from a spectral point of view, the weight matrix, CK, and NTK are almost invariant and static

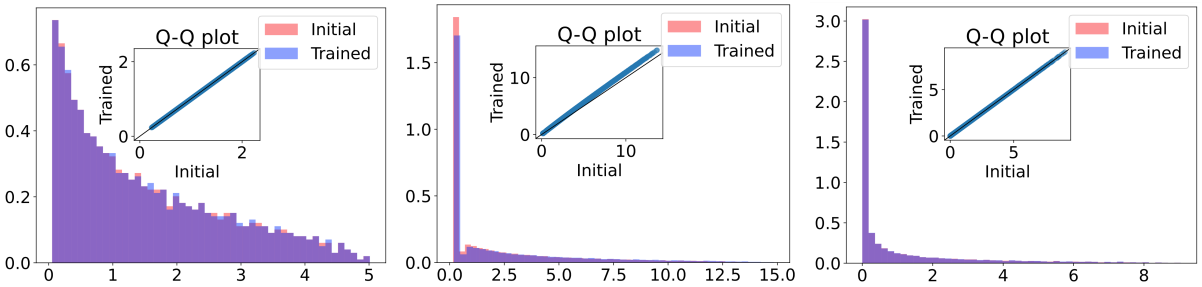


(a) CK spectra in Case 1. (b) NTK spectra in Case 1. (c) Losses and R^2 scores.

Figure 5.2. Performance of Case 1 in Table 5.1: (a) The initial and trained spectra of the first-hidden layer $\boldsymbol{W}$. (b) The initial and trained spectra of empirical NTK matrix defined by (1.1.5). Q-Q subplot shows these two spectra are almost the same. (c) Training and test losses and R^2 scores vs. epochs for GD.

through training. The initial spectrum of weight $\boldsymbol{W}_0$ converges to Marčenko–Pastur law; the initial spectrum of NTK under proportional limit has been studied in Chapter 2. Based on Figures 5.1(a) and 5.2, we can empirically verify that, globally, the spectra of $\boldsymbol{W}$, $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ are not changing over training as $n/d \to \gamma_1$ and $h/d \to \gamma_2$. This can be explained by the fact that NN trained by GD with a small learning rate will eventually converge to a global minimum point that is close to the initialization. Figure 5.2(c) demonstrates the global convergence for GD under the proportional regime, as proved in Theorem 103 from Section 5.4.1. In Section 5.4.1, by investigating the global convergence of GD, we prove this invariant-bulk phenomenon under some specific assumptions.

As a complement of Figure 5.1(b), Figure 5.3 exhibits the spectra of $\boldsymbol{W}$, $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ for Case 2 in Table 5.1. The phenomena are similar to Case 1. This observation provides evidence that all results in Section 5.4.1 can be extended to SGD training with sufficiently small learning rates, which is subject to future work. Analogously to GD case, we conjecture that the global convergence when training both layers of NN with SGD still holds in this proportional limit. In summary, from Figures 5.1(a), 5.2 and 5.3, one can observe the spectral distributions of the weight, CK and NTK matrices remain invariant and static during training in Cases 1&2, which indicates both cases still belong to the lazy regime. This spectral invariance impedes further feature learning during the training process.



(a) Weight spectra. (b) NTK spectra. (c) CK spectra.

Figure 5.3. Spectral properties for Case 2 in Table 5.1: (a) The initial and trained spectra of the first-hidden layer $\boldsymbol{W}$. (b) The initial and trained spectra of empirical NTK are defined by (1.1.5). (c) The initial and trained spectra of empirical CK defined by (1.1.4).

5.2.2 Emergence of Outliers and Spike Alignments

As suggested in Figure 5.1(b) for Case 3 in Table 5.1, the outlier eigenvalues may appear in the spectra of the trained weight matrix, CK, and NTK when NNs are optimized with large learning rates. Heuristically, this indicates that the NN is learning the feature from the teacher model f^*. We further explore this phenomenon in this section. Additionally, in the Appendix of [WES+23], one can find out more experiments on this phenomenon for $\boldsymbol{W}$, $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ through different training processes.

Spike alignments of weight matrices.

The differences between Cases 2&3 empirically validate the benefits of training with large learning rates [LWM19, Nak20, Lon21, BMR22, AVPVF23]. Inspired by [BES+22], we consider the alignment between the leading right singular vector $\boldsymbol{u}_1$ of $\boldsymbol{W}_t$ and the signal $\boldsymbol{\beta}$ in the teacher model defined by (5.2.3). For Case 3, a notable alignment appearing in Figure 5.4 after training suggests that $\boldsymbol{W}_t$ is capturing the feature $\boldsymbol{\beta}$ during training. Although this does not ensure NN will entirely beat the optimal kernel lower bound, this alignment reveals a non-negligible feature selection [BGL+21] via

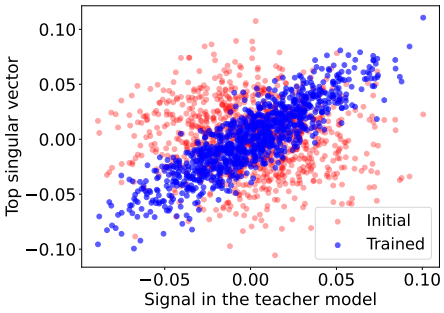


Figure 5.4. Alignment between teacher feature $\boldsymbol{\beta}$ and first PC $\boldsymbol{u}_1$ of the trained and initial weights, respectively, in Case 3 of Table 5.1.

large-stepsize training. This dynamical alignment along the task-relevant direction may further interpret the generalization of the NN, which has been proved in [BES+22] at the early stage of the training process. A similar phenomenon on the alignment between spike eigenvector of trained CK and data labels will be theoretically justified in Section 5.4.2.

Spikes of kernel matrices.

From [CSTEK01], the alignment of the kernel matrix with the training labels $\boldsymbol{y}$ is defined, called Kernel Target Alignment (KTA), as follows: when kernel $\boldsymbol{K}$ is either CK or NTK,

$$\mathrm{KTA} = \frac{\langle \boldsymbol{K}, \boldsymbol{y}^\top \boldsymbol{y} \rangle}{\|\boldsymbol{K}\|_F \|\boldsymbol{y}\|^2}. \tag{5.2.4}$$

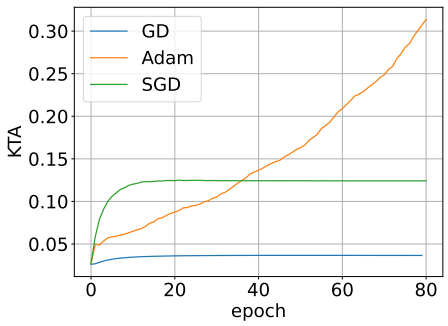
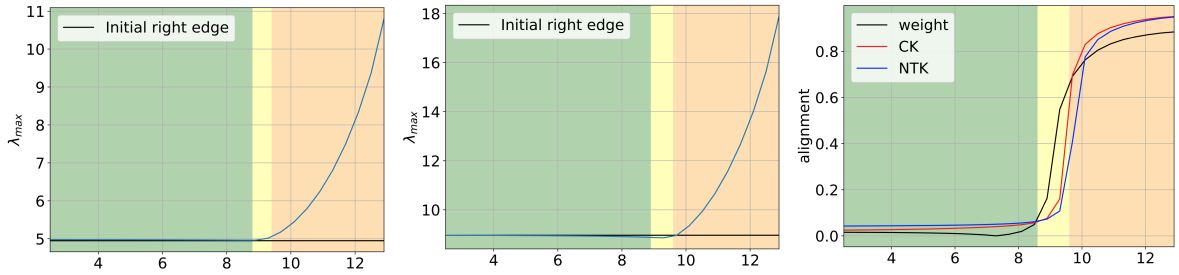


Figure 5.5. Evolution of KTA of CK defined by (5.2.4) with respect to training labels for Cases 1, 3&4 in Table 5.1. We normalize the epoch scales (x-axis) for better observations.

Analogously to [BGL+21, ABP22, SB21], Figure 5.5 depicts the evolution of KTA of CK in several cases. Based on Figure 5.6(c), when the spike appears outside the bulk (Case 3), its corresponding (leading) eigenvector $\boldsymbol{v}_1$ of kernel matrix naturally dominates the alignment with $\boldsymbol{y}$ (also see Figure 5.12 in Section 5.4.2), which is regarded as a kernel rotation during training in [OJMDF21]. Notice that this is not the common situation in Cases 1&2 of Table 5.1. On the other hand, KTA measures the alignment between $\boldsymbol{y}$ and the full eigenbasis of the kernel. These kernel alignments improve the speed of the convergence of training dynamics but may hurt or boost the generalization of the NNs [OJMDF21, SB21, BGL+21]. Moreover, Figure 5.5 indicates that Case 4 with heavy-tailed spectra after training with Adam [KB14] has a larger KTA than the other cases. In this case, the emergence of a heavy tail in the spectrum is closely related to a better generalization of the NN and more significant feature learning.

Transitions of the spike as a function of learning rate η.

From Case 2 to Case 3, we observe the emergence of outliers in the trained spectra when increasing the learning rate η. This indicates a transition of the emergence of the spike outside the bulk distribution. Figure 5.6, analogously to the well-known BBP transition by Baik, Ben Arous, and Péché in [BAP05] from the RMT community, shows there is a threshold (yellow region) for learning rate: the outliers only appear when η exceeds this threshold. We fix the same NN and dataset for all trials of training. In the green region, the largest eigenvalues are attached to the bulk (black horizontal lines) and the alignments are weak; in the orange one, outliers become apparent and the alignments become stronger. Here, the x-axis represents varying learning rates η. The flat black lines in Figures 5.6(a) and (b) are the right edges of the limiting spectra at initialization. Figure 5.6(c) records the angles between $\boldsymbol{\beta}$ and the leading eigenvector of $\boldsymbol{W}_t^\top \boldsymbol{W}_t/d$, and $\boldsymbol{y}$ and the leading eigenvectors of $\boldsymbol{K}_t^{\mathrm{CK}}$ and $\boldsymbol{K}_t^{\mathrm{NTK}}$ after training for different η. Similarly with [BGL+21], when η is sufficiently large (orange region), we obtain significant alignments which suggest potential feature learning. These transitions of leading eigenvalue and eigenvector alignment have been proved for $\boldsymbol{W}_t$ by [BES+22] for a different scenario. We apply NTK parameterization for our neural networks and train both layers until convergence, while [BES+22] considers the mean-field initialization and early stage of training dynamics of GD for the first layer.



(a) η v.s. $\lambda_{\max}(\boldsymbol{W}_t\boldsymbol{W}_t^\top/d)$. (b) η v.s. $\lambda_{\max}(\boldsymbol{K}_t^{\mathrm{CK}})$. (c) η v.s. Alignments.

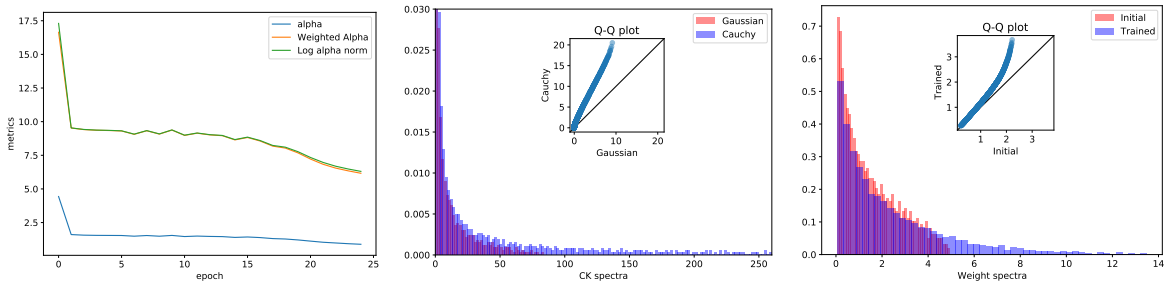
Figure 5.6. (a)-(c) Transitions of $\lambda_{\max}(\boldsymbol{W}_t\boldsymbol{W}_t^\top/d)$, $\lambda_{\max}(\boldsymbol{K}_t^{\mathrm{CK}})$ and alignments ($|\boldsymbol{\beta}^\top\boldsymbol{u}_1|/\|\boldsymbol{\beta}\|$ and $|\boldsymbol{y}^\top\boldsymbol{v}_1|/\|\boldsymbol{y}\|$ where $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the first singular vectors of $\boldsymbol{W}_t$ and either $\boldsymbol{K}_t^{\mathrm{CK}}$ or $\boldsymbol{K}_t^{\mathrm{NTK}}$, respectively) when increasing the learning rate η while training the NN with SGD until training loss is less than 10^{-5}.

Overall, based on our analysis in this section, the emergence of the outlier eigenvalue in Figures 5.1(b) shows the improvement over lazy training and potential feature learning via the training process, where the spectra possibly inherit the structures in teacher models (see Section 5.4.2). Comparing with Case 2, Case 3 of Table 5.1 suggests the importance of the large learning rate regime for training NNs [LWM19, Nak20, Lon21, BMR22, AVPVF23]. As a remark, our spectral results of Case 3 are consistent with the observations in [TSR22] through RMT hypothesis testing, where the majority of trained weight matrices remain random, and the learned feature may be contained in the outlier singular value and associated singular vector.

5.2.3 Phenomenon of Heavy-tailed Spectra

From Figures 5.1(c), Case 4 further exhibits heavy tails in the spectra of trained NNs, which thoroughly goes beyond the realm of the initial kernel machine. Notably, this phenomenon is not unique to Adam since heavy tails also occur with AdaGrad (see Appendix A of [WES+23]). Although all of these cases have the same initialization, different optimization methods eventually lead to various training trajectories and evolutions of the spectra of the weight and kernel matrices. To acquire feature learning, Cases 3&4 cause weights to deviate far from initialization. This section will provide more refined analyses of heavy tails regarding feature learning.

As mentioned in Section 5.1, [MM19, MM21] found a strong correlation between the heavy-tailed spectra of trained state-of-the-art models with better generalization. [MPM21] established several metrics, power α, weighted Alpha and Log α-norm, to measure how heavy the power-law tail is. Following the setting of Case 4 in Table 5.1, Figure 5.7(a) additionally presents the evolution of power α, weighted Alpha and Log α-norm during the training process. We can observe that in Figure 5.7(a) the tail in the spectrum of $\boldsymbol{W}_t$ is becoming heavier as the number of steps t is increasing, and the spectrum changes sharply at the early stage of training. Heavy-tailed spectra can be viewed as an extreme of "bulk+spikes", where a fraction of the eigenvalues move out of the initial bulk. In RMT, heavy-tailed spectra generally appear when the entries of the matrix are highly correlated [MM21]. This could heuristically explain heavy-tailed



(a) Evolution of metrics. (b) Heavy-tailed initial CK. (c) Bad example.

Figure 5.7. (a) The evolution of power α, weighted Alpha and Log α-norm during the training process in Case 4 of Table 5.1. (b) The CK spectra at two initializations for $\boldsymbol{W}$: standard Gaussian and Cauchy distributions. (c) Weight spectra at initial and after SGD training without good generalization error.

