UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Random Matrix Theory for Linearized Neural Networks

Permalink

https://escholarship.org/uc/item/7z1089cf

ISBN

9798186182559

Author

Cranston, Collin

Publication Date

2026-08-04

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Random Matrix Theory for Linearized Neural Networks

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

in

Mathematics, with Specialization in Statistics

by

Collin Cranston

Committee in charge:

Professor Todd Kemp, Chair
Professor Mikhail Belkin
Professor Alexander Cloninger
Professor Ioana Dumitriu

2026

Copyright

Collin Cranston, 2026

All rights reserved.

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

University of California San Diego

2026

DEDICATION

To my parents and my brother:
To my wonderful girlfriend, Katie;
For all their love and support;

EPIGRAPH

I am indebted to my father for living, but to my teacher for living well.

Alexander the Great

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables x

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xiv

Chapter 1 Introduction 1
1.1 Neural Tangent Model 3
1.1.1 Motivation and connection to Neural Networks 3
1.1.2 Kernel Matrices 7
1.2 Random Matrix Theory (RMT) 9
1.2.1 Empirical Spectral Distributions and Stieltjes Transform 9
1.2.2 Deterministic Equivalents 11
1.2.3 Marchenko-Pastur Law 12
1.2.4 Hermite Orthogonal Polynomials 13
1.3 Partial and Stochastic Differential Equations 15
1.3.1 Itô Calculus 15
1.3.2 Hamilton-Jacobi and Method of Characteristics 18
1.4 Contributions 20
1.5 Notation 22

Chapter 2 NTK Regression under dual Power-law model 23
2.1 Introduction 23
2.1.1 Related Works 25
2.1.2 Contributions 26
2.2 Main results 27
2.2.1 Problem setup 27
2.2.2 Gaussian Equivalence 29
2.2.3 Generalized Cross Validation (GCV) 30
2.2.4 Dual power-law model 32
2.2.5 Assumptions 33

2.2.6 Outline of approach . . . 34
2.2.7 Statement of main results . . . 36
2.2.8 Finite-width NTK scaling law under unnormalized power-law spectra . . 45
2.3 Proofs of Main Results . . . 49
2.3.0 Preliminaries . . . 49
2.3.1 Application: Anisotropic Sample Covariance . . . 51
2.3.2 Finite Dimension SDE . . . 60
2.3.3 Asymptotic PDE . . . 73
2.3.4 Hamilton-Jacobi System . . . 75
2.3.5 Proof of Theorem 2 . . . 81
2.4 Numerical Experiments . . . 97
2.4.1 Numerical root solver . . . 97

Chapter 3 Eigen-spike emergence for Nonlinear Data . . . 103
3.1 Introduction . . . 103
3.1.1 Related works . . . 105
3.1.2 Contributions . . . 107
3.1.3 Approach . . . 108
3.2 Main Results . . . 110
3.2.1 Notations and Preliminaries . . . 110
3.2.2 Statement of Main Results . . . 113
3.2.3 Proportional Limit with Finite SNR Case . . . 114
3.2.4 Large-SNR Proportional Limit Case . . . 119
3.2.5 Pretrained / Spiked Weight Matrix Case . . . 120
3.2.6 Quadratic Sample-size Case . . . 122
3.3 Proofs of Main Results . . . 124
3.3.1 Preliminaries . . . 124
3.3.2 Properties of XOR Data . . . 125
3.3.3 Properties of CK Matrix . . . 132
3.3.4 The z-transform and the T-transform . . . 140
3.3.5 Centered and Normalized GELU . . . 146
3.3.6 Quadratic Equivalence for Proportional Limit . . . 150
3.3.7 Proof Strategy for Main Results . . . 153
3.3.8 Proofs for Finite-SNR and Proportional Limit Regime . . . 155
3.3.9 Analysis of the Population Covariance Matrix . . . 156
3.3.10 Analysis of Spike in Population Covariance . . . 157
3.3.11 BBP Transition for Null Model . . . 160
3.3.12 BBP Transition for Rank-three Additive Deformation . . . 161
3.3.13 Proof of Theorem 5 . . . 195
3.3.14 Proof of Kernel Spectral Clustering . . . 196
3.3.15 Proof for Large-SNR Regime . . . 198
3.3.16 Proof for Spiked Weight Regime . . . 210
3.3.17 Step 1: A QE for spiked weights. . . . 212
3.3.18 Proof of Theorem 10 . . . 216

3.3.19 Proof of Quadratic Sample-size Regime 219
3.3.20 Analysis of the Population CK for Quadratic Sample-size Regime 220
3.3.21 Deterministic Equivalent of Spiked Kernel Model 227
3.3.22 BBP Transition for Quadratic Regime 234
3.4 Numerical Experiments 234
3.4.1 Baseline Case in Section 3.2.3 235
3.4.2 Knob I: Increase SNR 236
3.4.3 Knob II: Activation nonlinearity σ 237
3.4.4 Knob III: Weight structure 239

Bibliography 246

LIST OF FIGURES

Figure 2.1. Finite-width NTK Scaling-law 99

Figure 2.2. Excess risk of NTK regression at optimally-tuned ridge parameter 99

Figure 2.3. Optimally-tuned excess NTK risk against sample size risk under various regimes 100

Figure 2.4. Excess NTK risk under isotropic vs power-law data 100

Figure 2.5. Theory vs empirical for trace resolvents where *capacity* $\alpha < 1$ 101

Figure 2.6. Theory vs empirical for trace resolvents where *capacity* $\alpha > 1$ 101

Figure 2.7. Theory vs empirical for signal resolvent 101

Figure 3.1. Finite-SNR proportional regime CK Spectra: Failure case 235

Figure 3.2. Large-SNR proportional regime CK Spectra: Informative spikes 236

Figure 3.3. CK Spectra SNR Phase Transition 238

Figure 3.4. CK Spectra under single-parameter family of activation function 239

Figure 3.5. CK Spectra under Large-SNR regime with odd activation function 240

Figure 3.6. CK Spectra under spiked weights: informative task-aligned spike 241

Figure 3.7. CIFAR-2 Pretrained Model CK Spectra 242

Figure 3.8. Pretrained weights spectra phase transition over training 243

Figure 3.9. Pretrained weights eigenvector alignment phase transition over training . . 243

LIST OF TABLES

Table 3.1. Summary of phase transitions for CK outlier spikes 108

ACKNOWLEDGEMENTS

First and foremost I want to thank my PhD advisor Todd Kemp. I am deeply indebted to the support, guidance, and advice he has given me throughout my PhD studies. I was first introduced into the world of random matrices as an undergraduate attending one of his talks at UCLA. Since then he has helped shaped my research journey through his mathematical expertise and heartfelt words of encouragement. I mark myself incredibly fortunate to have an advisor who was so patient and genuinely invested in my growth as a mathematician. Second, I want to thank my collaborator and peer Zhichao Wang, who was, in all practical purposes, a co-advisor in my research journey. I extend my deepest gratitude to him for sharing his expertise in random matrix theory and deep learning theory through countless hours of meetings and discussions. He has left a profound impact on my research growth, and I am extremely privileged to have had the opportunity to work with him.

I also want to thank my committee members, Mikhail Belkin, Alexander Cloninger, and Ioana Dumitriu for their valued feedback and guidance throughout during my research journey. I extend particular thanks to Mikhail Belkin for his talks and discussions, which led me to the path of deep learning theory. I thank David Jekel for his mentorship during my time as an undergraduate at UCLA, and for introducing me to Free Probability. I also thank Dimitri Shlyakhtenko for directing my research interest into research in Free Probability and Random Matrix Theory, and for his guidance in my pursuit of graduate studies in mathematics.

To my parents for their everlasting love, support, and constant belief. To my brother, for our coffee runs and tennis matches which provided much needed balance to my life as a graduate student. To my girlfriend Katie Gulick, who has provided immeasurable amounts of support and affirmation from the beginning of my PhD. The completion of this thesis would not have been possible without their encouragement.

Chapter 2 is extracted from the paper “Collin Cranston, Zhichao Wang, Todd Kemp. NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods.” Currently under review. The dissertation author is the co-author of this paper.

Chapter 3 is extracted from the paper “Collin Cranston, Zhichao Wang, Todd Kemp, Michael W. Mahoney. Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data, Proceedings of the 39th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research (PMLR)*. arXiv:2605.29669” To appear. The dissertation author is co-author of this paper.

VITA

2019	Bachelor of Science, University of California, Los Angeles
2020–2026	Teaching Assistant, Department of Mathematics University of California San Diego
2023	Masters of Science, University of California San Diego
2023	Teaching Assistant, MSRI Machine Learning Summer School
2026	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

"Collin Cranston, Zhichao Wang, Todd Kemp. NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods." Currently under review.

"Collin Cranston, Zhichao Wang, Todd Kemp, Michael W. Mahoney. Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data, Proceedings of the 39th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research (PMLR)*. arXiv:2605.29669" To appear.

ABSTRACT OF THE DISSERTATION

Random Matrix Theory for Linearized Neural Networks

by

Collin Cranston

Doctor of Philosophy in Mathematics, with Specialization in Statistics

University of California San Diego, 2026

Professor Todd Kemp, Chair

Non-linear Random Matrix Theory (RMT) has recently emerged as a powerful paradigm for the theoretical understanding of deep learning theory. Throughout recent works, a universality principle, the *Gaussian Equivalence Theorem* (GET), has become an indispensable tool allowing for the behavior of complex nonlinear neural networks to be understood through tractable linear kernel models. This thesis contributes to this emerging field, first by using the GET universality principle to derive a novel scaling law in Neural Tangent Kernel (NTK) regression, and second by studying the implications of this idealized linear equivalence on a high-dimensional nonlinearly separable dataset. The main contributions can be summarized as follows.

We study finite-width NTK regression under a dual-power law model, in which the

training data covariance spectrum exhibits a power-law decay and the labels are generated with deterministic signal whose *source*-exponent measures the difficulty of the learning task. We introduce a technique novel to RMT in deep learning theory, combining partial differential equations (PDE) and stochastic matrix calculus. This technique allows for the extension of previous results beyond the isotropic Gaussian setting and a deterministic equivalent for the generalization risk of the NTK to be derived under general data assumptions. Using this deterministic equivalence, we obtain a novel scaling-law exponent for the excess risk of finite-width NTK regression.

Motivated by the question of whether the idealized Gaussian equivalence principle remains meaningful beyond the isotropic Gaussian setting, we study the Conjugate Kernel (CK) on a non-linearly separable dataset, which induces spiked eigenvalues that isolate from the bulk. We develop a *quadratic equivalent* to the CK that enables precise analysis of emergent informative spikes, eigenvalues whose eigenvectors asymptotically align with task-relevant directions. Using this quadratic equivalent, we derive phase transitions governing the emergence of such informative spikes in various settings that are relevant to machine learning practice, helping translate the power of deterministic equivalence tools to practical relevance in machine learning.

Chapter 1

Introduction

Modern deep learning models have achieved increasingly impressive benchmarks in a wide variety of applications, such as computer vision, video generation, and large language models. Spurred by the success of deep learning methods, significant effort has been made to bridge the growing gap between theoretical understanding and empirical observations, some of which seemingly contradict conventional statistical wisdom [BHMM19]. Random Matrix Theory (RMT) and Free Probability have emerged as powerful frameworks that provide insights into high-dimensional data structures and statistical properties of deep learning models.

In the era of *big data*, where both the number of training samples and the size of ambient dimension grow at proportional rates, RMT provides a powerful theoretical paradigm for the characterization of both the spectral properties of datasets and the asymptotic alignment of their eigenvectors to target learning directions. In the context of machine learning, nonlinear RMT has been utilized to demonstrable success in the study of nonlinear kernel matrices that arise from the propagation of data through neural networks. Moreover, RMT naturally extends to the notion of *over-parameterization* in machine learning models, wherein the number of model parameters exceeds the size of training dataset. A precise theoretical understanding of the powerful generalization ability of highly over-parameterized models, a phenomena often referred to as benign overfitting [BLLT20, CCBG22], remains elusive and is an active area of research in deep learning theory.

The study of the spectrum of nonlinear kernel matrices originated in the pioneering work of [EK10], in which a careful entry-wise Taylor expansion of inner-product kernel matrices on high-dimensional data revealed that a broad class of kernel matrices essentially become linear in the proportional asymptotic regime. Subsequent work [FW20, MM19b, PW17] has extended upon this linearization phenomena, leading to a universality principle for a wide class of nonlinear feature models and their associated learning problems, which have been shown to be asymptotically indistinguishable from that of linear Gaussian covariate models, in both training and test behavior. Much of the modern technical analysis of neural network performance in the proportional-width regime relies on this equivalence, which has been dubbed the Gaussian Equivalence Theorem (GET) [GLR$^+$22, HL22b].

In Chapter 2, we utilize this universality principle to study a tractable model that approximates the training dynamics of an ultra-wide neural network, under a dataset that exhibits a power-law spectrum decay. Indeed several interesting phenomena exhibited in deep neural networks, such as *double-descent* risk curves [BHMM19] and *neural scaling laws* [KMH$^+$20], can be observed in the analytically tractable problem of high-dimensional ridge regression on a suitable kernel model. We also introduce a novel technique, in the context of RMT in deep learning, drawing upon tools from partial differential equations (PDE) and stochastic matrix calculus, which we expect to generalize to greater applicability in linearized neural network models.

In Chapter 3, we test the robustness of this Gaussian equivalence under the study of a random two-layer neural network on a canonical nonlinearly separable dataset, the XOR. We use the location of spike eigenvalues, ones that asymptotically separate from the bulk of the spectrum, and the alignment of their corresponding eigenvectors to binary XOR-labels as a proxy for the learnability of the classification task. The remainder of this chapter is dedicated to introducing the background of RMT in application to deep learning theory, and the PDE and stochastic differential equations (SDE) methods used to derive the main results of this thesis.

1.1 Neural Tangent Model

1.1.1 Motivation and connection to Neural Networks

In this section, we introduce and motivate the neural tangent model [JGH18, MM23] through its connection to the training and test dynamics of neural networks. The primary focus of Chapter 2 is to study an analytically tractable model that approximates a two-layer neural network $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ with hidden-layer width N, which we define as

$$f_\theta(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \boldsymbol{w}^\top \sigma(\boldsymbol{W}_1 \boldsymbol{x}), \qquad \theta := \left(\boldsymbol{w}_1^{(1)}, \ldots, \boldsymbol{w}_N^{(1)}, w_1, \ldots, w_N\right) \in \mathbb{R}^{N(p+1)} \tag{1.1.1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is the input data, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the nonlinear activation function applied entry-wise which we assume to be differentiable almost everywhere. The first layer weights $\boldsymbol{W}_1 \in \mathbb{R}^{N \times p}$ are initialized with entries i.i.d. $\mathcal{N}(0, \frac{1}{p})$, while the $\frac{1}{\sqrt{N}}$ factor in (1.1.1) is due to LeCun/NTK normalization [JGH18, LBOM12] and ensures the model output stays $O(1)$ as the width increases. The second layer weights $\boldsymbol{w} \in \mathbb{R}^N$ are initialized as $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_W^2)$, where the variance σ_W^2 is a hyper-parameter that will tune the behavior of the underlying kernel models derived below. Let θ_0 denote the parametrized weights at their random initialization. Under certain settings, which have been dubbed *lazy training* [COB19], it has been observed that the weights θ stay close to their initialization θ_0 throughout training, and the network output is well approximated by a first-order Taylor expansion around θ_0. As a first step in the analysis therefore, we perform a first order Taylor expansion of f_θ around θ_0:

$$f_\theta(\boldsymbol{x}) \approx f(\boldsymbol{x}; \theta_0) + \left(\theta - \theta_0\right)^\top \nabla_\theta f(\boldsymbol{x}; \theta_0). \tag{1.1.2}$$

We refer to [BMR21, COB19] for a rigorous treatment of this expansion, and instead focus on the implications when it remains valid. A quick calculation of gradients with respect to first and

second layer weights results in the following:

$$f_\theta(\boldsymbol{x}) \approx f(\boldsymbol{x};\theta_0) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \Delta w_i \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1,0)}\rangle\big) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i^{(0)} \sigma'\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1,0)}\rangle\big)\langle \boldsymbol{x}, \Delta \boldsymbol{w}_i^{(1)}\rangle \quad (1.1.3)$$

where $\Delta \boldsymbol{w} = \boldsymbol{w} - \boldsymbol{w}_0$, and $\boldsymbol{w}^{(1,0)}$ and $\boldsymbol{w}_i^{(0)}$ denote the first and second layer weights respectively at their initialization. For simplicity we set $f(\boldsymbol{x}, \theta_0) = 0$, which is referred to as centering by [AP20]. This can be achieved by either doubling the number of neurons in the network, or by subtracting the initial network value throughout training, and does not affect the analysis made in subsequent chapters. The expansion (1.1.3) motivates the definition of the following model classes.

Random Features (RF) Model

By freezing the first layer weights at initialization, i.e $\Delta \boldsymbol{w}^{(1)} = 0$, the only learnable parameters are the second layer weights $\boldsymbol{w}$. In this case, (1.1.3) reduces to the Random Feature (RF) model

$$f_w^{\mathrm{RF}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1)}\rangle\big) = \boldsymbol{w}^\top \phi_{\mathrm{RF}}(\boldsymbol{x}) \quad (1.1.4)$$

where we define the feature map $\phi_{\mathrm{RF}} : \mathbb{R}^p \to \mathbb{R}^N$ with entries given by

$$(\phi_{\mathrm{RF}})_i(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1)}\rangle\big), \qquad \text{for } i \in [N] \quad (1.1.5)$$

Learning the second layer weights $\boldsymbol{w} \in \mathbb{R}^N$ through ridge regression yields Random Feature Ridge Regression (RFRR), which was originally introduced in [RR07] as a random approximation to kernel methods. As a statistical model at fixed width N, RFRR has garnered extensive independent research interest, in part due to its connection to neural networks. The asymptotic training error, and generalization properties of (1.1.4) have been, at this point, fully characterized in the isotropic dataset setting [MM19b, DL20, LLC18, ALP19], with extensions to more general datasets [WZ23, AZVP26, LGC$^+$21], and dimension-free deterministic equivalent

analysis as well [DLM24, SCDL23]. In particular, the RF model (1.1.4) exhibits both *double-descent* risk curves [MM19b], and scaling-law phenomena under power-law data [MRS22], and therefore marks an attractive analytically tractable model that approximates f_θ (1.1.1) (in the sense of (1.1.3)).

Neural Tangent Kernel (NTK) Model

By allowing the weights θ to change throughout training, but taking the linearization (1.1.3) as the model definition, we arrive at the Neural Tangent Kernel (NTK) model. Re-parameterizing by the weight displacements $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2) = \theta - \theta_0$, we define the NTK predictor as

$$f_\theta^{\mathrm{NTK}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} b_{1,i} \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1,0)} \rangle\big) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^{(0)} \sigma'\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1,0)} \rangle\big) \langle \boldsymbol{x}, \boldsymbol{b}_{2,i} \rangle \tag{1.1.6}$$

where $\boldsymbol{b}_1 \in \mathbb{R}^N$ and $\boldsymbol{b}_2 = \big(\boldsymbol{b}_{2,1}, \ldots, \boldsymbol{b}_{2,N}\big) \in \mathbb{R}^{Np}$ are learnable parameters. Here we define the NTK feature map $\phi_{\mathrm{NTK}} : \mathbb{R}^p \to \mathbb{R}^{Np}$ with rows given by

$$(\phi_{\mathrm{NTK}})_i(\boldsymbol{x}) = \frac{1}{\sqrt{Np}} \sigma'\big(\langle \boldsymbol{x}, \boldsymbol{w}_i^{(1,0)} \rangle\big) \boldsymbol{x}^\top, \qquad \text{for } i \in [N] \tag{1.1.7}$$

With a suitable re-definition of $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$, namely absorbing the second-layer weights into $\boldsymbol{b}_2$, the NTK predictor f_θ^{NTK} can be written in the following compact form:

$$f_\theta^{\mathrm{NTK}}(\boldsymbol{x}) = \boldsymbol{b}_1^\top \phi_{\mathrm{RF}}(\boldsymbol{x}) + \sqrt{p} \boldsymbol{b}_2^\top \phi_{\mathrm{NTK}}(\boldsymbol{x}) \tag{1.1.8}$$

The parameters $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^N \times \mathbb{R}^{Np}$ are learned through the NTK regression problem, with ridge parameter $\lambda > 0$

$$\widehat{\boldsymbol{b}}(\lambda) = \operatorname*{argmin}_{\boldsymbol{b}} \Big\{ \frac{1}{n} ||\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{b}||_2^2 + \lambda ||\boldsymbol{b}_1||_2^2 + \lambda ||\boldsymbol{b}_2||_2^2 \Big\}. \tag{1.1.9}$$

Here $\boldsymbol{y} \in \mathbb{R}^n$ are the training labels, and $\boldsymbol{\Phi} \in \mathbb{R}^{n \times N(p+1)}$ denotes the combined feature matrix

$$\boldsymbol{\Phi} := \begin{bmatrix} —\phi_{\mathrm{RF}}(\boldsymbol{x}_1)— & —\sqrt{p}\phi_{\mathrm{NTK}}(\boldsymbol{x}_1)— \\ —\phi_{\mathrm{RF}}(\boldsymbol{x}_2)— & —\sqrt{p}\phi_{\mathrm{NTK}}(\boldsymbol{x}_2)— \\ \vdots & \vdots \\ —\phi_{\mathrm{RF}}(\boldsymbol{x}_n)— & —\sqrt{p}\phi_{\mathrm{NTK}}(\boldsymbol{x}_n)— \end{bmatrix}. \tag{1.1.10}$$

The ridge solution (1.1.9) takes the following closed form

$$\widehat{\boldsymbol{b}}(\lambda) = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}. \tag{1.1.11}$$

With this, the NTK predictor at a test point $\boldsymbol{x} \in \mathbb{R}^p$ is given by

$$\widehat{f}_\lambda(\boldsymbol{x}) = \phi(x)^\top \widehat{\boldsymbol{b}}(\lambda) = \phi(x)^\top \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y} \tag{1.1.12}$$

where $\phi(\boldsymbol{x}) = \left(\phi_{\mathrm{RF}}(\boldsymbol{x}), \sqrt{p}\phi_{\mathrm{NTK}}(\boldsymbol{x})\right)$ is the combined feature vector.

Since its initial derivation as a bridge between wide neural networks trained in the lazy regime to kernel ridge regression (KRR) on a deterministic kernel [JGH18], the statistical properties of NTK regression and the spectral properties of its related kernel have been studied in various settings [FW20, NMM21, GDZ$^+$22, LHY25]. The generalization error of $\widehat{f}_\lambda$ was characterized in [AP20] in the isotropic setting, where non-monotonic risk curves of NTK regression were analyzed.

Kernel Ridge Regression (KRR) Model

In the infinite-width limit $N \to \infty$ the neural tangent model (1.1.12) simplifies to Kernel Ridge Regression (KRR) on a rotationally-invariant deterministic kernel. The specific kernel in this limit is often referred to as the NTK in literature, however throughout this thesis we refer to the RF and NTK models in their finite-width realizations, and instead refer to their deterministic limit kernels as the *infinite-width* NTK and RF models respectively.

The asymptotic performance of KRR has been studied in the proportional regime for general inner-product kernels [HMRT22, AKM$^+$17, MS24a], and the *infinite-width* NTK [AP20, WHS22]. Extensions past the proportional regime to polynomial-scaling regime [Mis22, HL22a], and beyond the isotropic setting to power-law data [WL25], have revealed several interesting phenomena unobserved in the isotropic proportional regime.

1.1.2 Kernel Matrices

In this section, we introduce the kernel matrices that are studied in this thesis. The two kernels of main interest, the Conjugate Kernel (CK) and the Neural Tangent Kernel (NTK), are motivated through the Neural Tangent model introduced in Section 1.1.1. We will let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ denote the training sample matrix.

Conjugate Kernel (CK)

The Conjugate Kernel (CK) is defined to be the empirical gram matrix of the hidden-layer activations in the network (1.1.1).

Definition 1 (Conjugate Kernel)**.** *We define the CK as*

$$\boldsymbol{K}^{\mathrm{CK}} = \frac{1}{N}\sigma(\boldsymbol{W}_1\boldsymbol{X})^\top \sigma(\boldsymbol{W}_1\boldsymbol{X}). \tag{1.1.13}$$

Furthermore, define the feedforward map $\boldsymbol{F} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}_1\boldsymbol{X})$*. Then the CK can be written as* $\boldsymbol{K}^{\mathrm{CK}} = \boldsymbol{F}^\top\boldsymbol{F}$*.*

The spectral distribution of the CK governs the training dynamics and generalization behavior of RFRR, both of which were computed in [MM19b] and [HL22b] for isotropic Gaussian input. The CK can be viewed as a scalable approximation to the infinite-width limit kernel, wherein regression on $\boldsymbol{K}^{\mathrm{CK}}$ approximates regression in the reproducing kernel Hilbert Space (RKHS) defined by limit kernel [RR07].

The spectrum of the CK has been studied in [ALP19, FW20, LC18, LLC18], where an

equivalence on the order of bulk statistics has been established between the CK and a linear Gaussian covariates model under suitable smoothness assumptions. We will explore both the consequences and failures of this equivalence in both Chapters 2 and 3.

Neural Tangent Kernel (NTK)

The Neural Tangent Kernel (NTK) is defined as the gram matrix of the Jacobian of the network (1.1.1) with respect to the parameter-weights θ.

Definition 2 (Neural Tangent Kernel)**.** *We define the NTK to have entries given by*

$$\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x},\boldsymbol{x}') := \Big(\nabla_\theta f_\theta(\boldsymbol{x})\Big)^\top \Big(\nabla_\theta f_\theta(\boldsymbol{x}')\Big). \tag{1.1.14}$$

Letting $\boldsymbol{J} = \nabla_\theta f_\theta(\boldsymbol{X})$, the NTK can be written as $\boldsymbol{K}^{\mathrm{NTK}} = \boldsymbol{J}^\top \boldsymbol{J}$.

Partitioning the gradient over first and second layer weights, we have

$$\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x},\boldsymbol{x}') = \underbrace{\nabla_{\boldsymbol{w}} f_\theta(\boldsymbol{x})^\top \nabla_{\boldsymbol{w}} f_\theta(\boldsymbol{x}')}_{\boldsymbol{K}_1(\boldsymbol{x},\boldsymbol{x}')} + \underbrace{\nabla_{\boldsymbol{W}_1} f_\theta(\boldsymbol{x})^\top \nabla_{\boldsymbol{W}_1} f_\theta(\boldsymbol{x}')}_{\boldsymbol{K}_2(\boldsymbol{x},\boldsymbol{x}')}. \tag{1.1.15}$$

The first and second layer kernels can be written as

$$\boldsymbol{K}_1(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{F}^\top \boldsymbol{F} = \boldsymbol{K}^{\mathrm{CK}}, \qquad \boldsymbol{K}_2(\boldsymbol{X},\boldsymbol{X}) = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \frac{(\boldsymbol{F}')^\top \mathrm{diag}(\boldsymbol{w})^2 \boldsymbol{F}'}{N} \tag{1.1.16}$$

where $\odot$ denotes the Hadamard (entry-wise) product. Here $\boldsymbol{F} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}_1\boldsymbol{X})$, $\boldsymbol{F}' = \frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}_1\boldsymbol{X})$ are the feed-forward feature maps of σ and σ' respectively. We recognize the first layer kernel $\boldsymbol{K}_1$ as the CK.

The NTK was introduced in [JGH18] to study the training dynamics of ultra-wide neural networks. Its origins follows a long line of works connecting infinite-width networks to kernel methods, originating with the observation that single-layer infinite-width networks converge to Gaussian processes [Nea95, Wil96]. This observation has since been extended to

deep infinite-width neural networks via the Neural Network Gaussian Process (NNGP) kernel [MHR$^+$18, LBN$^+$18]. Moving beyond random initialization, the NTK governs the optimization dynamics of ultra-wide networks trained under gradient descent. Indeed, under LeCun/NTK normalization the weights of an infinite-width network stay infinitesimally close to their random initialization throughout training. Specifically, under gradient-flow training dynamics with an ℓ_2-squared loss, i.e gradient descent with infinitesimal step size, the model $f_{\theta(t)}$ evolves as a linear-model in the tangent feature space, and satisfies the following ordinary differential equation (ODE) [DZPS19, LXS$^+$20]

$$\frac{\partial}{\partial t} f_{\theta(t)}(\boldsymbol{X}) = -\eta \boldsymbol{K}^{\mathrm{NTK}}(t) \cdot \big(f_{\theta(t)}(\boldsymbol{X}) - \boldsymbol{y}\big). \tag{1.1.17}$$

Here $\theta(t), \boldsymbol{K}^{\mathrm{NTK}}(t)$ denote the parameter weights, and the NTK at gradient-descent step t, and $\eta > 0$ denotes the learning rate parameter of gradient descent.

1.2 Random Matrix Theory (RMT)

In this section, we briefly introduce the background of the main tools from Random Matrix Theory (RMT) which are of interest to deep learning theory, and in particular the results presented in this thesis. This section is not a comprehensive review and we refer the reader to [CL22] for more background on RMT in machine learning.