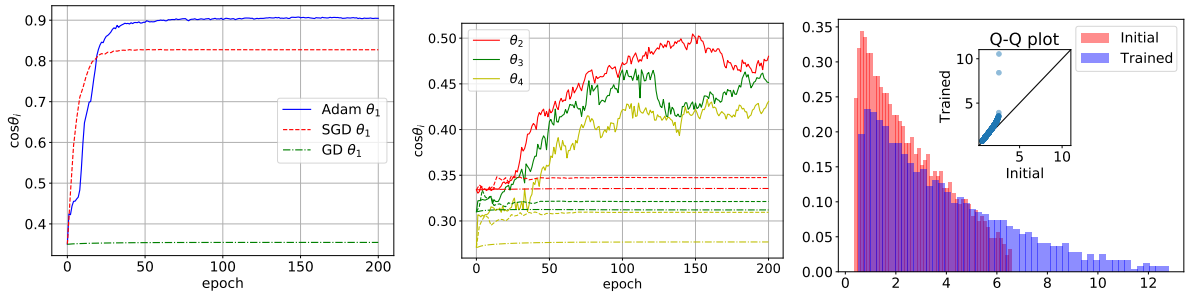
phenomena in the spectra since the entries of well-trained $\boldsymbol{W}_t$ should be strongly correlated.

We emphasize that heavy tails are not sufficient for good generalization, in general, [MPM21, MY23]. Figures 5.7(b) and (c) exhibit NNs with heavy-tailed weights but in the absence of good performance at initialization. Figure 5.7(b) exhibits the heavy-tail phenomenon even at random initialization with Cauchy distribution. In Figure 5.7(b), after training, the weight reveals a heavy tail but generalizes not as well as former examples, where the final test loss is 1.47504 and R^2 score is -0.48. In fact, it is the alignments between the features learned from the heavy-tailed part and the features in the teacher model that finally determine the generalization error of NNs. Unlike [MM19, MM21], we focus on the heavy-tailed phenomena for both weight and kernel matrices in a simpler model (5.2.1) and provide a connection between feature learning and heavy-tailed spectra, which opens an important avenue for further theoretical analysis.

Multiple-index examples for heavy-tailed spectra.

In Figure 5.8, we provide an example of when heavy tails indicate better generalizations. Consider the multiple-index teacher model (5.2.2) with $k = 5$ feature directions $\boldsymbol{\beta}_i$, and train NNs (5.2.1) with GD, SGD, and Adam to get invariant bulk, bulk with one spike and heavy tails, respectively, after training.

In this experiment, we consider $\sigma = \mathrm{ReLU}$, $n = 5000$, $h = 2500$ and $d = 1000$ for NN (1.1.3). Comparing with the teacher model (5.2.3) used in Table 5.1, we employ the multiple-

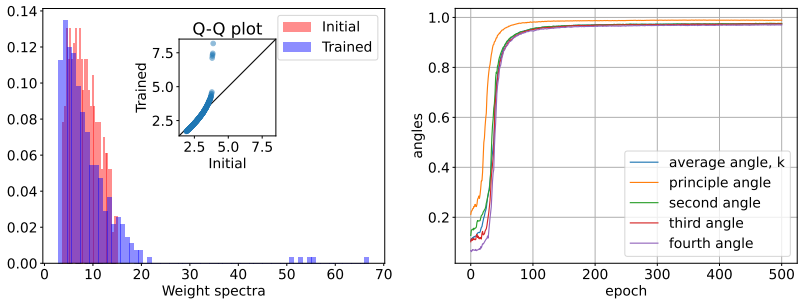


(a) Evolution of PC angle. (b) Alignment evolution. (c) Weight spectra.

Figure 5.8. (a) Evolutions of the first PC angle θ_1 between feature subspace U of the multiple-index model and the eigenspace spanned by top 100 of eigenvectors of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ during training with Adam, SGD, and GD. (b) Evolution of PC angles θ_i between feature subspace U and top 100 eigenspace of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$. (c) Initial and trained spectra for weight matrices when training with Adam (blue solid line in (a)).

index teacher model (5.2.2) with $k = 5$ and $\sigma^* = \sigma$. We trained this student-teacher model using GD ($\eta = 15$), SGD ($\eta = 7.25$ and batch size 8), and Adam ($\eta = 0.007$ and batch size 16) for training this NN, respectively. Similarly with Figure 5.1, correspondingly, we observe invariant spectrum, bulk with one spike, and heavy tails after training respectively. Heuristically, to learn this f^*, the weight $\boldsymbol{W}$ of NN should gradually align with the *feature subspace* $U = \text{span}\{\boldsymbol{\beta}_i\}_{i=1}^k$. Hence, to study feature learning, we can apply principle angles to measure the alignment between $\boldsymbol{W}$ and U. Consider the eigen-decomposition of $\boldsymbol{W}_t^\top \boldsymbol{W}_t = \sum_{i=1}^d \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Figure 5.8 shows the heavy-tailed part (the *eigenspace* $E := \text{span}\{\boldsymbol{v}_i\}_{i=1}^{100}$) is aligned with U after training, which shows how features are learned in the heavy-tailed spectra. Remarkably, the test errors for training processes with SGD and Adam are even smaller than $\|\mathrm{P}_{>1} f^*\|^2$ and $\|\mathrm{P}_{>2} f^*\|^2$, where $\mathrm{P}_{>1}$ denotes the orthogonal projection onto the nonlinear part of the function w.r.t. Gaussian measure. Thus, we experimentally showed that NNs with heavy-tailed spectra can obtain feature learning and generalize better than the other two cases.

In Figure 5.8(a), we present the evolutions of the top *principle angle* θ_1 between feature subspace $U = \text{span}\{\boldsymbol{\beta}_i\}_{i=1}^k$ and top 100 eigenspace of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ during different training processes with Adam (blue solid line), SGD (red dashed line) and GD (green dash-dot). The final test error is 0.33865 and the R^2 score is -0.71065 for GD. The test error is 0.10814 and the R^2 score is 0.45373 for SGD, where one spike emerges in the weight spectrum after training. The test error



(a) Initial v.s. Trained. (b) Evolution of PC angles.

Figure 5.9. (a) Initial and trained spectra for weight matrices when training with Adam. Five leading spikes emerge in this case. (b) Evolution of the angles between the first four PCs of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ and feature subspace U during training with Adam.

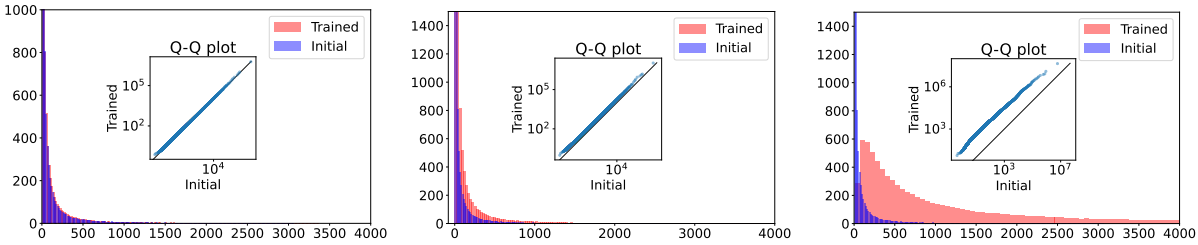
is 0.08672 and the R^2 score is 0.56195 for Adam. In Figure 5.8(a), we show the evolution of PC angles θ_i ($i = 2,3,4$) between feature subspace U of (5.2.2) and top 100 eigenspace of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ during training with Adam (solid line), SGD (dashed line) and GD (dash-dot). This eigenspace for the top 100 eigenvalues of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ corresponds to the heavy-tail part of the spectrum in $\boldsymbol{W}_t^\top \boldsymbol{W}_t$. Comparing with GD and SGD training processes, we observe strong alignments between feature space U and eigenspace w.r.t heavy tails in Adam case in Figures 5.8(a) and (b), which explains why Adam case (NNs with heavy-tailed spectra) generalizes better than the other two cases. This concludes that NNs with heavy-tailed spectra in Figure 5.8(c) can generalize better only when the teacher features from data are aligned with the heavy-tailed part of spectra. Suppose the feature dimension in the teacher model is high (i.e. the teacher model is more complicated and intrinsically high-dimensional). In that case, we expect a heavy-tailed weight spectrum of well-trained NN where the heavy-tailed part learns all the features in the teacher modes. This example explains why we can use the heavy tails to discriminate well-trained and poorly-trained large models [MPM21, MY23, YTH$^+$22].

Another example is exhibited in Figure 5.9. In this case, $k = 5$ in teacher model (5.2.2) and there are five leading outlier eigenvalues in the spectrum of the trained weight matrix with Adam optimizer, along with a heavy-tailed bulk in Figure 5.9(a). The test error is 0.01681 and R^2 score is 0.9154 for Adam. Figure 5.9(b) shows the evolution of the angles between the first

four PCs of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ and *feature subspace* $U = \text{span}\{\boldsymbol{\beta}_i\}_{i=1}^k$ of the multiple-index model (5.2.2) during training with Adam. Interestingly, Figure 5.9(b) justifies that the eigenspace of these five leading outlier eigenvalues in $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ is strongly aligned with features $\boldsymbol{\beta}_i$ for $1 \le i \le 5$. This indicates that heavy-tailed spectra with large spikes may have a correlation with feature learning and good generalizations.

5.3 Further Discussions on Real-world Dataset

Above, we empirically investigated how the spectra of $\boldsymbol{W}$, $\boldsymbol{K}^{\text{CK}}$, and $\boldsymbol{K}^{\text{NTK}}$ evolve under the LWR for an idealized student-teacher setting. Our work implies that understanding the relationship between feature learning and training processes requires understanding the evolution of the spectra of both weight and kernel matrices. In particular, we show that different training processes affect the eigenstructure of weight and kernel matrices. While synthetic data is easier to analyze theoretically, we also investigate these spectral properties on real-world data and more complicated tasks below. In practice, people mainly focus on analyzing spectra of the weight matrices in fully connected layers; we study the spectral properties of kernel matrices induced by the NNs, which contain abundant information [CHS20, Lon21, ABP22, SB21].

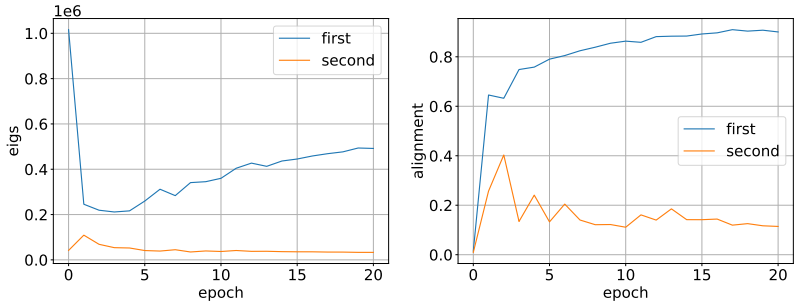


(a) GD. (b) SGD. (c) Adam.

Figure 5.10. Different NTK spectra for a small-CNN model on CIFAR-2. The subplots are Q-Q plots for the comparison between initial and trained spectra. Test accuracies: (a) 79%, (b) 84%, (c) 86.4%.

First, we show the spectra of $\boldsymbol{K}^{\text{NTK}}$ before and after different training processes for binary classification on CIFAR-2 through small CNNs in Figure 5.10. Similarly with Case 1, Figure 5.10(a) (especially in the Q-Q subplot) manifests the invariant spectral distribution of

NTK through GD training while SGD exhibits a heavier tail in NTK spectrum in Figure 5.10(b). This phenomenon is more evident when trained by Adam in Figure 5.10(c) with improved accuracy. Figure 5.10 suggests that our observations on synthetic data in Section 5.2 can be extended to real-world data and on more practical architectures. We note that there is a lack of the emergence of spikes after training because spikes already exist in the initial NTK spectrum for this complicated neural architecture on real-world datasets. Figure 5.10(a) also indicates that the spectral invariance of NTK through training will impede the feature learning and the NN does not generalize well in this training process.



(a) Eigenvalues v.s. epochs. (b) Alignments v.s. epochs.

Figure 5.11. We use SGD for fine-tuning the BERT model. (a) The evolution of first and second eigenvalues of empirical CK during fine-tuning. (b) The alignments of training labels with first and second eigenvectors of CK during fine-tuning.

We also investigate the spectral properties on the pre-trained model, BERT [DCLT18], with fine-tuning on Sentiment140 dataset of tweets[1] from [GBH09]. We fine-tune the BERT model for a binary classifier on Sentiment140 and capture the evolution of CK spectra, rather than the NTK due to the size of BERT, in Figure 5.11. The training accuracy is 95.90% and the test accuracy is 84%. A heavy-tailed CK spectrum with several spikes already exists in this pre-trained model. Unlike Figure 5.10 (and cases in Table 5.1) where the first spike of NTK becomes larger than at random initialization after training, in Figure 5.11(a), the leading eigenvalue first decreases and then increases. Moreover, similarly to Figure 5.6, our Figure 5.11(b) shows that the alignment of the first eigenvector of the CK and training labels becomes more apparent through

[1] https://www.kaggle.com/datasets/kazanova/sentiment140

fine-tuning with the leading eigenvalue decrease. Heuristically, this process seems to unlearn the features in the pre-trained model and, remarkably, learn new features on the new dataset in only a few epochs of fine-tuning. We believe that the evolutions of the kernel matrices and some spectral metrics are crucial for understanding feature learning through fine-tuning [WHS22]. A more comprehensive exploration of the evolutionary spectral properties of "foundation models" may help shed further light on these phenomena.

5.4 Theoretical Study of the Spectra in Trained NNs

Inspired by our empirical simulations in Section 5.2, we now theoretically analyze the trained weight and kernel matrices in two simple cases:

Invariant spectra through training processes.

We justify the invariant spectra after training with full-batch GD with small learning rates observed in Section 5.2.1. Following the global convergence of GD [OFLS19, OS19] and NTK theory [JGH18], we can prove that the spectra of NNs trained with full batch gradient descent (GD) are globally *invariant*, indicating that the NN is still close to a kernel machine.

Spiked weight and kernel matrices in early training.

It is known that NNs can learn useful representations that adapt to the learning problem, and outperform the random features model defined by randomly initialized weights [GMMM19, WLLM19, AAM22]. Recent works have shown that when the target function is low-dimensional, the gradient update with a large learning rate for two-layer NNs around initialization is *low-rank*, e.g., [BES$^+$22, DLS22, WES$^+$23], and hence the updated weight matrix $\boldsymbol{W}$ is well-approximated by a spiked random matrix model. Following the spiked sample covariance model we characterized in Chapter 3, we can show that finite steps of gradient descent updates produce a spiked structure in the pre-trained kernel model of NNs and this spike structure implies a useful representation learning in the dataset.

5.4.1 Invariant Bulk Distributions

Figures 5.1(a), 5.2 and 5.3 have already present that the bulk distributions of weight and kernel matrices in Cases 1&2 remain globally unchanged (invariant) over the training process. Now, by investigating the global convergence of GD, we prove this invariant-bulk phenomenon under certain assumptions.

Assumption 14 (Linear-Width Regime (LWR))**.** Assume that $\frac{n}{d} \to \gamma_1$ and $\frac{h}{d} \to \gamma_2$ as $n \to \infty$ where the aspect ratios $\gamma_1, \gamma_2 \in (0, \infty)$ are two fixed constants.

LWR stands as a pivotal setting grounded in high-dimensional statistics [AP20, MM22]. It offers valuable insights especially when addressing real-world datasets. This is in contrast to the infinite-width regime, in which we are already in the asymptotic limit for width at first. Hence, LWR is a better approximation of real-world datasets and practical neural networks compared with the infinite-width regime.

Assumption 15 (Activation function)**.** Suppose that the activation function $\sigma(x)$ is nonlinear and λ_σ-Lipschitz with $|\sigma'(x)|, |\sigma''(x)| \leq \lambda_\sigma$ for all $x \in \mathbb{R}$. Moreover, $\mathbb{E}[\sigma(z)] = 0$ for $z \sim \mathcal{N}(0,1)$.

For simplicity, we focus on analyzing the training process of the first-hidden layer with the second layer $\boldsymbol{v}$ fixed. Denote $f_{\boldsymbol{\theta}}(\boldsymbol{X})$ by $f_{\boldsymbol{W}}(\boldsymbol{X})$ in this case. At any time $t \in \mathbb{N}$, consider the gradient steps:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}} \mathscr{L}(\boldsymbol{W}_t). \tag{5.4.1}$$

Denote the CK and NTK at gradient step $t \in \mathbb{N}$ by $\boldsymbol{K}_t^{\mathrm{CK}} := \frac{1}{h}\sigma(\boldsymbol{W}_t\boldsymbol{x})^\top \sigma(\boldsymbol{W}_t\boldsymbol{x})$, and $\boldsymbol{K}_t^{\mathrm{NTK}} := \frac{1}{d}\boldsymbol{x}^\top\boldsymbol{x} \odot \frac{1}{h}\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{x}\right)^\top \mathrm{diag}(\boldsymbol{v}_t)^2 \sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{x}\right)$ respectively. First, we present an elaborate description of the changes in the weight, CK, and NTK at the *early phase* of the training (after any finite t steps) as follows.

Lemma 102 (Early phase)**.** *Under Assumptions 13, 14, and 15, we further assume that* $\|\boldsymbol{v}\|_\infty \leq 1$ *and* f^* *is a* λ_σ*-Lipschitz function. Given any fixed* $t \in \mathbb{N}$ *and learning rate* $\eta = \Theta(1)$*, after* t

gradient steps, the changes $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_F$, $\left\|\boldsymbol{K}_t^{CK}-\boldsymbol{K}_0^{CK}\right\|_F$, *and* $\left\|\boldsymbol{K}_t^{NTK}-\boldsymbol{K}_0^{NTK}\right\|$ *are all less than* $\frac{C}{n}$, *with probability at least* $1-4n\exp(-cn)$, *for some positive constants* $c,C>0$ *which only depend on step* t *and parameters* $\eta,\gamma_1,\gamma_2,\lambda_\sigma,\sigma_\varepsilon$.

Lemma 102 shows $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|$, $\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{\mathrm{CK}}\right\|$, and $\left\|\boldsymbol{K}_t^{\mathrm{NTK}}-\boldsymbol{K}_0^{\mathrm{NTK}}\right\|$ are asymptotically vanishing for any fixed time t. Therefore, all the eigenvalues/eigenvectors are asymptotically unchanged at the early phase of the training (see Corollary 108 in Section 5.5). Now we aim to analyze the spectra at the end of the training process (5.4.1). In this case, although we are unable to show the invariance for each eigenvalue, we can verify the invariance of the limiting bulk distributions for $\boldsymbol{K}_t^{\mathrm{CK}}$ and $\boldsymbol{K}_t^{\mathrm{CK}}$ for all t.

By Theorem 69, the smallest eigenvalue of $\boldsymbol{K}_0^{\mathrm{NTK}}$ has an asymptotic lower bound:

$$\lambda_{\min}(\boldsymbol{K}_0^{\mathrm{NTK}})\geq\left(a_\sigma-\sum_{k=0}^{2}\eta_k^2\right)(1-o_{d,\mathbb{P}}(1)),\tag{5.4.2}$$

where $a_\sigma:=\mathbb{E}[\sigma'(\xi)^2]$ and η_k is the k-th Hermite coefficient of σ'. Hence, we can claim there exists some constant $\alpha>0$ only dependent on σ such that $\lambda_{\min}(\boldsymbol{K}_0^{\mathrm{NTK}})\geq 4\alpha^2$ with high probability. Note that α is not vanishing since σ is nonlinear. With this lower bound, we obtain the following global convergence for (5.4.1) and norm control of $\boldsymbol{W}_t$ as $n/d\to\gamma_1$ and $h/d\to\gamma_2$.

Theorem 103 (Global convergence)**.** *Under the same assumptions of Lemma 102, we further assume* v_i*'s are independent and centered random variables in the second layer. For any* $\eta<\min\{\frac{\alpha^2 n}{2},\frac{n}{4\lambda_\sigma^2(1+\sqrt{\gamma_1})^2}\}$ *and all* $t\in\mathbb{N}$, *there exists some* $\gamma^*>0$ *such that, when* $\gamma_2\geq\gamma^*$, *the gradient steps* (5.4.1) *will satisfy*

$$\ell(\boldsymbol{W}_t)\leq\left(1-\frac{\eta\alpha^2}{2n}\right)^t\ell(\boldsymbol{W}_0),\tag{5.4.3}$$

$$\frac{1}{4}\alpha\|\boldsymbol{W}_0-\boldsymbol{W}_t\|_F+\ell(\boldsymbol{W}_t)\leq\ell(\boldsymbol{W}_0),\tag{5.4.4}$$

$$\sum_{t=0}^{\infty}\|\boldsymbol{W}_{t+1}-\boldsymbol{W}_t\|_F\leq\frac{4\ell(\boldsymbol{W}_0)}{\alpha},\tag{5.4.5}$$

with high probability, as $n/d \to \gamma_1$ *and* $h/d \to \gamma_2$. *Here, training loss* $\ell(\boldsymbol{W}) := \|\boldsymbol{y} - f_{\boldsymbol{W}}(\boldsymbol{x})\|$.

We apply the techniques and results by [OFLS19, OS19] to obtain Theorem 103. Notice that, unlike Lemma 102, the largest learning rate we can choose is of order $\Theta(n)$. As a byproduct, the Frobenius norm in (5.4.4) implies the following corollary for the invariance of limiting *bulk* distribution.

Corollary 104. *Under the same assumptions of Theorem 103, for all* $t \in \mathbb{N}$*, with high probability, there exists some constant* $R > 0$ *such that the changes* $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, $\left\|\boldsymbol{K}_t^{CK} - \boldsymbol{K}_0^{CK}\right\|_F$, *and* $\left\|\boldsymbol{K}_t^{NTK} - \boldsymbol{K}_0^{NTK}\right\|_F$ *are all less than R with high probability. This implies the limiting empirical spectra of* $\frac{1}{h}\boldsymbol{W}_t^\top \boldsymbol{W}_t$, $\boldsymbol{K}_t^{CK}$ *and* $\boldsymbol{K}_t^{NTK}$ *are the same as the limiting spectra of* $\frac{1}{h}\boldsymbol{W}_0^\top \boldsymbol{W}_0$, $\boldsymbol{K}_0^{CK}$ *and* $\boldsymbol{K}_0^{NTK}$ *respectively, almost surely as* $n/d \to \gamma_1$ *and* $h/d \to \gamma_2$.

Corollary 104 is empirically validated by Figure 5.14 in Section 5.5. In addition, based on Figure 5.3, one can further extend Corollary 104 to the SGD training process. The total path is $O(\sqrt{h})$ in (5.4.4) and (5.4.5), which is negligible compared with the Frobenius norm of initial weight matrix (which is of order $\Theta(h)$). Thus, gradient descent iterates (5.4.1) remain close to initialization and small perturbation of NTK ensures the smallest eigenvalue (5.4.2) of NTK is always lower bounded away from zero. Theorem 103, however, does not require that the NTK stays unchanged all the time. Moreover, Corollary 104 only shows the invariance of the bulk distribution, while the emergence of outliers cannot be excluded from this result. Though we have global convergence in general, we may still move out of the kernel regime. Global convergence does not indicate when the NN in LWR outperforms the kernel regime. Notice that [BMR21, Theorem 5.4] is not directly applicable to show that a network is still close to lazy training under the LWR. It requires deeper analysis to claim whether the NN still belongs to the kernel regime or already goes beyond in our case. As we will show in Section 5.2.2, this also relies on the magnitude of the learning rate for GD/SGD.

5.4.2 Feature Learning in CK Matrix After Finitely Many Steps of GD

The preceding section studied the spike eigenstructure of the CK induced by low-rank structure in the input data. Here, focusing on a two-layer model, we study an alternative setting where spiked structure arises instead in the weight matrix $\boldsymbol{W}$ from gradient descent training.

We consider an early training regime studied in [BES+22], with a width-N two-layer feedforward NN,

$$f_{\mathrm{NN}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle) = \frac{1}{\sqrt{N}} \sigma(\boldsymbol{x}^\top \boldsymbol{W}) \boldsymbol{a}. \tag{5.4.6}$$

Here $\boldsymbol{x} \in \mathbb{R}^d$ is the input, and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N] \in \mathbb{R}^{d \times N}$ and $\boldsymbol{a} \in \mathbb{R}^N$ are the network weights. For clarity of the subsequent discussion, we will transpose the notation for $\boldsymbol{X}$ and $\boldsymbol{W}$ from the preceding section, and incorporate a $1/\sqrt{d}$ scaling into $\boldsymbol{W}$ rather than into the input data $\boldsymbol{X}$.

Given are an input feature matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ and labels $\boldsymbol{y} \in \mathbb{R}^n$ for n samples, where $y_i = f_*(\boldsymbol{x}_i) + \text{noise}$. We consider the training of first-layer weights $\boldsymbol{W}$ to minimize the mean squared error

$$\mathscr{L}(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} (f_{\mathrm{NN}}(\boldsymbol{x}_i) - y_i)^2,$$

fixing the second-layer weight vector $\boldsymbol{a}$. From a random initialization $\boldsymbol{W}_0 \in \mathbb{R}^{d \times N}$, and over T steps with learning rates $\eta_1, \ldots, \eta_T$ scaled by $\sqrt{N}$, the gradient descent (GD) updates take the form

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \eta_{t+1} \sqrt{N} \cdot \boldsymbol{G}_t, \qquad \boldsymbol{G}_t = -\nabla \mathscr{L}(\boldsymbol{W}_t). \tag{5.4.7}$$

Of interest is the information about the label function f_* that is learned by $\boldsymbol{W}_{\text{trained}} \equiv \boldsymbol{W}_T$, which may be characterized by the spectral alignment of the CK matrix with the class label vector on independent test data $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$. This use of independent test data may be understood as a pre-training setup, also considered previously in [BES+22, MLHD23] and studied for real-world data in [WHS22].

It was shown in [BES+22] that in a training regime with initialization $\|\boldsymbol{W}_0\| \asymp 1$ such that $|f_{\mathrm{NN}}(\boldsymbol{x}_i)| \ll 1$ for each $i = 1,\dots,N$, and with learning rates $\eta_1,\dots,\eta_T \asymp 1$ for a fixed number T of GD steps, the weight matrix $\boldsymbol{W}$ undergoes a change during training that is $O(1)$ in operator norm and approximately rank-1,

$$\boldsymbol{W}_{\text{trained}} \approx \boldsymbol{W}_0 + \frac{\eta b_\sigma}{n} \boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{a}^\top \qquad \text{where} \qquad \eta = \sum_{t=1}^{T} \eta_t.$$

[BES+22, Conjecture 4] conjectured that for the CK matrix

$$\boldsymbol{K} = \frac{1}{N} \sigma(\tilde{\boldsymbol{X}}\boldsymbol{W}_{\text{trained}}) \sigma(\tilde{\boldsymbol{X}}\boldsymbol{W}_{\text{trained}})^\top \in \mathbb{R}^{n\times n} \tag{5.4.8}$$

defined by the pre-trained weights and test data $\tilde{\boldsymbol{X}}$, the resulting spike eigenvalue and the alignment of its spike eigenvector with the test labels $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$ are accurately predicted by a Gaussian equivalent model. Our main result of this section is an affirmative verification of this conjecture and precise characterization of the spike eigenstructure of $\boldsymbol{K}$, in the following representative setting.

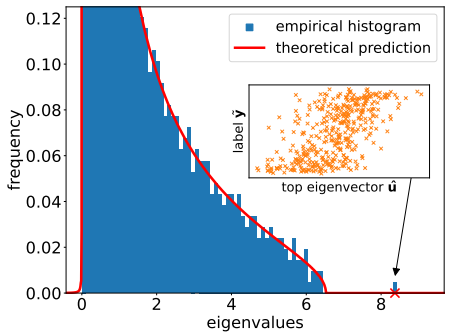
Assumption 16. For a two-layer NN in (5.4.6) with GD training defined by (5.4.7), we assume that

(a) (LWR) $n, d, N \to \infty$ such that $N/d \to \gamma_0 \in (0,\infty)$ and $N/n \to \gamma_1 \in (0,\infty)$.

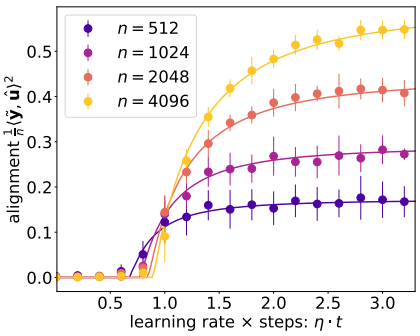
(b) Training features $\boldsymbol{X} = [\boldsymbol{x}_1,\dots,\boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ have entries $[\boldsymbol{X}]_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, training labels $\boldsymbol{y} \in \mathbb{R}^n$ have entries $y_i = \sigma_*(\boldsymbol{\beta}_*^\top \boldsymbol{x}_i) + \varepsilon_i$ where $\boldsymbol{\beta}_* \in \mathbb{R}^d$ is a deterministic unit vector and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2)$, and test data $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$ is an independent copy of $(\boldsymbol{X}, \boldsymbol{y})$.

(c) The NN activation $\sigma : \mathbb{R} \to \mathbb{R}$ and label function $\sigma_* : \mathbb{R} \to \mathbb{R}$ both satisfy Assumption 5, with $b_\sigma := \mathbb{E}[\sigma'(\xi)] \neq 0$ and $b_{\sigma_*} := \mathbb{E}[\sigma_*'(\xi)] \neq 0$.

(d) The weight initializations satisfy $[\boldsymbol{W}_0]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1/d)$ and $a_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1/N)$.



(a) Spectrum of the updated CK.



(b) Eigenvector alignment of the updated CK.

Figure 5.12. (a) We set $n = 2000, d = 1600, N = 2400, \eta \cdot t = 2$, and $\sigma = \sigma_* = \mathrm{erf}$. (b) We set $d = 2048, N = 1024, \eta = 0.2$, $\sigma = \tanh, \sigma_* = \mathrm{SoftPlus}$, and vary the sample size n and number of GD steps t; dots represent empirical simulations (over 10 runs) and solid curves are theoretical predictions from Theorem 105.