1.2.1 Empirical Spectral Distributions and Stieltjes Transform

A primary object of interest in RMT is the distribution of the eigenvalues of symmetric random matrices in high dimensions, and the weak convergence of their empirical spectral distribution (ESD). Throughout this section, let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ denote a symmetric random matrix with eigenvalues $\sigma(\boldsymbol{M}) = \{\lambda_1(\boldsymbol{M}) \leq \ldots \leq \lambda_n(\boldsymbol{M})\}$. The ESD of $\boldsymbol{M}$ is given by

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{M})}. \tag{1.2.1}$$

Of interest is the weak limit of (1.2.1), which we refer to as the limiting spectral measure of $\boldsymbol{M}$ and denote it by

$$\lim \operatorname{spec} \boldsymbol{M} = \mu. \tag{1.2.2}$$

This can be understood as the almost sure convergence of the following

$$\frac{1}{n}\sum_{i=1}^{n} f\big(\lambda_i(\boldsymbol{M})\big) \to \mathbb{E}_{\sigma\sim\mu}[f(\sigma)] \tag{1.2.3}$$

as $n \to \infty$, for every bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$. This is also referred to as convergence of the bulk of the spectrum of $\boldsymbol{M}$, and does not guarantee the convergence of the outlier eigenvalues of $\boldsymbol{M}$ to the support of μ.

A classical tool to characterize the limiting spectral measure is the Stieljtes transform. For a probability distribution μ on $\mathbb{R}$ with support $\operatorname{supp}(\mu)$, define the Stieltjes transform of μ for $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$

$$m_\mu(z) = \int \frac{1}{x-z}\mu(dx). \tag{1.2.4}$$

Observe that the ESD μ_n is also a probability measure, which motivates the empirical Stieltjes transform

$$m_{\mu_n}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i(\boldsymbol{M}) - z} \tag{1.2.5}$$

We also introduce the resolvent, which is another foundational object in RMT used to characterize the ESD of symmetric random matrices. Define the resolvent R of $\boldsymbol{M}$ as

$$R(z) = \big(\boldsymbol{M} - z\mathbf{I}_n\big)^{-1}, \quad z \in \mathbb{C} \setminus \sigma(\boldsymbol{M}) \tag{1.2.6}$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The ESD, resolvent, and limiting spectrum of $\boldsymbol{M}$ are related through the following observation.

$$m_{\mu_n}(z) = \frac{1}{n}\sum_{i=1}^{n} \int \frac{1}{x-z}\delta_{\lambda_i(\boldsymbol{M})}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i(\boldsymbol{M}) - z} = \frac{1}{n}\operatorname{Tr}\big(\boldsymbol{M} - z\mathbf{I}_n\big)^{-1}. \tag{1.2.7}$$

The pointwise almost sure convergence of $m_{\mu_n} \to m_\mu$ over the upper-half complex plane $\mathscr{H}^+$ is equivalent to the weak convergence of μ_n to μ, i.e that $\lim \operatorname{spec} \boldsymbol{M} = \mu$ (Theorem 8.11 in [Kem11]). The Stilejtes transform thereby uniquely encodes the weak convergence of μ_n to μ. Moreover, the Stieltjes-inversion formula allows one to recover the density g of the limiting measure μ, should it exist

$$g(x) = \frac{1}{\pi} \lim_{y \to 0^+} \operatorname{Im}\Big[m_\mu(x+iy)\Big]. \tag{1.2.8}$$

1.2.2 Deterministic Equivalents

While the classical resolvent methods introduced in Section 1.2.1 allow for the study of asymptotic eigenvalue behavior, they do not directly provide information about their eigenvectors. Indeed, by characterizing only the normalized trace of the resolvent in (1.2.7), one loses information of $\boldsymbol{M}$ encoded in its resolvent. Here we introduce the modern approach of characterizing the asymptotic spectral behavior of $\boldsymbol{M}$ through *deterministic equivalents*, which are finite dimensional deterministic matrices that approximate $\boldsymbol{M}$ asymptotically (in the sense of scalar functionals). In particular, we say $\overline{\boldsymbol{M}} \in \mathbb{R}^{n \times n}$ is a deterministic equivalent of $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, and write $\boldsymbol{M} \asymp \overline{\boldsymbol{M}}$, if the following holds uniformly over a class of deterministic test matrices $\boldsymbol{A}$ of bounded operator norm and unit-norm vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$

$$\frac{1}{n} \operatorname{tr}\Big(\boldsymbol{A}(\boldsymbol{M} - \overline{\boldsymbol{M}})\Big) \to 0, \qquad \boldsymbol{a}^\top \big(\boldsymbol{M} - \overline{\boldsymbol{M}}\big)\boldsymbol{b} \to 0 \tag{1.2.9}$$

almost surely, as $n \to \infty$. We note that (1.2.9) is just one way to define the notion of deterministic equivalence, which does not admit a standard definition in the literature. One advantage of (1.2.9) is it naturally encodes traces and bilinear forms of $\boldsymbol{M}$, which are essential to studying its asymptotic spectral distribution and eigenvectors. The notion of deterministic equivalence originated in [Gir01], and will be of particular use for deriving scaling-laws in Chapter 2.

1.2.3 Marchenko-Pastur Law

As an example of the resolvent method and deterministic equivalence, we consider the foundational Marchenko-Pastur law [MP67], which characterizes the spectra of sample covariance matrices. Let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ have i.i.d $\mathcal{N}(0,1)$ entries, and let $R(z) = \left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\mathbf{I}_p\right)^{-1}$ denote the resolvent of the sample covariance $\boldsymbol{S}_p = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{p\times p}$. Then, as $n, p \to \infty$ at rate $\frac{p}{n} \to \gamma \in (0,\infty)$

$$R(z) \asymp \overline{R}(z), \qquad \overline{R}(z) = m(z)\mathbf{I}_p \tag{1.2.10}$$

where $m(z)$ is given by the unique analytic solution in $\mathscr{H}^+$ to the Marchenko-Pastur equation

$$z\gamma m(z)^2 - (1-\gamma-z)m(z) + 1 = 0. \tag{1.2.11}$$

Here $m(z)$ is the Stieltjes transform of the Marchenko-Pastur distribution $\rho_\gamma^{\mathrm{MP}}$, which has explicit density

$$d\rho_\gamma^{\mathrm{MP}}(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{\gamma x}\cdot \mathbf{1}_{x\in[\lambda_-,\lambda_+]}dx, \qquad \lambda_\pm = (1\pm\sqrt{\gamma})^2 \tag{1.2.12}$$

when $\gamma \le 1$, with an additional point mass $(1-\gamma^{-1})$ at 0 when $\gamma > 1$. In particular, the ESD $\mu_{\boldsymbol{S}_p}$ of the sample covariance $\boldsymbol{S}_p$ converges weakly to $\rho_\gamma^{\mathrm{MP}}$ with probability one.

A line of works culminating in [SB95] extends this beyond i.i.d Gaussian entries to general sample covariance matrices of $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} \in \mathbb{R}^{p\times n}$, where $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is a deterministic positive semi-definite matrix with limiting spectral measure $\lim \operatorname{spec}\boldsymbol{\Sigma} = \nu$ as $n \to \infty$. Let $\widetilde{R}(z) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - z\mathbf{I}_n\right)^{-1}$ denote the resolvent of the $n\times n$ sample covariance $\boldsymbol{S}_n = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$; then

$$\begin{aligned} R(z) \asymp \overline{R}(z) &= -\frac{1}{z}(\widetilde{m}_p(z)\boldsymbol{\Sigma} + \mathbf{I}_p)^{-1} \\ \widetilde{R}(z) \asymp \overline{\widetilde{R}}(z) &= \widetilde{m}_p(z) \end{aligned} \tag{1.2.13}$$

where $\widetilde{m}_p(z)$ is the unique solution in $\mathscr{H}^+$ to the general Marchenko-Pastur equation

$$\widetilde{m}_p(z) = \left(-z + \frac{1}{n} \operatorname{tr} \mathbf{\Sigma} \left(\mathbf{I}_p + \widetilde{m}_p(z) \mathbf{\Sigma} \right)^{-1} \right)^{-1}. \tag{1.2.14}$$

In particular, the empirical spectra measures $\mu_{\boldsymbol{S}_p}$ and $\mu_{\boldsymbol{S}_n}$ of $\boldsymbol{S}_p$ and $\boldsymbol{S}_n$ respectively converge weakly to μ and $\tilde{\mu}$, which are characterized by their Stieltjes transforms $m(z)$ and $\widetilde{m}(z)$.

$$m(z) = \frac{1}{\gamma} \widetilde{m}(z) + \frac{1-\gamma}{\gamma z}, \qquad \widetilde{m}(z) = \left(-z + \gamma \int \frac{\lambda d\nu(\lambda)}{1 + \lambda \widetilde{m}(z)} \right)^{-1} \tag{1.2.15}$$

The limiting spectral measures $\mu, \tilde{\mu}$ can also be characterized through free multiplicative convolution $\boxtimes$, which was introduced in [Voi87] and allows for the asymptotic study of products of free random matrices,

$$\mu = \rho_\gamma^{\mathrm{MP}} \boxtimes \nu, \qquad \tilde{\mu} = \gamma \mu + (1-\gamma) \delta_0. \tag{1.2.16}$$

In particular, the Stieltjes-transform $m(z)$ satisfies the following self-consistent Marchenko-Pastur equation.

$$m(z) = \int \frac{1}{\lambda \left(1 - \gamma - \gamma m(z) \right) - z} d\nu(\lambda). \tag{1.2.17}$$

The main tool for analyzing free convolution is the S-transform [Voi87], which has recently been used to study the asymptotics of high-dimensional regression in [AZVP26]. We refer the reader to the book [MS17] for more background on Free Probability and the S-transform.

1.2.4 Hermite Orthogonal Polynomials

We next introduce Hermite orthogonal polynomials, which are an essential tool in the asymptotic analysis of nonlinear kernel matrices. The defining property of Hermite polynomials is that they form an orthogonal basis of $L^2(\mathbb{R}, \Gamma)$, where Γ is the standard Gaussian probability distribution.

Definition 3 (Normalized Hermite Polynomials)**.** *The $r-$th normalized Hermite polynomial is*

given by

$$h_r(x) = \frac{1}{\sqrt{r!}}(-1)^r e^{x^2/2}\frac{d^r}{dx^r}e^{-x^2/2}. \tag{1.2.18}$$

The $\{h_r\}_{r=0}^{\infty}$ *form an orthonormal basis of* $L^2(\mathbb{R},\Gamma)$*, where* Γ *denotes the standard Gaussian distribution. Consequently, 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*

$$\zeta_r(\sigma) = \int_{-\infty}^{\infty} \sigma(x) h_r(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

Hermite expansions of the entry-wise nonlinear functions in kernel matrices have been widely used to characterize both the spectra of these kernels [FW20, WZ21, WWF24], and also the performance of kernel and random feature models [MM19b, PW17, WZ23, NM20]. As an illustrative example, consider the infinite-width limit of the CK (1.1.13)

$$\boldsymbol{K}_{\infty}^{\mathrm{CK}} = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^{\top}\boldsymbol{X})^{\top}\sigma(\boldsymbol{w}^{\top}\boldsymbol{X})], \qquad \boldsymbol{w} \sim \mathcal{N}(0, \mathbf{I}_d). \tag{1.2.19}$$

Assuming σ is Lipschitz, and centered and normalized to unit second moment with respect to the standard Gaussian, a Hermite expansion of σ leads to the following:

$$\left(\boldsymbol{K}_{\infty}^{\mathrm{CK}}\right)_{\alpha\beta} = \sum_{k=0}^{\infty} \zeta_k(\sigma_\alpha)\zeta_k(\sigma_\beta)\left(\frac{\boldsymbol{x}_\alpha^{\top}\boldsymbol{x}_\beta}{||\boldsymbol{x}_\alpha||||\boldsymbol{x}_\beta||}\right)^k \tag{1.2.20}$$

where $\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \in \mathbb{R}^p$ are columns of $\boldsymbol{X}$ and ζ_k are coefficients defined by

$$\zeta_k(\sigma_\alpha) = \mathbb{E}[\sigma(\xi \cdot ||\boldsymbol{x}_\alpha||)h_k(\xi)], \quad \zeta_k(\sigma_\beta) = \mathbb{E}[\sigma(\xi \cdot ||\boldsymbol{x}_\beta||)h_k(\xi)] \tag{1.2.21}$$

for $\alpha, \beta \in [n]$. Further assumptions on the approximate orthonormality of the columns of $\boldsymbol{X}$ lead

to control of the higher-order terms in the Hermite expansion (1.2.20) [WZ21, WWF24].

1.3 Partial and Stochastic Differential Equations

We introduce tools from partial differential equations (PDE) and stochastic differential equations (SDE) that underpin the derivations made in Chapter 2. Section 1.3.1 first provides the stochastic matrix calculus machinery required to derive a PDE that models the asymptotics of NTK regression. The necessary PDE tools, introduced in Section 1.3.2, consist of Method of Characteristics applied to first order nonlinear PDEs of Hamilton-Jacobi type.

1.3.1 Itô Calculus

A central technical tool in Chapter 2 is stochastic matrix calculus, which we use to derive a SDE governing the log-determinant potential of the NTK (1.1.14) with Brownian motion embedded entries. We provide a very brief background on Itô-calculus in this section, and refer the reader to [Øks98] for a more thorough treatment.

Throughout, fix a filtered probability space $\left(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P}\right)$ and let $\left(B_t\right)_{t\geq 0}$ denote standard adapted Brownian motion. We begin by defining an Itôprocess, also known as a stochastic integral.

Definition 4 (Itô-process)**.** *An Itô process is a real-valued stochastic process* $\left(X_t\right)_{t\geq 0}$ *adapted to* $(\mathscr{F}_t)$*, written as*

$$X_t = X_0 + \int_0^t U_s ds + \int_0^t V_s dB_s \tag{1.3.1}$$

where U, V *are progressively measurable processes satisfying*

$$\int_0^t |U_s| ds < \infty, \quad \int_0^t V_s^2 ds < \infty \qquad a.s \tag{1.3.2}$$

for every $t \geq 0$*. We adopt the stochastic differential notation and write* (1.3.1) *as*

$$dX_t = U_t dt + V_t dB_t. \tag{1.3.3}$$

U, V in (1.3.1) are referred to as the drift and diffusion coefficients, respectively. We next define the quadratic variation process, which measures the path oscillation of a stochastic process and is useful in motivating Itô's Formula.

Definition 5 (Quadratic Variation)**.** *Let* $\left(X_t\right)_{t\geq 0}$ *be a stochastic process, and let* $\Pi = \{0 = t_0 < t_1 < \cdots < t_{m(\pi)} = t\}$ *be a partition of* $[0,t]$*. The* quadratic variation *of* X *is the process given by*

$$[X]_t := \lim_{\text{mesh}(\pi)\to 0} \sum_i \Big(X_{t_{i+1}} - X_{t_i}\Big)^2 \tag{1.3.4}$$

for all $t \geq 0$*, whenever this limit* (1.3.4) *holds in probability. Moreover, if* X, Y *are two stochastic processes on the same probability space, define the* quadratic covariation process $\langle X, Y\rangle = \Big(\langle X, Y\rangle_t\Big)_{t\geq 0}$ *by*

$$\begin{aligned} \langle X, Y\rangle &:= \lim_{\text{mesh}(\pi)\to 0} \sum_i \big(X_{t_{i+1}} - X_{t_i}\big)\big(Y_{t_{i+1}} - Y_{t_i}\big) \\ &= \frac{1}{4}\Big(\big[X+Y\big] - \big[X-Y\big]\Big) = \frac{1}{2}\Big(\big[X+Y\big] - [X] - [Y]\Big) \end{aligned} \tag{1.3.5}$$

provided the quadratic variation processes on the right of (1.3.5) *exist, in the sense of the limit* (1.3.4).

The Itô integral $\int_0^t V_s dB_s$ in (1.3.1) differs from a classical Riemann-Stieltjes integral in that the Brownian motion integrator B_t has non-trivial quadratic variation. Indeed, an alternate characterization of standard Brownian motion is a local martingale B_t whose quadratic variation process is $[B]_t = t$. This fact is responsible for the second-order correction in Itô's formula, which is the foundational technical tool in Itô Calculus and stated below.

Proposition 6 (Itô's Formula)**.** *Let* X_t *be an Itô-process, i.e* $dX_t = U_t dt + V_t dB_t$*, and let* $f \in C^2(\mathbb{R})$ *be a twice continuously differentiable function. Then* $Y_t = f(X_t)$ *is also an Itô-process satisfying the following*

$$dY_t = \frac{\partial f}{\partial x}(X_t)dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(X_t)(dX_t)\cdot(dX_t) \tag{1.3.6}$$

where $(dX_t) \cdot (dX_t) = d[X]_t$ *is computed using the following rules*

$$dt \cdot dt = dt \cdot dB_t = dB_t \cdot dt = 0, \quad dB_t \cdot dB_t = dt. \tag{1.3.7}$$

Itô's Formula is the stochastic analogue of the chain rule, and can be viewed as a second-order Taylor expansion in which the second-order term survives due to the square Brownian motion increment $(dB_t)^2$ scaling on the same order as the first-order term dt. Using the shorthand notation $dX_t = U_t dt + V_t dB_t$ and the stochastic rules (1.3.7), Itô's formula can be written as

$$dY_t = \Big(\frac{\partial f}{\partial x}(X_t)U_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(X_t)V_t^2\Big)dt + \frac{\partial f}{\partial x}(X_t)V_t dB_t. \tag{1.3.8}$$

Here Y_t is an Itô process with drift $f'(X_t)U_t + \frac{1}{2}f''(X_t)V_t^2$ and diffusion $f'(X_t)V_t$. While the Martingale term $f'(X_t)V_t dB_t$ vanishes in expectation for all t, the extra Itô correction $\frac{1}{2}f''(X_t)V_t^2 dt$ remains and is the distinguishing feature of stochastic calculus.

For the matrix-valued stochastic calculus setting of Chapter 2, we require the multi-dimensional formulation of Itô's Formula, in which a vector-valued Itô process is driven by a vector of independent Brownian motions.

Proposition 7 (Multi-Dimensional Itô's Formula [Øks98])**.** *Let*

$$\boldsymbol{B}(t) = (B_1(t), \cdots, B_m(t))$$

be an $m-$*dimensional vector of independent Brownian motions adapted to* $\big(\mathscr{F}_t\big)_{t\geq 0}$*. Define the vector-valued Itô process* $d\boldsymbol{V}_t = \boldsymbol{U}_t dt + \boldsymbol{H}_t d\boldsymbol{B}_t$*, where* $\boldsymbol{U}_t \in \mathbb{R}^m$ *and* $\boldsymbol{H}_t \in \mathbb{R}^{n\times m}$ *are progressively measurable with* L^2 *entries. Suppose* $g : \mathbb{R}^m \to \mathbb{R}$ *is a twice continuously differentiable function, then* $\boldsymbol{Y}_t = g(\boldsymbol{V}_t)$ *is a vector-valued Itô Process*

$$d\boldsymbol{Y}_t = \sum_{i=1}^{m} \frac{\partial g}{\partial x_i}(\boldsymbol{V}_t) \cdot dV_{i,t} + \frac{1}{2}\sum_{i,j=1}^{m} \frac{\partial^2 g}{\partial x_i \partial x_j}(\boldsymbol{V}_t)\, dV_{i,t} \cdot dV_{j,t} \tag{1.3.9}$$

where the quadratic covariation $dV_{i,t} \cdot dV_{j,t} = d\langle V_{i,t}, V_{j,t} \rangle$ is computed according to the following:

$$dB_i \cdot dB_j = \delta_{ij} dt, \text{ and } dB_i \cdot dt = dt \cdot dB_i = 0. \tag{1.3.10}$$

1.3.2 Hamilton-Jacobi and Method of Characteristics

In Chapter 2, we derive a deterministic first-order nonlinear PDE of Hamilton-Jacobi type, and analyze its solution through the method of characteristics. Here we provide a brief introduction to Hamilton-Jacobi PDEs, referring the reader to the book [Eva10] for a more thorough background.

We study first order Hamilton-Jacobi PDEs of the form

$$\begin{cases} \partial_t S(t, \boldsymbol{u}) + H(\boldsymbol{u}, \nabla_{\boldsymbol{u}} S) = 0 \\ S(0, \boldsymbol{u}_0) = f \end{cases} \tag{1.3.11}$$

where $S = S(t, \boldsymbol{u})$ is the unknown function of independent variables $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$, and $\nabla_{\boldsymbol{u}} = \left(\partial_{u_1}, \ldots, \partial_{u_n}\right)$. The initial function $f : \mathbb{R}^n \to \mathbb{R}$ and the Hamiltonian $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ are both given. The dependence of the nonlinear H on $\nabla_{\boldsymbol{u}} S$ makes (1.3.11) a genuinely nonlinear PDE, and explicit closed-form solutions are generally unavailable.

A powerful technique arises, however, for first-order Hamilton-Jacobi PDEs. Namely the *method of characteristics*, which reduces the PDE (1.3.11) into a system of ordinary differential equations (ODEs) in $\mathbb{R}^n$ called *characteristic equations*. The solutions to these ODEs are called *characteristic curves*, which we denote by $\boldsymbol{u}(t) = (u_1(t), \ldots, u_n(t)) \in \mathbb{R}^n$. Auxiliary momentum variables $\boldsymbol{p}(t)$ are identified with the derivatives $\nabla_{\boldsymbol{u}} S(t, \boldsymbol{u})$ along the curve $\boldsymbol{u}(t)$. With this identification, differentiating (1.3.11) in $\boldsymbol{u}$, and applying the chain-rule results in Hamiltonian equations, which govern the evolution of particles. In the following, $\boldsymbol{u}$ represent the location of the particles and $\boldsymbol{p}$ their momentum. Hamiltonian equations are defined to be the following

system of ODEs that arise from the characteristic equations of H.

$$\begin{cases} \frac{\partial \boldsymbol{u}}{\partial t} = \partial_{\boldsymbol{p}} H(\boldsymbol{u}, \boldsymbol{p}) \\ \frac{\partial \boldsymbol{p}}{\partial t} = -\partial_{\boldsymbol{u}} H(\boldsymbol{u}, \boldsymbol{p}) \end{cases} \tag{1.3.12}$$

with initial conditions $\boldsymbol{u}(0) = \boldsymbol{u}_0$ chosen freely, which in turn determine the initial momentum conditions $\boldsymbol{p}(0) = \nabla_{\boldsymbol{u}} f(\boldsymbol{u}_0)$. When H does not depend on t, as is the case in Chapter 2, the Hamiltonian is conserved along characteristic curves.

$$H\big(\boldsymbol{u}(t), \boldsymbol{p}(t)\big) = H\big(\boldsymbol{u}_0, \boldsymbol{p}_0\big) \tag{1.3.13}$$

In this case, the solution $S(t, \boldsymbol{u})$ to the PDE (1.3.11) is given by the following

$$S\big(t, \boldsymbol{u}(t)\big) = S(0, \boldsymbol{u}_0) - H(\boldsymbol{u}_0, \boldsymbol{p}_0)t + \int_0^t \boldsymbol{p}(s) \cdot \frac{d\boldsymbol{u}}{ds} ds \tag{1.3.14}$$

The method of characteristics thereby reduces solving (1.3.11) to a two-step procedure:

(i) Solve the coupled system of ODEs given by Hamilton's equations (1.3.12) together with (1.3.14) for arbitrary initial conditions.

(ii) Given a target point $(\boldsymbol{u}^*, t^*) \in \mathbb{R}^n \times \mathbb{R}$, identify the unique initial conditions $\boldsymbol{u}_0 \in \mathbb{R}^n$ that ensure the characteristic curves emanating from $(\boldsymbol{u}_0, \boldsymbol{p}_0)$ satisfy $\big(\boldsymbol{u}(t^*); \boldsymbol{u}_0\big) = \boldsymbol{u}^*$.

This procedure yields a classical smooth solution to (1.3.11) for all time $t \geq 0$, provided the characteristic curves $\boldsymbol{u}$ originating from distinct initial values $\boldsymbol{u}_0$ do not intersect. In this case, the mapping $\boldsymbol{u}_0 \mapsto \big(\boldsymbol{u}(t^*); \boldsymbol{u}_0\big)$ is a diffeomorphism, wherein the inverse function theorem guarantees the inversion required in Step **(ii)**. Otherwise if characteristic curves cross, the gradient $\nabla_{\boldsymbol{u}} S$ becomes multi-valued, and one must pass to the theory of *viscosity* solutions to solve the PDE (1.3.11). Fortunately, the Hamilton-Jacobi PDE derived in Chapter 2 has characteristic curves that yield local diffeomorphisms, and we work exclusively with classical smooth solutions.

1.4 Contributions

Recent works have used Random Matrix Theory and Free Probability to analyze the spectral properties of kernel models associated to Neural Networks [ALP19, FM19, FW20, LC18], and the performance of their associated learning problems [AP20, WZ23, MM19b, MMM22]. Through these works, a *universality* principle has emerged in which the kernel matrices and their learning problems are asymptotically equivalent to that of a linear Gaussian covariate model [HL22b, MM19b]. This thesis contributes to the growing field of nonlinear RMT in deep learning theory by utilizing this *Gaussian Equivalence* universality for a precise study of NTK regression under anisotropic power-law training data. We then further probe the robustness of this equivalence under nonlinear data and examine its implications in the context of applicability to machine learning practitioners. We summarize the contributions of this thesis as follows.

NTK Regression under dual power-law model in Chapter 2

In Chapter 2, we study NTK regression under a dual power-law model, which we define as follows. The input data is given by $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$, where $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and the population covariance follows a power-law decay $\boldsymbol{\Sigma}_{jj} = j^{-\alpha}$ for some $\alpha \in (0,\infty)$ called the *capacity* exponent. The training labels are generated via a deterministic signal $\boldsymbol{y} = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta}_j = j^{-r}$ for some $r \in (0,\infty)$ called the *source* exponent, and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2)$ is the label noise. The study of of NTK regression under power-law spectral decay data is motivated by two empirical observations. First, the covariance and learned-feature spectra of real datasets, such as images and text representations, often exhibit power-law decay [Rud97]. Moreover, the decay exponent provides a controllable knob on the effective dimensionality of the dataset [SK20]. Second, a remarkable recent body of work has established *neural scaling laws*; wherein the performance of large models has been shown to exhibit power-law scaling in training-set size, model parameters, and compute [KMH$^+$20, PPXP25, HBM$^+$22]. A leading hypothesis is the model performance exponent is governed by the decay exponent in the training data. While

theoretical explorations of this hypothesis have mostly focused on high-dimensional linear ridge and random feature regression models, we derive a *novel* scaling law for the excess risk of finite-width NTK regression. Our work introduces a new technique involving PDE and SDE methods to derive an asymptotic deterministic PDE whose characteristic curves yield closed form expressions for the trace resolvents that form deterministic equivalents to the prediction risk of KRR on the finite-width NTK. This technique extends previous works past the isotropic Gaussian setting allowing for general data assumptions, and gives deterministic equivalents for both standard and signal-weighted resolvents. Specializing to the dual-power law model, we use these deterministic equivalents to establish a scaling law for optimally-tuned finite-width NTK model, and derive a closed-form excess-risk exponent in terms of *source-capacity* parameters (α, r).

Spike eigenvalue analysis under nonlinear data in Chapter 3

In Chapter 3, we probe the *Gaussian Equivalence* universality principle [HL22b, MM19b, GLR$^+$22, BPH23] under various knobs that are commonly used in machine learning practice: sample complexity, signal-to-noise-ratio (SNR), choice of activation function, and the use of pretrained features. While the Gaussian Equivalence Theorem (GET) is a highly useful theoretical tool, as it allows for the translation of the performance of complex nonlinear neural networks into tractable kernel models, it leaves open the question whether the idealized linear equivalence remains meaningful when dealing with high-dimensional nonlinearly separable data. Motivated by this question, we study the CK under a canonical nonlinearly separable high dimensional dataset: the XOR. We use the *informative* spike eigenvalues of the CK and their asymptotic alignment with XOR-labels as a proxy for nonlinear learnability. To study this, we develop a RMT framework for the CK that is based on a *quadratic* equivalent. Namely, by writing the XOR dataset $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{Z}$ into a rank-2 perturbation $\boldsymbol{M}$ of noise $\boldsymbol{Z}$, we derive

$$\sigma(\boldsymbol{W}\boldsymbol{X}) \approx \underbrace{\sigma(\boldsymbol{W}\boldsymbol{Z})}_{\text{uninformative bulk}} + \underbrace{\boldsymbol{T}_1}_{\text{linear spikes}} + \underbrace{\boldsymbol{T}_2}_{\text{quadratic spikes}} \tag{1.4.1}$$

where only the eigenvectors corresponding to the spiked eigenvalues of $\boldsymbol{T}_2$ asymptotically align with XOR-labels. Based on this quadratic equivalent (QE), we derive precise BBP (Baik–Ben Arous–Péché) phase transitions for each of the four settings: linear-width regime, large SNR regime, pretrained features, and quadratic sample size regime. We conclude with an experiment with pretrained features from a deep neural network on CIFAR-10 training data, exhibiting a phase transition in the steps of gradient descent wherein the XOR task becomes learnable via the principal components of the pretrained CK.