(e) The number of iterations T and learning rates $\eta_1, \ldots, \eta_T$ are fixed independently of n, d, N.

Notice that the initialization in Assumption 16(d) is different from the initialization for (5.2.1) in Section 5.4.1. Under these assumptions, the following theorem characterizes the spike eigenvalue of the CK matrix and the alignment between the corresponding eigenvector and the test labels, as a function of the learning rate η_t and the number of gradient descent steps T.

Theorem 105. *Suppose that Assumption 16 holds, and set* $\eta = \sum_{t=1}^{T} \eta_t$. *Define*

$$\theta_1 = b_\sigma \eta \cdot \sqrt{(\gamma_1/\gamma_0)(1+\sigma_\varepsilon^2) + b_{\sigma_*}^2}, \qquad \theta_2 = b_\sigma b_{\sigma_*} \eta. \tag{5.4.9}$$

Let $z(\cdot)$ *and* $\varphi(\cdot)$ *be defined by* (3.2.6) *for* $\ell = 1$ *with* γ_1 *and* $\nu_0 = b_\sigma^2 \otimes \rho_{\gamma_0}^{\mathrm{MP}} \oplus (1 - b_\sigma^2)$, *and set*

$$\lambda_1 = b_\sigma^2 \frac{(1+\theta_1^2)(\gamma_0+\theta_1^2)}{\theta_1^2} + 1 - b_\sigma^2.$$

Then $\boldsymbol{K}$ *defined by* (5.4.8) *has a spike eigenvalue if and only if* $\theta_1 > \gamma_0^{1/4}$ *and* $z'(-1/\lambda_1) > 0$. *In this case,* $\lambda_{\max}(\boldsymbol{K}) \to \gamma_1 z(-1/\lambda_1)$ *a.s., and the leading unit eigenvector* $\widehat{\boldsymbol{u}} \in \mathbb{R}^n$ *of* $\boldsymbol{K}$ *satisfies*

$$\frac{1}{\sqrt{n}} |\tilde{\mathbf{y}}^\top \widehat{\boldsymbol{u}}| \to b_\sigma b_{\sigma_*} \frac{\sqrt{z(-1/\lambda_1)\varphi(-1/\lambda_1)}}{\lambda_1} \cdot \frac{\theta_2 \sqrt{(\theta_1^4 - \gamma_0)(\gamma_0 + \theta_1^2)}}{\theta_1^3} > 0 \ \text{a.s.} \tag{5.4.10}$$

Numerical illustration.

Figure 5.12 empirically validates the predictions of Theorem 105, for a two-layer NN trained with a small number of GD steps. Figure 5.12(a) shows that one spike eigenvalue emerges over training in the test-data CK, the location of which is accurately predicted by Theorem 105; moreover, the leading eigenvector $\widehat{\boldsymbol{u}}$ aligns with the labels $\tilde{\boldsymbol{y}}$. This is quantified in Figure 5.12(b), where above a phase transition threshold, the alignment $\langle \widehat{\boldsymbol{u}}, \tilde{\boldsymbol{y}} \rangle^2$ (predicted by (5.4.10)) increases with the learning rate or number of GD steps; in addition, alignment also increases with the training set size n. Then, compared with random initialization ($\eta = 0$), this illustrates that training improves the NN representation, and the test-data CK contains information on the label function f_*.

5.5 Proofs of Results in Section 5.4.1

Recall the definition of the entry-wise 2-∞ matrix norm $\|\boldsymbol{M}\|_{2,\infty}$ in Section 1.4. For any matrix $\boldsymbol{M} \in \mathbb{R}^{N \times d}$, notice that

$$\|\boldsymbol{M}\|_{2,\infty} \le \|\boldsymbol{M}\| \le \|\boldsymbol{M}\|_F. \tag{5.5.1}$$

5.5.1 GD Analysis at Early Phase

From (5.4.1), the GD process with learning rate $\eta > 0$ can be written by

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \eta \cdot \boldsymbol{G}_t, \text{ where} \tag{5.5.2}$$

$$\boldsymbol{G}_t = \frac{1}{n\sqrt{dh}} \left[\left(\boldsymbol{v} \left(\boldsymbol{y} - \frac{1}{\sqrt{h}} \boldsymbol{v}^\top \sigma(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d}) \right) \right) \odot \sigma'(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d}) \right] \boldsymbol{X}^\top, \tag{5.5.3}$$

for $t \in \mathbb{N}$, where $\boldsymbol{y} \in \mathbb{R}^{1 \times n}$. Following [BES+22, Appendix B], in this section, we prove the control for gradient step $\boldsymbol{G}_t$. For simplicity, denote $f_t(\boldsymbol{X}) := f_{\boldsymbol{\theta}_t}(\boldsymbol{X}) = \frac{1}{\sqrt{h}} \boldsymbol{v}^\top \sigma(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d})$ for $t \in \mathbb{N}$.

Lemma 106. *Under the same assumptions as in Lemma 102, we have*

$$\mathbb{P}\Big(\Big\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\Big\| \geq C\sqrt{n}\Big) \leq 2e^{-cn},$$
$$\mathbb{P}\big(\|\boldsymbol{y}\| \geq C\sqrt{n}\big) \leq 2e^{-cn},$$

for some constants $C, c > 0$ only depending on σ_ε, λ_σ, γ_1, and γ_2.

Proof. Due to Lemma 17, we can directly obtain that

$$\mathbb{P}\left(\Big\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\Big\| \geq C'(\sqrt{n}+\sqrt{h})\sqrt{\frac{h}{d}}\right) \leq 2e^{-cn}.$$

Here we use the fact that both $\boldsymbol{W}_0$ and $\boldsymbol{X}$ are i.i.d. Gaussian random matrices. Then by Assumption 14, we conclude that we control $\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})$. Recall that Assumption 13 implies that $\boldsymbol{y} = f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon}$. Hence, by Lipschitz Gaussian concentration inequality [Ver18, Theorem 5.2.2], each entry of $f^*(\boldsymbol{X})$ has independent sub-Gaussian coordinates, whence we can get $\|f^*(\boldsymbol{X})\| \leq C\sqrt{n}$ with probability at least $1-2ne^{-cn}$ for some constants $c, C > 0$. On the other hand, $[\boldsymbol{\varepsilon}]_i = \varepsilon_i$ are i.i.d. centered sub-Gaussian noises with variance σ_ε^2. By [Ver18, Theorem 3.1.1], we have

$$\mathbb{P}\big(\|\boldsymbol{\varepsilon}\| \leq 2\sigma_\varepsilon\sqrt{n}\big) \geq 1 - 2\exp\Big(-\frac{cn}{K^4}\Big),$$

where the constant K is the sub-Gaussian norm defined by $K = \max_{i\in[n]} \|\varepsilon_i\|_{\psi_2}$. Hence, combining all things together, we obtain the second inequality of this lemma. □

Lemma 107. *Under the assumptions of Lemma 102, given any fixed $t \in \mathbb{N}$ and learning rate $\eta = \Theta(1)$, the weight matrix after t gradient steps $\boldsymbol{W}_t$ defined in* (5.5.2) *satisfies*

$$\mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \geq \frac{C}{\sqrt{n}}\right) \leq \exp(-cn), \tag{5.5.4}$$

for some positive constants $c, C > 0$ only depending on t, η, σ_ε, λ_σ, γ_1 and γ_2.

Proof. Denote $\sigma_\perp(x) = \sigma(x) - \mu_1 x$ which is the nonlinear part of σ and $\mu_1 = \mathbb{E}[z\sigma(z)]$. Thus, $\mathbb{E}[\sigma_\perp(z)z] = 0$ for $z \sim \mathcal{N}(0,1)$. Based on this, we can further decompose the gradient $\boldsymbol{G}_t$ into

$$\boldsymbol{G}_t = \underbrace{\frac{\mu_1}{n\sqrt{dh}}\boldsymbol{v}(\boldsymbol{y} - f_t(\boldsymbol{X}))\boldsymbol{X}^\top}_{\boldsymbol{A}^t} + \underbrace{\frac{1}{n\sqrt{dh}}\Big(\boldsymbol{v}(\boldsymbol{y} - f_t(\boldsymbol{X})) \odot \sigma_\perp'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\Big)\boldsymbol{X}^\top}_{\boldsymbol{B}^t}. \tag{5.5.5}$$

At first, consider $t = 0$ in (5.5.2) and bound the spectral norm of $\boldsymbol{W}_1$. By assumption, we know $\|\boldsymbol{v}\| \le \sqrt{h}$. Due to Corollary 7.3.3 in [Ver18], we have

$$\mathbb{P}\bigg(\frac{1}{\sqrt{d}}\|\boldsymbol{X}\| \ge 2\bigg(1 + \sqrt{\frac{n}{d}}\bigg)\bigg) \le 2\exp(-cn). \tag{5.5.6}$$

Therefore, by (5.5.5), we can control $\boldsymbol{A}^0$ and $\boldsymbol{B}^0$ separately. Notice that, as a rank-one matrix,

$$\begin{aligned}\Big\|\boldsymbol{A}^0\Big\| = \Big\|\boldsymbol{A}^0\Big\|_F &\le \frac{\mu_1}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{1}{\sqrt{n}}(\|\boldsymbol{y}\| + \|f_0(\boldsymbol{X})\|)\frac{\|\boldsymbol{v}\|}{\sqrt{h}} \\ &\le \frac{\mu_1}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{\|\boldsymbol{v}\|}{\sqrt{h}}\frac{1}{\sqrt{n}}\bigg(\|\boldsymbol{y}\| + \frac{\|\boldsymbol{v}\|}{\sqrt{h}}\Big\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\Big\|\bigg).\end{aligned}$$

Hence, by Lemma 106 and (5.5.6), one can easily claim that $\|\boldsymbol{A}^0\| \le C/\sqrt{n}$ with probability at least $1 - e^{-cn}$ for some constants $c, C > 0$. On the other hand, since $\boldsymbol{v}(\boldsymbol{y} - f_t(\boldsymbol{X}))$ is rank-one and $\sigma_\perp' = \sigma' - \mu_1$ with $|\sigma'(x)| \le \lambda_\sigma$, we can similarly obtain

$$\begin{aligned}\big\|\boldsymbol{B}^0\big\|_F &\le \frac{1}{n\sqrt{dh}}\Big\|\boldsymbol{v}(\boldsymbol{y} - f_t(\boldsymbol{X})) \odot \sigma_\perp'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\Big\|_F\|\boldsymbol{X}\| \\ &\le \frac{1}{n\sqrt{hd}}\|\boldsymbol{X}\|(\|\boldsymbol{y}\| + \|f_0(\boldsymbol{X})\|)\|\boldsymbol{v}\|\max_{i,j}\Big|\sigma_\perp'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\Big|_{i,j} \\ &\le \frac{\mu_1 + \lambda_\sigma}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{\|\boldsymbol{v}\|}{\sqrt{h}}\frac{1}{\sqrt{n}}\bigg(\|\boldsymbol{y}\| + \frac{\|\boldsymbol{v}\|}{\sqrt{h}}\Big\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\Big\|\bigg).\end{aligned}$$

As $\boldsymbol{A}^0$, we can apply Lemma 106 and (5.5.6) again to conclude (5.5.4) for $t = 1$.

For general t, we apply induction. We assume that after the t-th gradient step with $\eta = \Theta(1)$, Eq. (5.5.4) holds for some constants $C, c > 0$. Following [BES+22, Lemma 16], we

now show that the similar high-probability statement also holds for $\boldsymbol{W}_{t+1}$ (for some different constants c', C'). Firstly, following the same argument as [OS20, Setion 6.6.1], we know that

$$\begin{aligned}\|f_t(\boldsymbol{X})\| \leq & \|f_0(\boldsymbol{X})\| + \|f_t(\boldsymbol{X}) - f_0(\boldsymbol{X})\| \\ \leq & \|f_0(\boldsymbol{X})\| + \frac{\lambda_\sigma}{\sqrt{h}}\|\boldsymbol{v}\|\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F. \end{aligned} \tag{5.5.7}$$

Note that $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F = O(1/\sqrt{n})$ with high probability by the induction hypothesis. Hence, by Lemma 106 and (5.5.6), we have $\|f_t(\boldsymbol{X})\| \leq C\sqrt{n}$ with high probability. Indeed, the difference between $f_t(\boldsymbol{X})$ and $f_0(\boldsymbol{X})$ is significantly negligible comparing with the initial value $f_0(\boldsymbol{X})$. Similarly with $\boldsymbol{A}_0$, $\boldsymbol{A}^t$ satisfies

$$\left\|\boldsymbol{A}^t\right\| = \left\|\boldsymbol{A}^t\right\|_F \leq \frac{\mu_1}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{1}{\sqrt{n}}(\|\boldsymbol{y}\| + \|f_t(\boldsymbol{X})\|)\frac{\|\boldsymbol{v}\|}{\sqrt{h}}.$$

Analogously for $\boldsymbol{B}^t$, we have

$$\left\|\boldsymbol{B}^t\right\|_F \leq \frac{\mu_1 + \lambda_\sigma}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{\|\boldsymbol{v}\|}{\sqrt{h}}\frac{1}{\sqrt{n}}(\|\boldsymbol{y}\| + \|f_t(\boldsymbol{X})\|).$$

Thus, Lemma 106, (5.5.6), and (5.5.7) ensure that

$$\mathbb{P}\left(\left\|\boldsymbol{A}^t\right\|_F \geq \frac{C'}{\sqrt{n}}\right) \leq \exp\left(-c'n\right),\ \mathbb{P}\left(\left\|\boldsymbol{B}^t\right\|_F \geq \frac{C'}{\sqrt{n}}\right) \leq \exp\left(-c'n\right),$$

for constants $c', C' > 0$. Since $\|\boldsymbol{W}_{t+1} - \boldsymbol{W}_0\|_F \leq \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F + \eta\left\|\boldsymbol{A}^t\right\|_F + \eta\|\boldsymbol{B}^t\|_F$, by induction hypothesis, we can conclude that (5.5.4) holds for the $(t+1)$-th step with some constants $C, c > 0$, which are different from the constants at the t-th step. □

As a corollary, by (5.5.1), we can also deduce the following norm bounds:

$$\mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\| \geq \frac{C}{\sqrt{n}}\right) \leq \exp(-cn), \quad \mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty} \geq \frac{C}{\sqrt{n}}\right) \leq \exp(-cn).$$

Lemma 107 and the above bounds are empirically verified by Figure 5.13(a) for $t = 3$. Not only upper bounds, this simulation also shows that at early phase $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|$, $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, and $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$ are all of the same $\Theta(1/\sqrt{n})$ order.

As a remark, from the bound of the second term of (5.5.7), we can deduce that the change of the output of the NN satisfies

$$|f_t(\boldsymbol{X}) - f_0(\boldsymbol{X})| \leq \frac{C}{\sqrt{n}},$$

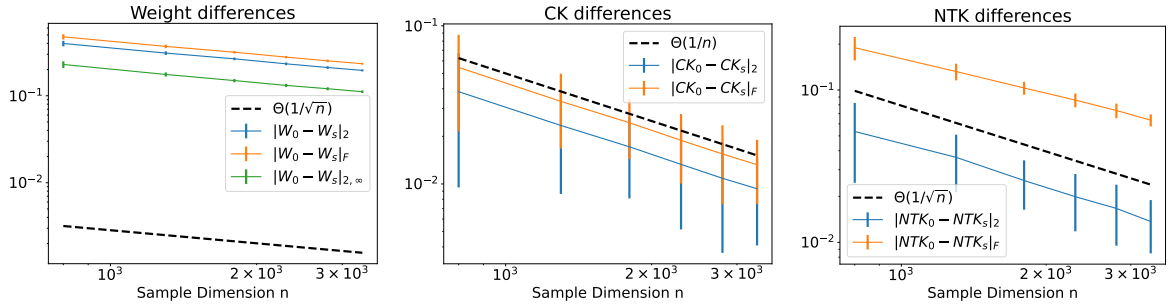
for some t-dependent constant $C > 0$, any $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{1})$ and any finite time t. In other words, when $\eta = \Theta(1)$, the change of the output of the NN at the early phase (i.e. $t = \Theta(1)$) is negligible and its order is $O(\frac{1}{\sqrt{n}})$.

5.5.2 Proof of Lemma 102

In this section, we complete the proof of Lemma 102. We first mention the empirical validation of Lemma 102 in Figure 5.13. Here $\sigma_\varepsilon = 0.2$, activation σ is a normalized ReLU and the target function σ^* is normalized tanh. Fix $d/n = 0.6$ and $N/n = 1.2$ as n is increasing. At each dimension, we take 25 trials to average. Notice that the changes in Frobenius norm for $\boldsymbol{W}$ and $\boldsymbol{K}^{\mathrm{CK}}$ are exactly $\Theta(1/\sqrt{n})$ and $\Theta(1/n)$, respectively. The operator norm of $\boldsymbol{K}^{\mathrm{NTK}}$ matches with Lemma 102, while the Frobenius norm of the change decays slower than the rate $\Theta(1/n)$. Additionally, in the simulation, we use $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{1})$, which indicates that our assumption for $\boldsymbol{v}$ in Lemma 102 can be weakened.

Proof of Lemma 102. Lemma 107 directly validates the control of $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$. By virtue of this result, we now present estimates for CK and NTK. Based on [OS20, Section 6.6.1], we have

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\| \leq \frac{\lambda_\sigma}{\sqrt{d}}\|\boldsymbol{X}\|\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F. \tag{5.5.8}$$



(a) Early phase for $\boldsymbol{W}_t$. (b) Early phase for $\boldsymbol{K}^{\mathrm{CK}}$. (c) Early phase for $\boldsymbol{K}^{\mathrm{NTK}}$.

Figure 5.13. Empirical validations for Lemma 102 and Lemma 107 at $t = 3$. (a) Norms of the changes for $\boldsymbol{W}_3 - \boldsymbol{W}_0$. (b) Norms of the changes for $\boldsymbol{K}_3^{\mathrm{CK}} - \boldsymbol{K}_0^{\mathrm{CK}}$. (c) Norms of the changes for $\boldsymbol{K}_3^{\mathrm{NTK}} - \boldsymbol{K}_0^{\mathrm{NTK}}$.

We apply the mean value theorem to obtain this inequality. Recall the operator norm bound for Gaussian random matrix $\boldsymbol{X}$ in (5.5.6). We know $\|\boldsymbol{X}/\sqrt{d}\| \lesssim 1 + \sqrt{\gamma_1}$ with high probability as $n/d \to \gamma_1$. Hence, with the help of Lemma 107, we can claim

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\| \leq C\lambda_\sigma(1+\sqrt{\gamma_1})/\sqrt{n},$$

with probability at least $1 - \exp(-cn)$, for any fixed finite $t \in [n]$. Similarly, we can control the change in the Frobenius norm as follows:

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\|_F^2 \leq \frac{\lambda_\sigma^2}{d}\|\boldsymbol{X}\|^2\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F^2 \leq C\lambda_\sigma^2(1+\sqrt{\gamma_1})^2/n, \tag{5.5.9}$$

with probability at least $1 - \exp(-cn)$. Therefore, we can control the change in the CK matrix in the Frobenius norm by the following inequalities:

$$\begin{aligned}
&\left\|\boldsymbol{K}_t^{\mathrm{CK}} - \boldsymbol{K}_0^{CK}\right\|_F \\
&\leq \frac{1}{h}\left(\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| + \left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}\right\|\right) \cdot \left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F \\
&\quad + \frac{1}{h}\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| \cdot \left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F.
\end{aligned}$$

Therefore, by (5.5.8), (5.5.9) and Lemma 106, we can claim that there exist constants $c, C > 0$

such that with probability at least $1-\exp(-cn)$, $\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{CK}\right\|_F$ is upper bounded by C/n in the LWR.

Now we consider the change in the NTK matrix during training. Since the empirical NTK can be decomposed into two parts, one of which is exactly the CK, it suffices to consider the change of the first part of the empirical NTK. Recall that

$$\boldsymbol{K}_t := \frac{1}{d}\boldsymbol{X}^\top\boldsymbol{X}\odot\frac{1}{h}\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{X}\right)^\top \mathrm{diag}(\boldsymbol{v}_t)^2\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{X}\right).$$

Following the notation in [OS20], we denote $\mathscr{J}(\boldsymbol{W}_t) := [\mathscr{J}(\boldsymbol{W}_1^t),\ldots,\mathscr{J}(\boldsymbol{W}_N^t)] \in \mathbb{R}^{n\times hd}$ with $\mathscr{J}(\boldsymbol{W}_i) := \frac{v_i}{\sqrt{h}}\mathrm{diag}(\sigma'(\boldsymbol{X}^\top\boldsymbol{W}_i/\sqrt{d}))\boldsymbol{X}^\top/\sqrt{d} \in \mathbb{R}^{n\times d}$. Hence, $\boldsymbol{K}_t = \mathscr{J}(\boldsymbol{W}_t)\mathscr{J}(\boldsymbol{W}_t)^\top$ and

$$\begin{aligned}\|\boldsymbol{K}_t-\boldsymbol{K}_0\| &= \left\|\mathscr{J}(\boldsymbol{W}_t)\mathscr{J}(\boldsymbol{W}_t)^\top-\mathscr{J}(\boldsymbol{W}_0)\mathscr{J}(\boldsymbol{W}_0)^\top\right\| \\ &\le 2\|\mathscr{J}(\boldsymbol{W}_0)\|\|\mathscr{J}(\boldsymbol{W}_t)-\mathscr{J}(\boldsymbol{W}_0)\|+\|\mathscr{J}(\boldsymbol{W}_t)-\mathscr{J}(\boldsymbol{W}_0)\|^2. \end{aligned} \tag{5.5.10}$$

By [OS20, Lemma 6.6], we know $\|\mathscr{J}(\boldsymbol{W}_0)\|^2 = \left\|\boldsymbol{K}_0^{\mathrm{NTK}}\right\|$ is upper bounded by some constant $C>0$ with high probability. Then, we apply the inequalities from Lemma 6.5 of [OS20] to obtain that

$$\begin{aligned}&\|\mathscr{J}(\boldsymbol{W}_t)-\mathscr{J}(\boldsymbol{W}_0)\|^2 \\ &\le \left\|\left(\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})-\sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right)^\top\frac{\mathrm{diag}(\boldsymbol{v})}{\sqrt{h}}\right\|^2\left(\max_{i\in[n]}\left\|\boldsymbol{X}_i/\sqrt{d}\right\|^2\right) \\ &\le \frac{1}{h}\|\boldsymbol{v}\|_\infty^2\left\|\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})-\sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|^2\left(\max_{i\in[n]}\left\|\boldsymbol{X}_i/\sqrt{d}\right\|^2\right) \\ &\le \frac{\lambda_\sigma^2}{h}\frac{\|\boldsymbol{X}\|^2}{d}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_F^2\left(\max_{i\in[n]}\left\|\boldsymbol{X}_i/\sqrt{d}\right\|^2\right), \end{aligned} \tag{5.5.11}$$

where the last inequality is due to the mean value theorem, the uniform bound on σ'', and the

assumption on the second layer $\boldsymbol{v}$. Notice that Gaussian random vectors satisfy

$$\mathbb{P}\left(\max_{i\in[n]}\frac{1}{d}\|\boldsymbol{X}_i\|^2\geq 2\right)\leq 2ne^{-cn}, \tag{5.5.12}$$

as $n/d\to\gamma_1$ and $h/d\to\gamma_2$. Thus, with (5.5.6) and Lemma 107, we obtain

$$\mathbb{P}\left(\|\mathscr{J}(\boldsymbol{W}_t)-\mathscr{J}(\boldsymbol{W}_0)\|\geq\frac{C\lambda_\sigma(1+\gamma_1)}{n}\right)\leq 4ne^{-cn},$$

where constant C relies on the number of steps t. Hence, by (5.5.10), we finally bound in norm the difference between the initial and the trained NTK matrices at the early phase (t is finite). □

Corollary 108. *For any fixed $t\in\mathbb{N}$, $i\in[d]$ and $k\in[n]$, denote λ_i^t, ν_k^t and μ_k^t the i-th, and k-th eigenvalues of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{CK}$ and $\boldsymbol{K}_t^{NTK}$, respectively. Then, under the assumptions of Lemma 102, we have*

$$|\lambda_i^t-\lambda_i^0|,\ |\nu_k^t-\nu_k^0|,\ |\mu_k^t-\mu_k^0|\to 0,$$

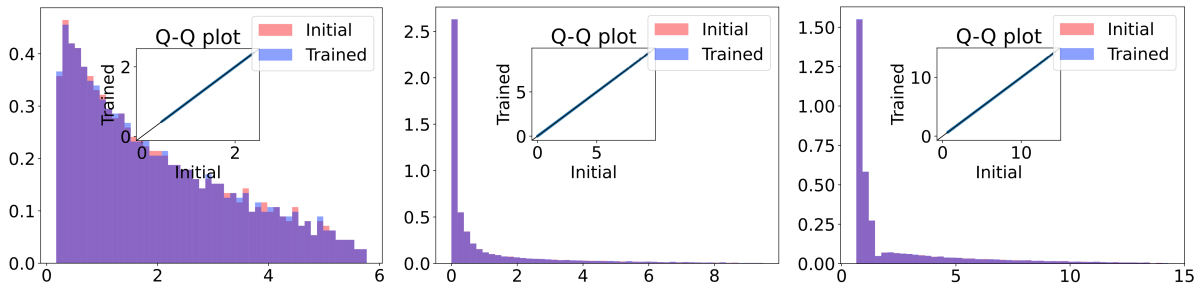
almost surely in LWR. Consequently, the eigenvalues of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{CK}$ and $\boldsymbol{K}_t^{NTK}$ are the same as corresponding the eigenvalues of initial $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{CK}$ and $\boldsymbol{K}_0^{NTK}$, respectively.

This corollary is a direct outcome of Weyl's inequality from Theorem A.46 in [BS10]. Consequently, this corollary concludes that for any fixed $t\geq 0$, almost surely, the limiting spectra of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{\mathrm{CK}}$ and $\boldsymbol{K}_t^{\mathrm{NTK}}$ are the same as those of $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{\mathrm{CK}}$ and $\boldsymbol{K}_0^{\mathrm{NTK}}$ in LWR. This corollary claims that not only does the bulk of distributions stay identical to the initialization, but also that any eigenvalues stay the same as at the initialization. This shows that the smallest eigenvalue of $\boldsymbol{K}_t^{\mathrm{NTK}}$ has the same lower bound as $\boldsymbol{K}_0^{\mathrm{NTK}}$ in the early phase of training.

5.5.3 Global Convergence for GD Under Linear-Width Regime

In this section, we study the final stage of (5.4.1) as training loss is approaching zero and prove Theorem 103. Figure 5.14 shows that the spectra are unchanged globally, even after

training in this case. Here the training loss is less than 10^{-5}, $h = 3000, n = 2000$, and $d = 1000$. The final R^2 score is 0.55964 and the test loss is 0.44724. The activation is a normalized ReLU, and the target is Sigmoid. In Corollary 104, we confirm this observation for the weight, CK, and NTK matrices via Frobenius norm control. In the simulation, the second layer is initialized as $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{1})$, which is more general than our assumption on $\boldsymbol{v}$ in Theorem 103.



(a) Weight Spectra. (b) CK Spectra. (c) NTK Spectra.

Figure 5.14. The initial and trained spectra with GD only for the first layer: (a) Weight spectra. (b) $\boldsymbol{K}^{\mathrm{CK}}$ spectra. (c) $\boldsymbol{K}^{\mathrm{NTK}}$ spectra. Here Q-Q subplots indicate the invariant spectra of weight and kernel matrices.

Proof of Theorem 103. Recall the Jacobian matrix $\mathscr{J}(\boldsymbol{W})$ defined in the proof of Lemma 102, and the definition of α based on (5.4.2) in Section 5.2. Denote the event

$$\mathscr{A} := \left\{ \|\boldsymbol{X}\| \le 2(1+\sqrt{\gamma_1})\sqrt{d},\ \max_{i\in[n]} \|\boldsymbol{X}_i\|^2 \le 2d, \sigma_{\min}(\mathscr{J}(\boldsymbol{W}_0)) \ge 2\alpha \right\}.$$