1.5 Notation

Throughout this thesis, unless explicitly stated otherwise, any limit will denote *with high probability*. We will use upper case trace of a square $n \times n$ matrix $\boldsymbol{A}$ to denote $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \boldsymbol{A}_{ii}$ while lower case trace denotes the normalized trace, i.e $\mathrm{tr}(\boldsymbol{A}) = \frac{1}{n}\mathrm{Tr}$. Throughout, $||\cdot||$ will denote the Euclidean ℓ_2 norm of a vector, and operator norm for matrices $||\boldsymbol{M}|| = \sup_{||v||=1} ||\boldsymbol{M}\boldsymbol{v}||$. We will let $||\cdot||_F$ denote the Frobenius norm, i,e $||\boldsymbol{M}||_F = \sqrt{\mathrm{Tr}(\boldsymbol{M}^*\boldsymbol{M})}$, where $\boldsymbol{M}^*$ denotes the conjugate transpose of $\boldsymbol{M}$.

Let $\boldsymbol{A} \odot \boldsymbol{B}$ denote the Hadamard (entry-wise) product of matrices, i.e $(\boldsymbol{A} \odot \boldsymbol{B})_{ij} = \boldsymbol{A}_{ij} \odot \boldsymbol{B}_{ij}$ for $i \in [n]$, where $[n] = \{i = 1, \ldots, n\}$. Let $\mathbb{E}_{\boldsymbol{w}}[\cdot]$ denote expectation with respect to the random vector $\boldsymbol{w}$ only, and let $o_{n,\mathbb{P}}(\cdot)$ denote little-o in probability as $n \to \infty$. Define the distance between a point and set as $\mathrm{dist}(x, A) = \inf\{|x - y| : y \in A\}$. For a probability measure μ on $\mathbb{R}$, define its support $\mathrm{supp}(\mu)$ as

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

Then define the $\varepsilon-$neighborhood of the support as

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

Chapter 2

NTK Regression under dual Power-law model

2.1 Introduction

Understanding the generalization performance of over-parameterized neural networks (NNs) remains an active and open area of research in deep learning theory. An especially productive line of works proceed by *linearizing* NNs around their random initialization, leading to the Neural Tangent Kernel (NTK) [JGH18] and Random Feature (RF) [RR07] models. The NTK has garnered particular attention due to its connection to the training dynamics of wide NNs under gradient descent. Under certain settings, which have been termed the *lazy training* regime [COB19], the training dynamics and generalization risk of ultra-wide NNs are well approximated by those of kernel ridge regression (KRR) on the NTK [JGH18, LXS$^+$20, AP20]. In the high-dimensional regime where the number of training samples n, ambient features p, and width of the network N grow proportionally, Random Matrix Theory (RMT) and Free Probability have been used to provide sharp asymptotic characterization of KRR risk and kernel spectra, for both RF [MM19b, MMM22, WZ23] and the NTK [AP20, FW20, MJBM22].

Much of the existing analysis relies upon the assumption of isotropic features, that is data drawn from a distribution with identity covariance. With significant effort, the generalization risk of KRR has been explicitly characterized in the isotropic feature setting for various Kernel models. [MM19b] derived formula for exact asymptotic risk curves of RF regression, and

identified both the *double*-descent peak location and shape. [AP20] then extended this analysis to explicitly characterize the non-monotonic *triple*-descent risk curves of NTK regression, and identified two over-parameterization transitional phases. However the isotropic setting fails to capture empirical phenomena observed in real-world data, where singular values of inputs from image, text, and audio representations roughly follow a power-law decay [Rud97]. Moreover, recent work on *neural scaling laws* has shown the exponent controlling this decay governs the risk curve rate, location of *double*-descent peaks, and optimal tuning of regularization [KMH$^+$20, BAP24, MRS22, PPXP25]. It remains an open question to determine the exact class of models that support scaling laws.

This motivates a precise study of NTK regression under the anisotropic power law setting. A central question we aim to answer in this chapter is the following: *What is scaling limit for finite-width NTK regression under power-law dataset?*

The primary challenge in extending the previous findings on the isotropic dataset lies in the inability of the employed techniques to readily generalize to anisotropic features. The majority of analysis relies first upon a *Gaussian Equivalence* universality result [HL22b], which translates the non-linear feature model to an equivalent Gaussian covariates model with matching first and second moments. The asymptotic risk of the Gaussian equivalent model is explicitly determined by several specific random matrix functionals. Classical RMT and Free Probability tools such as linear pencil methods from operated-valued free probability [AP20] and the replica method from statistical physics [BCP21], can then be applied to derive asymptotic formulae for these functionals.

Though our current approach relies on the Gaussian Equivalence Theorem (GET), formally stated as Conjecture (**C1**) in Section 2.2.5, we derive a new RMT method that extends to NTK regression for general datasets, including anisotropic power-law structured features and labels. Our method is motivated by [DHK22] which derives the large-n limit density of eigenvalues of certain matrix-valued stochastic processes. This result was further developed for non-commutative stochastic calculus in [JKN26]. We embed the Gaussian entries of the

NTK kernel with independent Brownian motions, and derive a finite-n Stochastic PDE using entry-wise Itô calculus. The primary technical tool in our analysis is the time-parameterized log determinant potential, $\mathscr{F}_n(t,\kappa,\gamma) = \frac{1}{n}\log\det \boldsymbol{K}(t,\alpha,\gamma)$, where $\boldsymbol{K}(t,\alpha,\gamma)$ is the Brownian-motion embedded NTK matrix with auxiliary scalar parameters α and γ corresponding to the additive shifts of the sample covariance matrix and resolvent shift in the NTK. Our main technical result is the derivation of a deterministic, first-order Hamilton-Jacobi (HJ) PDE whose characteristic curve solutions explicitly determine the asymptotic risk of NTK regression in the dual power-law covariance and signal decay setting.

2.1.1 Related Works

Random feature models were introduced as a scalable approximation to kernel methods [RR07]. NTK is a special case of random feature model, derived by wide neural networks trained in the lazy regime to kernel regression at initialization [JGH18, COB19, LXS$^+$20]. In high-dimensional proportional limits, RMT has yielded precise spectral and risk characterizations for random feature regression and kernel methods, including double-descent phenomena and kernel-matrix concentration [MM19b, MMM22, ALP19, JSS$^+$20, WZ23]. For NTK regression at linear width, [AP20] and [FW20] derived sharp descriptions of the NTK spectrum and generalization behavior, while related work analyzes spectra of nonlinear kernel matrices and random feature maps [EK10, LC18, WZ21, MJBM22]. Our work differs from this line by studying anisotropic, dual power-law data with a deterministic source vector, and by deriving the ordinary and signal-weighted resolvent traces through a Hamilton-Jacobi characteristic system.

A common route for analyzing nonlinear random feature and NTK models is to replace nonlinear features by a Gaussian surrogate with matching low-order moments. This Gaussian-equivalence principle underlies several precise asymptotic analyses of random features and kernel matrices [MM19b, Péc19, HL22b, LY22, HL22a]. Recent deterministic-equivalent results provide characterizations of for kernel ridge regression and random feature ridge regression, including non-asymptotic test-error and GCV guarantees for KRR [MS24a] and dimension-free

deterministic equivalents for random feature regression [DLM24]. These works typically express risk through kernel or feature-map eigenvalues and self-consistent equations. Our contribution is complementary: for the Gaussian-equivalent finite-width NTK/RF model, we obtain a scalar fixed point for the resolvent traces and introduce a source-dependent signal resolvent for general anisotropic data via PDE method.

Classical kernel learning theory relates KRR rates to source and capacity assumptions, formulated through the eigendecay of the kernel integral operator and the alignment of the target function with its eigenbasis [CDV07a, CLKZ22]. Related spectral learning-curve analyses show that kernel regression and wide networks can exhibit rates governed by the data/task spectrum [BCP21]. More recent neural scaling laws [KMH$^+$20] have been proved for random feature models, gradient dynamics, and compute constraints [WHS22, MRS22, BAP24, PPXP25]. Closest to our power-law data setting, [WL25] study high-dimensional KRR with anisotropic Gaussian data whose covariance follows a power law, showing how the data spectrum is transferred to the kernel spectrum and affects excess risk. In contrast, we keep the finite-width random-feature/NTK structure and analyze how both the capacity exponent of the input covariance and the source exponent of the teacher function.

2.1.2 Contributions

We summarize our main contributions as follows. We derive an asymptotic deterministic PDE whose characteristic curves yield closed form expressions for the trace resolvents that form deterministic equivalents to the prediction risk of KRR on the finite-width NTK. This PDE holds under general data assumptions, and gives deterministic equivalents for both standard and signal-weighted resolvents.

Specializing to the dual-power law model, we use these deterministic equivalents to establish a scaling law for optimally-tuned finite-width NTK model, and derive a closed-form excess-risk exponent in terms of source and capacity parameters (α, r).

2.2 Main results

2.2.1 Problem setup

Consider a two-layer neural network $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ with hidden-layer width N

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\boldsymbol{w}^\top \sigma(\boldsymbol{W}_1 \boldsymbol{x}), \qquad \boldsymbol{\theta} := \left(\boldsymbol{w}_1^{(1)}, \ldots, \boldsymbol{w}_N^{(1)}, w_1, \ldots, w_N\right) \in \mathbb{R}^{N(p+1)} \tag{2.2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is the input data, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the non-linear activation function applied entry-wise. The first layer weights $\boldsymbol{W}_1 \in \mathbb{R}^{N\times p}$ are initialized as i.i.d. $\mathcal{N}(0, \frac{1}{p})$, while the $\frac{1}{\sqrt{N}}$ factor in (2.2.1) is due to the LeCun/NTK normalization [JGH18, LBOM12], which ensures the model output remains $O(1)$ as the width grows. The second layer weights $\boldsymbol{w} \in \mathbb{R}^N$ are initialized as $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_W^2)$, where the variance σ_W^2 is a hyperparameter that will tune the NTK behavior. Let $\boldsymbol{\theta}_0$ denote the weights at their random initialization. For simplicity we set $f_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}_0) = 0$, which is referred to as *centering* by [AP20]. This either amounts to doubling the number of neurons in the network or subtracting the initial network value throughout training, and does not alter our analysis. Under the *lazy training regime* [COB19, LXS+20], the weights $\boldsymbol{\theta}$ remain close to initialization $\boldsymbol{\theta}_0$ for large N, and the network output is well approximated by a first-order Taylor expansion around $\boldsymbol{\theta}_0$. While this connection motivates our study of the NTK, our primary focus is on NTK regression as a problem of independent interest, especially in the fast-growing literature of *neural scaling laws*.

NTK is the defined as the gram matrix of the Jacobian of the network with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{K}^{\text{NTK}}(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}'; \boldsymbol{\theta}_0). \tag{2.2.2}$$

Following [AP20], we can partition into first and second layer weights

$$\boldsymbol{K}^{\text{NTK}}(\boldsymbol{x}, \boldsymbol{x}') = \underbrace{\nabla_{\boldsymbol{w}} f_{\boldsymbol{\theta}}(\boldsymbol{x}; \theta_0)^\top \nabla_{\boldsymbol{w}} f_{\boldsymbol{\theta}}(\boldsymbol{x}'; \theta_0)}_{\boldsymbol{K}_1(\boldsymbol{x}, \boldsymbol{x}')} + \underbrace{\nabla_{\boldsymbol{W}_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}; \theta_0)^\top \nabla_{\boldsymbol{W}_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}'; \boldsymbol{\theta}_0)}_{\boldsymbol{K}_2(\boldsymbol{x}, \boldsymbol{x}')}$$

The first and second layer kernels can be written as

$$\boldsymbol{K}_1(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{F}^\top \boldsymbol{F}, \qquad \boldsymbol{K}_2(\boldsymbol{X},\boldsymbol{X}) = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \frac{(\boldsymbol{F}')^\top \mathrm{diag}(\boldsymbol{w})^2 \boldsymbol{F}'}{N}. \tag{2.2.3}$$

where $\boldsymbol{F} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}_1\boldsymbol{X})$ and $\boldsymbol{F}' = \frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}_1\boldsymbol{X})$ are the feed-forward feature maps of σ and σ' respectively. We study the problem of learning an unknown target function $f^* : \mathbb{R} \to \mathbb{R}$ from n training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$, where

$$y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2) \tag{2.2.4}$$

The predictor of the kernel ridge regression problem on the NTK is given by

$$\widehat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{k}_{\mathrm{NTK}}(\boldsymbol{x})^\top \left(\boldsymbol{K}^{\mathrm{NTK}} + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{y} \tag{2.2.5}$$

where $\boldsymbol{k}_{\mathrm{NTK}}(\boldsymbol{x}) \in \mathbb{R}^n$ has entries $\boldsymbol{K}^{\mathrm{NTK}}(\boldsymbol{x}, \boldsymbol{x}_\alpha)$ for $\alpha \in [n]$. Note the ridge parameter $\lambda > 0$ corresponds to an ℓ^2 penalty on the weights $\boldsymbol{\theta}$ towards their initialization. The training error and prediction risk of estimator (2.2.5) are given by

$$\mathrm{E}_{\mathrm{train}}(\lambda) = \frac{1}{n}||\boldsymbol{y} - \widehat{f}_\lambda(\boldsymbol{x})||_2^2, \qquad \mathscr{R}(\lambda) = \mathbb{E}_{\boldsymbol{x}}\big[\big(\widehat{f}_\lambda(\boldsymbol{x}) - \boldsymbol{y}\big)^2\big]. \tag{2.2.6}$$

Here the expectation in $\mathscr{R}(\lambda)$ is over an independent test point $\boldsymbol{x}$, conditional on the training dataset and randomness from the weights $\boldsymbol{\theta}$. We note the risk is thereby a random variable, however it will concentrate around its mean in the asymptotic, proportional regime. Deriving asymptotic deterministic equivalents to the two random variables (2.2.6) constitutes the principal technical goal of this chapter.

2.2.2 Gaussian Equivalence

The training error and prediction risk (2.2.6) involve large random matrices with non-linear dependencies, which fall outside the immediate scope of classical RMT tools. To make analysis tractable, we invoke the Gaussian Equivalence Theorem [HL22b, Péc19, MM19b], a universality result that reduces the KRR problem (2.2.5) to regression on a linear Gaussian surrogate model with matching first and second moments.

As an illustrative example, consider the non-linear feature matrix $\boldsymbol{F} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}_1\boldsymbol{X})$. By matching first and second order mixed moments of entries, we have the following approximation

$$\sigma(\boldsymbol{W}_1\boldsymbol{X}) \approx b_1\boldsymbol{W}_1\boldsymbol{X} + b_2\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \tag{2.2.7}$$

for a suitable choice of b_1, b_2 which arise from a Hermite polynomial expansion of σ. The approximation $\approx$ holds on the order of bulk statistics.

When the input data is given by $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$ for $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ independent from $\boldsymbol{W}_1$, a straight-forward central limit theorem argument shows the entries of $\boldsymbol{W}_1\boldsymbol{X}$ are marginally normal, with mean zero and variance $\tau^2 := \frac{1}{p}\operatorname{Tr}\boldsymbol{\Sigma}$. Therefore, (2.2.7) arises from a Hermite polynomial expansion of σ against $\mathcal{N}(0,\tau^2)$. We define these Hermite coefficients for σ, σ' as follows.

$$b_1 = \frac{1}{\tau^2}\mathbb{E}[g\cdot\sigma(g)] = \mathbb{E}[\sigma'(g)], \qquad b_2^2 = \mathbb{E}[\sigma(g)^2] - \tau^2 b_1^2$$

where $g \sim \mathcal{N}(0,\tau^2)$. Further, define the following

$$c_1 = \mathbb{E}[\sigma'(g)^2], \qquad \rho = \mathbb{E}[\sigma(g)^2] = b_2^2 + \tau^2 b_1^2$$

We note that $\tau = \tau(p)$ is p-dependent, and therefore so too are the Hermite coefficients b_1, b_2. While this is non-standard, our goal is to derive deterministic equivalents for functionals of $\boldsymbol{K}^{\text{NTK}}$, and our analysis directly encodes the p-dependence in b_1, b_2. One can avoid dimension-

dependent Hermite coefficients by normalizing $\boldsymbol{\Sigma}$ to have unit trace.

We make the following simplifications on the two kernels $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, which preserve bulk statistics asymptotically. Specifically, for $\boldsymbol{K}_2$ one has

$$\boldsymbol{K}_2 \approx b_1^2 \sigma_W^2 p^{-1} \boldsymbol{X}^\top \boldsymbol{X} + \tau^2 \sigma_W^2 \big(c_1 - b_1^2\big) \mathbf{I}_n. \tag{2.2.8}$$

We collect the substitutions (2.2.7) and (2.2.8) into the *linearized* kernel $\boldsymbol{K}_{\text{lin}}$.

$$\boldsymbol{K}_{\text{lin}} = \boldsymbol{F}_{\text{lin}}^\top \boldsymbol{F}_{\text{lin}} + \sigma_W^2 \Big(p^{-1} b_1^2 \boldsymbol{X}^\top \boldsymbol{X} + \tau^2 (c_1 - b_1^2) \mathbf{I}_n \Big) + z \mathbf{I}_n, \quad \boldsymbol{F}_{\text{lin}} = \frac{1}{\sqrt{N}} \big(b_1 \boldsymbol{W}_1 \boldsymbol{X} + b_2 \boldsymbol{\Theta} \big) \tag{2.2.9}$$

For simplicity, we drop the subscript *lin* and directly study the linearized model $\boldsymbol{K}$ under the substitutions.

2.2.3 Generalized Cross Validation (GCV)

The full prediction risk formula (2.2.6) involves functionals of random matrices $\boldsymbol{W}_1, \boldsymbol{\Theta}$, the resolvent $\boldsymbol{K}^{-1}$, and the population covariance $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is typically unknown, we introduce the GCV metric, which is an estimator accessible from training data alone:

$$\text{GCV}(\lambda) = \frac{\text{E}_{\text{train}}(\lambda)}{\Big(\lambda \, \text{tr}\big(\boldsymbol{K}^{\text{NTK}} + \lambda \mathbf{I}_n \big)^{-1} \Big)^2}. \tag{2.2.10}$$

Initially derived as a rotationally-invariant variant of leave-one-out cross validation (LOOCV) by [GHW79], the GCV has been shown to be a consistent estimator of prediction risk across a wide variety of high-dimensional ridge regression [WHS22, HMRT22, PRT22], and KRR models [AP20, JSS$^+$20]. It has also been called the Kernel Alignment Risk Estimator (KARE) in the setting of KRR [JSS$^+$20]. Most relevant to our setting, [AP20] proved the asymptotic equivalence of the GCV to the prediction risk of the NTK in the isotropic Gaussian setting. We conjecture, and empirically verify in Figures 2.2 and 2.3, this equivalence extends to the

anisotropic power-law setting studied here. More numerical experiments are presented in Section 2.4.

The training error $\mathrm{E}_{\text{train}}(\lambda)$ concentrates around its expectation over label noise. A quick calculation gives the following

$$\mathrm{E}_{\text{train}}(\lambda) = \lambda^2\Big(\sigma_\varepsilon^2 \operatorname{tr}\big(\boldsymbol{K}+\lambda \mathbf{I}_n\big)^{-2} + \operatorname{tr}\Big(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{K}+\lambda \mathbf{I}_n)^{-2}\Big)\Big) + o_{n,\mathbb{P}}(1). \tag{2.2.11}$$

This motivates the definition of the following trace functionals

$$m_1^{(n)}(z) = \operatorname{tr}\big(\boldsymbol{K}+z\mathbf{I}_n\big)^{-1}, \qquad m_2^{(n)}(z) = \operatorname{tr}\big(p^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{K}+z\mathbf{I}_n)^{-1}\big) \tag{2.2.12}$$

$$m_\beta^{(n)}(z) = \frac{1}{n}\big(\boldsymbol{\beta}^\top \boldsymbol{X}^\top (\boldsymbol{K}+z\mathbf{I}_n)^{-1}\boldsymbol{X}\boldsymbol{\beta}\big). \tag{2.2.13}$$

Under an isotropic signal vector $\boldsymbol{\beta} \sim \mathcal{N}(0, p^{-1}\mathbf{I}_p)$, an application of the Hanson-Wright concentration inequality gives $m_\beta^{(n)} = m_2^{(n)} + o_{n,\mathbb{P}}(1)$ as $n \to \infty$. However, the structured signal case of Assumption (**A0**) necessitates the separate treatment of the quadratic form m_β. Using (2.2.11), we define the GCV in terms of derivatives of these trace functionals.

$$\mathrm{GCV}(\lambda) = \frac{-\big(\sigma_\varepsilon^2 \partial_\lambda m_1^{(n)}(\lambda) + \partial_\lambda m_\beta^{(n)}(\lambda)\big)}{\big(m_1^{(n)}(\lambda)\big)^2} \tag{2.2.14}$$

The GCV can be understood as a correction to train error to account for model complexity, wherein the denominator of (2.2.10) accounts for the train-test gap. For some intuition behind this, write

$$\lambda \operatorname{tr}\big(\boldsymbol{K}^{\mathrm{NTK}}+\lambda \mathbf{I}_n\big)^{-1} = 1 - \operatorname{tr}\Big(\boldsymbol{K}^{\mathrm{NTK}}\big(\boldsymbol{K}^{\mathrm{NTK}}+\lambda \mathbf{I}_n\big)^{-1}\Big) = 1 - \mathrm{df}_{\boldsymbol{K}^{\mathrm{NTK}}}(\lambda)$$

Here $\mathrm{df}_{\boldsymbol{K}^{\mathrm{NTK}}}(\lambda) := \operatorname{tr}\Big(\boldsymbol{K}^{\mathrm{NTK}}\big(\boldsymbol{K}^{\mathrm{NTK}}+\lambda \mathbf{I}_n\big)^{-1}\Big)$ provides an effective degrees of freedom of a

linear estimator on the NTK matrix, and is closely related to the Signal Capture Threshold (SCT) introduced in [JSS+20]. We refer to Section 7.6 of [HTF09] for further reading on this measure.

2.2.4 Dual power-law model

In order to study NTK under a dataset that allows for scaling laws, we restrict to the power-law covariance. Our methodology however extends to very general data assumptions on $\boldsymbol{X}$. We make the following dual power-law assumption on the train data and train labels.

(**A0**) **Dual power-law model.** The training data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are drawn i.i.d from $\mathcal{N}(0, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \operatorname{diag}\big(\sigma_1, \cdots, \sigma_p\big), \qquad \sigma_j = j^{-\alpha}, \quad j = 1, \ldots, p \tag{2.2.15}$$

for some fixed $\alpha \in (0, \infty)$. The labels are generated as a linear power-teacher.

$$\boldsymbol{y}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \beta_j = j^{-r}, \quad j = 1, \ldots, p \tag{2.2.16}$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ is independent of $\boldsymbol{x}$ and $r > 0$ is a fixed constant.

The exponents α and r are referred to as *capacity* and *source* in power-law literature [CDV07b, CLKZ22]. Roughly speaking α controls the effective dimensionality of the data, with a larger α causing the data to concentrate into lower effective dimensions. The source r controls the difficulty of the learning problem, as it measures the alignment of the signal along the top eigenvectors of $\boldsymbol{\Sigma}$. We make further restrictions on r in the following section. These are to ensure the GET approximations in Section 2.2.2 are still valid.

A quick remark about the choice of scaling on the power-law covariance eigenvalues. Following classic *source-capacity* convention [CDV07b] we study NTK regression under population eigenvalues $\sigma_j = j^{-\alpha}$, which we will refer to as *unscaled* power-law. While extreme eigenvalues of the unscaled power-law remain $O(1)$ in the proportional limit, the bulk eigenvalues collapse to a delta point mass at zero. [WL25] study a *trace*-scaled power-law in the context of KRR,

i.e population covariance $\Sigma_{jj} = c_p(\alpha) j^{-\alpha}$, where $c_p(\alpha) = \left(\sum_{k=1}^{\alpha} k^{-\alpha}\right)^{-1}$ ensures $\mathrm{Tr}\boldsymbol{\Sigma} = 1$. For $\alpha > 1$ these two normalizations are equivalent up to a scalar.

2.2.5 Assumptions

We make the following general assumptions throughout the rest of this chapter.

(**A1**) **Proportional Regime.** $n, p, N \to \infty$ at rates $\frac{n}{p} \to \phi$ and $\frac{N}{p} \to \psi$ for some $\phi, \psi \in (0, \infty)$.

(**A2**) **Capacity Exponent.** We take $\boldsymbol{\Sigma}_{jj} = j^{-\alpha}$ with $\alpha > 0$.

(**A3**) **Source Exponent.** The ground truth vector satisfies $\beta_j = j^{-r}$ with $r > \frac{1}{2}$.

(**A4**) **Nonlinearity.** The non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz, centered with respect to $\mathcal{N}(0, \tau^2)$, and has finite Gaussian moments of all order.

Assumption (**A3**) ensures $||\boldsymbol{\beta}||_2 < \infty$ as $p \to \infty$, which would otherwise cause the GCV consistency conjecture stated below to be violated.

We state the two ansatz, whose formal verification is delegated to future work but are expected to hold under the current setting and are verified empirically in Section 2.4 and by Figures 2.2 and 2.3.

(**C1**) **Generalized Gaussian Equivalence.** The GET holds under the power-law covariance:

$$\left|\mathscr{R}(\lambda) - \mathscr{R}_{\mathrm{GE}}(\lambda)\right| = o_{n,\mathbb{P}}(1) \tag{2.2.17}$$

pointwise for each fixed $\lambda \geq 0$. Here $\mathscr{R}_{\mathrm{GE}}(\lambda)$ denotes prediction risk of KRR on the linearized NTK kernel (2.2.9).

(**C2**) **GCV Estimator.** The GCV estimator (2.2.10) is a uniformly consistent estimator of NTK risk $\mathscr{R}_\lambda$ over any bounded, compact interval away from 0.

$$\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left|\mathscr{R}(\lambda) - \mathrm{GCV}(\lambda)\right| = o_{n,\mathbb{P}}(1) \tag{2.2.18}$$

We further conjecture (2.2.18) extends to $\lambda_{\min} = 0$ for the NTK, as was shown in [MS24a] for KRR in the the benign-overfitting regime.

Conjectures (**C1**)- (**C2**) extend to NTK under power-law covariance a well developed line of work. Gaussian equivalence for random feature and NTK models have been established in [HXAP20, GLR$^+$22, MM19b, MS22], and consistency of the GCV estimator for KRR was established in [AP20, MS24a, JSS$^+$20, WHS22]. Furthermore, uniform consistency of the GCV over ridge penalty has been established in [PWRT21, MS24a, XMRH21]. While we make no statement on the rate of convergence of (2.2.18), [MS24a] proved a bound at rate $\mathscr{O}_{n,\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ uniformly over $\lambda \in [0, \lambda_{\max}]$ under general assumptions. We conjecture the analogues for the finite-width NTK and verify them empirically in Section 2.4.

2.2.6 Outline of approach

We outline our novel technical approach , which we expect to generalize to deriving deterministic equivalents for a large class of random matrix functionals that characterize the risk of linearized NNs and associated kernel models. The key component in our analysis is the log-determinant Brownian-motion embedded kernel matrix $\boldsymbol{K}(t, \boldsymbol{u})$:

$$\mathscr{F}_n(t, \boldsymbol{u}) = \frac{1}{n} \log\det \boldsymbol{K}(t, \boldsymbol{u}) \tag{2.2.19}$$

where $\boldsymbol{u} = (\boldsymbol{u}_j)$ is a finite number of additional scalar auxiliary parameters alongside a time parameter t for the Brownian motions of entries of $\boldsymbol{K}(t, \boldsymbol{u})$. The motivation of (2.2.19) is to recover the trace functionals discussed in Section 2.2.3 as partial derivatives of $\mathscr{F}_n$:

$$\partial_{\boldsymbol{u}_1} \mathscr{F}_n = \operatorname{tr}\big(\boldsymbol{K}(t, \boldsymbol{u}_1)^{-1}\big) =: m_1^{(n)}(t, \boldsymbol{u}_1), \quad \partial_{\boldsymbol{u}_2} \mathscr{F}_n = \operatorname{tr}\big(\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{K}(t, \boldsymbol{u}_2)^{-1}\big) =: m_2^{(n)}(t, \boldsymbol{u}_2) \tag{2.2.20}$$

for appropriately chosen $\boldsymbol{u}_1, \boldsymbol{u}_2$. At time $t = 0$ the log-potential $\mathscr{F}_n(0, \boldsymbol{u}_0)$ and its partial derivatives are pure linear kernels in the sample covariance $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$. These trace functionals

have deterministic equivalents through the anisotropic generalization of the Marchenko-Pastur Theorem [SB95]. Using Itô-matrix calculus and PDE methods, we evolve the entries of Gaussian matrices $\boldsymbol{K}$ as i.i.d Brownian motions, and derive a deterministic Hamilton-Jacobi type PDE in the proportional limit. We outline these steps as follows.

Step 1: Finite Dimensional Stochastic PDE.