By (5.5.6), (5.5.12) and Theorem 69, we have $\mathbb{P}(\mathscr{A}) \ge 1 - 2e^{-cn} - 2ne^{-cn} - n^{-7/3}$ for some constant $c > 0$ and all large n in LWR. In the following, conditionally on event $\mathscr{A}$, we will apply Theorem 6.10 of [OS20] to obtain the global convergence. Conditionally on $\mathscr{A}$, Lemma 6.6 of [OS20] implies

$$\|\mathscr{J}(\boldsymbol{W})\| \le \lambda_\sigma \|\boldsymbol{v}\|_\infty \left\| \boldsymbol{X}/\sqrt{d} \right\| \le 2\lambda_\sigma(1+\sqrt{\gamma_1}), \tag{5.5.13}$$

for any $\boldsymbol{W}$. Define $\beta = 2\lambda_\sigma(1+\sqrt{\gamma_1})$. Moreover, in terms of (5.5.11), we can verify the Lipschitz

property for the Jacobian matrix as follows: conditionally on $\mathscr{A}$,

$$\left\| \mathscr{J}(\tilde{\boldsymbol{W}}) - \mathscr{J}(\boldsymbol{W}) \right\| \leq \frac{2\beta}{\sqrt{h}} \left\| \tilde{\boldsymbol{W}} - \boldsymbol{W} \right\|_F, \tag{5.5.14}$$

for any $\tilde{\boldsymbol{W}}, \boldsymbol{W} \in \mathbb{R}^{h \times d}$. Therefore, conditionally on $\mathscr{A}$, $\mathscr{J}(\boldsymbol{W})$ is a L-Lipschitz function with respect to $\boldsymbol{W}$ where $L := \frac{2\beta}{\sqrt{h}}$. To complete the proof, it suffices to investigate the smallest singular value of $\mathscr{J}(\boldsymbol{W})$ when $\boldsymbol{W}$ is in the vicinity of $\boldsymbol{W}_0$. Recall $\ell(\boldsymbol{W}) = \|\boldsymbol{y} - f_{\boldsymbol{W}}(\boldsymbol{X})\|$. Notice that for any unit vector $\boldsymbol{u} \in \mathbb{R}^n$, we have $\boldsymbol{u}^\top f_{\boldsymbol{W}_0}(\boldsymbol{X}) = \frac{1}{\sqrt{h}} \sum_{i=1}^h v_i \sigma(\boldsymbol{W}_i^\top \boldsymbol{X}/\sqrt{d})\boldsymbol{u}$, where $\boldsymbol{W}_i^\top$ is the i-th row of $\boldsymbol{W}_0$ for $i \in [N]$. Consider event $\mathscr{B} := \left\{ \left\| \sigma(\boldsymbol{W}_0 \boldsymbol{X}/\sqrt{d}) \right\| \leq C\sqrt{n} \right\}$ for some universal constant $C > 0$. Lemma 106 proves $\mathbb{P}(\mathscr{B}) \geq 1 - 2e^{-cn}$. By the assumption of $\boldsymbol{v}$, we know each entry v_i is a sub-Gaussian random variable with a sub-Gaussian norm at most 1. Then, according to Hoeffding's inequality, conditionally on the event $\mathscr{B}$, we have

$$\mathbb{P}\left(\left| \frac{1}{\sqrt{h}} \sum_{i=1}^h v_i \sigma(\boldsymbol{W}_i^\top \boldsymbol{X}/\sqrt{d})\boldsymbol{u} \right| \geq t \right) \leq 2\exp\left(-ct^2\right),$$

for every $t \geq 0$ and some constant $c > 0$. Let $t = 2\sqrt{n}$. Considering an $\frac{1}{4}$-net $\mathscr{N}$ of the unit sphere $\mathbb{S}^{n-1}$, we can get

$$\mathbb{P}\left(\left\| f_{\boldsymbol{W}_0}(\boldsymbol{X}) \right\| \geq \sqrt{n} \right) \leq \mathbb{P}\left(2 \max_{\boldsymbol{u} \in \mathscr{N}} \left| \boldsymbol{u}^\top f_{\boldsymbol{W}_0}(\boldsymbol{X}) \right| \geq \sqrt{n} \right) \leq 9^n 2\exp(-cn) \leq 2e^{-c'n}, \tag{5.5.15}$$

for some constant $c' > 0$. Hence, based on Lemma 106 and (5.5.15), we can obtain $\ell(\boldsymbol{W}_0) \leq C_0\sqrt{n}$ with high probability for some universal constant $C_0 > 0$. Let us denote this event as $\mathscr{C} : \{\ell(\boldsymbol{W}_0) \leq C_0\sqrt{n}\}$. Define $R := 4\ell(\boldsymbol{W}_0)/\alpha$. For any $\boldsymbol{W}$ in a ball of radius R centered at $\boldsymbol{W}_0$, we have $\|\boldsymbol{W}_0 - \boldsymbol{W}\|_F \leq R$ and $\|\mathscr{J}(\boldsymbol{W}) - \mathscr{J}(\boldsymbol{W}_0)\| \leq LR$, conditionally on event $\mathscr{A}$. Thus, by (5.5.14), on event $\mathscr{A} \cap \mathscr{C}$, the smallest singular value $\sigma_{\min}(\mathscr{J}(\boldsymbol{W}))$ of the Jacobian matrix

$\mathscr{J}(\boldsymbol{W})$ can be bounded by

$$\begin{aligned}\sigma_{\min}(\mathscr{J}(\boldsymbol{W})) &\geq \sigma_{\min}(\mathscr{J}(\boldsymbol{W}_0)) - \|\mathscr{J}(\boldsymbol{W}) - \mathscr{J}(\boldsymbol{W}_0)\| \\ &\geq 2\alpha - LR \geq 2\alpha - \frac{8\beta}{\alpha}\frac{\ell(\boldsymbol{W}_0)}{\sqrt{h}} \geq 2\alpha - \frac{8C\beta}{\alpha}\sqrt{\frac{\gamma_1}{\gamma_2}},\end{aligned}$$

for some universal constant $C > 0$ and sufficiently large n, d, h. Notice that here constants C, β, and α do not rely on γ_2. Therefore, there exists a sufficiently large $\gamma^* > 0$ such that for all $\gamma_2 \geq \gamma^*$, we have $2\alpha - \frac{8C\beta}{\alpha}\sqrt{\frac{\gamma_1}{\gamma_2}} \geq \alpha$. In other words, when h is sufficiently large but still in the same order as n and d, for all $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \leq R$, we have $\sigma_{\min}(\mathscr{J}(\boldsymbol{W})) \geq \alpha$ conditionally on $\mathscr{C} \cap \mathscr{A}$. Combining with (5.5.13) and (5.5.14), conditionally on $\mathscr{C} \cap \mathscr{A}$, all the assumptions of Theorem 6.10 by [OS20] are satisfied when $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \leq R$. Therefore, when the learning rate $\frac{\eta}{n} \leq \frac{1}{\beta^2}\min\{1, \frac{4\alpha}{LR}\}$, we can get (5.4.3)-(5.4.5) for all $t \in \mathbb{N}$, conditionally on $\mathscr{C} \cap \mathscr{A}$. Both events $\mathscr{A}$ and $\mathscr{C}$ occur with high probability and only depend on initialization $\boldsymbol{W}_0$, $\boldsymbol{X}$ and $\boldsymbol{y}$. Hence we complete the proof of this theorem. Notice that since $\gamma_2 \geq \gamma^*$ is sufficiently large, $\frac{4\alpha}{LR} \geq \frac{\alpha^2}{2C\beta}\sqrt{\frac{\gamma_2}{\gamma_1}} > 1$. Therefore, it suffices to require $\eta \leq n/\beta^2$ to conclude that (5.4.3), (5.4.4) and (5.4.5) hold with high probability. This completes the proof. Moreover, (5.4.5) further shows that for all $t \in \mathbb{N}$,

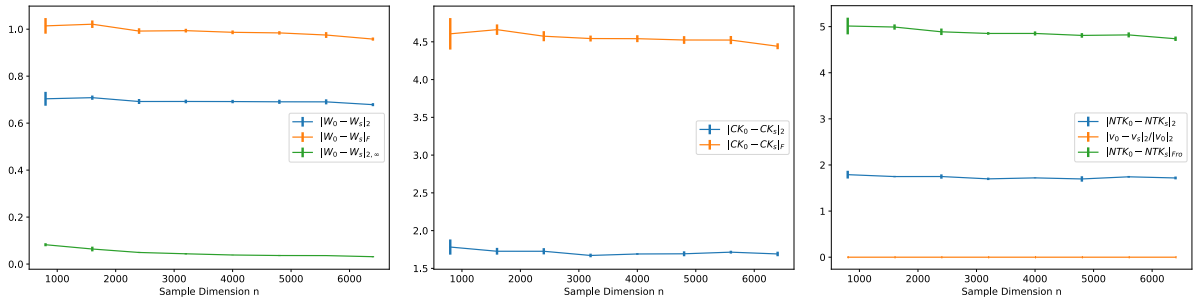
$$\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F \leq R \leq C\sqrt{n} + o_{d,\mathbb{P}}(1), \tag{5.5.16}$$

where we again apply Lemma 106 in the following way:

$$\ell(\boldsymbol{W}_0) \leq C\sqrt{n} + o_{d,\mathbb{P}}(1),$$

for some constant $C > 0$ only depending on $\gamma_1, \gamma_2, \sigma_\varepsilon, \sigma$ and σ^*. □

As a corollary, (5.4.5) controls the deviation of the final step weight from the initial weight. In Figure 5.15, we empirically verify this result. For instance, Figure 5.15(a) shows that $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|, \frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, and $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$ are $\Theta(1)$ when trainable parameters are



(a) Changes for $\frac{1}{\sqrt{d}}\boldsymbol{W}$. (b) Changes for CK. (c) Changes for NTK.

Figure 5.15. The change for the weight, CK, and NTK matrices when training NN. (a) The changes $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|$, $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_F$, and $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_{2,\infty}$. (b) The changes $\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{\mathrm{CK}}\right\|$ and $\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{\mathrm{CK}}\right\|_F$. (b) The changes $\|\boldsymbol{v}_t-\boldsymbol{v}_0\|$, $\left\|\boldsymbol{K}_t^{\mathrm{NTK}}-\boldsymbol{K}_0^{\mathrm{NTK}}\right\|$ and $\left\|\boldsymbol{K}_t^{\mathrm{NTK}}-\boldsymbol{K}_0^{\mathrm{NTK}}\right\|_F$.

convergent, where $\boldsymbol{W}_t$ represents the weight matrix after training. This implies that the final $\boldsymbol{W}_t$ is still close to the initial weight $\boldsymbol{W}_0$, even after training. Here, we fix $d/n=1.2$ and $h/n=0.6$ when n is increasing. Here, σ is normalized ReLU and the target is normalized tanh. The largest $n=6400$ and the learning rate $\eta=5.0$ for all training processes. We train each neural network until the training losses approach zero. Each experiment repeats 4 times. Next, we prove this observation and Corollary 104.

Proof of Corollary 104. Based on (5.5.16), we know $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_0-\boldsymbol{W}_t\|_F\leq C_0$ holds with high probability for some universal constant $C_0>0$. Conditionally on this event, we can then estimate changes in CK and NTK after training. The method is analogous to Lemma 102. For CK, we employ Lemma 106 and (5.5.9) to get

$$
\begin{aligned}
&\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{CK}\right\|_F\\
&\leq\frac{2}{h}\left(\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})-\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|+\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}\right\|\right)\cdot\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})-\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F\\
&\lesssim\frac{2\lambda_\sigma^2(1+\sqrt{\gamma_1})^2}{h}\|\boldsymbol{W}_0-\boldsymbol{W}_t\|_F^2+\frac{2C\sqrt{n}\lambda_\sigma(1+\sqrt{\gamma_1})}{h}\|\boldsymbol{W}_0-\boldsymbol{W}_t\|_F\\
&\lesssim\frac{2\lambda_\sigma(1+\sqrt{\gamma_1})C_0}{\gamma_2^2}(\lambda_\sigma(1+\sqrt{\gamma_1})C_0+C\sqrt{\gamma_1})=O_{d,\mathbb{P}}(1).
\end{aligned}
$$

This shows control of the change for the CK matrix after training, compared with the initial CK.

Let us denote $\boldsymbol{W}_i^t \in \mathbb{R}^{1\times d}$ as the i-th row of $\boldsymbol{W}_t$, and $\boldsymbol{X}_j$ as the j-th column of $\boldsymbol{X}$. Additionally, by Assumption 15, we know that

$$|\sigma'(x)-\sigma'(y)| \le \lambda_\sigma |x-y|, \tag{5.5.17}$$

for any $x, y \in \mathbb{R}$. For NTK, by modifying (5.5.11), one can deduce that

$$\begin{aligned}
&\|\mathscr{J}(\boldsymbol{W}_t)-\mathscr{J}(\boldsymbol{W}_0)\|_F^2 = \sum_{i=1}^h \left\|\mathscr{J}(\boldsymbol{W}_i^t)-\mathscr{J}(\boldsymbol{W}_i^0)\right\|_F^2 \\
&\overset{(i)}{\le} \frac{1}{h}\sum_{i=1}^h \left\|\operatorname{diag}\Big(\sigma'(\boldsymbol{W}_i^t\boldsymbol{X}/\sqrt{d}-\sigma'(\boldsymbol{W}_i^0\boldsymbol{X}/\sqrt{d})\Big)\right\|_F^2 \left\|\frac{\boldsymbol{X}}{\sqrt{d}}\right\|^2 \\
&\overset{(ii)}{\le} \frac{(1+\sqrt{\gamma_1})^2}{h}\sum_{i=1}^h \left\|\operatorname{diag}\Big(\sigma'(\boldsymbol{W}_i^t\boldsymbol{X}/\sqrt{d}-\sigma'(\boldsymbol{W}_i^0\boldsymbol{X}/\sqrt{d})\Big)\right\|_F^2 + o_{d,\mathbb{P}}(1) \\
&\overset{(iii)}{\le} \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^2}{h}\sum_{i=1}^h\sum_{j=1}^n \left(\frac{1}{\sqrt{d}}(\boldsymbol{W}_i^t-\boldsymbol{W}_i^0)\boldsymbol{X}_j\right) + o_{d,\mathbb{P}}(1) \\
&\overset{(iv)}{\le} \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^4}{h}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_F^2 + o_{d,\mathbb{P}}(1) \le \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^4 C_0^2}{\gamma_2} + o_{d,\mathbb{P}}(1),
\end{aligned}$$

where (i) is because of [Ver18, Exercise 6.3.3] and the assumption on $\boldsymbol{v}$, (ii) is due to (5.5.6), (iii) is due to the definition of Frobenius norm and (5.5.17), and (iv) is due to [Ver18, Exercise 6.3.3] and (5.5.6). As a result, from (5.5.10), we can finally conclude that $\left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{\mathrm{CK}}\right\|_F = O_{d,\mathbb{P}}(1)$ as $n/d \to \gamma_1$ and $h/d \to \gamma_2$.

As for the limiting spectra of weight and kernel matrices, since we know that

$$\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t-\boldsymbol{W}_0\|_F,\ \left\|\boldsymbol{K}_t^{\mathrm{CK}}-\boldsymbol{K}_0^{\mathrm{CK}}\right\|_F,\ \left\|\boldsymbol{K}_t^{\mathrm{NTK}}-\boldsymbol{K}_0^{\mathrm{NTK}}\right\|_F = O_{d,\mathbb{P}}(1),$$

we can automatically apply Corollary A.41 of [BS10]. This directly implies that the limiting empirical spectra of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{\mathrm{CK}}$ and $\boldsymbol{K}_t^{\mathrm{NTK}}$ are the same as the limiting spectra of $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{\mathrm{CK}}$ and $\boldsymbol{K}_0^{\mathrm{NTK}}$, respectively, as $n/d \to \gamma_1$ and $h/d \to \gamma_2$ (see Figure 5.14). □

5.6 Proofs for Spiked Eigenstructure of the Trained CK

In this section, we prove Theorem 105. The proof is an application of Theorem 40 in Chapter 3 as in the one-layer setting of the preceding section, but now reversing the roles of $\boldsymbol{X}$ and $\boldsymbol{W}$. We abbreviate

$$\boldsymbol{W} = \boldsymbol{W}_{\text{trained}}, \qquad \boldsymbol{Y} = \frac{1}{\sqrt{N}}\sigma(\tilde{\boldsymbol{X}}\boldsymbol{W}), \qquad \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{Y}^\top$$

where $\boldsymbol{K}$ is the CK matrix of interest. In contrast to the preceding section, the theorem requires characterizing the *left* spike singular vector of $\boldsymbol{Y}$, and we will do so using Theorem 40(c). Notice that Theorem 34 studied the spikes in dataset $\boldsymbol{X}$, while here we

We first recall the following approximation of $\boldsymbol{W}$ from [BES+22].