$\mathscr{F}_n(t,\boldsymbol{u})$ is C^2 in the entries of $\boldsymbol{K}(t,\boldsymbol{u})$, and therefore amenable to Stochastic calculus via a multi-dimensional variant of Itô's formula (2.3.2). With some matrix calculus and bookkeeping, we derive a SDE of the form

$$d\mathscr{F}_n(t,\boldsymbol{u}) = d\mathscr{M}_n(t,\boldsymbol{u}) + \mathscr{D}_n(t,\boldsymbol{u})dt + E_n(t) \tag{2.2.21}$$

where the Martingale terms $\mathscr{M}_n(t)$ and remainder $E_n(t)$ will vanish uniformly in $t \in [0,1]$. Details of this derivation are presented in Section 2.3.2. The next step involves a careful tracking of vanishing order terms, and then differentiating the left over terms in (2.2.21) with respect to the time parameter t.

Step 2: Asymptotic Deterministic PDE.

The martingale terms in (2.2.21) satisfy $\mathscr{M}_n(t,\boldsymbol{u}) \to 0$ uniformly in probability, while the remainder $E_n(t) \to 0$ uniformly in t as well. The drift terms $\mathscr{D}_n(t,\boldsymbol{u})$ can be realized as partial derivatives of $\mathscr{F}_n(t,\boldsymbol{u})$ with respect to the spatial parameters $\boldsymbol{u}$, after interchanging limit and derivative. By controlling the vanishing order terms, and then differentiating the surviving terms in (2.2.21) we have

$$\partial_t \mathscr{F}_\infty(t,\boldsymbol{u}) + H(\boldsymbol{u}, \partial_{\boldsymbol{u}} \mathscr{F}_\infty) = 0 \tag{2.2.22}$$

where H is a function of the spatial parameters $\boldsymbol{u}_j$ and partial derivatives of $\mathscr{F}_\infty$ with respect to $\boldsymbol{u}_j$ only. In our setting H is a polynomial in its arguments, and therefore the PDE (2.2.22) has a unique analytic solution due to the C-K Theorem, given the initial condition $\mathscr{F}_\infty(0)$.

Step 3: Hamilton-Jacobi PDE.

The PDE (2.2.22) is a first-order, nonlinear PDE, which can prove to be difficult to solve analytically in general. However, we recognize (2.2.22) as a Hamilton-Jacobi type PDE of the form

$$\partial_t \mathscr{F}_\infty(t, \boldsymbol{u}) + H\big(\boldsymbol{u}, p_{\boldsymbol{u}}\big) = 0 \tag{2.2.23}$$

where H is the Hamiltonian function obtained from replacing partial derivatives $\partial_{\boldsymbol{u}_j}$ with momentum variables p_j in (2.2.22). The remainder of our analysis is built upon the solution to (2.2.23), and in particular the solutions to its characteristic curves.

Step 4: Solution via Method of Characteristics

Hamilton's equations for H give a system of coupled ODEs

$$\dot{\boldsymbol{u}}_j = \partial_{p_j} H, \qquad \dot{p}_j = -\partial_{\boldsymbol{u}_j} H \tag{2.2.24}$$

where $\dot{x} = \partial_t x$. The initial conditions to (2.2.24) are given by the limit of $\mathscr{F}_n(0, \boldsymbol{u})$ and its partial derivatives along $\boldsymbol{u}_j$, which are given by the general Marchenko-Pastur Theorem.

The motivation behind (2.2.24) is that the momentum variables p_j will correspond to derivatives of S_t along the curves $\boldsymbol{u}_j(t)$. Therefore, an apt choice parameters $\boldsymbol{u}_j(t)$ and their initial conditions $\boldsymbol{u}_j(0)$ will allow the recovery of the target NTK traces, as in (2.2.20).

2.2.7 Statement of main results

The results presented in this section rest on a closed-form characterization of three NTK random matrix functionals that jointly determine asymptotic training and test error of NTK regression. The first two are the resolvent traces, characterized in the isotropic setting by linear pencil methods in [AP20], and for general data assumptions for the RF kernel model in [ALP19] using resolvent methods. The third is a novel object, which we call the signal resolvent. We work throughout in the proportional limit, and take in probability limits. Under the Assumptions

(**A0**)-(**A4**), each functional below will concentrate onto a deterministic equivalent quantity. We define the following two traces

$$\begin{aligned} m_1^{\infty}(z) &= \lim_{n\to\infty} \operatorname{tr}\big(\boldsymbol{K}^{\mathrm{NTK}}(\kappa_{\mathrm{target}}, \gamma_{\mathrm{target}})^{-1}\big) \\ m_2^{\infty}(z) &= \lim_{n\to\infty} \operatorname{tr}\Big(\frac{1}{p}\boldsymbol{X}^{\top}\boldsymbol{X}\big(\boldsymbol{K}^{\mathrm{NTK}}(\kappa_{\mathrm{target}}, \gamma_{\mathrm{target}})^{-1}\big)\Big) \end{aligned} \tag{2.2.25}$$

where $\kappa_{\mathrm{target}} = b_1^2\sigma_W^2$ and $\gamma_{\mathrm{target}} = z + \tau^2\sigma_W^2(c_1 - b_1^2)$ are chosen to match the scalar parameters in (2.2.9) for $z \in \mathscr{H}^+$. We next define the signal resolvent to be

$$m_{\beta}^{(n)}(z) = \frac{1}{n}\boldsymbol{\beta}^{\top}\boldsymbol{X}^{\top}(\boldsymbol{K}^{\mathrm{NTK}}(\kappa_{\mathrm{target}}, \gamma_{\mathrm{target}})^{-1}\boldsymbol{X}\boldsymbol{\beta} \tag{2.2.26}$$

with limit $m_{\beta}^{\infty}(z) = \lim_{n\to\infty} m_{\beta}^{(n)}(z)$. The signal-resolvent m_{β} measures the strength of alignment of the signal $\boldsymbol{\beta}$ with the resolvent of the NTK. Theorem 2 makes this connection explicit, expressing $m_{\beta}^{(n)}$ as a discrete integral against the signal-weighted measure $\nu_{\beta} = \sum_j \beta_j^2 \delta_{\sigma_j}$.

A quick remark about reduction to isotropic labels. In the case that $\boldsymbol{\beta}$ is drawn uniform randomly on the sphere, such as in [ALP19, MM19b], m_{β} reduces to m_2 by an application of Hanson-Wright concentration inequality (Theorem 6.2.1 in [Ver18]). This reduction fails, however, under the structured label assumption made in (2.2.16).

SDE and PDE

Our first main results are the explicit form of the finite-dimensional SDE and asymptotic HJ PDE introduced in Section 2.2.6. Write

$$\mathscr{F}_n(t, \kappa, \gamma) = \frac{1}{n}\log\det\boldsymbol{K}(t, \kappa, \gamma).$$

Proposition 8 (Finite SDE)**.** *Under Assumptions (**A1**) and (**A4**), the potential $\mathscr{F}_n$ satisfies the*

Itô SDE

$$d\mathscr{F}_n(t) = \left[\left(b_1^2 \partial_\kappa \mathscr{F} + b_2^2 \partial_\gamma \mathscr{F}\right) \left((1 - \frac{\phi}{\psi}) + \frac{\phi}{\psi} \left(\kappa \partial_\kappa \mathscr{F} + \gamma \partial_\gamma \mathscr{F}\right) \right) \right] \cdot dt + d\mathscr{M}_n(t) + E_n(t) \tag{2.2.27}$$

where we adopt the Stochastic differential notation $d\boldsymbol{Y}_t = f \cdot d\boldsymbol{V}_t + hdt$*, which is shorthand for*

$$\boldsymbol{Y}(t) - \boldsymbol{Y}(0) = \int_0^t f \cdot d\boldsymbol{V}_s + \int_0^t h(s) ds.$$

The remainder term satisfies $\sup_{t\in[0,1]} |E_n(t)| = O(n^{-1})$, and the Martingale terms $\mathscr{M}_n(t) \to 0$ in probability uniformly over $t \in [0,1]$. As a consequence, the in-probability limit potential $\mathscr{F}_\infty(t) := \lim_{n\to\infty} \mathscr{F}_n(t)$ satisfies the following deterministic Hamilton-Jacobi (HJ) PDE.

Proposition 9. *Under Assumptions* ***(A1)*** *and* ***(A4)****, the limiting potential* $\mathscr{F}_\infty$ *satisfies the following deterministic PDE*

$$\partial_t \mathscr{F}_\infty = \left(b_1^2 \partial_\kappa \mathscr{F}_\infty + b_2^2 \partial_\gamma \mathscr{F}_\infty\right) \left[(1 - \frac{\phi}{\psi}) + \frac{\phi}{\psi} \left(\kappa \partial_\kappa \mathscr{F}_\infty + \gamma \partial_\gamma \mathscr{F}_\infty\right) \right] \tag{2.2.28}$$

The passage from the finite SDE (2.2.27) to the PDE (2.2.28), closing the non-linear drift and interchanging limits with spatial derivatives, is discussed in the remark below and justified in Section 2.3. We solve the PDE (2.2.28) via the Method of Characteristics. The initial condition is given by the $t = 0$ potential, which is a pure linear kernel in the sample covariance $\boldsymbol{S}_p := \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$

$$\mathscr{F}(0, \kappa_0, \gamma_0) = \lim_{n\to\infty} \frac{1}{n} \log\det\left(\kappa_0 p^{-1} \boldsymbol{X}^\top \boldsymbol{X} + \gamma_0 \mathbf{I}_n\right). \tag{2.2.29}$$

where (κ_0, γ_0) are initial values of the $\big(\kappa(t), \gamma(t)\big)$ characteristic curves, chosen to reach NTK specific target values at $t = 1$.

$$\big(\kappa(1), \gamma(1)\big) = \left(b_1^2 \sigma_W^2, z + \tau^2 \sigma_W^2 (c_1 - b_1^2)\right)$$

Differentiating $\mathscr{F}_n(t,\kappa,\gamma)$ in κ and γ at $t=0$ yields the initial momenta

$$\partial_\kappa \mathscr{F}_n(0,\kappa_0,\gamma_0) = \lim_{n\to\infty} \frac{1}{n} \mathrm{Tr}\Big(p^{-1}\boldsymbol{X}^\top \boldsymbol{X} \big(\kappa_0 p^{-1}\boldsymbol{X}^\top \boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1}\Big) \tag{2.2.30}$$

$$\partial_\gamma \mathscr{F}_n(0,\kappa_0,\gamma_0) = \lim_{n\to\infty} \frac{1}{n} \mathrm{Tr}\big(\kappa_0 p^{-1}\boldsymbol{X}^\top \boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1}. \tag{2.2.31}$$

We first summarize the explicit solution to the PDE (2.2.28) in the following Theorem for general data assumptions, and then explain their connection to limit trace resolvents m_1^∞, m_2^∞, and m_β^∞. Following this theorem, we state the power-law specific trace resolvent characterization, which then allows for the derivation of scaling law in NTK regression.

General Data Assumptions

For the next Theorem, we make the following general assumption on the dataset $\boldsymbol{X}$ and isotropic signal $\boldsymbol{\beta} \sim \mathscr{N}(0, p^{-1}\mathbf{I}_p)$.

(**B1**) **Data (general).** Assume the columns of $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ satisfy, for every $\alpha \neq \beta \in [n]$

$$\big|||\boldsymbol{x}_\alpha||^2 - 1\big| \le \varepsilon_n, \qquad \big|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\big| \le \varepsilon_n$$

in high probability, where $\varepsilon_n n^{\frac{1}{4}} \to 0$. Moreover, assume the ESD of $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ converges weakly to a probability measure μ on $[0,\infty)$.

(**B2**) **Labels (isotropic)**: Assume labels are generated $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} \sim \mathscr{N}(0, p^{-1}\mathbf{I}_p)$ and $\mathrm{Var}(\varepsilon)_i \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0,\sigma_\varepsilon^2)$. Let $\boldsymbol{y} = (y_1,\dots,y_n) \in \mathbb{R}^n$ denote the training labels vector.

Assumption (**B1**) imposes an approximate orthonormality on the columns of $\boldsymbol{X}$, and ensures it has a limiting ESD. The same assumption was made in [FW20], while a similar assumption was made in [ALP19].

Before presenting the main result of this section, we introduce formulas and terms that will be used to characterize the prediction error of the NTK regression under the general assumption (**B1**).

Definition 10. *Denote the sample covariance matrix by* $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$*, and let* $\mu = \text{limspec}\big(\boldsymbol{S}_p\big)$ *be limiting spectral measure of* $\boldsymbol{S}_p$*. For parameters* $b, s > 0$*, define the following Stieltjes-type integrals*

$$G_k(b,s) = \int \frac{x^k \mu(dx)}{bx+s}, \quad H_k(b,s) = \int \frac{x^k \mu(dx)}{\big(bx+s\big)^2} \qquad k \in \mathbb{N} \tag{2.2.32}$$

and the function

$$\eta(b,s) = 1 - \frac{\phi}{\psi}\Big(b_2^2 G_0(b,s) + b_1^2 G_1(b,s)\Big) \tag{2.2.33}$$

For positive real numbers (κ, γ) *define* $\eta^* > 0$ *to be the unique positive solution to the fixed-point equation*

$$\eta^* = 1 - \frac{\phi}{\psi}\Big[b_2^2 G_0\big(\kappa_0(\eta^*), \gamma_0(\eta^*)\big) + b_1^2 G_1\big(\kappa_0(\eta^*), \gamma_0(\eta^*)\big)\Big] \tag{2.2.34}$$

where

$$\kappa_0 = \kappa_0(\eta) = b_1^2 + \frac{\kappa}{\eta} \tag{2.2.35}$$

$$\gamma_0 = \gamma_0(\eta) = b_1^2 + \frac{\gamma}{\eta} \tag{2.2.36}$$

Moreover, for the solutions κ_0, γ_0*, define the quantities*

$$G_k = G_k(\kappa_0, \gamma_0), \quad H_k = H_k(\kappa_0, \gamma_0) \tag{2.2.37}$$

$$\mathscr{H}_0 = b_2^2 H_0 + b_1^2 H_1, \quad \mathscr{H}_1 = b_2^2 H_1 + b_1^2 H_2 \tag{2.2.38}$$

With these formulas, we can introduce our main theoretical result, which is a novel characterization of the two trace functionals m_1 and m_2.

Theorem 1. *Define* $\kappa_{\text{target}} = b_1^2 \sigma_W^2$ *and* $\gamma_{\text{target}} = z + \tau^2 \sigma_W^2 (c_1 - b_1^2)$ *and let* $\eta^* > 0$ *be the unique*

positive solution to (2.2.34)*, where* κ_0, γ_0 *are defined as*

$$\kappa_0 = b_1^2 + \frac{\kappa_{\text{target}}}{\eta^*}, \qquad \gamma_0 = b_2^2 + \frac{\gamma_{\text{target}}}{\eta^*}. \tag{2.2.39}$$

Then under Assumptions ***(B1)****, and* ***(A1)*** *-* ***(A4)****, we have*

$$m_1(z) = \frac{G_0(\kappa_0, \gamma_0)}{\eta^*}, \qquad m_2(z) = \frac{G_1(\kappa_0, \gamma_0)}{\eta^*}. \tag{2.2.40}$$

Moreover, if the resolvent $\boldsymbol{R}(z) := \left(\frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X} + z\mathbf{I}_n\right)^{-1}$ *admits a deterministic equivalent* $\overline{\boldsymbol{R}}(z)$ *then*

$$m_1(z) \asymp \frac{1}{\kappa_0 \phi \eta^*} \overline{\boldsymbol{R}}\big(-\frac{\gamma_0}{\kappa_0 \phi}\big), \qquad m_2(z) \asymp \frac{1}{\kappa_0 \eta^*}\Big(1 - \frac{\gamma_0}{\kappa_0 \phi}\overline{\boldsymbol{R}}\big(-\frac{\gamma_0}{\kappa_0 \phi}\big)\Big). \tag{2.2.41}$$

Remark The fixed point equation (2.2.34) can be solved numerically using a standard root-finding algorithm, e.g brentq in SciPY [Bre73]. In the isotropic setting where μ is the Marchenko-Pastur law with aspect ratio ϕ, the fixed point equations become quartic with unique physical root. Moreover, by restricting to the isotropic setting the two trace functionals m_1, m_2 satisfy the coupled polynomial equations derived in [AP20] for $z \geq 0$. Note that by sending $\sigma_W \to 0$, which can be achieved by setting the initial condition $\kappa_0 = b_1^2$, we recover the pure RF kernel results of [ALP19] for m_1 and m_2 exactly.

Infinite Width KRR Limit

As an immediate application, we consider the infinite width $N \to \infty$ limiting kernel

$$\boldsymbol{K}^\infty := \mathbb{E}_{\boldsymbol{w}}\big[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})\big], \qquad \boldsymbol{w} \sim \mathcal{N}\big(0, p^{-1}\mathbf{I}_d\big). \tag{2.2.42}$$

To achieve this, we send $\frac{\phi}{\psi} = \frac{n}{N} \to 0$ by taking $\psi \to \infty$. The correct interpretation of this is the width N grows faster than any constant times n, p. The fixed-point solutions from the characteristic curves defined in Theorem 1 simplify to linear equations, and become explicitly invertible. By sending $\frac{\phi}{\psi} \to 0$ and $\sigma_W^2 \to 0$, we have the following fully explicit asymptotic

prediction risk for KRR.

Proposition 11 (KRR Risk)**.** *Let* $\hat{f}_\lambda^{\boldsymbol{K}}(\boldsymbol{x}) = \boldsymbol{K}^\infty(\boldsymbol{x},\boldsymbol{X})\big(\boldsymbol{K}^\infty + \lambda \mathbf{I}_n\big)^{-1}\boldsymbol{y}$ *denote the optimal kernel ridge predictor onto the kernel* $\boldsymbol{K}^\infty$*. Under Assumptions (**B1**)-(**B2**), and (**A1**), (**A4**), the asymptotic prediction risk* $\mathscr{R}^{\boldsymbol{k}}$ *of* $\hat{f}_\lambda^{\boldsymbol{K}}$ *is given by*

$$\mathscr{R}^{\boldsymbol{K}}(\lambda) = \frac{H_1(\lambda) + \sigma_\varepsilon^2 H_0(\lambda)}{G_0(\lambda)^2} \tag{2.2.43}$$

where

$$G_k(\lambda) = \int \frac{x^k \mu(dx)}{b_1^2 x + b_2^2 + \lambda}, \quad H_k(\lambda) = \int \frac{x^k \mu(dx)}{(b_1^2 x + b_2^2 + \lambda)^2}, \qquad k = 0, 1. \tag{2.2.44}$$

Proposition 11 is a limiting case of Theorem 2. The asymptotic risk (2.2.43) is fully explicit in terms of Stieltjes-type integrals against $\mu = \lim \operatorname{spec}(\boldsymbol{X}^\top \boldsymbol{X}/p)$, and requires no fixed-point or root solver.

Power-law Assumption

The following result is an application of Theorem 1 on the dual-power law model Assumption (**A0**), as well as a characterization of the signal-resolvent m_β.

Theorem 2 (Power-law Resolvent Traces)**.** *Suppose Assumptions (**A0**)-(**A4**) hold. Define the following*

$$D_z = z + \tau^2 \sigma_W^2 (c_1 - b_1^2). \tag{2.2.45}$$

For $\eta > 0$*, define the rational functions*

$$s(\eta) = \frac{b_1^2(\eta + \sigma_W^2)}{\eta\big(D_z - b_2^2\sigma_W^2\big)}\big(\phi\eta - \psi(1-\eta)(\eta + \sigma_W^2)\big), \qquad \zeta(\eta) = \frac{b_2^2\eta + D_z}{b_1^2\big(\eta + \sigma_W^2\big)}. \tag{2.2.46}$$

Let $\eta^* > 0$ be the unique positive solution to the fixed point equation

$$\Psi(\eta^*) = F_\alpha\big(s(\eta^*)\big) + \zeta(\eta^*)s(\eta^*) - \phi = 0, \qquad F_\alpha(s) = \frac{1}{p}\sum_{j=1}^{p} \frac{s}{s + j^\alpha}. \tag{2.2.47}$$

Then we have

$$m_1(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} s(\eta^*), \qquad m_2(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_\alpha\big(s(\eta^*)\big). \tag{2.2.48}$$

Moreover,

$$m_\beta(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_r(s(\eta^*)), \qquad F_r(s) = \sum_{j=1}^{p} \frac{s j^{-(\alpha+2r)}}{s j^{-\alpha} + 1}. \tag{2.2.49}$$

Remark. Several remarks are in order. Theorem 2 characterizes the resolvent traces as the momentum curves of the HJ PDE (2.2.28) $m_1 = p_\gamma, m_2 = p_\kappa$, with initial momenta $p_\gamma(0), p_\kappa(0)$ given by the deterministic equivalents of the sample covariance resolvent at $t = 0$. The signal weighted resolvent m_β arises as a rank-one augmented PDE, and inherits the same fixed-point scalar equation η^*.

The fixed point equation (2.2.47) has a unique physical root $\eta > 0$, corresponding to $s(\eta) > 0$, found via Newton's method with details deferred to Section 2.4. Theorem 2 thereby reduces the joint asymptotic problem of computing m_1, m_2 and m_β to a single scalar-root finding one, a strict simplification of the coupled-polynomial system in [AP20], and self-consistent equations of [MM19b].

The full rigorous justification of the identifications $m_1^\infty = m_1, m_2^\infty = m_2$ and $m_\beta^\infty = m_\beta$, and more generally passing solutions from the finite-dimensional Itô SDE to the limiting PDE, is of independent mathematical interest and beyond the scope of this work. We state this conjecture as an ansatz in Section 2.3, and refer the reader to the works by [JKN26] for a treatment of this conjecture in the square, self-adjoint matrix setting. From here on, we will refer to the relevant traces as m_1, m_2, m_β with the understanding the ansatz is assumed.

Train Error and Prediction Risk

Under Ansatz (**C1**) and (**C2**), the asymptotic training error and risk of NTK regression are fully determined by m_1, m_β along with the derivatives m_1', m_β' with respect to ridge parameter λ. Implicit differentiation of the fixed-point equation in Theorem 2 yield closed-form expressions for the derivatives.

Proposition 12 (Asymptotic Train Error and Prediction Risk)**.** *Under assumptions (**A0**)-(**A4**), suppose the GET Conjecture (**C1**) holds as well. Then, the training error is given by*

$$\mathrm{E}_{\text{train}}(\lambda) = -\lambda^2\big[\partial_\lambda m_\beta(\lambda) + \sigma_\varepsilon^2 \partial_\lambda m_1(\lambda)\big] + o_{n,\mathbb{P}}(1) \tag{2.2.50}$$

*Under the GCV consistency Conjecture (**C2**) the prediction risk admits a deterministic equivalent with bias-variance decomposition*

$$\mathscr{R}(\lambda) \asymp \mathscr{B}(\lambda;\alpha,r) + \mathscr{V}(\lambda;\alpha,r) \tag{2.2.51}$$

where the bias and variance terms are given by

$$\mathscr{B}(\lambda;\alpha,r) = -\frac{\partial_\lambda m_\beta(\lambda)}{m_1(\lambda)^2}, \qquad \mathscr{V}(\lambda;\alpha,r) = -\frac{\sigma_\varepsilon^2 \partial_\lambda m_1(\lambda)}{m_1(\lambda)^2} \tag{2.2.52}$$

where

$$\begin{aligned} m_1'(\lambda) &= m_1(\lambda)\Big(\frac{s'}{s} - \frac{1}{\eta^* + \sigma_W^2}\frac{\partial \eta^*}{\partial \lambda}\Big) \\ m_2'(\lambda) &= m_2(\lambda)\Big(\frac{F_\alpha'(s)s'}{F_\alpha(s)} - \frac{1}{\eta^* + \sigma_W^2}\frac{\partial \eta^*}{\partial \lambda}\Big) \\ m_\beta'(\lambda) &= m_\beta(\lambda)\Big(\frac{F_r'(s)s'}{F_r(s)} - \frac{1}{\eta^* + \sigma_W^2}\frac{\partial \eta^*}{\partial \lambda}\Big) \end{aligned} \tag{2.2.53}$$

and

$$\begin{aligned} s'(\lambda) = \frac{b_1^2}{\eta^*\big(D_\lambda - b_2^2\sigma_W^2\big)}\frac{\partial \eta^*}{\partial \lambda}\Big\{ &-\frac{\sigma_W^2}{\eta^*}\big(\phi\eta^* - \psi(1-\eta^*)(\eta^* + \sigma_W^2)\big) \\ &+(\eta^* + \sigma_W^2)\big(\phi + \psi(2\eta^* + \sigma_W^2 - 1)\big)\Big\} - \frac{s(\lambda)}{D_\lambda - b_2^2\sigma_W^2} \end{aligned} \tag{2.2.54}$$

Finally,

$$\frac{\partial \eta^*}{\partial \lambda} = -\frac{\big(F'_\alpha(s)+\zeta\big)s_\ell + \zeta_\ell s}{\big(F'_\alpha(s)+\zeta\big)s_e + \zeta_e s} \tag{2.2.55}$$

where we define the auxiliary terms

$$\begin{aligned} s_\ell &:= -\frac{s}{D_\lambda - b_2^2\sigma_W^2}, \quad \zeta_\ell := \frac{1}{b_1^2(\eta^*+\sigma_W^2)}, \quad \zeta_e := \frac{b_2^2\sigma_W^2 - D_\lambda}{b_1^2(\eta^*+\sigma_W^2)^2} \\ s_e &:= \frac{-b_1^2\sigma_W^2\big(\phi\eta^* - \psi(1-\eta^*)(\eta^*+\sigma_W^2)\big) + \eta^* b_1^2(\eta^*+\sigma_W^2)\big(\phi+\psi(2\eta^*+\sigma_W^2-1)\big)}{(\eta^*)^2\big(D_\lambda - b_2^2\sigma_W^2\big)} \end{aligned} \tag{2.2.56}$$

By optimizing the GCV-DE (2.2.51) in λ, we derive an implicit equation for the optimal ridge parameter λ^*. Note this is only valid under the further conjecture of the uniform consistency of the GCV estimator extending to $\lambda \in [0, \lambda_{\max}]$, as made in Conjecture (**C2**).

Proposition 13 (Optimal Ridge Parameter)**.** *Under the setup of Theorem 2, assume the GCV uniform consistency conjecture (**C2**) extends to $\lambda_{\min} = 0$. Then the optimal ridge parameter $\lambda^* := \operatorname{argmin}_{\lambda\geq 0}\mathscr{R}(\lambda)$ satisfies the following implicit equation*

$$\big(\sigma_\varepsilon^2 m_1''(\lambda^*) + m_\beta''(\lambda^*)\big)m_1(\lambda^*) - 2\big(\sigma_\varepsilon^2 m_1'(\lambda^*) + m_\beta'(\lambda^*)\big)m_1'(\lambda^*) = 0 \tag{2.2.57}$$

While Proposition 13 does not admit a closed form optimal ridge parameter, (2.2.57) can be solved numerically using standard bracketing methods.

2.2.8 Finite-width NTK scaling law under unnormalized power-law spectra

Throughout this subsection, for two positive sequences a_n, b_n we write $a_n \asymp b_n$ if there exist constants $0 < c < C < \infty$, independent of n, p, N and of the ridge parameter in the scaling window, such that $cb_n \leq a_n \leq Cb_n$.

We use the same Gaussian-equivalent NTK notation as in Section 2.2.2, but for unnormalized power-law spectra we replace the LeCun factor $p^{-1/2}$ by the trace normalization $T_p^{-1/2}$,

where

$$T_p := \operatorname{Tr}\boldsymbol{\Sigma} = \sum_{j=1}^{p} j^{-\alpha}.$$

Equivalently, every occurrence of $p^{-1}X^\top X$ in the HJ potential is replaced by

$$S_p := T_p^{-1}\boldsymbol{X}^\top \boldsymbol{X}.$$

This normalization keeps the pre-activations non-degenerate when $\sigma_j = j^{-\alpha}$.

Assumption 1 (Unnormalized dual power law)**.** *Let*

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_j), \qquad \sigma_j = j^{-\alpha}, \qquad \alpha > 1.$$

The training inputs are $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$*, with* $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p\times n}$*. The teacher is linear,* $f_\star(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$*, with*

$$\boldsymbol{\beta}_j = c_{p,\alpha,r} j^{-r}, \qquad c_{p,\alpha,r}^{-2} = \sum_{k=1}^{p} k^{-(\alpha+2r)}.$$

Thus $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$*. We assume* $\alpha + 2r > 1$ *to ensure* $f^* \in L^2$*.*

Let

$$\bar{\sigma}_{p,j} := \frac{\sigma_j}{T_p}, \qquad c_{\mathrm{NTK}} := b_1^2(1 + \sigma_W^2) > 0,$$

where b_1 and σ_W^2 are the constants from the Gaussian-equivalent NTK. For an effective ridge $\mu > 0$, define the effective dimension and the source-bias profile by

$$\mathscr{N}_p(\mu) := \sum_{j=1}^{p} \frac{c_{\mathrm{NTK}}\bar{\sigma}_{p,j}}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j} + \mu}, \tag{2.2.58}$$

$$\mathscr{B}_p(\mu) := \sum_{j=1}^{p} \sigma_j \boldsymbol{\beta}_j^2 \left(\frac{\mu}{c_{\mathrm{NTK}}\bar{\sigma}_{p,j} + \mu} \right)^2. \tag{2.2.59}$$

The constants c_{NTK} and T_p only affect multiplicative constants and not the exponents below.

Assumption 2 (Local HJ/GCV deterministic equivalent)**.** *Let $R_{n,N,p}(\lambda) = \mathbb{E}_x[(\hat{f}_\lambda(x) - f_\star(x))^2]$ denote the finite-n,p,N prediction risk in the sense of Section 2.2.1, i.e. without adding an independent test-label noise. Let*

$$\mu_\lambda := \lambda + \lambda_{\text{diag}}$$

be the effective ridge appearing in the HJ fixed point, where λ_{diag} is the deterministic diagonal shift in the Gaussian-equivalent NTK. Assume either that the diagonal-corrected kernel is used, so that $\lambda_{\text{diag}} = 0$, or that $\lambda_{\text{diag}} = o(\mu_{\lambda_\star})$ in the scaling window.

There exist constants $C_B, C_V \in (0,\infty)$, depending only on the NTK constants and the noise variance, such that uniformly for $\mu_\lambda \downarrow 0$ with $\mathscr{N}_p(\mu_\lambda) = o(n \wedge N)$ and $\mu_\lambda^{-1/\alpha} = o(p)$,

$$R_{n,N,p}(\lambda) = C_B \mathscr{B}_p(\mu_\lambda) + C_V \left(\frac{1}{n} + \frac{1}{N}\right) \mathscr{N}_p(\mu_\lambda) + o\left(\mathscr{B}_p(\mu_\lambda) + \left(\frac{1}{n} + \frac{1}{N}\right) \mathscr{N}_p(\mu_\lambda)\right). \quad (2.2.60)$$

Theorem 3 (Finite-width NTK scaling under unnormalized power-law spectra)**.** *Assume Assumptions 1 and 2. Define*

$$q_{\alpha,r} := \min\left\{2, \frac{\alpha + 2r - 1}{\alpha}\right\}, \qquad \delta_{n,N} := \frac{1}{n} + \frac{1}{N}.$$

Suppose $\alpha + 2r \neq 2\alpha + 1$, $\delta_{n,N} \to 0$, and

$$p \gg \delta_{n,N}^{-1/(\alpha q_{\alpha,r} + 1)}. \quad (2.2.61)$$

Then the optimally tuned finite-width NTK risk satisfies

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \delta_{n,N}^{\frac{\alpha q_{\alpha,r}}{\alpha q_{\alpha,r} + 1}}. \quad (2.2.62)$$

Moreover, any ridge parameter satisfying

$$\mu_{\lambda_\star} \asymp \delta_{n,N}^{\frac{\alpha}{\alpha q_{\alpha,r} + 1}} \quad (2.2.63)$$

attains the rate in (2.2.62).

In the non-saturated source regime $\alpha + 2r < 2\alpha + 1$,

$$q_{\alpha,r} = \frac{\alpha + 2r - 1}{\alpha}, \qquad \inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \delta_{n,N}^{\frac{\alpha+2r-1}{\alpha+2r}}, \qquad \mu_{\lambda_\star} \asymp \delta_{n,N}^{\frac{\alpha}{\alpha+2r}}. \tag{2.2.64}$$

In the saturated source regime $\alpha + 2r > 2\alpha + 1$,

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \delta_{n,N}^{\frac{2\alpha}{2\alpha+1}}, \qquad \mu_{\lambda_\star} \asymp \delta_{n,N}^{\frac{\alpha}{2\alpha+1}}. \tag{2.2.65}$$

Finally, if $N \asymp n^\ell$ *for some* $\ell > 0$*, then*

$$\delta_{n,N} \asymp n^{-\min\{1,\ell\}},$$

and hence

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp n^{-\theta_{\mathrm{NTK}}(\alpha,r,\ell)}, \qquad \theta_{\mathrm{NTK}}(\alpha, r, \ell) = \min\{1, \ell\} \frac{\alpha q_{\alpha,r}}{\alpha q_{\alpha,r} + 1}. \tag{2.2.66}$$

Remark (Logarithmic boundary). At the boundary $\alpha + 2r = 2\alpha + 1$,

$$\mathscr{B}_p(\mu) \asymp \mu^2 \log(1/\mu), \qquad \mathscr{N}_p(\mu) \asymp \mu^{-1/\alpha}.$$

Consequently,

$$\mu_{\lambda_\star} \asymp \left(\frac{\delta_{n,N}}{\log(1/\delta_{n,N})} \right)^{\alpha/(2\alpha+1)},$$

and

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \delta_{n,N}^{2\alpha/(2\alpha+1)} \left(\log \frac{1}{\delta_{n,N}} \right)^{1/(2\alpha+1)}.$$

2.3 Proofs of Main Results

2.3.0 Preliminaries

We first introduce the following auxiliary results, which will be used throughout the proofs presented in this chapter. The first is the ansatz introduced in Step 2. The log-determinant potential $\mathscr{F}_n(t)$ satisfies the finite SDE (2.2.21), plus vanishing-order terms. This is shown explicitly through the Itô calculations handled in Section 2.3.2. The ansatz is that the limit log-determinant $\mathscr{F}_\infty(t)$ is also solution to the deterministic limit PDE (2.2.22). We summarize this below.

(**C3**) **Ansatz.** The limit log-determinant potential $\mathscr{F}_\infty(t,\boldsymbol{u})$ is a solution to the deterministic PDE

$$\partial_t \mathscr{F}_\infty(t,\boldsymbol{u}) + H\left(\boldsymbol{u}, \partial_{\boldsymbol{u}} \mathscr{F}_\infty\right) \tag{2.3.1}$$

We will assume (**C3**) holds throughout the proofs, without explicitly stating it. The main technical tool needed to derive (2.2.21) is the multi-dimensional Itô's formula.

Proposition 14 (Multi-Dimensional Itô's Formula [Øks98])**.** *Let*

$$\boldsymbol{B}(t) = (B_1(t), \ldots, B_m(t))$$

be an $m-$dimensional vector of independent Brownian motions adapted to $\left(\mathscr{F}_t\right)_{t\geq 0}$. Define the vector-valued Itô process $d\boldsymbol{V}_t = \boldsymbol{U}_t dt + \boldsymbol{H}_t d\boldsymbol{B}_t$, where $\boldsymbol{U}_t \in \mathbb{R}^m$ and $\boldsymbol{H}_t \in \mathbb{R}^{n\times m}$ are progressively measurable with L^2 entries. Suppose $g : \mathbb{R}^m \to \mathbb{R}$ is a twice continuously differentiable function, then $\boldsymbol{Y}_t = g(\boldsymbol{V}_t)$ is a vector-valued Itô Process

$$d\boldsymbol{Y}_t = \sum_{i=1}^{m} \frac{\partial g}{\partial x_i}(\boldsymbol{V}_t) \cdot dV_{i,t} + \frac{1}{2} \sum_{i,j=1}^{m} \frac{\partial^2 g}{\partial x_i \partial x_j}(\boldsymbol{V}_t)\, dV_{i,t} \cdot dV_{j,t} \tag{2.3.2}$$

where

$$dB_i \cdot dB_j = \delta_{ij} dt, \text{ and } dB_i \cdot dt = dt \cdot dB_i = 0 \tag{2.3.3}$$

Note that the "stochastic differential" notation $d\boldsymbol{Y}_t = \boldsymbol{f} \cdot d\boldsymbol{V}_t + h\,dt$ as above, which we use throughout the calculations below, is shorthand for the stochastic integral equation

$$\boldsymbol{Y}(t) - \boldsymbol{Y}(0) = \int_0^t \boldsymbol{f} \cdot d\boldsymbol{V}_s + \int_0^t h\,ds \tag{2.3.4}$$

The next lemma is necessary to handle the martingale terms $\mathscr{M}_n(t)$ in Step 2.

Lemma 15. *Let $\{\mathscr{M}_n(t), t \in [0,T]\}$ be a sequence of continuous L^2-martingales with $\mathscr{M}_n(0) = 0$. Suppose the quadratic variation $\langle \mathscr{M} \rangle_t$ satisfies*

$$\big\langle \mathscr{M}_n \big\rangle_t \leq C\big(\frac{T}{n^2}\big) \tag{2.3.5}$$

almost surely, for some absolute constant $C > 0$. Then the following holds

$$\sup_{t \in [0,T]} \big|\mathscr{M}_n(t)\big| \xrightarrow{\mathbb{P}} 0. \tag{2.3.6}$$

In other words, $\mathscr{M}_n(t) \to 0$ in probability, uniformly over $t \in [0,T]$.

Proof of Lemma 15. This is a direct consequence of Doob's L^2 maximal inequality. Indeed, since $\mathscr{M}_n(t)$ is square integrable with $\mathscr{M}_n(0) = 0$, $|\mathscr{M}_n(t)|^2$ is a sub-Martingale and $\mathbb{E}[|\mathscr{M}_n(t)|^2] = \mathbb{E}[\langle \mathscr{M}_n \rangle_t]$. By Doob's maximal inequality, for any $\delta > 0$

$$\mathbb{P}\Big(\sup_{t \in [0,T]} |\mathscr{M}_n(t)| > \delta\Big) \leq \frac{\mathbb{E}[|\mathscr{M}_n(t)|^2]}{\delta^2} = \frac{\mathbb{E}[\langle \mathscr{M}_n \rangle_T]}{\delta^2} \leq C\Big(\frac{T}{\delta^2 n^2}\Big) \tag{2.3.7}$$

Therefore, setting $\delta = n^{-1+\varepsilon}$ for any $\varepsilon > 0$ concludes the proof.

□

As an aside, throughout the proofs we will impose the unnormalized scaling $[\boldsymbol{W}_1]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ on the first layer weight matrix $\boldsymbol{W}_1$, and track the $p^{-\frac{1}{2}}$ scaling as an explicit factor throughout all calculations. For notational simplicity, we also drop the $_1$ subscript and refer to $\boldsymbol{W}$ as the first layer weights.

2.3.1 Application: Anisotropic Sample Covariance

In order to introduce the HJ PDE approach, and verify its validity, we first demonstrate the approach on the simpler case of the log-determinant potential of the sample covariance $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$. In the case that $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$, with $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we recover the Stieltjes transform of the general Marchenko-Pastur law (Theorem 2.6 in [CL22]), as well as a population covariance resolvent functional. We make this data assumption $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$ explicit in the following assumption.

(**B3**) **Data (Anisotropic Gaussian).** Assume $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} \in \mathbb{R}^{p\times n}$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is symmetric, non-negative definite and $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Further, assume the ESD of $\boldsymbol{\Sigma}$ converges to a probability measure $\nu_{\boldsymbol{\Sigma}}$ on $[0,\infty)$.

The Marchenko-Pastur theorem holds for more general assumptions on $\boldsymbol{Z}$, but we restrict to the i.i.d Gaussian entry setting in order to embed $\boldsymbol{Z} \mapsto \boldsymbol{Z}(t)$ with Brownian motion entries.

Define the empirical Stieltjes transform $\tau^{(n)}(z) : \mathscr{H}^+ \to \mathbb{C}$ and its companion transform $\widetilde{\tau}^{(n)} : \mathscr{H}^+ \to \mathbb{C}$ as

$$\tau_1^{(n)}(z) = \frac{1}{p}\operatorname{Tr}\big(p^{-1}\boldsymbol{X}\boldsymbol{X}^\top + z\mathbf{I}_p\big)^{-1}, \qquad \widetilde{\tau}_1^{(n)}(z) = \frac{1}{n}\operatorname{Tr}\big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + z\mathbf{I}_n\big)^{-1} \tag{2.3.8}$$

Moreover, define the population-resolvent $\tau_2^{(n)}(z) : \mathscr{H}^+ \to \mathbb{C}$

$$\tau_2^{(n)}(z) = \frac{1}{p}\operatorname{Tr}\Big(\boldsymbol{\Sigma}\big(p^{-1}\boldsymbol{X}\boldsymbol{X}^\top + z\mathbf{I}_p\big)^{-1}\Big) \tag{2.3.9}$$

We state the main result of this section, which explicitly characterizes the limit of the trace

functionals $\tau_1^{(n)}$, $\widetilde{\tau}_1{}^{(n)}$, and $\tau_2^{(n)}$.

Proposition 16. *Let $n, p \to \infty$ with $n/p = \phi$, and define for $z > 0$*

$$\tau_1(z) = \lim_{n\to\infty} \mathbb{E}[\tau_1^{(n)}(z)], \qquad \tau_2(z) = \lim_{n\to\infty} \mathbb{E}[\tau_n^{(2)}(z)] \tag{2.3.10}$$

*Then under Assumption (**B3**),*

$$\tau_1(z) = \frac{1-\phi}{z} + s_\phi(z), \qquad \tau_2(z) = \frac{\phi}{z s_\phi(z)} - 1 \tag{2.3.11}$$

where $s_\phi(z)$ solves the fixed point equation

$$\phi - z s_\phi(z) = F_{\mathbf{\Sigma}}\big(s_\phi(z)\big), \qquad F_{\mathbf{\Sigma}}(s) = \int \frac{s\sigma}{s\sigma+1} d\nu_{\mathbf{\Sigma}}(\sigma) \tag{2.3.12}$$

Moreover,

$$\widetilde{\tau}_1(z) = \phi^{-1} s_\phi(z) \tag{2.3.13}$$

Note that (2.3.13) follows immediately from the companion Stieltjes transform relationship

$$\tau_1(z) = \phi \widetilde{\tau}_1(z) + \frac{1-\phi}{z} \tag{2.3.14}$$

Remark. As an immediate corollary, we specialize to the *scaled* power-law $\mathbf{\Sigma}_{jj} = (\frac{j}{p})^{-\alpha}$ for $\alpha \in (0,1)$. The scaling is chosen here to ensure the limit ESD $\nu_{\mathbf{\Sigma}}$ is a non-degenerate measure on the real line. Indeed, a straightforward Riemann-sum argument results in the fixed point equation

$$\phi - z s_\phi(z) = F_\alpha\big(s_\phi(z)\big), \qquad F_\alpha(s) = \int_0^1 \frac{s}{s + x^\alpha} dx \tag{2.3.15}$$

As a sanity check, we can send $\alpha \to 0^+$ in (2.3.15) to recover the isotropic Marchenko-

Pastur law. In this case, $F_0(s) = \frac{s}{s+1}$, and the fixed point equation (2.3.15) becomes

$$\phi - z s_\phi = \frac{s_\phi}{1 + s_\phi} \tag{2.3.16}$$

Equivalently,

$$z s_\phi^2(z) + (z + (1-\phi)) s_\phi(z) - \phi = 0 \tag{2.3.17}$$

Substituting $s_\phi(z) = \phi \widetilde{\tau}_1(z)$ recovers the usual Marchenko-Pastur fixed point equation. As a quick aside, while the *scaled* power-law covariance $\mathbf{\Sigma}$ does not satisfy the uniformly bounded operator norm condition of Assumption (**B3**), the proof below only requires $\limsup_{p\to\infty} \frac{1}{p} \operatorname{Tr} \mathbf{\Sigma} < \infty$, which is satisfied when $\alpha < 1$. This, in addition to the numerical experiments in Section 2.4, provides evidence towards the validity of ansatz (**C3**).

Proof of Proposition 16. To prove Proposition 16, we follow the steps outlined in Section 2.3.0. Step 1 involves embedding the Gaussian matrix $\boldsymbol{Z} \leftrightarrow \boldsymbol{Z}(t)$ with i.i.d Brownian motion entries, and applying Itô's formula (Proposition 14).

For a parameter $\omega > 0$, define the time-parameterized log-determinant potential of the sample covariance as

$$\mathscr{G}_p(t, \omega, z) = p^{-1} \log\det\Big(\mathbf{\Sigma}^{\frac{1}{2}} \big(\frac{\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top}{p} + \omega \mathbf{I}_p \big) \mathbf{\Sigma}^{\frac{1}{2}} + z \mathbf{I}_p \Big). \tag{2.3.18}$$

The choice to include this auxiliary parameter $\omega > 0$ is motivated by the observation

$$\frac{\partial \mathscr{G}_p}{\partial \omega}\Big|_{\omega=0} = \operatorname{tr}\big(\mathbf{\Sigma} \boldsymbol{R}(z) \big), \qquad \boldsymbol{R}(z) := \big(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{p} + z \mathbf{I}_p \big)^{-1}. \tag{2.3.19}$$

Indeed, this will allow the recovery of τ_2 as a partial derivative with respect to ω. Define the limit log-determinant as

$$\mathscr{G}_\infty(t, \omega, z) = \lim_{p\to\infty} p^{-1} \mathbb{E}\Big[\log\det\big(\mathbf{\Sigma}^{\frac{1}{2}} \big(\frac{\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top}{p} + \omega \mathbf{I}_p \big) \mathbf{\Sigma}^{\frac{1}{2}} + z \mathbf{I}_p \big) \Big] \tag{2.3.20}$$

With the following partial derivatives

$$\frac{\partial \mathscr{G}_\infty}{\partial \omega} = \lim_{n\to\infty} \mathbb{E}\Big[\mathrm{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(t,z)\big)\Big] \tag{2.3.21}$$

$$\frac{\partial \mathscr{G}_\infty}{\partial z} = \lim_{n\to\infty} \mathbb{E}\Big[\mathrm{tr}\big(\boldsymbol{R}_\omega(t,z)\big)\Big] \tag{2.3.22}$$

where we define

$$\boldsymbol{R}_\omega(t,z) := \Big(\boldsymbol{\Sigma}^{\frac{1}{2}}\big(p^{-1}\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top + \omega \mathbf{I}_p\big)\boldsymbol{\Sigma}^{\frac{1}{2}} + z\mathbf{I}_p\Big)^{-1}. \tag{2.3.23}$$

Let $\boldsymbol{C} = (c_{ij})$ be an auxiliary matrix of the same shape as $\boldsymbol{Z}(t)$, and define the function $f : \mathbb{R}^{p\times n} \to \mathbb{R}$ by

$$f\big(\mathrm{vec}(c_{ij})\big) = p^{-1}\log\det\Big(\boldsymbol{\Sigma}^{\frac{1}{2}}\big(\frac{\boldsymbol{C}\boldsymbol{C}^\top}{p} + \omega\mathbf{I}_p\big)\boldsymbol{\Sigma}^{\frac{1}{2}} + z\mathbf{I}_p\Big) \tag{2.3.24}$$

with $f\big(\mathrm{vec}(\boldsymbol{Z}_{ij}(t))\big) = \mathscr{G}_p(t,\omega,z)$. We proceed by deriving a finite dimensional SDE that is satisfied by (2.3.18).

Step One: Finite Dimensional Stochastic PDE

The first step is to compute first and second order partial derivatives of f with respect to the entries of c_{ij}. For notation, let $\boldsymbol{R}_\omega(z)$ denote (2.3.23) with the auxiliary matrix $\boldsymbol{C}$ in place of $\boldsymbol{Z}(t)$.

We compute first order partial derivatives with respect to the entries c_{ij} of $\boldsymbol{C}$. We will make repeated use of the following matrix calculus rules. For a symmetric matrix $\boldsymbol{Y}$, whose entries possibly depend on some variable x

$$\frac{\partial}{\partial x}\log\det\big(\boldsymbol{Y}+\lambda\mathbf{I}\big) = \mathrm{Tr}\Big(\big(\boldsymbol{Y}+\lambda\mathbf{I}\big)^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\Big) \tag{2.3.25}$$

$$\frac{\partial}{\partial \lambda}\log\det\big(\boldsymbol{Y}+\lambda\mathbf{I}\big) = \mathrm{Tr}\big(\boldsymbol{Y}+\lambda\mathbf{I}\big)^{-1} \tag{2.3.26}$$

Moreover, if $\boldsymbol{Y}$ is invertible.

$$\frac{\partial}{\partial x}\boldsymbol{Y}^{-1} = -\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\boldsymbol{Y}^{-1} \tag{2.3.27}$$

First, by applying the cyclic property of the trace and some algebra, we find that

$$\frac{\partial f}{\partial c_{ij}} = 2p^{-1}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{C}\boldsymbol{E}_{ji}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big) \tag{2.3.28}$$

Collecting all Itô integral terms as a martingale, define

$$\mathscr{M}_n^0(t) = \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{\partial f}{\partial c_{ij}}(\boldsymbol{Z}_{ij}(t))\cdot d\boldsymbol{Z}_{ij,t} = \frac{2}{p^2}\sum_{ij}\Big(\boldsymbol{\Sigma}^{\frac{1}{2}}R_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}_{ij}(t)\Big)_{ij} \tag{2.3.29}$$

We verify explicitly that $\mathscr{M}_n^0(t) \xrightarrow{\mathbb{P}} 0$ uniformly over $t \in [0,1]$ in Proposition 18 and Lemma 15.

Next, we compute second order partial derivatives for the drift term in Itô's formula. For notation, define the matrix

$$M_2^{(c)} = \boldsymbol{C}\boldsymbol{E}_{ji} \tag{2.3.30}$$

Applying the third derivative rule (2.3.27) to the resolvent $\boldsymbol{R}_\omega$,

$$\frac{\partial}{\partial c_{ij}}\boldsymbol{R}_\omega(z) = -\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\big(\frac{\boldsymbol{E}_{ij}\boldsymbol{C}^\top}{p} + \frac{\boldsymbol{C}\boldsymbol{E}_{ji}}{p}\big)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z) \tag{2.3.31}$$

Next, by differentiating (2.3.28) with respect to c_{ij}, we have

$$\begin{aligned}\frac{\partial^2 f}{\partial c_{ij}\partial c_{ij}} = &-2p^{-1}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}p^{-1}\big(\boldsymbol{M}_2^{(c)} + (\boldsymbol{M}_2^{(c)})^\top\big)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}_2^{(c)}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big)\\ &+2p^{-1}\operatorname{tr}\big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{E}_{ii}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big)\end{aligned} \tag{2.3.32}$$

We now sum each term in (2.3.32) over the entries c_{ij} of $\boldsymbol{C}$. First,

$$\begin{aligned}&\sum_{i=1}^{p}\sum_{j=1}^{n}2p^{-2}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}_2^{(c)}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}_2^{(c)}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big)\\ &\qquad = 2p^{-1}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\frac{\boldsymbol{C}\boldsymbol{C}^\top}{p}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}\Big).\end{aligned} \tag{2.3.33}$$

Observe (2.3.33) is the normalized trace of matrices that are each normalized to have $O(1)$

entries. Therefore, the p^{-1} factor will cause this term to vanish in the asymptotic limit. Indeed, for a symmetric psd matrix $\boldsymbol{A} \in \mathbb{R}^n$, the resolvent bound gives

$$\|\boldsymbol{A} + z\mathbf{I}_n\| \leq \frac{1}{\operatorname{dist}\big(-\mathrm{z}, \operatorname{spec}(\boldsymbol{A})\big)} \leq \frac{1}{|\operatorname{Re}(z)|} \tag{2.3.34}$$

Evaluating (2.3.33) at $\boldsymbol{C} = \boldsymbol{Z}(t)$, and applying the resolvent bound we have

$$\|2p^{-1}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\frac{\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top}{p}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}\| \leq \frac{2p^{-1}}{z^2}\|\boldsymbol{\Sigma}\|^2\|p^{-1}\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top\| = O(p^{-1}) \tag{2.3.35}$$

where the last equality follows from $\|p^{-1}\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top\| = t \cdot O(1)$ with high probability, uniformly over $t \in [0,1]$.

Next,

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=1}^{n} 2p^{-2}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{M}_2^{(c)})^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}_2^{(c)}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big) \\ = 2\operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big)\cdot\operatorname{tr}\Big(\boldsymbol{\Sigma}^{\frac{1}{2}}\frac{\boldsymbol{C}\boldsymbol{C}^\top}{p}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{R}_\omega(z)\Big)\end{aligned} \tag{2.3.36}$$

Finally,

$$\sum_{i=1}^{p}\sum_{j=1}^{n} 2p^{-1}\operatorname{tr}\Big(\boldsymbol{R}_\omega(z)\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{E}_{ii}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big) = 2\frac{n}{p}\operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big) \tag{2.3.37}$$

With some further algebra, namely adding and subtracting both $\omega\mathbf{I}_p$ and $z\mathbf{I}_p$ in (2.3.36), we can express the finite$-p$ SDE as follows

$$\begin{aligned}d\mathscr{G}_p(t) = \mathscr{M}_n^0(t) + \Bigg[-\operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big)\Big(1 - \omega\operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big) - z\operatorname{tr}\big(\boldsymbol{R}_\omega(z)\big)\Big) \\ + \frac{n}{p}\operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big) + O(p^{-1})\Bigg]\cdot dt.\end{aligned} \tag{2.3.38}$$

Recognizing $\frac{\partial}{\partial\omega}\mathscr{G}_p(t,\omega,z) = \operatorname{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}_\omega(z)\big)$ and $\frac{\partial}{\partial z}\mathscr{G}_p(t,\omega,z) = \operatorname{tr}\boldsymbol{R}_\omega(z)$ allows us to build the asymptotic PDE, which we accomplish in the next step.

Step 2: Asymptotic Deterministic PDE Sending $\mathscr{M}_n^{(0)}(t) \to 0$ in probability, we then differen-

tiate (2.3.38) with respect to t. Controlling the vanishing-order terms, we obtain the following deterministic PDE

$$\frac{\partial \mathscr{G}}{\partial t} = -\frac{\partial \mathscr{G}}{\partial \omega}\big(1 - \omega \frac{\partial \mathscr{G}}{\partial \omega} - z\frac{\partial \mathscr{G}}{\partial z}\big) + \phi \frac{\partial \mathscr{G}}{\partial \omega}. \tag{2.3.39}$$

Under the ansatz (**C3**), the solution to (2.3.39) will recover the two limit traces τ_1, τ_2.

Step 3: Hamilton Jacobi Solution

We solve this PDE using Hamilton-Jacobi Theory and the Method of Characteristics. First, write our PDE (2.3.39) in the form

$$\frac{\partial \mathscr{G}}{\partial t} = \omega\big(\frac{\partial \mathscr{G}}{\partial \omega}\big)^2 + z\big(\frac{\partial \mathscr{G}}{\partial z} \cdot \frac{\partial \mathscr{G}}{\partial \omega}\big) - (1-\phi)\frac{\partial \mathscr{G}}{\partial \omega} \tag{2.3.40}$$

Define the Hamiltonian to be the right hand side of (2.3.40) with an overall negative sign, and partial derivatives with respect to ω, z replaced with corresponding momentum variables p_ω, p_z.

$$H(\omega, z, p_\omega, p_z) = -\big(\omega p_\omega^2 + z(p_z \cdot p_\omega)\big) + (1-\phi)p_\omega \tag{2.3.41}$$

Hamilton's equations define the following system of coupled ODEs

$$\frac{\partial \omega}{\partial t} = \frac{\partial H}{\partial p_\omega} = -2\omega p_\omega - z p_z + (1-\phi) \tag{2.3.42}$$

$$\frac{\partial z}{\partial t} = \frac{\partial H}{\partial p_z} = -z p_\omega \tag{2.3.43}$$

$$\frac{\partial p_\omega}{\partial t} = -\frac{\partial H}{\partial \omega} = p_\omega^2 \tag{2.3.44}$$

$$\frac{\partial p_z}{\partial t} = -\frac{\partial H}{\partial z} = p_z p_\omega \tag{2.3.45}$$

with arbitrary initial conditions $\omega(0) = \omega_0$ and $z(0) = z_0$ for ω and z. These determine the initial momenta through the limits (2.3.21) and (2.3.22). Under Assumption (**B3**), weak convergence

of the ESD of $\mathbf{\Sigma}$ implies

$$p_\omega(0) = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \frac{\sigma_i}{\omega_0\sigma_i + z_0} = \int \frac{\sigma d\nu_{\mathbf{\Sigma}}(\sigma)}{\omega_0\sigma + z_0} \tag{2.3.46}$$

$$p_z(0) = \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\kappa_0\sigma_i + z_0} = \int \frac{d\nu_{\mathbf{\Sigma}}(\sigma)}{\omega_0\sigma + z_0} \tag{2.3.47}$$

We make note of the following identity, which is immediate from the integral representations in (2.3.62) and (2.3.63).

$$\omega_0 p_\omega(0) + z_0 p_z(0) = 1. \tag{2.3.48}$$

The characteristic curve for p_ω is uncoupled, and can be solved immediately. We use its solution to solve the characteristic curve for p_z as well.

$$p_\omega(t) = \frac{p_\omega(0)}{1 - p_\omega(0)t} \tag{2.3.49}$$

$$p_z(t) = \frac{p_z(0)}{1 - p_\omega(0)t} \tag{2.3.50}$$

Note along characteristic curves, we have the following constant of motion.

$$\frac{\partial}{\partial t}(zp_Z(t)) = (-zp_\omega)p_z(t) + z(p_z p_\omega) = 0 \tag{2.3.51}$$

Thus, $zp(z) \equiv z_0 p_z(0)$ along characteristic curves. This allows us to solve the characteristic curve for $z(t)$ as well.

$$z(t) = z_0(1 - p_\omega(0)t) \tag{2.3.52}$$

On the other hand, plugging this constant of motion into (2.3.42) allows us to solve the resulting ODE using an integrating factor.

$$\omega(t) = (1 - tp_\omega(0))\Big[\omega_0\big(1 - tp_\omega(0)\big) + \big((1-\phi) - z_0 p_z(0)\big)t\Big] \tag{2.3.53}$$

Note that taking $\omega_0 = \phi t$ and applying (2.3.48), we have $\omega(t) = 0$. Moreover by (2.3.23), sending $\omega(t) \to 0$ is necessary to recover the two traces

$$\tau_1(z) = \lim_{n\to\infty} \mathbb{E}\big[\mathrm{tr}\big(\boldsymbol{R}(z)\big)\big], \qquad \tau_2(z) = \lim_{n\to\infty} \mathbb{E}\big[\mathrm{tr}\big(\boldsymbol{\Sigma}\boldsymbol{R}(z)\big)\big]. \tag{2.3.54}$$

Under $\omega(t) = \phi t$, the initial momentum variables become

$$\begin{aligned} p_\omega(0) &= \int \frac{\sigma d\nu_{\boldsymbol{\Sigma}}(\sigma)}{\phi t\sigma + z_0} \\ p_z(0) &= \int \frac{d\nu_{\boldsymbol{\Sigma}}(\sigma)}{\phi t + z_0}. \end{aligned} \tag{2.3.55}$$

Here z_0 is defined implicitly through the characteristic equation for $p(z)$, again with $\omega_0 = \phi t$.

$$z(t) = z_0\Big(1 - \int \frac{\sigma d\nu_{\boldsymbol{\Sigma}}(\sigma)}{\phi t\sigma + z_0}\Big) \tag{2.3.56}$$

Now plug in $t = 1$. For $z_0 > 0$ denote $s_\phi = \frac{\phi}{z_0}$, which is implicitly a function of z through $z_0 = z_0(z)$. We therefore take $s_\phi(z) = s_\phi(z_0)$ as the solution to the following

$$\phi - s_\phi(z)z = \int \frac{s_\phi(z)\sigma}{s_\phi(z)\sigma + 1} d\nu_{\boldsymbol{\Sigma}}(\sigma). \tag{2.3.57}$$

Under this, the initial momenta can be expressed as

$$p_\omega(0) = 1 - \frac{s_\phi(z)z}{\phi} \tag{2.3.58}$$

$$p_z(0) = \frac{s_\phi(z)}{\phi}\big(1 - \phi + s_\phi(z)z\big) \tag{2.3.59}$$

Plugging these into (2.3.55), we finally recover the two sample covariance trace functionals

$$\tau_1(z) = \frac{(1-\phi)}{z} + s_\phi(z) \tag{2.3.60}$$

$$\tau_2(z) = \frac{\phi}{s_\phi(z)z} - 1 \tag{2.3.61}$$

□

For the power-law covariance $\mathbf{\Sigma}_{jj} = (\frac{j}{p})^{-\alpha}$ specifically, (2.3.46) and (2.3.47) become

$$p_\omega(0) = \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \frac{1}{\alpha_0 + z_0(j/p)^\alpha} = \int_0^1 \frac{dx}{\omega_0 + z_0 x^\alpha} \tag{2.3.62}$$

$$p_z(0) = \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \frac{(j/p)^\alpha}{\alpha_0 + z_0(j/p)^\alpha} = \int_0^1 \frac{x^\alpha dx}{\omega_0 + z_0 x^\alpha} \tag{2.3.63}$$

The second equality arises from recognizing the two sums in (2.3.62) and (2.3.63) as Riemann-sums for $x_j = \frac{j}{p}$ on $[0,1]$.