Proposition 109. *Under Assumption 16, set $\eta = \sum_{t=1}^{T} \eta_t$, and let θ_1, θ_2 be as defined in* (5.4.9). *Then*

$$\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\| \prec N^{-1/2} \quad \textit{where} \quad \widetilde{\boldsymbol{W}} = \boldsymbol{W}_0 + \frac{b_\sigma \eta}{n}\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{a}^\top. \tag{5.6.1}$$

The largest singular value $s_{\max}(\boldsymbol{W})$ falls outside the limit of its empirical singular value distribution if and only if $\theta_1 > \gamma_0^{1/4}$, in which case $s_{\max}(\boldsymbol{W})$ and its unit-norm left singular vector $\boldsymbol{u}(\boldsymbol{W})$ satisfy

$$s_{\max}(\boldsymbol{W}) \to s_1 := \sqrt{\frac{(1+\theta_1^2)(\gamma_0+\theta_1^2)}{\theta_1^2}}, \quad |\boldsymbol{u}(\boldsymbol{W})^\top \boldsymbol{\beta}_*|^2 \to \frac{\theta_2^2}{\theta_1^2}\left(1 - \frac{\gamma_0+\theta_1^2}{\theta_1^2(\theta_1^2+1)}\right) \quad a.s. \tag{5.6.2}$$

Proof. Notice that each gradient update matrix $\boldsymbol{G}_t$ of (5.4.7) takes the form

$$\boldsymbol{G}_t = \frac{1}{n}\boldsymbol{X}^\top\left[\left(\frac{1}{\sqrt{N}}\left(\boldsymbol{y} - \frac{1}{\sqrt{N}}\sigma(\boldsymbol{X}\boldsymbol{W}_t)\boldsymbol{a}\right)\boldsymbol{a}^\top\right) \odot \sigma'(\boldsymbol{X}\boldsymbol{W}_t)\right],$$

From the proof of [BES+22, Lemma 16], for each $t = 1, \ldots, T$, this matrix $\boldsymbol{G}_t$ satisfies the same

rank-one approximation

$$\left\| \sqrt{N}\boldsymbol{G}_t - \frac{b_\sigma}{n}\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{a}^\top \right\| \prec N^{-1/2}.$$

This implies (5.6.1) in light of (5.4.7), and the statements of (5.6.2) then follow from [BES+22, Theorem 3]. □

We denote the columns of $\boldsymbol{W} \equiv \boldsymbol{W}_{\text{trained}} \in \mathbb{R}^{d\times N}$ and of the initialization $\boldsymbol{W}_0 \in \mathbb{R}^{d\times N}$ by

$$\boldsymbol{w}_i \in \mathbb{R}^d,\ \boldsymbol{w}_{i,0} \in \mathbb{R}^d \text{ for } i \in [N]$$

respectively. Fixing a large constant $B > 0$ and small constant $\varepsilon > 0$, define the event

$$\mathscr{E}(\boldsymbol{W}) = \Big\{ \|\boldsymbol{W}\| < B,\ |\boldsymbol{w}_i^\top \boldsymbol{w}_j| < n^{-1/2+\varepsilon} \text{ and } |\|\boldsymbol{w}_i\| - 1| < n^{-1/2+\varepsilon} \text{ for all } i \neq j \in [N] \Big\}.$$

Lemma 110. *Under Assumption 16, for some sufficiently large constant $B > 0$ and any fixed $\varepsilon > 0$, $\mathscr{E}(\boldsymbol{W})$ holds almost surely for all large n and any fixed $T \in \mathbb{N}$.*

Proof. By the assumption $[\boldsymbol{W}_0]_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/d)$, it is immediate to check that $\mathscr{E}(\boldsymbol{W}_0)$ holds almost surely for all large n. To show that $\mathscr{E}(\boldsymbol{W})$ holds, we apply the approximation (5.6.1). Here, under Assumption 16, we have by standard tail bounds for Gaussian vectors and matrices that

$$\mathbf{1}\{\|\boldsymbol{X}^\top \boldsymbol{y}\| > Cn\} \le \mathbf{1}\{\|\boldsymbol{X}\| \cdot (\lambda_{\sigma_*}\|\boldsymbol{X}\boldsymbol{\beta}_*\| + \|\boldsymbol{\varepsilon}\|) > Cn\} \prec 0, \qquad \mathbf{1}\{\|\boldsymbol{a}\| > C\} \prec 0$$

for a sufficiently large constant $C > 0$, and also $\|\boldsymbol{a}\|_\infty \prec N^{-1/2}$. Then this implies

$$\max_{1\le i\le N} \|\boldsymbol{w}_i - \boldsymbol{w}_{i,0}\| \le \max_{1\le i\le N} \|\widetilde{\boldsymbol{w}}_i - \boldsymbol{w}_{i,0}\| + \|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\| \prec N^{-1/2} \tag{5.6.3}$$

and $\mathbf{1}\{\|\boldsymbol{W} - \boldsymbol{W}_0\| > C'\} \prec 0$ for a constant $C' > 0$. Then $\mathscr{E}(\boldsymbol{W})$ also holds almost surely for all

large n, as claimed. □

Analogous to the argument of Section 3.7.1, we may now condition on $\boldsymbol{W}$, i.e. we assume that $\boldsymbol{W}$ is deterministic and satisfies $\mathscr{E}(\boldsymbol{W})$ for all large n, and we write $\mathbb{E}$ for the expectation over only the randomness of the new data $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$. Defining

$$\boldsymbol{G} = \sqrt{\frac{N}{n}}(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}) \in \mathbb{R}^{n \times N}, \quad \boldsymbol{u} = \frac{1}{\sqrt{n}}\tilde{\boldsymbol{y}} \in \mathbb{R}^n \quad \text{where} \quad \boldsymbol{Y} = \frac{1}{\sqrt{N}}\sigma(\tilde{\boldsymbol{X}}\boldsymbol{W}), \tag{5.6.4}$$

observe that $[\boldsymbol{u}, \boldsymbol{G}] \in \mathbb{R}^{n \times (N+1)}$ has centered i.i.d. rows with respect to the randomness of $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$. We will write $\mathbb{E}_{\boldsymbol{x}}$ for the expectation with respect to a standard Gaussian vector $\boldsymbol{x} \sim \mathscr{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

Lemma 111. *Suppose $\boldsymbol{W}$ satisfies $\mathscr{E}(\boldsymbol{W})$ for all large n. Then*

$$\|\mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}^\top \boldsymbol{W})]\| \to 0, \qquad \|\mathbb{E}\boldsymbol{Y}\| \to 0, \qquad \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{lin}}\| \to 0 \tag{5.6.5}$$

where

$$\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}^\top \boldsymbol{W})^\top \sigma(\boldsymbol{x}^\top \boldsymbol{W})] - \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}^\top \boldsymbol{W})]^\top \mathbb{E}_{\boldsymbol{x}}[\sigma(\boldsymbol{x}^\top \boldsymbol{W})] \tag{5.6.6}$$

$$\boldsymbol{\Sigma}_{\text{lin}} := b_\sigma^2(\boldsymbol{W}^\top \boldsymbol{W}) + (1 - b_\sigma^2)\boldsymbol{I}_N. \tag{5.6.7}$$

Proof. The proof is the same as Lemmas 58 and 59 in Chapter 3. We ignore the details for simplicity. □

Proof of Theorem 105. We condition on $\boldsymbol{W}$ satisfying $\mathscr{E}(\boldsymbol{W})$ for all large n, and we apply Theorem 40(c) for $[\boldsymbol{u}, \boldsymbol{G}] \in \mathbb{R}^{(n+1) \times N}$ (exchanging n and N). It may be checked that Assumption 6 holds for $[\boldsymbol{u}, \boldsymbol{G}]$ by the same argument as in Lemma 62.

By the convergence $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{lin}}\| \to 0$ in Lemma 111 and Proposition 109, if $\theta_1 > \gamma_0^{1/4}$,

then Assumption 7 holds for $\boldsymbol{\Sigma}$ with $r = 1$ and

$$\nu = b_\sigma^2 \otimes \rho_{\gamma_0}^{\mathrm{MP}} \oplus (1 - b_\sigma^2), \qquad \lambda_1 = b_\sigma^2 \frac{(1+\theta_1^2)(\gamma_0+\theta_1^2)}{\theta_1^2} + (1-b_\sigma^2) \notin \mathrm{supp}(\nu),$$

where $\rho_{\gamma_0}^{\mathrm{MP}}$ is the standard Marčenko-Pastur limit for the empirical eigenvalue distribution of $\boldsymbol{W}^\top \boldsymbol{W}$, hence ν is the limit empirical eigenvalue distribution of $\boldsymbol{\Sigma}$, and λ_1 is the limit of $\lambda_{\max}(\boldsymbol{\Sigma})$. If instead $\theta_1 \le \gamma_0^{1/4}$, then Assumption 7 holds with $r = 0$.

Then Theorem 40(a,c) characterizes the outlier eigenvalue and eigenvector of $\boldsymbol{G}\boldsymbol{G}^\top$, showing:

- $\boldsymbol{G}\boldsymbol{G}^\top$ has a spike eigenvalue if and only if $\theta_1 > \gamma_0^{1/4}$ and $z'(-1/\lambda_1) > 0$, where $z(\cdot)$ is defined by (1.2.6) with $\gamma = \gamma_1$ and the measure ν given above. In this case, $\lambda_{\max}(\boldsymbol{G}\boldsymbol{G}^\top) \to z(-1/\lambda_1)$ almost surely.
- When $\theta_1 > \gamma_0^{1/4}$ and $z'(-1/\lambda_1) > 0$, letting $\boldsymbol{u}(\boldsymbol{G}), \boldsymbol{v}(\boldsymbol{\Sigma})$ be the leading unit-norm left singular vector of $\boldsymbol{G}$ and leading unit-norm eigenvector of $\boldsymbol{\Sigma}$, almost surely

$$|\boldsymbol{u}^\top \boldsymbol{u}(\boldsymbol{G})| - \frac{\sqrt{z(-1/\lambda_1)\varphi(-1/\lambda_1)}}{\lambda_1} \cdot \left| \mathbb{E}_{\boldsymbol{x}}[\sigma_*(\boldsymbol{\beta}_*^\top \boldsymbol{x})\sigma(\boldsymbol{x}^\top \boldsymbol{W})]\boldsymbol{v}(\boldsymbol{\Sigma}) \right| \to 0.$$

  where $\varphi(\cdot)$ is defined by (3.3.6) also with $\gamma = \gamma_1$ and the above measure ν.

By an application of Weyl's inequality and the Davis-Kahan Theorem as in the proof of Theorem 34, this implies for $\boldsymbol{K} = \boldsymbol{Y}\boldsymbol{Y}^\top$ that if $\theta_1 > \gamma_0^{1/4}$ and $z'(-1/\lambda_1) > 0$, then its leading eigenvalue $\lambda_{\max}(\boldsymbol{K})$ and unit eigenvector $\widehat{\boldsymbol{u}}$ satisfy

$$\begin{aligned} &\lambda_{\max}(\boldsymbol{K}) \to \gamma_1^{-1} z(-1/\lambda_1), \\ &|\boldsymbol{u}^\top \widehat{\boldsymbol{u}}| - \frac{\sqrt{z(-1/\lambda_1)\varphi(-1/\lambda_1)}}{\lambda_1} \cdot \left| \mathbb{E}_{\boldsymbol{x}}[\sigma_*(\boldsymbol{\beta}_*^\top \boldsymbol{x})\sigma(\boldsymbol{x}^\top \boldsymbol{W})]\boldsymbol{v}(\boldsymbol{W}) \right| \to 0, \end{aligned} \tag{5.6.8}$$

where $\boldsymbol{v}(\boldsymbol{W})$ is the leading unit eigenvector of $\boldsymbol{\Sigma}_{\mathrm{lin}}$, i.e. the leading right singular vector of $\boldsymbol{W}$. If $\theta_1 \le \gamma_0^{1/4}$ or $z'(-1/\lambda_1) \le 0$, then all eigenvalues of $\boldsymbol{K}$ converge to the support of its limiting

empirical eigenvalue law.

Finally, in the case of $\theta_1 > \gamma_0^{1/4}$ and $z'(-1/\lambda_1) > 0$, we may conclude the proof by showing

$$\left\| \mathbb{E}_{\boldsymbol{x}}[\sigma_*(\boldsymbol{\beta}_*^\top \boldsymbol{x})\sigma(\boldsymbol{x}^\top \boldsymbol{W})] - b_\sigma b_{\sigma_*} \boldsymbol{\beta}_*^\top \boldsymbol{W} \right\| \to 0 \text{ a.s.} \tag{5.6.9}$$

For each column $i \in [N]$, we have from (5.6.3) that

$$\|\boldsymbol{w}_i - \boldsymbol{w}_{i,0}\| \prec N^{-1/2},$$

where $\boldsymbol{w}_{i,0} \sim \mathcal{N}(0, d^{-1}\boldsymbol{I})$ and $\boldsymbol{\beta}_*$ is deterministic. Hence $(\boldsymbol{w}_i, \boldsymbol{\beta}_*)$ satisfy the approximate orthonormality conditions $|\|\boldsymbol{w}_i\| - 1| \prec N^{-1/2}$, $\|\boldsymbol{\beta}_*\| - 1 = 0$, and $|\boldsymbol{w}_i^\top \boldsymbol{\beta}_*| \prec N^{-1/2}$. Then Lemma 16 implies

$$\left| \mathbb{E}_{\boldsymbol{x}}[\sigma_*(\boldsymbol{\beta}_*^\top \boldsymbol{x})\sigma(\boldsymbol{x}^\top \boldsymbol{w}_i)] - b_\sigma b_{\sigma_*} \boldsymbol{\beta}_*^\top \boldsymbol{w}_i \right| \prec N^{-1}.$$

(We note that Lemma 16(a) assumes $\sigma = \sigma_*$, but the proof is identical for $\sigma \neq \sigma_*$ both satisfying Assumption 5.) Applying this to each coordinate $i \in [N]$ yields (5.6.9). Observe that $\boldsymbol{\beta}_*^\top \boldsymbol{W} \boldsymbol{v}(\boldsymbol{W}) = s_{\max}(\boldsymbol{W}) \cdot \boldsymbol{\beta}_*^\top \boldsymbol{u}(\boldsymbol{W})$ where $s_{\max}(\boldsymbol{W})$ and $\boldsymbol{u}(\boldsymbol{W})$ are the leading singular value and left singular vector of $\boldsymbol{W}$, and recall from the definitions (5.6.4) that $\boldsymbol{u} = \frac{1}{\sqrt{n}}\tilde{\boldsymbol{y}}$. Then we can apply (5.6.9) and Proposition 109 to (5.6.8) to conclude that

$$\frac{1}{\sqrt{n}} |\tilde{\boldsymbol{y}}^\top \widehat{\boldsymbol{u}}| \to b_\sigma b_{\sigma_*} \frac{\sqrt{z(-1/\lambda_1)\varphi(-1/\lambda_1)}}{\lambda_1} \cdot \frac{\theta_2 \sqrt{(\theta_1^4 - \gamma_0)(\gamma_0 + \theta_1^2)}}{\theta_1^3} > 0 \text{ a.s.}$$

□

5.7 Acknowledgment

Chapter 5 is extracted from a combination of the paper, "Zhichao Wang, Andrew Engel, Anand D. Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in Neural Information Processing Systems* 36 (2024)", and the preprint, "Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. arXiv:2402.10127 (2024)". The thesis author is the co-author of these two papers.