2.3.2 Finite Dimension SDE

In this section, we derive a finite n, p, N Stochastic Differential Equation (SDE) for the log-determinant potential of the NTK. Following the Gaussian equivalent simplifications derived in 2.2.2, we follow the same general steps as in Section 2.3.1 and derive a SDE that models the trajectory of the time-parameterized NTK viewed as a function of multi-dimensional Brownian Motions.

First, $\boldsymbol{X}$ is independent from the other random matrices $\boldsymbol{W}$ and $\boldsymbol{\Theta}$. Therefore we condition on $\boldsymbol{X}$ and treat it as deterministic with respect to the Itô calculations made below. For a matrix $\boldsymbol{A}$ with Gaussian entries $\boldsymbol{A}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, let $\boldsymbol{A}(t)$ denote the matrix of the same shape with i.i.d

Brownian motion entries $\boldsymbol{A}_{ij}(t) \overset{\text{i.i.d.}}{\sim} \text{BM}(0,1)$. We then define $\boldsymbol{K}(t,\kappa,\gamma)$

$$\begin{aligned} \boldsymbol{K}(t,\kappa,\gamma) &= \boldsymbol{F}(t)^\top \boldsymbol{F}(t) + \kappa p^{-1} \boldsymbol{X}^\top \boldsymbol{X} + \gamma \mathbf{I}_n, \\ \boldsymbol{F}(t) &= \frac{1}{\sqrt{N}} (b_1 p^{-\frac{1}{2}} \boldsymbol{W}(t) \boldsymbol{X} + b_2 \boldsymbol{\Theta}(t)) \end{aligned} \tag{2.3.64}$$

where we introduce parameters κ, γ, with the explicit goal of substituting $\kappa_{\text{target}} = \sigma_W^2 b_1^2$ and $\gamma_{\text{target}} = z + \tau^2 \sigma_W^2 (c_1 - b_1^2)$. For fixed n, p, N, define the following

$$\mathscr{F}_n(t,\kappa,\gamma) = n^{-1} \log\det \boldsymbol{K}(t,\kappa,\gamma). \tag{2.3.65}$$

The main goal of this section is to prove the following proposition.

Proposition 17 (Finite dimension SDE)**.** *The potential $\mathscr{F}_n = \mathscr{F}_n(t,\kappa,\gamma)$ satisfies the following Itô SDE*

$$\begin{aligned} d\mathscr{F}_n = &\Bigg[b_1^2 (1 - \frac{n}{N}) \partial_\kappa \mathscr{F}_n + b_2^2 (1 - \frac{n}{N}) \partial_\gamma \mathscr{F}_n && (2.3.66) \\ &+ \frac{n}{N} \Big(\kappa b_1^2 \big(\partial_\kappa \mathscr{F}_n\big)^2 + \gamma b_2^2 \big(\partial_\gamma \mathscr{F}_n\big)^2 + (\kappa b_1^2 + \gamma b_2^2) \Big(\partial_\kappa \mathscr{F}_n \cdot \partial_\gamma \mathscr{F}_n \Big) \Big) \Bigg] \cdot dt && (2.3.67) \\ &+ d\mathscr{M}_n(t) + E_n(t) \end{aligned}$$

where $\mathscr{M}_n(t)$ is a collection of Martingale terms, which will vanish in probability uniformly over $t \in [0,1]$, and the remainder term satisfies $\sup_{t\in[0,1]} |E_n(t)| = O(n^{-1})$.

In our setting, we take $\boldsymbol{V}_t = \text{vec}\big(\boldsymbol{W}_{ij}(t), \boldsymbol{\Theta}_{ij}(t)\big)$ to be the vector of the $m = Np + Nn$ independent Brownian motions corresponding to entries of $\boldsymbol{W}(t)$ and $\boldsymbol{\Theta}(t)$. Applying the Stochastic Calculus rules (2.3.3), Itô's formula simplifies to the following.

$$d\boldsymbol{Y}_t = \sum_{c_{ij} \in [\boldsymbol{W}, \boldsymbol{\Theta}]} \frac{\partial g}{\partial c_{ij}} (\boldsymbol{V}_t) \cdot d\boldsymbol{V}_{c_{ij},t} + \frac{1}{2} \sum_{c_{ij} \in [\boldsymbol{W}, \boldsymbol{\Theta}]} \frac{\partial^2 g}{\partial c_{ij} \partial c_{ij}} (\boldsymbol{V}_t) \cdot dt \tag{2.3.68}$$

where the summations are taken over all entries of $\boldsymbol{W}, \boldsymbol{\Theta}$.

Proof of Proposition 17. Let w_{ij} and θ_{ij} denote the (i,j)-th entries of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ respectively. Then define the function $g : \mathbb{R}^m \to \mathbb{R}$ by $g\big(\mathrm{vec}(w_{ij}), \mathrm{vec}(\theta_{ij})\big) = \mathscr{F}_n$. In other words, g is a function of all $m = N(p+n)$ entries of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$.

To apply Itô's formula to $g(\boldsymbol{V}_t)$, we define the matrices $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{N\times p}$ and $\boldsymbol{B} = (b_{ij}) \in \mathbb{R}^{N\times n}$ as auxiliaries for $\boldsymbol{W}$ and $\boldsymbol{\Theta}$ respectively. Then by applying (2.3.68) to $g(\boldsymbol{V}_t)$, we obtain

$$\begin{aligned} d\mathscr{F}_n = &\Big[\sum_{i,j} \frac{\partial g}{\partial a_{ij}}(\boldsymbol{V}_t)\cdot d\boldsymbol{W}_{ij,t} + \frac{\partial g}{\partial b_{ij}}(\boldsymbol{V}_t)\cdot d\boldsymbol{B}_{ij,t}\Big] \\ &+\frac{1}{2}\sum_{i,j}\Big[\frac{\partial^2 g}{\partial a_{ij}\partial a_{ij}}(\boldsymbol{V}_t)\cdot dt + \frac{\partial^2 g}{\partial b_{ij}\partial b_{ij}}(\boldsymbol{V}_t)\cdot dt\Big]. \end{aligned} \tag{2.3.69}$$

Here the partial derivatives of g are taken with respect to the entries a_{ij}, b_{ij} of $\boldsymbol{A}$ and $\boldsymbol{B}$, and then evaluated at the entries of $\boldsymbol{W}(t)$ and $\boldsymbol{\Theta}(t)$ respectively. Furthermore, the summations in (2.3.69) are understood to be over all entries of the corresponding matrix.

To compute partial derivatives, we will make repeated use of the matrix derivative rules (2.3.25)-(2.3.27). For convenience, we restate them here. Let $\boldsymbol{Y}$ be a symmetric matrix, whose entries possibly depend on x, and let $\lambda > 0$.

$$\frac{\partial}{\partial x}\log\det(\boldsymbol{Y}+\lambda\mathbf{I}) = \mathrm{Tr}\Big((\boldsymbol{Y}+\lambda\mathbf{I})^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\Big) \tag{2.3.70}$$

$$\frac{\partial}{\partial \lambda}\log\det(\boldsymbol{Y}+\lambda\mathbf{I}) = \mathrm{Tr}(\boldsymbol{Y}+\lambda\mathbf{I})^{-1} \tag{2.3.71}$$

Moreover, if $\boldsymbol{Y}$ is invertible

$$\frac{\partial}{\partial x}\boldsymbol{Y}^{-1} = -\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial x}\boldsymbol{Y}^{-1} \tag{2.3.72}$$

Observe that by (2.3.70) and (2.3.71)

$$\frac{d}{d\kappa}\mathscr{F}_n(t,\kappa,\gamma) = \mathrm{tr}\big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{K}(t,\kappa,\gamma)^{-1}\big), \qquad \frac{d}{d\gamma}\mathscr{F}_n(t,\kappa,\gamma) = \mathrm{tr}\,\boldsymbol{K}(t,\kappa,\gamma)^{-1}. \tag{2.3.73}$$

Throughout the proof, let $\boldsymbol{E}_{ij}$ denote the matrix that is 0 everywhere except in the $(i,j)-th$

entry where it is 1. Further, let $\mathrm{Tr}(\cdot)$ denote the trace operator, and $\mathrm{tr}(\boldsymbol{M}) = \frac{1}{n}\mathrm{Tr}(\boldsymbol{M})$ denote the normalized trace of a square matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ by its dimension. For notational simplicity, we define the parameters

$$a = b_1^2(Np)^{-1}, \quad b = b_1 b_2 N^{-1} p^{-\frac{1}{2}}, \text{ and } \quad c = b_2^2 N^{-1} \tag{2.3.74}$$

Then denote the following

$$g\big(\mathrm{vec}(a_{ij}), \mathrm{vec}(b_{ij})\big) = n^{-1}\log\det\Big(\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma\mathbf{I}_n\Big) \tag{2.3.75}$$

where we define

$$\boldsymbol{\Phi} = a(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X} + b\Big((\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{B} + \boldsymbol{B}^\top(\boldsymbol{A}\boldsymbol{X})\Big) + c\boldsymbol{B}^\top\boldsymbol{B}. \tag{2.3.76}$$

The first step is to compute the partial derivatives $\frac{\partial g}{\partial a_{ij}}$ and $\frac{\partial g}{\partial b_{ij}}$.

Step 1: Compute first order partial derivatives By (2.3.70), we have

$$\frac{\partial g}{\partial a_{ij}} = \mathrm{tr}\big((\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma\mathbf{I}_n\big)^{-1}\frac{\partial\boldsymbol{\Phi}}{\partial a_{ij}}\big) \tag{2.3.77}$$

where

$$\frac{\partial\boldsymbol{\Phi}}{\partial a_{ij}} = a\big(\boldsymbol{X}^\top\boldsymbol{E}_{ji}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{E}_{ij}\boldsymbol{X}\big) + b\big(\boldsymbol{X}^\top\boldsymbol{E}_{ji}\boldsymbol{B} + \boldsymbol{B}^\top\boldsymbol{E}_{ij}\boldsymbol{X}\big) \tag{2.3.78}$$

By a similar calculation, we have

$$\frac{\partial g}{\partial b_{ij}} = \mathrm{tr}\big((\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma\mathbf{I}_n\big)^{-1}\frac{\partial\boldsymbol{\Phi}}{\partial b_{ij}}\big) \tag{2.3.79}$$

where

$$\frac{\partial \boldsymbol{\Phi}}{\partial b_{ij}} = b\big((\boldsymbol{AX})^\top \boldsymbol{E}_{ij} + \boldsymbol{E}_{ji}\boldsymbol{AX}\big) + c\big(\boldsymbol{E}_{ji}\boldsymbol{B} + \boldsymbol{B}^\top \boldsymbol{E}_{ij}\big). \tag{2.3.80}$$

We now collect terms

$$\mathscr{M}_n^{(1)}(t) := \sum_{a_{ij}\in[\boldsymbol{A}]} \frac{\partial g}{\partial a_{ij}}(\boldsymbol{V}(t)) \cdot d\boldsymbol{W}_{ij}(t) \tag{2.3.81}$$

$$\mathscr{M}_n^{(2)}(t) := \sum_{b_{ij}\in[\boldsymbol{B}]} \frac{\partial g}{\partial b_{ij}}(\boldsymbol{V}(t)) \cdot d\boldsymbol{\Theta}_{ij}(t) \tag{2.3.82}$$

Observe both $\mathscr{M}_1^{(n)}(t)$ and $\mathscr{M}_2^{(n)}(t)$ are finite sums of Itô-integrals against Brownian motions. It is classical that Itô-integrals against Brownian Motions are martingales, provided that the integrand is square integrable and adapted.

We will verify both $\mathscr{M}_1^{(n)}(t)$ and $\mathscr{M}_2^{(n)}(t)$ are indeed square integrable, and therefore $\mathscr{M}_n(t) := \mathscr{M}_n^{(1)}(t) + \mathscr{M}_n^{(2)}(t)$ will vanish in probability uniformly over $t \in [0,1]$, as shown in Proposition 18.

Step 2: Compute second order partial derivatives

Following the discussion above, the second order partial derivative terms in (2.3.69) are the only terms that will survive in the asymptotic limit. We proceed by differentiating both (2.3.77) and (2.3.79). For ease of notation, let us further define

$$\boldsymbol{\Omega}(\kappa,\gamma) := \big(\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma \mathbf{I}_n\big)^{-1} \tag{2.3.83}$$

By repeatedly applying (2.3.72) to (2.3.77), we have

$$\frac{\partial^2 g}{\partial a_{ij}\partial a_{ij}} = -\operatorname{tr}\left(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}\right) + \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial^2 \boldsymbol{\Phi}}{\partial a_{ij}\partial a_{ij}}\Big) \tag{2.3.84}$$

The second term is relatively straight-forward. Using $\boldsymbol{E}_{ji}\boldsymbol{E}_{ij} = \boldsymbol{E}_{jj}$, we find that

$$\frac{\partial^2 \boldsymbol{\Phi}}{\partial a_{ij}\partial a_{ij}} = 2a\boldsymbol{X}^\top \boldsymbol{E}_{jj}\boldsymbol{X} \tag{2.3.85}$$

In this case, the second term can be written as

$$\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial^2 \boldsymbol{\Phi}}{\partial a_{ij}\partial a_{ij}}\Big) = 2a\operatorname{tr}\Big(\boldsymbol{X}^\top \boldsymbol{E}_{jj}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.86}$$

The first term in (2.3.84) is more involved. To avoid cumbersome notation, we define

$$\boldsymbol{M}_1^{(a)} := \boldsymbol{X}^\top \boldsymbol{E}_{ji}\boldsymbol{A}\boldsymbol{X}, \qquad \boldsymbol{M}_2^{(a)} = \boldsymbol{X}^\top \boldsymbol{E}_{ji}\boldsymbol{B} \tag{2.3.87}$$

We can then write $\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}$ as

$$\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}} = a\big(\boldsymbol{M}_1^{(a)} + (\boldsymbol{M}_1^{(a)})^\top\big) + b\big(\boldsymbol{M}_2^{(a)} + (\boldsymbol{M}_2^{(a)})^\top\big) \tag{2.3.88}$$

Then, expanding both $\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}$ terms, and making use of various properties of the trace, we arrive at

$$-\operatorname{tr}\left(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial a_{ij}}\right) = 2a^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\Big) \tag{2.3.89}$$

$$+2a^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_1^{(a)})^\top\Big) + 2b^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\Big) \tag{2.3.90}$$

$$+2b^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(a)})^\top\Big) + 4ab\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\Big) \tag{2.3.91}$$

$$+4ab\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(a)})^\top\Big) \tag{2.3.92}$$

For second order partials with respect to b_{ij}, we perform a similar computation.

$$\frac{\partial^2 g}{\partial b_{ij}\partial b_{ij}} = -\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial b_{ij}}\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial \boldsymbol{\Phi}}{\partial b_{ij}}\Big) + \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial^2 \boldsymbol{\Phi}}{\partial b_{ij}\partial b_{ij}}\Big) \tag{2.3.93}$$

Again, the second term simplifies. Indeed, note that

$$\frac{\partial^2 \mathbf{\Phi}}{\partial b_{ij} \partial b_{ij}} = 2c \boldsymbol{E}_{jj} \tag{2.3.94}$$

Thus,

$$\operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\frac{\partial^2 \mathbf{\Phi}}{\partial b_{ij} \partial b_{ij}}\Big) = 2c \operatorname{tr}\Big(\boldsymbol{E}_{jj}\mathbf{\Omega}(\kappa,\gamma)\Big) \tag{2.3.95}$$

For notational brevity, let us also define the following matrices.

$$\boldsymbol{M}_1^{(z)} = \boldsymbol{E}_{ji}\boldsymbol{A}\boldsymbol{X}, \qquad \boldsymbol{M}_2^{(z)} = \boldsymbol{E}_{ji}\boldsymbol{B} \tag{2.3.96}$$

In this case, $\frac{\partial \mathbf{\Phi}}{\partial b_{ij}}$ can be written as

$$\frac{\partial \mathbf{\Phi}}{\partial b_{ij}} = b\big(\boldsymbol{M}_1^{(z)} + (\boldsymbol{M}_1^{(z)})^\top\big) + c\big(\boldsymbol{M}_2^{(z)} + (\boldsymbol{M}_2^{(z)})^\top\big) \tag{2.3.97}$$

We again expand both $\frac{\partial \mathbf{\Phi}}{\partial b_{ij}}$ in the first term of (2.3.93).

$$-\operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\frac{\partial \mathbf{\Phi}}{\partial b_{ij}}\mathbf{\Omega}(\kappa,\gamma)\frac{\partial \mathbf{\Phi}}{\partial b_{ij}}\Big) = 2b^2 \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\Big) \tag{2.3.98}$$

$$+2b^2 \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\mathbf{\Omega}(\kappa,\gamma)(\boldsymbol{M}_1^{(z)})^\top\Big) + 2c^2 \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\Big) \tag{2.3.99}$$

$$+2c^2 \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\mathbf{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(z)})^\top\Big) + 4bc \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\Big) \tag{2.3.100}$$

$$+4bc \operatorname{tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\mathbf{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(z)})^\top\Big) \tag{2.3.101}$$

Step 3: Construct the PDE

We are now ready to construct the PDE (2.3.66). Our approach can be outlined as follows: First sum all terms in (2.3.84) and (2.3.93) over the entries of $\boldsymbol{A}, \boldsymbol{B}$. We then collect any non-vanishing terms, and recognize them as partial derivatives with respect to the parameters (κ, γ) via (2.3.73). Finally, we recover the resulting SDE (2.3.66).

For the following calculations, we will make repeated use of the following Stochastic

Calculus rules. For an $n \times p$ Brownian motion matrix $\boldsymbol{B}$, and matrices $\boldsymbol{M}$ and $\boldsymbol{N}$ of appropriate sizes

$$(d\boldsymbol{B})^\top \boldsymbol{M}^\top \boldsymbol{M}(d\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{M}^\top \boldsymbol{M})\mathbf{I} \cdot dt \tag{2.3.102}$$

$$(d\boldsymbol{B})\boldsymbol{N}(d\boldsymbol{B}) = (d\boldsymbol{B})^\top \boldsymbol{N}(d\boldsymbol{B})^\top = \boldsymbol{N}^\top \cdot dt \tag{2.3.103}$$

We begin by summing each term in (2.3.89) over the indices $i \in [N]$, $j \in [p]$. By applying the element-wise versions of (2.3.102) and (2.3.103), and by the linearity of trace we have

$$\begin{aligned} &\sum_{i=1}^{N}\sum_{j=1}^{p} 2a^2 \,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\Big) \\ &\qquad = 2b_1^4 N^{-1}\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)(Np)^{-1}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X}\Big) \end{aligned} \tag{2.3.104}$$

Here we used $a = b_1^2(Np)^{-1}$. Note (2.3.104) is the normalized trace of matrices that are normalized to have order $O(1)$ entries, with an additional $O(N^{-1})$ prefactor. Therefore, an application of the resolvent bound (2.3.34) and a similar argument made in the proof of Proposition 16 case shows that this term vanishes uniformly over $t \in [0,1]$ as $n \to \infty$ under Assumption (**A1**). In fact, we will see that only products of two-trace terms will survive in the large n limit. These will correspond to terms computed via the rule (2.3.102), and are the finite-dimensional analogue of the following Free Brownian Motion stochastic calculus rule

$$(d\boldsymbol{b})\boldsymbol{a}(d\boldsymbol{b}) = 0 \tag{2.3.105}$$

and refer the reader to [DHK22] for more details on stochastic calculus. We proceed similarly with the other terms. Using the rule (2.3.102), we collect all non-vanishing order terms from

(2.3.89) as follows.

$$\sum_{i=1}^{N}\sum_{j=1}^{p} 2a^2 \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_1^{(a)})^\top\Big) \\ = 2b_1^4\frac{n}{N}\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\cdot\operatorname{tr}\Big((Np)^{-1}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.106}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p} 2b^2 \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(a)})^\top\Big) \\ = 2b_1^2b_2^2\frac{n}{N}\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\cdot\operatorname{tr}\Big(N^{-1}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.107}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p} 4ab \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(a)})^\top\Big) \\ = 4b_1^3b_2\frac{n}{N}\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\cdot\operatorname{tr}\Big(p^{-\frac{1}{2}}N^{-1}\boldsymbol{B}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.108}$$

After some algebra, namely adding and subtracting $\kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ and $\gamma\mathbf{I}_n$, these terms sum to the following expression.

$$2b_1^2\frac{n}{N}\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\Big[1-\kappa\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)-\gamma\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\Big)\Big] \tag{2.3.109}$$

We also sum (2.3.86) over $i\in[N], j\in[p]$.

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\frac{\partial^2\boldsymbol{\Phi}}{\partial a_{ij}\partial a_{ij}}\Big) = 2b_1^2\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.110}$$

On the other hand, we collect all vanishing order terms in (2.3.89) using (2.3.103).

$$\sum_{i=1}^{N}\sum_{j=1}^{p} 2b^2 \operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\Big) \\ = 2b_1^2b_2^2N^{-1}\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)N^{-1}\boldsymbol{B}^\top\boldsymbol{B}\Big) \tag{2.3.111}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p} 4ab\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\Big) \\ = 4b_1^3 b_2 N^{-1}\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)p^{-\frac{1}{2}}N^{-1}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{B}\Big) \tag{2.3.112}$$

Indeed, each term in (2.3.111) and (2.3.112) is the normalized trace of matrices normalized to have $O(1)$ entries. By the resolvent bound (2.3.34), and similar argument made in the sample covariance case, the N^{-1} prefactor forces both terms to be $O(N^{-1})$ in operator norm, uniformly in t. We collect these vanishing order terms under the remainder $E_n^{(1)}(t)$ In summary,

$$\frac{1}{2}\Big[\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2 g}{\partial a_{ij}\partial a_{ij}}(\boldsymbol{A},\boldsymbol{B})\Big] = -b_1^2\frac{n}{N}\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\times \tag{2.3.113}$$

$$\Big[1-\kappa\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)-\gamma\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\Big)\Big] \tag{2.3.114}$$

$$+b_1^2\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)+E_n^{(1)}(t) \tag{2.3.115}$$

We proceed with a similar computation for $\frac{\partial^2 g}{\partial b_{ij}\partial b_{ij}}$. First we collect all non-vanishing order terms from summing (2.3.98) over $i\in[N], j\in[n]$.

$$\sum_{i=1}^{N}\sum_{j=1}^{n} 2b^2\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(\gamma)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_1^{(z)})^\top\Big) \\ =2b_1^2b_2^2\frac{n}{N}\,\mathrm{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\cdot\mathrm{tr}\Big((Np)^{-1}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.116}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{n} 2c^2\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(z)})^\top\Big) \\ = 2b_2^4\frac{n}{N}\,\mathrm{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\cdot\mathrm{tr}\Big(N^{-1}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.117}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{n} 4bc\,\mathrm{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(z)}\boldsymbol{\Omega}(\kappa,\gamma)(\boldsymbol{M}_2^{(z)})^\top\Big) \\ = 4b_1b_2^3\frac{n}{N}\,\mathrm{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\cdot\mathrm{tr}\Big(N^{-1}p^{-\frac{1}{2}}\boldsymbol{B}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.118}$$

Again, after some algebra these terms can be combined as follows.

$$2\frac{n}{N}b_2^2\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\Big[1-\kappa\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)-\gamma\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\Big] \tag{2.3.119}$$

On the other hand, the rest of the terms from (2.3.98) are order $O(N^{-1})$, which we collect under $E_n^{(2)}(t)$.

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{n}2b_2^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(\gamma)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(\gamma)}\Big)\\ =2b_1^2b_2^2N^{-1}\operatorname{tr}\Big((Np)^{-1}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)^2\Big)\end{aligned} \tag{2.3.120}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{n}2c^2\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(z)}\Big)=2b_2^4N^{-1}\operatorname{tr}\Big(N^{-1}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\Omega}(\kappa,\gamma)^2\Big) \tag{2.3.121}$$

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{n}4bc\operatorname{tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_1^{(a)}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{M}_2^{(a)}\Big)\\ =4b_1b_2^3N^{-1}\operatorname{tr}\Big(N^{-1}p^{-\frac{1}{2}}(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{B}\boldsymbol{\Omega}(\kappa,\gamma)^2\Big)\end{aligned} \tag{2.3.122}$$

Finally, the sum of the last term from (2.3.93) over the indices of $\boldsymbol{B}$ is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{n}2c\operatorname{tr}\Big(\boldsymbol{E}_{jj}\boldsymbol{\Omega}(\kappa,\gamma)\Big)=2b_2^2\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big) \tag{2.3.123}$$

Collecting all terms from the above computations, we have

$$\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{n}\frac{\partial^2 g}{\partial b_{ij}\partial b_{ij}}(\boldsymbol{A},\boldsymbol{B})\right]=-\frac{n}{N}b_2^2\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\left[1-\kappa\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big)\right. \tag{2.3.124}$$

$$\left.-\gamma\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)\right]+b_2^2\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big)+E_n^{(2)}(t) \tag{2.3.125}$$

Now, we substitute $\boldsymbol{V}_t=\Big(\operatorname{vec}(\boldsymbol{W}_{ij}(t)),\operatorname{vec}(\boldsymbol{\Theta}_{ij}(t))\Big)$ into (2.3.113) and (2.3.124). We also apply the the partial derivatives 2.3.73

$$\frac{\partial}{\partial\kappa}\mathscr{F}_n=\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\Big),\qquad\frac{\partial}{\partial\gamma}\mathscr{F}_n=\operatorname{tr}\big(\boldsymbol{\Omega}(\kappa,\gamma)\big) \tag{2.3.126}$$

With some further algebra we arrive at the following SDE

$$d\mathscr{F}_n = \mathscr{M}_n(t) + \Bigg[b_1^2(1-\frac{n}{N})\frac{\partial}{\partial\kappa}\mathscr{F}_n + b_2^2(1-\frac{n}{N})\frac{\partial}{\partial\gamma}\mathscr{F}_n \tag{2.3.127}$$

$$+\frac{n}{N}\left(\kappa b_1^2\big(\frac{\partial}{\partial\kappa}\mathscr{F}_n\big)^2 + \gamma b_2^2\big(\frac{\partial}{\partial\gamma}\mathscr{F}_n\big)^2 + (\kappa b_2^2+\gamma b_1^2)\big(\frac{\partial}{\partial\kappa}\mathscr{F}_n\cdot\frac{\partial}{\partial\gamma}\mathscr{F}_n\big)\right)\Bigg]\cdot dt + E_n(t) \tag{2.3.128}$$

We complete the proof of Proposition 17 by controlling $\mathscr{M}_n(t)\to 0$ in probability uniformly in $t\in[0,1]$, differentiating the result with respect to t, and passing through the limit to recover the PDE (2.3.66). □

In order to build the deterministic asymptotic PDE, we first verify the Martingale terms $\mathscr{M}_n(t)$ vanish uniformly in probability. We state this as a proposition.