Bibliography

[AAM22] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[ABP22] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

[Ada15] Radoslaw Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

[ADH+19a] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[ADH+19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2010.

[AKM+17] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262. PMLR, 2017.

[ALP22] Ben Adlam, Jake A Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3434–3457. PMLR, 2022.

[AP20] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[AS17] Guillaume Aubrun and Stanisław J Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.

[ASS20] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[Aub12] Guillaume Aubrun. Partial transposition of random states and non-centered semicircular distributions. *Random Matrices: Theory and Applications*, 1(02):1250001, 2012.

[AVPVF23] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[AW15] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, 2015.

[AZLL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[Bac13] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209. PMLR, 2013.

[Bac17] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[BAGHJ23] Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. High-dimensional sgd aligns with emerging outlier eigenspaces. *arXiv preprint arXiv:2310.03010*, 2023.

[BAGJ23] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Communications on Pure and Applied Mathematics*, 2023.

[Bao12] Zhigang Bao. Strong convergence of esd for the generalized sample covariance matrices when $p/n \to 0$. *Statistics & Probability Letters*, 82(5):894–901, 2012.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[Bar51] Maurice S Bartlett. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111, 1951.

[BCP20] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[BEK+14] Alex Bloemendal, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability*, 19(none):1 – 53, 2014.

[BES+22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *arXiv preprint arXiv:2205.01445*, 2022.

[BES+23] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[BG10] Florent Benaych-Georges. On a surprising relation between the marchenko-pastur law, rectangular and square free convolutions. *Annales de l'IHP Probabilités et statistiques*, 46(3):644–652, 2010.

[BGL+21] Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 2269–2277. PMLR, 2021.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[BGN12] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

[BKYY16] Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1-2):459–552, 2016.

[BL20] Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2020.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BM19] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12873–12884, 2019.

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[BMR22] Gaspard Beugnot, Julien Mairal, and Alessandro Rudi. On the benefits of large learning rates for kernel methods. In *Conference on Learning Theory*, pages 254–282. PMLR, 2022.

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

[BP22] Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*, 2022.

[BPH23] David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. *arXiv preprint arXiv:2302.06210*, 2023.

[BS98] Zhi-Dong Bai and Jack W Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

[BS06] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

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

[BX22] Zhigang Bao and Xiaocong Xu. Extreme eigenvalues of log-concave ensemble. *arXiv preprint arXiv:2212.11634*, 2022.

[BY88] Zhidong Bai and Y. Q. Yin. Convergence to the semicircle law. *The Annals of Probability*, pages 863–875, 1988.

[BY12] Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

[BZ10] ZD Bai and LX Zhang. The limiting spectral distribution of the product of the Wigner matrix and a nonnegative definite matrix. *Journal of Multivariate Analysis*, 101(9):1927–1949, 2010.

[Cap13] Mireille Capitaine. Additive/multiplicative free subordination property and limiting eigenvectors of spiked additive deformations of wigner matrices and spiked sample covariance matrices. *Journal of Theoretical Probability*, 26:595–648, 2013.

[Cap18] Mireille Capitaine. Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. In *Séminaire de Probabilités XLIX*, pages 119–164. Springer, 2018.

[CBG16] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

[CH23] Benoit Collins and Tomohiro Hayase. Asymptotic freeness of layerwise jacobians caused by invariance of multilayer perceptron: The haar orthogonal case. *Communications in Mathematical Physics*, 397(1):85–109, 2023.

[Cha22] Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

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

[Cho23] Clément Chouard. Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks. *arXiv preprint arXiv:2306.05850*, 2023.

[CHS20] Shuxiao Chen, Hangfeng He, and Weijie Su. Label-aware neural tangent kernel: Toward better generalization and local elasticity. *Advances in Neural Information Processing Systems*, 33, 2020.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

[CP12] Binbin Chen and Guangming Pan. Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, 18(4):1405–1420, 2012.

[CP15] Binbin Chen and Guangming Pan. CLT for linear spectral statistics of normalized sample covariance matrices with the dimension much larger than the sample size. *Bernoulli*, 21(2):1089–1133, 2015.

[CS09] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[CSTEK01] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 14, 2001.

[CT18] Djalil Chafaï and Konstantin Tikhomirov. On the convergence of the extremal eigenvalues of empirical covariance matrices with dependence. *Probability Theory and Related Fields*, 170(3-4):847–889, 2018.

[CYZ18] Benoît Collins, Zhi Yin, and Ping Zhong. The PPT square conjecture holds generically for some classes of independent states. *Journal of Physics A: Mathematical and Theoretical*, 51(42):425301, 2018.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[DGA20] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2020.

[Dic16] Lee H Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

[DK70] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[DKL$^+$23] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[DLL$^+$19a] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of International Conference on Machine Learning 36*, pages 1675–1685, 2019.

[DLL$^+$19b] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2019.

[DLMY23] Sofiia Dubova, Yue M Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime. *arXiv preprint arXiv:2310.18280*, 2023.

[DLS22] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[dRBK20] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In *International Conference on Machine Learning*, 2020.

[DW18] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[DZPS19a] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[DZPS19b] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[EK10] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[EKY13] László Erdős, Antti Knowles, and Horng-Tzer Yau. Averaging fluctuations in resolvents of random band matrices. In *Annales Henri Poincaré*, volume 14, pages 1837–1926. Springer, 2013.

[FDP+20] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

[Fel23a] Michael J Feldman. Spectral properties of elementwise-transformed spiked matrices. *arXiv preprint arXiv:2311.02040*, 2023.

[Fel23b] Michael J Feldman. Spiked singular values and vectors under extreme aspect ratios. *Journal of Multivariate Analysis*, 196:105187, 2023.

[FJ22] Zhou Fan and Iain M Johnstone. Tracy-widom at each edge of real covariance and manova estimators. *The annals of applied probability: an official journal of the Institute of Mathematical Statistics*, 32(4):2967, 2022.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

[FW20] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[GKK+23] Alice Guionnet, Justin Ko, Florent Krzakala, Pierre Mergny, and Lenka Zdeborová. Spectral phase transitions in non-linear wigner spiked models. *arXiv preprint arXiv:2310.14055*, 2023.

[GKZ19] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv preprint arXiv:1912.01599*, 2019.

[GLK+20] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

[GLR+21] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. *Proceedings of Machine Learning Research vol*, 145:1–46, 2021.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019.

[GMMM20] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

[GMMM21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[GSd+19] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

[Gui09] Alice Guionnet. *Large random matrices*, volume 1957. Springer Science & Business Media, 2009.

[GZR22] Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *J. Mach. Learn. Res*, 23:1–65, 2022.

[HJ22] Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

[HK21] Tomohiro Hayase and Ryo Karakida. The spectrum of Fisher information of deep networks achieving dynamical isometry. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342. PMLR, 2021.

[HL20] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[HN20] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[HXAP20] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 17116–17128. Curran Associates, Inc., 2020.

[HY19] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*, 2019.

[HY20] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International conference on machine learning*, pages 4542–4551. PMLR, 2020.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[JGH19] Arthur Jacot, Franck Gabriel, and Clément Hongler. The asymptotic spectrum of the hessian of dnn throughout training. *arXiv preprint arXiv:1910.02875*, 2019.

[Jia04] Tiefeng Jiang. The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā: The Indian Journal of Statistics*, pages 35–48, 2004.

[JNG+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[Joh90] C.R. Johnson. *Matrix Theory and Applications*. AMS Short Course Lecture Notes. American Mathematical Society, 1990.

[Joh01] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[JSF+20] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020.

[JSS+20] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.

[KAA19] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041, 2019.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KI20] Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study. In *International Conference on Artificial Neural Networks*, pages 168–179. Springer, 2020.

[KO20] Ryo Karakida and Kazuki Osawa. Understanding approximate fisher information for fast convergence of natural gradient descent in wide neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[KR19] Shiva Prasad Kasiviswanathan and Mark Rudelson. Restricted isometry property under high correlations. *arXiv preprint arXiv:1904.05510*, 2019.

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[KWLS21] Stefani Karp, Ezra Winston, Yuanzhi Li, and Aarti Singh. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. *Advances in Neural Information Processing Systems*, 34, 2021.

[KY17] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017.

[LBD+20] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[LBN+17] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[LBN+18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

[LC18a] Zhenyu Liao and Romain Couillet. The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, pages 3072–3081. PMLR, 2018.

[LC18b] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pages 3063–3071. PMLR, 2018.

[LC19] Zhenyu Liao and Romain Couillet. On inner-product kernels of high dimensional data. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 579–583. IEEE, 2019.

[LCKS91] Yann Le Cun, Ido Kanter, and Sara A Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18):2396, 1991.

[LCM20] 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 *34th Conference on Neural Information Processing Systems*, 2020.

[LD21] Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

[LGC+21a] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34, 2021.

[LGC+21b] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34, 2021.

[LHAR22] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Evolution of neural tangent kernels under benign and adversarial training. *Advances in Neural Information Processing Systems*, 35:11642–11657, 2022.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[LLS21] Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.

[LM20] Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*, 2020.

[LMZ20] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

[LNR22] Mufan Li, Mihai Nica, and Dan Roy. The neural covariance sde: Shaped infinite depth-and-width networks at initialization. *Advances in Neural Information Processing Systems*, 35:10795–10808, 2022.

[Lon21] Philip M Long. Properties of the after kernel. *arXiv preprint arXiv:2105.10585*, 2021.

[LR20] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[LRZ19] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv preprint arXiv:1908.10292*, 2019.

[LRZ20] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

[LV07] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*, volume 1. Springer, 2007.

[LWM19] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11674–11685, 2019.

[LXS+19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, pages 8570–8581, 2019.

[LY16] Zeng Li and Jianfeng Yao. Testing the sphericity of a covariance matrix when the dimension is much larger than the sample size. *Electronic Journal of Statistics*, 10(2):2973–3010, 2016.

[LY22] Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv preprint arXiv:2205.06308*, 2022.

[LZB22] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

[MBD+21] James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

[MHR+18] Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[MHWSE23] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[MLHD23] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

[MM19] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[MM21] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[MM22] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[MMM21] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

[MP67a] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[MP67b] V.A. Marčenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sb*, 1:457–483, 01 1967.

[MPM21] Charles H Martin, Tongsu Serena Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):1–13, 2021.

[MS22] Andrea Montanari and Basil Saeed. Universality of empirical risk minimization. *arXiv preprint arXiv:2202.08832*, 2022.

[MY23] Xuran Meng and Jianfeng Yao. Impact of classification difficulty on the weight matrices spectra in deep learning and application to early-stopping. *Journal of Machine Learning Research*, 24(28):1–40, 2023.

[MZ20] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826v1*, 2020.

[Nak20] Preetum Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. *OPT2020 Workshop*, 2020.

[Nea95] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1995.

[Ngu21] Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning*, pages 8056–8062. PMLR, 2021.

[NM20] Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

[NMM21] Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR, 2021.

[NS06] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

[OFLS19] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

[OJMDF21] Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. What can linearized neural networks actually say about generalization? *Advances in Neural Information Processing Systems*, 34, 2021.

[OJMMF20] Guillermo Ortiz-Jiménez, Apostolos Modas, Seyed-Mohsen Moosavi, and Pascal Frossard. Neural anisotropy directions. *Advances in Neural Information Processing Systems*, 33:17896–17906, 2020.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

[Pau07] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[PB17] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806, 2017.

[PC22] Bartlomiej Polaczyk and Jacek Cyranka. Improved overparametrization bounds for global convergence of sgd for shallow neural networks. *Transactions on Machine Learning Research*, 2022.

[Péc06] Sandrine Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134:127–173, 2006.

[Péc19] S Péché. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

[PLR$^+$16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, pages 3360–3368, 2016.

[PSG17] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

[PSG18] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1924–1932. PMLR, 2018.

[PW17] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[PW18] Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, pages 5410–5419, 2018.

[PZA+21] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

[QLY23] Jiaxin Qiu, Zeng Li, and Jianfeng Yao. Asymptotic normality for eigenvalue statistics of a general sample covariance matrix when $p/n \to \infty$ and applications. *The Annals of Statistics*, 51(3):1427–1451, 2023.

[RGKZ21] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pages 8936–8947. PMLR, 2021.

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1177–1184, 2007.

[RR08] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[RR17] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[SB95] Jack W Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

[SB21] Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training. *arXiv preprint arXiv:2105.14301*, 2021.

[SC95] Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

[SCDL23] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. *arXiv preprint arXiv:2302.00401*, 2023.

[SEG+17] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *International Conference on Learning Representations*, 2017.

[SGGSD17] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

[Sil85] Jack W Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability*, pages 1364–1368, 1985.

[Sil95] Jack W Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

[SV13] Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013.

[SY19] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

[Tao12] Terence Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Society, 2012.

[TAP21] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[TSR22] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5):054124, 2022.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[Voi87] Dan Voiculescu. Multiplication of certain non-commuting random variables. *Journal of Operator Theory*, pages 223–235, 1987.

[VW15] Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms*, 47(4):792–821, 2015.

[WES+23] Zhichao Wang, Andrew William Engel, Anand Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[WHS22] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, pages 23549–23588. PMLR, 2022.

[Wig55] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

[Wil97] Christopher KI Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, pages 295–301, 1997.

[Wis28] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

[WLLM19] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2019.

[WP14] Lili Wang and Debashis Paul. Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$. *Journal of Multivariate Analysis*, 126:25–52, 2014.

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

[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*, volume 206 of *Proceedings of Machine Learning Research*, pages 8463–8493. PMLR, 25–27 Apr 2023.

[WZ24] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability*, 34(2):1896–1947, 2024.

[XBSD+18] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402. PMLR, 2018.

[XHM+22] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

[Xie13] Junshan Xie. Limiting spectral distribution of normalized sample covariance matrices with $p/n \to 0$. *Statistics & Probability Letters*, 83:543–550, 2013.

[XPS19] Lechao Xiao, Jeffrey Pennington, and Samuel S Schoenholz. Disentangling trainability and generalization in deep learning. *arXiv preprint arXiv:1912.13053*, 2019.

[Yan19] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[Yas16] Pavel Yaskov. Controlling the least eigenvalue of a random Gram matrix. *Linear Algebra and its Applications*, 504:108–123, 2016.

[YBM21] Zitong Yang, Yu Bai, and Song Mei. Exact gap between generalization error and uniform convergence in random feature models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11704–11715. PMLR, 18–24 Jul 2021.

[YH20] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[YS19a] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

[YS19b] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[YTH+22] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E Gonzalez, Kannan Ramchandran, Charles H Martin, and Michael W Mahoney. Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data. *arXiv preprint arXiv:2202.02842*, 2022.

[YXZ22] Long Yu, Jiahui Xie, and Wang Zhou. Testing Kronecker product covariance matrices for high-dimensional matrix-variate data. *Biometrika*, 11 2022. asac063.

[YZB15] Jianfeng Yao, Shurong Zheng, and Zhidong Bai. Large sample covariance matrices and high-dimensional data analysis. *Cambridge UP, New York*, 2015.

[ZNB22] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022.