Proposition 18. *Let $\mathscr{M}_n(t)$ denote the Martingale terms in Propositions 16 and 17. Then the following holds*

$$\sup_{t\in[0,1]}\big|\mathscr{M}_n(t)\big| = \mathscr{O}_{d,\mathbb{P}}\big(n^{-1}\big) \tag{2.3.129}$$

In particular, $\mathscr{M}_n(t)\xrightarrow{\mathbb{P}}0$ uniformly over $t\in[0,1]$.

Proof of Proposition 18. We provide the full proof for the martingale terms in the PDE of main interest (2.3.66). The sample covariance case (2.3.29) follows from a similar, but simpler calculation and application of Lemma 15.

Recall the two Martingales (2.3.81) and (2.3.82)

$$\mathscr{M}_1^{(n)}(t) := \sum_{a_{ij}\in[\boldsymbol{A}]}\frac{\partial g}{\partial a_{ij}}(\boldsymbol{V}(t))\cdot d\boldsymbol{W}_{ij}(t), \quad \mathscr{M}_2^{(n)}(t) := \sum_{b_{ij}\in[\boldsymbol{B}]}\frac{\partial g}{\partial b_{ij}}(\boldsymbol{V}(t))\cdot d\boldsymbol{\Theta}_{ij}(t) \tag{2.3.130}$$

Both $\mathscr{M}_1^{(n)}(t)$ and $\mathscr{M}_2^{(n)}(t)$ are continuous L^2-martingales, with $\mathscr{M}_1^{(n)}(0)=\mathscr{M}_2^{(n)}(0)=0$. We compute the quadratic variation of $\mathscr{M}_1^{(n)}(t)$, as the computation is similar for $\mathscr{M}_2^{(n)}(t)$. Note the

entries $\boldsymbol{W}_{ij}(t)$ are standard BM's. Therefore, the quadratic variation of $\mathscr{M}_1^{(n)}(t)$ is given by

$$\left\langle \mathscr{M}_1^{(n)} \right\rangle_t = \sum_{(i,j)\in[N]\times p} \int_0^t \left(\partial_{a_{ij}} g\right)^2 ds \tag{2.3.131}$$

where $\partial_{a_{ij}} g$ denotes the partials $\frac{\partial g}{\partial a_{ij}}$ evaluated at $\boldsymbol{V}(s) = \big(\boldsymbol{W}_{ij}(s), \boldsymbol{\Theta}_{ij}(s)\big)$. We recall the first order partials $\frac{\partial g}{\partial a_{ij}}$

$$\frac{\partial g}{\partial a_{ij}} = 2a \operatorname{tr}\Big(\boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{E}_{ij} \boldsymbol{X} \boldsymbol{\Omega}(\kappa,\gamma)\Big) + 2b \operatorname{tr}\Big(\boldsymbol{B}^\top \boldsymbol{E}_{ij} \boldsymbol{X} \boldsymbol{\Omega}(\kappa,\gamma)\Big) \tag{2.3.132}$$

$$= \frac{2}{n}\Big[a\big(\boldsymbol{A}\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top\big)_{ij} + b\big(\boldsymbol{B}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top\big)_{ij}\Big] \tag{2.3.133}$$

Summing and squaring (2.3.133) over $(i,j) \in [N]\times[p]$ results in a Frobenius norm. Indeed evaluate (2.3.133) at $\boldsymbol{A} \mapsto \boldsymbol{W}(s)$ and $\boldsymbol{B} \mapsto \boldsymbol{\Theta}(s)$, and recall $\boldsymbol{F}(s) = \frac{1}{\sqrt{N}}\big(p^{-\frac{1}{2}} b_1 \boldsymbol{W}(s)\boldsymbol{X} + b_2 \boldsymbol{\Theta}(s)\big)$. Writing $a\boldsymbol{W}(s)\boldsymbol{X} + b\boldsymbol{\Theta}(s) = \frac{b_1}{\sqrt{Np}}\boldsymbol{F}(s)$, we then have

$$\begin{aligned}
\sum_{ij}\left(\partial_{a_{ij}} g\right)^2 &= \frac{4b_1^2}{n^2 Np}\left\|\boldsymbol{F}(s)\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top\right\|_F^2 \\
&= \frac{4b_1^2}{n^2 Np}\operatorname{Tr}\Big(\boldsymbol{F}(s)\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{F}(s)^\top\Big) \\
&= \frac{4b_1^2}{n^2 Np}\operatorname{Tr}\Big(\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Omega}(\kappa,\gamma)\boldsymbol{\Phi}(s)\Big)
\end{aligned} \tag{2.3.134}$$

where $\|\boldsymbol{A}\|_F = \sqrt{\operatorname{Tr}\big(\boldsymbol{A}^\top\boldsymbol{A}\big)}$ denotes the Frobenius matrix norm, and the last equality follow from the cyclic property of the trace and $\boldsymbol{\Phi}(s) = \boldsymbol{F}(s)^\top\boldsymbol{F}(s)$. We control the right hand side of (2.3.134) in two steps. First apply monotonicity of the trace under the Loewner partial ordering on symmetric real matrices, which we denote $\boldsymbol{A} \preceq \boldsymbol{B} \Leftrightarrow \boldsymbol{B}-\boldsymbol{A}$ is psd. Moreover, for two symmetric matrices $\boldsymbol{A} \prec \boldsymbol{B}$ and a psd $\boldsymbol{B} \succeq 0$ of appropriate sizes, we have $\operatorname{Tr}\big(\boldsymbol{A}\boldsymbol{C}\big) \le \operatorname{Tr}\big(\boldsymbol{B}\boldsymbol{C}\big)$. Indeed,

$$\operatorname{Tr}\Big(\big(\boldsymbol{B}-\boldsymbol{A}\big)\boldsymbol{C}\Big) = \operatorname{Tr}\Big(\boldsymbol{C}^{\frac{1}{2}}\big(\boldsymbol{B}-\boldsymbol{A}\big)\boldsymbol{C}^{\frac{1}{2}}\Big) \ge 0 \implies \operatorname{Tr}\big(\boldsymbol{B}\boldsymbol{C}\big) \ge \operatorname{Tr}\big(\boldsymbol{A}\boldsymbol{C}\big)$$

Next observe $\mathbf{\Phi}(s) \preceq \mathbf{\Omega}^{-1}$ uniformly in $s \in [0,t]$, as $\mathbf{\Omega}^{-1} = \mathbf{\Phi}(s) + \kappa p^{-1} \boldsymbol{X}^\top \boldsymbol{X} + \gamma \mathbf{I}_n$ is a sum of symmetric psd matrices plus a resolvent shift. Using $\mathbf{\Omega} \boldsymbol{X}^\top \boldsymbol{X} \mathbf{\Omega} = \big(\boldsymbol{X}\mathbf{\Omega}\big)^\top \big(\boldsymbol{X}\mathbf{\Omega}\big) \succeq 0$ and monotonicity of the trace, we then have

$$\mathrm{Tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top \boldsymbol{X}\mathbf{\Omega}(\kappa,\gamma)\mathbf{\Phi}(s)\Big) \leq \mathrm{Tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top \boldsymbol{X}\Big) \tag{2.3.135}$$

Similarly, we have

$$p^{-1}\kappa \boldsymbol{X}^\top \boldsymbol{X} \preceq \mathbf{\Omega}^{-1} \implies \boldsymbol{X}^\top \boldsymbol{X} \preceq \kappa^{-1} p \mathbf{\Omega}^{-1}$$

Combining these,

$$\mathrm{Tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top \boldsymbol{X}\mathbf{\Omega}(\kappa,\gamma)\mathbf{\Phi}(s)\Big) \leq \frac{p}{\kappa}\mathrm{Tr}\big(\mathbf{\Omega}(\kappa,\gamma)^{-1}\mathbf{\Omega}(\kappa,\gamma) = \frac{np}{\kappa} \tag{2.3.136}$$

Plugging this back into (2.3.134), we find

$$\begin{aligned} \big\langle \mathscr{M}_1^{(n)} \big\rangle_t &= \frac{4b_1^2}{n^2 N p}\int_0^t \mathrm{Tr}\Big(\mathbf{\Omega}(\kappa,\gamma)\boldsymbol{X}^\top \boldsymbol{X}\mathbf{\Omega}(\kappa,\gamma)\mathbf{\Phi}(s)\Big) \\ &<= \frac{4b_1^2}{n^2 N p}\int_0^t \frac{np}{\kappa} ds = \frac{4b_1^2}{Nn} t \leq C\big(\frac{T}{n^2}\big) \end{aligned} \tag{2.3.137}$$

uniformly for $t \in [0,T]$, where $C = \frac{4b_1^2 \phi}{\psi}$. A similar calculation applied to $\mathscr{M}_2^{(n)}$ shows

$$\big\langle \mathscr{M}_2^{(n)} \big\rangle_t \leq \tilde{C}\big(\frac{T}{n^2}\big) \tag{2.3.138}$$

for an absolute constant $\tilde{C} > 0$. We finally invoke Lemma 15 to conclude the proof. □

2.3.3 Asymptotic PDE

Define $\mathscr{F}_\infty(t,\kappa,\gamma) = \lim_{n\to\infty} \mathscr{F}_n(t,\kappa,\gamma)$, where the in-probability limit is taken under the proportional regime of Assumption (**A1**). In this section, we derive a PDE that uniquely

determines $\mathscr{F}_\infty(t,\kappa,\gamma)$ under the ansatz (**C3**). We summarize this result in the following proposition.

Proposition 19 (Asymptotic PDE)**.** *Suppose the general data assumption Assumption (**B1**), and Assumptions (**A1**), (**A4**) hold. Then $\mathscr{F}_\infty = \mathscr{F}_\infty(t,\kappa,\gamma)$ is the unique analytic solution to the following first-order, non-linear PDE*

$$\frac{\partial}{\partial t}\mathscr{F}_\infty = b_1^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial\kappa}\mathscr{F}_\infty + b_2^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial\gamma}\mathscr{F}_\infty \tag{2.3.139}$$

$$+\frac{\phi}{\psi}\left[\kappa b_1^2\big(\frac{\partial}{\partial\kappa}\mathscr{F}_\infty\big)^2+\gamma b_2^2\big(\frac{\partial}{\partial\gamma}\mathscr{F}_\infty\big)^2+(\kappa b_2^2+\gamma b_1^2)\big(\frac{\partial}{\partial\gamma}\mathscr{F}_\infty\cdot\frac{\partial}{\partial\kappa}\mathscr{F}_\infty\big)\right] \tag{2.3.140}$$

with the initial condition

$$\mathscr{F}_\infty(0,\kappa_0,\gamma_0) = \lim_{n\to\infty}\mathscr{F}_n(0,\kappa_0,\gamma_0) \tag{2.3.141}$$

Two remarks before the proof. The right hand side of PDE (2.3.139) is an analytic function in $\mathscr{F}_\infty, \partial_\kappa\mathscr{F}_\infty, \partial_\gamma\mathscr{F}_\infty$, and the Cauchy data (2.2.29) is analytic in (κ,γ). The Cauchy-Kovalevskaya (C-K) Theorem therefore guarantees the unique analytic solution to (2.3.139), given the initial conditions $\mathscr{F}_\infty(0) = \lim_{n\to\infty}\mathscr{F}_n(0,\kappa_0,\gamma_0)$. Next observe the partial derivatives of the finite order quantities (2.3.65) are given by

$$\begin{aligned}\frac{\partial}{\partial\gamma}\mathscr{F}_n(t,\kappa,\gamma) &= \operatorname{tr}\big(\boldsymbol{K}(t,\kappa,\gamma)^{-1}\big)\\ \frac{\partial}{\partial\kappa}\mathscr{F}_n(t,\kappa,\gamma) &= \operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{K}(t,\kappa,\gamma)^{-1}\Big)\end{aligned} \tag{2.3.142}$$

Both functions in (2.3.142) are jointly analytic in (κ,γ), when $\gamma = z+\tau^2\sigma_W^2\big(c_1-b_1^2\big)$ and $z\in\mathscr{H}^+$. As a consequence of Vitali's theorem from Complex Analysis, we can commute derivative and limit

$$m_1^\infty(t,\kappa,\gamma) = \lim_{n\to\infty}\frac{\partial}{\partial\gamma}\mathscr{F}_n(t,\kappa,\gamma), \qquad m_2^\infty(t,\kappa,\gamma) = \lim_{n\to\infty}\frac{\partial}{\partial\kappa}\mathscr{F}_n(t,\kappa,\gamma) \tag{2.3.143}$$

where the convergence is locally uniformly in probability, with

$$m_i^{\infty}(z) = m_i^{\infty}(1, \kappa_{\text{target}}, \gamma_{\text{target}}), \qquad i = 1,2$$

recovering the target NTK traces of Section 2.2.7.

Proof of Proposition 19. The PDE in Proposition 19 follows from first controlling the Martingale terms $\mathscr{M}_n$ of the finite-dimensional SDE (2.3.38) via Proposition 18. Writing the finite-dimensional SDE

$$d\mathscr{F}_n(t) = \mathscr{D}_n(t)dt + \mathscr{M}_n(t) + E_n(t)$$

By Proposition 18, $\mathscr{M}_n(t) \to 0$ in probability uniformly in $t \in [0,1]$. By the proof of Proposition 17, the remainder term converges to 0 uniformly in $t \in [0,1]$. Furthermore by (2.3.143), the gradients converge locally uniformly in probability. Since $\mathscr{D}_n(t)$ is a polynomial in $(\kappa, \gamma, \partial_\kappa, \partial_\gamma)$ the limits pass through the products. Bounded convergence then yields, the limit

$$d\mathscr{F}_\infty(t,\kappa,\gamma) = \Bigg[b_1^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial\kappa}\mathscr{F}_\infty + b_2^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial\gamma}\mathscr{F}_\infty \tag{2.3.144}$$

$$+ \frac{\phi}{\psi}\left(\kappa b_1^2\big(\frac{\partial}{\partial\kappa}\mathscr{F}_\infty\big)^2 + \gamma b_2^2\big(\frac{\partial}{\partial\gamma}\mathscr{F}_\infty\big)^2 + (\kappa b_2^2 + \gamma b_1^2)\big(\frac{\partial}{\partial\gamma}\mathscr{F}_\infty \cdot \frac{\partial}{\partial\kappa}\mathscr{F}_\infty\big)\right)\Bigg] \cdot dt \tag{2.3.145}$$

Differentiating (2.3.144) with respect to t recovers (2.3.139), concluding the proof. □

2.3.4 Hamilton-Jacobi System

The PDE (2.3.139) is a first-order non-linear PDE, which prove difficult to solve analytically in general. However, the solution of (2.3.139) can be analyzed by recognizing the PDE as a Hamilton-Jacobi (HJ) PDE, and applying the Method of Characteristics. We refer the reader to the book [Eva10], Chapter 3 for further background.

Before stating the main result of this section, we define some terms and set up the

Hamilton-Jacobi approach. Consider our PDE

$$\begin{aligned}\frac{\partial}{\partial t}\mathscr{F}_\infty &= b_1^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial \kappa}\mathscr{F}_\infty + b_2^2(1-\frac{\phi}{\psi})\frac{\partial}{\partial \gamma}S \\ &+ \frac{\phi}{\psi}\left[\kappa b_1^2\big(\frac{\partial}{\partial \kappa}\mathscr{F}_\infty\big)^2 + \gamma b_2^2\big(\frac{\partial}{\partial \gamma}\mathscr{F}_\infty\big)^2 + (\kappa b_2^2+\gamma b_1^2)\big(\frac{\partial}{\partial \gamma}\mathscr{F}_\infty \cdot \frac{\partial}{\partial \kappa}\mathscr{F}_\infty\big)\right]\end{aligned} \tag{2.3.146}$$

We define the Hamiltonian to be the right hand side of (2.3.146) with an overall negative sign and the replacement of each partial derivative $\partial_\kappa, \partial_\gamma$ with the corresponding momentum variables p_κ, p_γ.

$$\begin{aligned}H(\kappa,\gamma,p_\kappa,p_\gamma) &= -b_1^2(1-\frac{\phi}{\psi})p_\kappa - b_2^2(1-\frac{\phi}{\psi})p_\gamma \\ &- \frac{\phi}{\psi}\left[\kappa b_1^2\big(p_\kappa\big)^2 + \gamma b_2^2\big(p_\gamma\big)^2 + (\kappa b_2^2+\gamma b_1^2)\big(p_\gamma \cdot p_\kappa\big)\right]\end{aligned} \tag{2.3.147}$$

We now formulate (2.3.146) as a HJ PDE.

$$\frac{\partial}{\partial t}\mathscr{F}_\infty + H(\kappa,\gamma,p_\kappa,p_\gamma) = 0 \tag{2.3.148}$$

The main goal of this section is to derive the solution to (2.3.148) using the Method of Characteristics, under the general data assumption Assumption (**B1**).

Proposition 20 (HJ Solution)**.** *The functions $\gamma(t), \kappa(t), p_\kappa(t)$ and $p_\gamma(t)$ are given by the solutions to Hamilton's equations, which are the following coupled ordinary differential equations (ODEs).*

$$\frac{\partial \gamma}{\partial t} = \frac{\partial H}{\partial p_\gamma} = -b_2^2(1-\frac{\phi}{\psi}) - \frac{\phi}{\psi}\big[2b_2^2\gamma p_\gamma + (b_1^2\gamma + b_2^2\kappa)p_\kappa\big] \tag{2.3.149}$$

$$\frac{\partial \kappa}{\partial t} = \frac{\partial H}{\partial p_\kappa} = -b_1^2(1-\frac{\phi}{\psi}) - \frac{\phi}{\psi}\big[2b_1^2\kappa p_\kappa + (b_1^2\gamma + b_2^2\kappa)p_\gamma\big] \tag{2.3.150}$$

$$\frac{\partial p_\gamma}{\partial t} = -\frac{\partial H}{\partial \gamma} = \frac{\phi}{\psi}p_\gamma\big[b_2^2 p_\gamma + b_1^2 p_\kappa\big] \tag{2.3.151}$$

$$\frac{\partial p_\kappa}{\partial t} = -\frac{\partial H}{\partial \kappa} = \frac{\phi}{\psi}p_\kappa\big[b_2^2 p_\gamma + b_1^2 p_\kappa\big] \tag{2.3.152}$$

The initial conditions for κ and γ are chosen arbitrarily $\kappa(0) = \kappa_0, \gamma(0) = \gamma_0$. With these, the initial momenta are given by

$$p_\gamma(0) = \int \frac{\mu(dx)}{\kappa_0 x + \gamma_0}, \tag{2.3.153}$$

$$p_\kappa(0) = \int \frac{x\mu(dx)}{\kappa_0 x + \gamma_0} \tag{2.3.154}$$

where $\mu = \lim \operatorname{spec}(\boldsymbol{S}_p)$ is the limiting spectral measure of the sample covariance matrix $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$.

Define $p_0 = b_2^2 p_\gamma(0) + b_1^2 p_\kappa(0)$. Then for the initial conditions $\big(\gamma_0, \kappa_0, p_\gamma(0), p_\kappa(0)\big)$, the solutions to Hamilton's equations are given by

$$p_\gamma(t) = \frac{p_\gamma(0)}{1 - \frac{\phi}{\psi} p_0 t} \tag{2.3.155}$$

$$p_\kappa(t) = \frac{p_\kappa(0)}{1 - \frac{\phi}{\psi} p_0 t} \tag{2.3.156}$$

$$\gamma(t) = (\gamma(0) - b_2^2 t)(1 - \frac{\phi}{\psi} p_0 t) \tag{2.3.157}$$

$$\kappa(t) = (\kappa(0) - b_1^2 t)(1 - \frac{\phi}{\psi} p_0 t) \tag{2.3.158}$$

Finally, the solution $\mathscr{F}(t, \kappa, \gamma)$ to (2.3.148) *is given by*

$$\mathscr{F}(t, \kappa, \gamma) = \mathscr{F}(0, \kappa_0, \gamma_0) - p_0 t + \frac{\phi - \psi}{\phi}\Big(\ln\big|p_0^{-1} - \frac{\phi}{\psi} t\big| + \ln\big|p_0\big|\Big) \tag{2.3.159}$$

Proof of Proposition 20.

For notational simplicity, we introduce the notation $\dot{\boldsymbol{x}}(t) = \frac{\partial \boldsymbol{x}}{\partial t}$ in this proof. The characteristic curves (2.3.149)- (2.3.152) are a direct consequence of differentiating the Hamiltonian (2.3.147) with respect to $p_\gamma, p_\kappa, \gamma, \kappa$ respectively. These form a system of four coupled ODEs, which can be solved using elementary methods. We outline the calculations of the solutions of each ODE as follows.

The initial conditions for κ_0, γ_0 are to be chosen freely. These inform the initial conditions for the momentum curves (2.3.153), (2.2.30), which are a direct consequence of the weak convergence of the ESD of $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$ to the measure μ.

First, the momentum characteristic ODEs (2.3.151) and (2.3.152) uncouple from the characteristic curves for$(\kappa(t), \gamma(T))$, and can be solved directly.

Define the following affine transformation of p_γ, and p_κ as $p(t) = b_2^2 p_\gamma(t) + b_1^2 p_\kappa(t)$. By applying a linear transformation to (2.3.151) and (2.3.152) we find $p(t)$ satisfies the following ODE

$$\frac{\partial p}{\partial t} = \frac{\phi}{\psi} p(t)^2 \tag{2.3.160}$$

This is a separable first order ODE, and can be solved using elementary methods.

$$p(t) = \big(p_0^{-1} - \frac{\phi}{\psi} t\big)^{-1} \tag{2.3.161}$$

On the other hand, by adding (2.3.151) to (2.3.152) we have

$$\frac{\partial}{\partial t}\big(p_\gamma(t) + p_\kappa(t)\big) = \frac{\phi}{\psi} p(t)\big[p_\gamma(t) + p_\kappa(t)\big] \tag{2.3.162}$$

Substituting the solution for $p(t)$, this ODE is also straightforward to solve.

$$p_\gamma(t) + p_\kappa(t) = \frac{p_\gamma(0) + p_\kappa(0)}{p(0)} \big(p(0)^{-1} - \frac{\phi}{\psi} t\big)^{-1} \tag{2.3.163}$$

Equations (2.3.161) and (2.3.163) define a linear system for $p_\gamma(t), p_\kappa(t)$.

$$\begin{pmatrix} 1 & 1 \\ b_2^2 & b_1^2 \end{pmatrix} \begin{pmatrix} p_\gamma(t) \\ p_\kappa(t) \end{pmatrix} = p_0^{-1} \big(p_0^{-1} - \frac{\phi}{\psi} t\big)^{-1} \begin{pmatrix} p_\gamma(0) + p_\kappa(0) \\ p_0 \end{pmatrix} \tag{2.3.164}$$

We can invert (2.3.164), and perform some linear algebra to recover the solutions (2.3.155) and (2.3.156).

Plugging in the solutions for p_γ, p_κ into the coupled ODEs (2.3.149) and (2.3.150), we can define the system as a first-order matrix ODE.

$$\frac{\partial}{\partial t}\boldsymbol{x}(t) = -\frac{\phi}{\psi}\boldsymbol{P}(t)\boldsymbol{x}(t) - (1-\frac{\phi}{\psi})\begin{pmatrix} b_2^2 \\ b_1^2 \end{pmatrix} \tag{2.3.165}$$

where $\boldsymbol{x}(t) = \big(\kappa(t), \gamma(t)\big)^\top$, and

$$\boldsymbol{P}(t) = \begin{pmatrix} b_2^2 p_\gamma(t) + p(t) & b_2^2 p_\kappa(t) \\ b_1^2 p_\gamma(t) & b_1^2 p_\kappa(t) + p(t) \end{pmatrix} \tag{2.3.166}$$

$\boldsymbol{P}(t)$ diagonalizes, so we can write

$$\boldsymbol{P}(t) = \boldsymbol{V}(t)\begin{pmatrix} p(t) & 0 \\ 0 & 2p(t) \end{pmatrix}\boldsymbol{V}(t)^{-1}, \qquad \boldsymbol{V}(t) = \begin{pmatrix} -b_1^2 p_\kappa(t) & b_2^2 p_\gamma(t) \\ b_1^2 p_\gamma(t) & b_1^2 p_\gamma(t) \end{pmatrix} \tag{2.3.167}$$

In order to uncouple the ODE (2.3.165), define the transformation $\boldsymbol{y}(t) = \boldsymbol{V}(t)^{-1}\boldsymbol{x}(t)$. By differentiating and some algebra, we observe $\boldsymbol{y}(t)$ satisfies the following ODE

$$\frac{\partial}{\partial t}\boldsymbol{y}(t) = -\begin{pmatrix} p(t) & 0 \\ 0 & 2p(t) \end{pmatrix}\boldsymbol{y}(t) - \boldsymbol{V}(t)^{-1}\big(\frac{\partial}{\partial t}\boldsymbol{V}(t)\big)\boldsymbol{y} \tag{2.3.168}$$

We claim the second term of (2.3.168) is also diagonal, and therefore the ODE for $\boldsymbol{y}(t)$ uncouples. To see this, write out the product $\boldsymbol{V}(t)^{-1}\big(\frac{\partial}{\partial t}\boldsymbol{V}(t)\big)$

$$\frac{1}{b_1^2 p_\gamma(t) p(t)}\begin{pmatrix} b_1^4 p_\gamma(t)\big(\frac{\partial p_\kappa(t)}{\partial t}\big) + b_1^2 b_2^2 p_\gamma(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) & -b_1^2 b_2^2 p_\gamma(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) + b_1^2 b_2^2 p_\gamma(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) \\ -b_1^4 p_\gamma(t)\big(\frac{\partial p_\kappa(t)}{\partial t}\big) + b_1^4 p_\kappa(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) & b_1^2 b_2^2 p_\gamma(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) + b_1^4 p_\kappa(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) \end{pmatrix}$$

The $(1,2)-$ entry of this matrix immediately cancels to zero. On the other hand, entry $(2,1)$

vanishes as a consequence of the following relation

$$p_\gamma(t)\big(\frac{\partial p_\kappa(t)}{\partial t}\big) = p_\gamma(t)p_\kappa(t)p(t) = p_\kappa(t)\big(\frac{\partial p_\gamma(t)}{\partial t}\big) \tag{2.3.169}$$

After some further simplification to the diagonal entries, we find (2.3.168) simplifies to following

$$\dot{\boldsymbol{y}}(t) = \begin{pmatrix} -2\frac{\phi}{\psi}p(t) & 0 \\ 0 & -3\frac{\phi}{\psi}p(t) \end{pmatrix}\boldsymbol{y}(t) + \begin{pmatrix} 0 \\ (1-\frac{\phi}{\psi})\big(p_\gamma(t)\big)^{-1} \end{pmatrix} \tag{2.3.170}$$

This matrix ODE (2.3.170) decouples into two ODEs

$$\frac{\partial y_1(t)}{\partial t} = -2\frac{\phi}{\psi}p(t)y_1(t) \tag{2.3.171}$$

$$\frac{\partial y_2(t)}{\partial t} = -3\frac{\phi}{\psi}p(t)y_2(t) - \big(p_\gamma(t)\big)^{-1}(1-\frac{\phi}{\psi}) \tag{2.3.172}$$

with initial conditions given by $\boldsymbol{y}(0) = \boldsymbol{V}(0)^{-1}\boldsymbol{x}(0)$. The first ODE (2.3.171) is separable, and can immediately be solved.

$$y_1(t) = \frac{1}{p(0)^2}y_1(0)\big(p_0^{-1} - \frac{\phi}{\psi}t\big)^2 \tag{2.3.173}$$

The second ODE (2.3.172) can be solved using an integrating factor $u(t) = \big(p_0^{-1} - \frac{\phi}{\psi}t\big)^{-1}$, which results in

$$y_2(t) = \frac{\phi-\psi}{\phi}\big(\frac{p(0)}{p_\gamma(0)}\big)\big(p_0^{-1} - \frac{\phi}{\psi}t\big)^2 + \Big(p_0^3 y_2(0) + \frac{\phi-\psi}{\phi}\frac{p_0^2}{p_\gamma(0)}\Big)\big(p(0)^{-1} - \frac{\phi}{\psi}t\big)^3 \tag{2.3.174}$$

Translating back to $\boldsymbol{x}-$ coordinates, and performing some algebra, we recover the solutions

(2.3.157) and (2.3.158)

$$\boldsymbol{x}(t) = \boldsymbol{V}(t)\boldsymbol{y}(t) = \begin{pmatrix} (\gamma(0) - b_2^2 t)(1 - \frac{\phi}{\psi} p_0 t) \\ (\kappa(0) - b_1^2 t)(1 - \frac{\phi}{\psi} p_0 t) \end{pmatrix} \tag{2.3.175}$$

Finally, the solution (2.3.159) is a consequence of Proposition 6.3 in [DHK22] along with the observation of that the initial value of the Hamiltonian is $H_0 = -b_2^2 p_\gamma(0) - b_1^2 p_\kappa(0)$.

□

2.3.5 Proof of Theorem 2

In this section, we Prove Theorem 2, which we state below for convenience with a general population covariance $\boldsymbol{\Sigma}$.

Theorem 4 (Anisotropic Resolvent Traces)**.** *Suppose Assumptions (**B3**) and (**A1**)-(**A4**) hold, and let σ_j denote the eigenvalues of $\boldsymbol{\Sigma}$. Define the following*

$$D_z = z + \tau^2 \sigma_W^2 (c_1 - b_1^2) \tag{2.3.176}$$

For $\eta > 0$, define the rational functions

$$s(\eta) = \frac{b_1^2(\eta + \sigma_W^2)}{\eta\big(D_z - b_2^2\sigma_W^2\big)}\big(\phi\eta - \psi(1-\eta)(\eta + \sigma_W^2)\big), \qquad \zeta(\eta) = \frac{b_2^2\eta + D_z}{b_1^2\big(\eta + \sigma_W^2\big)} \tag{2.3.177}$$

Let $\eta^* > 0$ be the unique positive solution to the fixed point equation

$$\Psi(\eta^*) = F_{\boldsymbol{\Sigma}}\big(s(\eta^*)\big) + \zeta(\eta^*)s(\eta^*) - \phi = 0, \qquad F_{\boldsymbol{\Sigma}}(s) = \frac{1}{p}\sum_{j=1}^{p} \frac{s\sigma_j}{s\sigma_j + 1} \tag{2.3.178}$$

Then we have

$$m_1(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} s(\eta^*), \qquad m_2(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_{\boldsymbol{\Sigma}}\big(s(\eta^*)\big) \tag{2.3.179}$$

Moreover,

$$m_\beta(z) \asymp \frac{1}{\phi b_1^2(\eta^* + \sigma_W^2)} F_\beta(s(\eta^*)), \qquad F_\beta(s) = \sum_{j=1}^{p} \beta_j^2 \frac{s\sigma_j}{s\sigma_j + 1} \tag{2.3.180}$$

Proof of Theorem 4. The deterministic equivalents (2.3.179) are an application of Theorem 1 under the more specific data assumption (**A0**), with some minor modifications. The signal-resolvent m_β is then derived using an augmented HJ PDE, which arises from a rank-one perturbation of the kernel $\boldsymbol{K}(t,\gamma,\kappa)$ by $\boldsymbol{X}^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{X}$. Details of the derivation of m_β are provided in Section 2.3.5.

Recall the initial momentum conditions of the HJ (2.3.148) are given by

$$p_\gamma(0) = \lim_{n\to\infty} \operatorname{tr}\big(\kappa_0 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1} \tag{2.3.181}$$

$$p_\kappa(0) = \lim_{n\to\infty} \operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big(\kappa_0 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1} \tag{2.3.182}$$

Re-write (2.3.182) as

$$\operatorname{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big(\kappa_0 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1}\Big) = \frac{1}{\kappa_0} - \frac{\gamma_0}{\kappa_0}\big(\kappa_0 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0 \mathbf{I}_n\big)^{-1} \tag{2.3.183}$$

We recognize both (2.3.181) and (2.3.182) as trace functionals of the resolvent of the sample covariance $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ to μ. In the proof of Theorem 1, we invoked weak convergence of the ESD of $\boldsymbol{S}_p$ to the limiting measure μ, as both integrands in (2.3.181) and (2.3.182) are bounded and continuous away from the spectrum of $\boldsymbol{S}_p$.

However under the specific data assumption $\boldsymbol{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$ with $\boldsymbol{Z}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we can apply the general Marchenko-Pastur theorem (see Theorem 2.6 of [CL22]), which gives the

deterministic equivalent for the resolvent of the gram matrix $\widetilde{R}(z) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - z\mathbf{I}_n\right)^{-1}$ as the following.

$$\widetilde{R}(z) \asymp \widetilde{m}_p(z)\mathbf{I}_n, \qquad \widetilde{m}_p(z) = \left(-z + \frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I}_p + \widetilde{m}_p(z)\boldsymbol{\Sigma}\right)^{-1}\right)^{-1} \tag{2.3.184}$$

Writing $\left(\kappa_0\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0\mathbf{I}_n\right)^{-1} = \frac{1}{\kappa_0\phi}\widetilde{R}\left(-\frac{\gamma_0}{\kappa_0\phi}\right)$, (2.3.184) implies

$$\left(\kappa_0\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0\mathbf{I}_n\right)^{-1} \asymp \frac{s}{\phi\kappa_0}\mathbf{I}_n, \qquad s := \widetilde{m}_p(-\frac{\gamma_0}{\kappa_0\phi}) \tag{2.3.185}$$

where $s = s(\kappa_0, \gamma_0)$ solves the fixed-point equation

$$\frac{1}{s} = \frac{\gamma_0}{\kappa_0\phi} + \frac{1}{\phi p}\sum_{j=1}^{p}\frac{\sigma_j}{1+s\sigma_j} \tag{2.3.186}$$

Re-arranging (2.3.186) gives

$$\phi = \frac{\gamma_0}{\kappa_0}s + \frac{1}{p}\sum_{j=1}^{p}\frac{s\sigma_j}{s\sigma_j+1} \tag{2.3.187}$$

In particular, the equivalence $\asymp$ implies the almost sure convergence of the following

$$\left|\mathbb{E}[\operatorname{tr}\left(\kappa_0\boldsymbol{S}_p + \gamma_0\mathbf{I}_p\right)^{-1}] - \frac{s}{\phi\kappa_0}\right| \to 0 \tag{2.3.188}$$

$$\left|\mathbb{E}\left[\operatorname{tr}\left(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\left(\kappa_0 p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0\mathbf{I}_n\right)^{-1}\right)\right] - \frac{1}{\kappa_0}\left(1 - \frac{s\gamma_0}{\phi\kappa_0}\right)\right| \to 0 \tag{2.3.189}$$

Therefore, we can write

$$p_\gamma(0) \asymp \frac{s}{\phi\kappa_0}, \qquad p_\kappa(0) \asymp \frac{1}{\kappa_0} - \frac{s\gamma_0}{\phi\kappa_0^2} \tag{2.3.190}$$

where $s = s(\kappa_0, \gamma_0)$ solves (2.3.187).

Working with the finite$-p$ fixed point equation (2.3.187) rather than the limiting integral against $\nu_{\boldsymbol{\Sigma}}$ is advantageous for two reasons. First, it allows for a more explicit tracking of the

capacity exponent α in the formulas for m_1 and m_2. Second, the limit ESD of the unscaled power-law $\sigma_j = j^{-\alpha}$ collapses to a point mass at zero, and therefore loses the bulk structure that will determine the test error scaling.

Recall from the proof of Theorem 1, the solutions to the momentum curves p_γ, p_κ take the form

$$p_\gamma(t) = \frac{p_\gamma(0)}{1-\frac{\phi}{\psi}p(0)t}, \qquad p_\kappa(t) = \frac{p_\gamma(0)}{1-\frac{\phi}{\psi}p(0)t} \tag{2.3.191}$$

where

$$\kappa(t) = (\kappa_0 - b_1^2)(1-\frac{\phi}{\psi}p(0)t), \qquad \gamma(t) = (\gamma_0 - b_2^2)(1-\frac{\phi}{\psi}p(0)t) \tag{2.3.192}$$

Solving for target values $\kappa(1) = \kappa_{\text{target}} = b_1^2\sigma_W^2$ and $\gamma(1) = \gamma_{\text{target}} = z + \tau^2\sigma_W^2(c_1 - b_1^2)$, we choose initial values

$$\kappa_0 = b_1^2 + \frac{\kappa_{\text{target}}}{\eta}, \quad \gamma_0 = b_2^2 + \frac{\gamma_{\text{target}}}{\eta}, \qquad \eta := 1 - \frac{\phi}{\psi}p(0) \tag{2.3.193}$$

Now, under (2.3.190)

$$p(0) = b_2^2 p_\gamma(0) + b_1^2 p_\kappa(0) \asymp b_2^2\frac{s}{\phi\kappa_0} + b_1^2\big(\frac{1}{\kappa_0} - \frac{s\gamma_0}{\phi\kappa_0^2}\big) \tag{2.3.194}$$

Therefore, we define $\eta^* > 0$ to be the solution to the fixed point equation,

$$\eta = 1 - \frac{\phi}{\psi}\Big(b_2^2\frac{s}{\phi\kappa_0(\eta)} + \frac{b_1^2}{\kappa_0(\eta)}\big(1 - \frac{s\gamma_0(\eta)}{\phi\kappa_0(\eta)}\big)\Big), \qquad s = s\big(\kappa_0(\eta), \gamma_0(\eta)\big) \tag{2.3.195}$$

Isolating for $s = s(\eta)$ in (2.3.195), we have

$$s(\eta) = \frac{b_1^2(\eta+\sigma_W^2)}{\eta(D_z - b_2^2\sigma_W^2)}\big(\phi\eta - \psi(1-\eta)(\eta+\sigma_W^2)\big) \tag{2.3.196}$$

where $D_z = z + \tau^2\sigma_W^2(c_1 - b_1^2)$. Next, defining $\zeta(\eta) = \frac{\gamma_0(\eta)}{\kappa_0(\eta)}$, we have

$$\zeta(\eta) = \frac{\gamma_0(\eta)}{\kappa_0(\eta)} = \frac{b_2^2\eta + D_z}{b_1^2(\eta + \sigma_W^2)} \tag{2.3.197}$$

The fixed-point definition of s (2.3.187) then becomes

$$\zeta(\eta)s(\eta) + F_{\boldsymbol{\Sigma}}(s(\eta)) - \phi = 0, \qquad F_{\boldsymbol{\Sigma}}(s) = \frac{1}{p}\sum_{j=1}^{p}\frac{s\sigma_j}{s\sigma_j + 1} \tag{2.3.198}$$

which recovers the fixed-point equation $\Psi(\eta^*)$. To conclude the proof, we simply recognize

$$\kappa_0\eta^* = \phi b_1^2(\eta^* + \sigma_W^2) \tag{2.3.199}$$

□

As an immediate corollary when $\sigma_j = j^{-\alpha}$ is the power-law, and $\beta_j = j^{-r}$, we denote the two functionals $F_{\boldsymbol{\Sigma}}$ and $F_{\boldsymbol{\beta}}$ by F_α and F_β respectively, which are defined as

$$F_\alpha(s) = \frac{1}{p}\sum_{j=1}^{p}\frac{s}{s + j^\alpha}, \qquad F_r(s) = \sum_{j=1}^{p}\frac{sj^{-(\alpha+2r)}}{sj^{-\alpha} + 1}. \tag{2.3.200}$$

Thereby recovering the forms of m_1, m_2, m_β presented in Theorem 2.

Signal-resolvent m_β

In this section, we derive a deterministic equivalent for the Signal-resolvent $m_\beta^{(n)}(z)$. We state the main result of this section in the following proposition.

Proposition 21. *For $\boldsymbol{\beta} \in \mathbb{R}^p$ deterministic, such that $||\boldsymbol{\beta}|| = O(1)$, define*

$$m_\beta^{(n)}(z) = \frac{1}{n}\boldsymbol{\beta}^\top \boldsymbol{X}\left(\boldsymbol{K}^{\mathrm{NTK}} + z\mathbf{I}_n\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\beta} \tag{2.3.201}$$

Then under the set up of Theorem 4

$$m_\beta^{(n)}(z) \asymp \frac{1}{\phi\kappa_0\eta^*}F_\beta(s), \qquad F_\beta(s) = \sum_{j=1}^{p}\beta_j^2\frac{s\sigma_j}{s\sigma_j+1} \tag{2.3.202}$$

Proof of Proposition 21. Let $\xi > 0$, then consider the augmented log-determinant potential

$$\widetilde{\mathscr{F}_n}(t,\lambda,\kappa,\xi) = n^{-1}\log\det\Big(\boldsymbol{K}(t,\kappa,\gamma) + \xi\boldsymbol{X}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{X}\Big) \tag{2.3.203}$$

Note that setting $\xi = 0$ we recover our original log-determinant potential $\mathscr{F}_n$. Moreover, by taking partial derivatives

$$\begin{aligned}\partial_\xi\widetilde{\mathscr{F}_n} &= \operatorname{tr}\Big(\beta^\top\boldsymbol{X}\big(\boldsymbol{K}^{\mathrm{NTK}}(t,\kappa,\gamma) + \xi\boldsymbol{X}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\beta}\Big)\\ \partial_\xi\widetilde{\mathscr{F}_n}\big|_{\xi=0} &= m_\beta^{(n)}(t,\gamma,\kappa)\end{aligned} \tag{2.3.204}$$

where $m_\beta^{(n)}(t,\gamma,\kappa) = \frac{1}{n}\boldsymbol{\beta}^\top\boldsymbol{X}\boldsymbol{K}^{\mathrm{NTK}}(t,\kappa,\gamma)^{-1}\boldsymbol{\beta}^\top\boldsymbol{X}$. This motivates applying our HJ PDE approach to derive a limiting deterministic PDE for the augmented log-determinant $\widetilde{\mathscr{F}_n}$. The characteristic curves for ξ and its momentum variable $\widetilde{p}_\xi$ can then be used to recover the limit $m_\beta^\infty(z) = \lim_{n\to\infty}\mathbb{E}[m_\beta^{(n)}(z)]$.

Define $\boldsymbol{u} = \boldsymbol{X}^\top\boldsymbol{\beta} \in \mathbb{R}^n$, then $\xi\boldsymbol{X}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{X} = \xi\boldsymbol{u}\boldsymbol{u}^\top$ is a rank-one perturbation of $\boldsymbol{K}(t,\kappa,\gamma)$ in (2.3.203). This, combined with the fact that $\xi\boldsymbol{u}\boldsymbol{u}^\top$ is deterministic with respect to the Itô-calculations on the Brownian motion-embedded matrices made in the proof of Proposition 17 will greatly simplify the computations that follow. Indeed, by the matrix-determinant lemma we have

$$\det\big(\boldsymbol{K}(t,\kappa,\gamma) + \xi\boldsymbol{u}\boldsymbol{u}^\top\big) = \big(1 + \xi\boldsymbol{u}^\top\boldsymbol{K}(t,\kappa,\gamma)^{-1}\boldsymbol{u}\big)\det\big(\boldsymbol{K}(t,\kappa,\gamma)\big) \tag{2.3.205}$$

Taking the normalized log of (2.3.205) then results in

$$\widetilde{\mathscr{F}_n}(t,\kappa,\gamma) = \mathscr{F}_n(t,\kappa,\gamma) + \frac{1}{n}\log\big(1 + \xi\boldsymbol{u}^\top\boldsymbol{K}(t,\kappa,\gamma)^{-1}\boldsymbol{u}\big). \tag{2.3.206}$$

Denote the quadratic form by $\boldsymbol{Q} = \boldsymbol{Q}(t,\kappa,\gamma) = \boldsymbol{u}^\top \boldsymbol{K}(t,\kappa,\gamma)\boldsymbol{u}$. The momentum variables of the augmented potential $\widetilde{\mathscr{F}_n}$ are related to the corresponding original momentum variables of $\mathscr{F}_n$ as follows.

$$\widetilde{p}_\gamma = p_\gamma + \frac{1}{n}\frac{\xi\,\partial_\gamma \boldsymbol{Q}}{1+\xi\boldsymbol{Q}} \tag{2.3.207}$$

$$\widetilde{p}_\kappa = p_\kappa + \frac{1}{n}\frac{\xi\,\partial_\kappa \boldsymbol{Q}}{1+\xi\boldsymbol{Q}} \tag{2.3.208}$$

$$\widetilde{p}_\xi = \frac{1}{n}\frac{\boldsymbol{Q}}{1+\xi\boldsymbol{Q}} \tag{2.3.209}$$

Therefore, at $\xi = 0$ we recover the original momentum variables p_γ, p_κ

$$\widetilde{p}_\gamma|_{\xi=0} = p_\gamma, \quad \widetilde{p}_\kappa|_{\xi=0} = p_\kappa, \quad \widetilde{p}_\xi|_{\xi=0} = \frac{1}{n}\boldsymbol{Q} \tag{2.3.210}$$

Now define $p_\xi(t) = \widetilde{p}_\xi(t)|_{\xi=0}$. We proceed by finding the initial condition

$$p_\xi(0) = \lim_{n\to\infty} p_\xi^{(n)}(0)$$

where .

$$p_\xi^{(n)}(0) = \frac{1}{n}\operatorname{tr}\boldsymbol{\beta}^\top \boldsymbol{X}\big(\kappa_0\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X} + \gamma_0\mathbf{I}_n\big)^{-1}\boldsymbol{X}^\top\boldsymbol{\beta} \tag{2.3.211}$$

Recalling $\boldsymbol{S}_p = \frac{1}{p}\boldsymbol{X}\boldsymbol{X}^\top$, the push-through identity gives us

$$p_\xi^{(n)}(0) = \frac{p}{n}\boldsymbol{\beta}^\top \boldsymbol{S}_p\big(\kappa_0\boldsymbol{S}_p + \gamma_0\mathbf{I}_p\big)^{-1}\boldsymbol{\beta}. \tag{2.3.212}$$

Now, apply the algebraic identity $\boldsymbol{S}_p = \kappa_0^{-1}\Big((\kappa_0\boldsymbol{S}_p + \gamma_0\mathbf{I}_p) - \gamma_0\mathbf{I}_p\Big)$.

$$p_\xi^{(n)}(0) = \frac{p}{n}\kappa_0^{-1}\Big(||\boldsymbol{\beta}||^2 - \gamma_0\boldsymbol{\beta}^\top\big(\kappa_0\boldsymbol{S}_p + \gamma_0\mathbf{I}_p\big)^{-1}\boldsymbol{\beta}\Big) \tag{2.3.213}$$

Note $\boldsymbol{S}_p$ is the sample covariance of the matrix $\boldsymbol{X}$ with i.i.d $\mathscr{N}(0,\boldsymbol{\Sigma})$ columns. To determine

the asymptotic limit of the quadratic form $\boldsymbol{\beta}^\top \big(\kappa_0 \boldsymbol{S}_p + \gamma_0 \mathbf{I}_p\big)^{-1} \boldsymbol{\beta}$, we can again appeal to the deterministic equivalent form of the general Marchenko-Pastur Theorem. This time we use the $p \times p$ the resolvent $R(z) = \big(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\mathbf{I}_p\big)^{-1}$, which admits a deterministic equivalent

$$R(z) \asymp \bar{R}(z) := \big(\frac{\boldsymbol{\Sigma}}{1+\delta(z)} - z\mathbf{I}_p\big)^{-1}, \qquad \delta(z) = \frac{1}{n}\operatorname{Tr}\big(\boldsymbol{\Sigma}\bar{R}(z)\big). \tag{2.3.214}$$

Writing $\big(\kappa_0 \boldsymbol{S}_p + \gamma_0 \mathbf{I}_p\big)^{-1} = \frac{1}{\kappa_0\phi} R\big(-\frac{\gamma_0}{\kappa_0\phi}\big)$, this becomes

$$\big(\kappa_0 \boldsymbol{S}_p + \gamma_0 \mathbf{I}_p\big)^{-1} \asymp \frac{1}{\gamma_0}\big(s\boldsymbol{\Sigma} + \mathbf{I}_p\big)^{-1}, \qquad s := \frac{\kappa_0\phi}{\gamma_0(1+\delta)} \tag{2.3.215}$$

where $s = s(\kappa_0, \gamma_0)$ solves the same fixed point defined in (2.3.187).

$$\frac{1}{s} = \frac{\gamma_0}{\kappa_0\phi} + \frac{1}{n}\sum_{j=1}^{p} \frac{\sigma_j}{1+s\sigma_j} \tag{2.3.216}$$

In particular, this implies for deterministic $\boldsymbol{a} \in \mathbb{R}^n$ with $\|\boldsymbol{a}\|_2 = O(1)$,

$$\boldsymbol{a}^\top \big(\kappa_0 + \gamma_0 \mathbf{I}_p\big)^{-1} \boldsymbol{a} - \gamma_0^{-1} \boldsymbol{a}^\top \big(s\boldsymbol{\Sigma} + \mathbf{I}_p\big)^{-1} \boldsymbol{a} \to 0 \qquad \text{a.s} \tag{2.3.217}$$

We apply this to the deterministic power-signal $\boldsymbol{\beta}$, which satisfies the asymptotic bounded norm condition when $r > \frac{1}{2}$. Using the fact that $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_j)$ is diagonal, the deterministic equivalent term in (2.3.217) becomes

$$\gamma_0^{-1} \boldsymbol{\beta}^\top \big(s\boldsymbol{\Sigma} + \mathbf{I}_p\big)^{-1} \boldsymbol{\beta} = \gamma_0^{-1} \sum_{j=1}^{p} \beta_j^2 \frac{1}{s\sigma_j + 1} \tag{2.3.218}$$

Substituting this back into (2.3.213),

$$p_\xi(0) \asymp \frac{1}{\phi\,\kappa_0}\Big(\sum_{j=1}^{p} \beta_j^2 \frac{s\sigma_j}{s\sigma_j + 1}\Big) \tag{2.3.219}$$

In order to derive the full momentum curve $\widetilde{p}_\xi$, we must now derive a limiting deterministic PDE for the augmented potential $\widetilde{\mathscr{F}_n}$ proceeding in the manner as in Proposition 17. However, as discussed at the beginning of this section, the computation is significantly simplified by the fact that the rank-one perturbation $\xi \boldsymbol{u}\boldsymbol{u}^\top$ is deterministic with respect to the Itô calculations.

To avoid redundancy, we will trace the derivation of the finite$-n$ PDE in the proof of Proposition 17 and make note of where the rank-one perturbation arises, as the calculations are almost identical.

Defining, as in Proposition 17 the function $\widetilde{g}\big(\mathrm{vec}(a_{ij}),\mathrm{vec}(b_{ij})\big)$ as

$$\widetilde{g}\big(\mathrm{vec}(a_{ij}),\mathrm{vec}(b_{ij})\big) = \frac{1}{n}\log\det\Big(\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma\mathbf{I}_n + \xi\boldsymbol{u}\boldsymbol{u}^\top\Big) \tag{2.3.220}$$

where we recall the definition of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = a(\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{A}\boldsymbol{X} + b\Big((\boldsymbol{A}\boldsymbol{X})^\top\boldsymbol{B} + \boldsymbol{B}^\top(\boldsymbol{A}\boldsymbol{X})\Big) + c\boldsymbol{B}^\top\boldsymbol{B} \tag{2.3.221}$$

with

$$a = b_1^2(Np)^{-1}, \quad b = b_1b_2N^{-1}p^{-\frac{1}{2}}, \quad c = b_2^2N^{-1} \tag{2.3.222}$$

Since $\xi\boldsymbol{u}\boldsymbol{u}^\top$ has no dependence on a_{ij} or b_{ij}, the partial derivatives $\frac{\partial\widetilde{g}}{\partial a_{ij}}, \frac{\partial\widetilde{g}}{\partial b_{ij}}, \frac{\partial^2\widetilde{g}}{\partial a_{ij}\partial a_{ij}}, \frac{\partial^2\widetilde{g}}{\partial b_{ij}\partial b_{ij}}$ are computed identically to that of g. Therefore, by defining

$$\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi) = \Big(\boldsymbol{\Phi} + \kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \gamma\mathbf{I}_n + \xi\boldsymbol{u}\boldsymbol{u}^\top\Big)^{-1} \tag{2.3.223}$$

we have

$$\begin{aligned}\frac{1}{2}\Big[\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2\widetilde{g}}{\partial a_{ij}\partial a_{ij}}(\boldsymbol{A},\boldsymbol{B})\Big] = -\frac{n}{N}b_1^2\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big)\,\mathrm{tr}\Big(\boldsymbol{\Phi}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big)\\ + b_1^2\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big) + O(N^{-1})\end{aligned}$$

Therefore, applying the same algebraic trick as in (2.3.109) and (2.3.119), namely adding and subtracting each of $\kappa p^{-1}\boldsymbol{X}^\top\boldsymbol{X}, \gamma\mathbf{I}_n$, and now $\xi\boldsymbol{u}\boldsymbol{u}^\top$, we have

$$\frac{1}{2}\Big[\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2\widetilde{g}}{\partial a_{ij}\partial a_{ij}}(\boldsymbol{A},\boldsymbol{B})\Big] = -b_1^2\frac{n}{N}\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big)\Bigg[1-\kappa\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big) \\ -\gamma\mathrm{tr}\big(\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\big)-\frac{1}{n}\xi\boldsymbol{u}^\top\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\boldsymbol{u}\Bigg]+b_1^2\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big)+O(N^{-1})$$

By an identical calculation to $\frac{\partial^2\widetilde{g}}{\partial b_{ij}\partial b_{ij}}$, we also have

$$\frac{1}{2}\Bigg[\sum_{i=1}^{N}\sum_{j=1}^{n}\frac{\partial^2\widetilde{g}}{\partial b_{ij}\partial b_{ij}}(\boldsymbol{A},\boldsymbol{B})\Bigg] = -\frac{n}{N}b_2^2\,\mathrm{tr}\big(\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\big)\Bigg[1-\kappa\,\mathrm{tr}\Big(p^{-1}\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\Big) \\ -\gamma\mathrm{tr}\big(\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\big)-\frac{1}{n}\xi\boldsymbol{u}^\top\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\boldsymbol{u}\Bigg]+b_2^2\,\mathrm{tr}\big(\widetilde{\boldsymbol{\Omega}}(\kappa,\gamma,\xi)\big)+O(N^{-1})$$

We can therefore form the finite$-n$ SDE for $\widetilde{\mathscr{F}_n}(t,\kappa,\gamma,\xi)$.

$$d\widetilde{\mathscr{F}_n} = \Bigg[]b_1^2(1-\frac{\phi}{\psi})\partial_\kappa\widetilde{\mathscr{F}_n}+b_2^2(1-\frac{\phi}{\psi})\partial_\gamma\widetilde{\mathscr{F}_n} \tag{2.3.224}$$

$$+\frac{\phi}{\psi}\Big(b_1^2\kappa(\partial_\kappa\widetilde{\mathscr{F}_n})^2+b_2^2\gamma(\partial_\gamma\widetilde{\mathscr{F}_n})^2+(b_1^2\gamma+b_2^2\kappa)(\partial_\gamma\widetilde{\mathscr{F}_n}\partial_\kappa\widetilde{\mathscr{F}_n})\Big) \tag{2.3.225}$$

$$+\frac{\phi}{\psi}\big(b_1^2\partial_\gamma\widetilde{\mathscr{F}_n}+b_2^2\partial_\kappa\widetilde{\mathscr{F}_n}\big)\xi\partial_\xi\widetilde{\mathscr{F}_n}\Bigg]\cdot dt+\widetilde{\mathscr{M}_n}(t)+\tilde{E}_n(t) \tag{2.3.226}$$

where $\widetilde{\mathscr{M}_n}(t)$ are the Martingale terms that will also vanish in probability thanks to Lemma 15. Similarly, the remainder terms $\tilde{E}_n(t)$ will vanish uniformly in t. Note that the form of (2.3.224) matches the original SDE, except for the $\xi\partial_\xi\widetilde{\mathscr{F}_n}$ term. Passing through the limit, using the same

arguments made in Section 2.3.3 we arrive at

$$\partial_t \widetilde{\mathscr{F}_\infty} = b_1^2(1-\frac{\phi}{\psi})\partial_\kappa \widetilde{\mathscr{F}_n} + b_2^2(1-\frac{\phi}{\psi})\partial_\gamma \widetilde{\mathscr{F}_n} + \frac{\phi}{\psi}\Bigg[b_1^2\kappa(\partial_\kappa \widetilde{\mathscr{F}_n})^2 + b_2^2\gamma(\partial_\gamma \widetilde{\mathscr{F}_n})^2 \tag{2.3.227}$$

$$+(b_1^2\gamma + b_2^2\kappa)(\partial_\gamma \widetilde{\mathscr{F}_n}\partial_\kappa \widetilde{\mathscr{F}_n})\Bigg] + \frac{\phi}{\psi}\big(b_1^2\partial_\gamma \widetilde{\mathscr{F}_n} + b_2^2\partial_\kappa \widetilde{\mathscr{F}_n}\big)\xi\,\partial_\xi \widetilde{\mathscr{F}_n} \tag{2.3.228}$$

We can write this as a HJ PDE for the augmented momentum variables $\widetilde{p}_\gamma, \widetilde{p}_\kappa$ and $\widetilde{p}_\xi$

$$\partial_t \widetilde{\mathscr{F}_\infty} + \widetilde{H}(\kappa,\gamma,\xi,\widetilde{p}_\kappa,\widetilde{p}_\gamma,\widetilde{p}_\xi) = 0 \tag{2.3.229}$$

where $\widetilde{H}$ is given by the right-hand side of (2.3.227), with an overall negative sign and momentum variables replacing the corresponding partial derivatives with respect to κ,γ,ξ. We can now define the characteristic curve ODEs of the augmented HJ (2.3.229). For the momentum characteristic ODEs, define $\widetilde{p}(t) = b_2^2\widetilde{p}_\gamma(t) + b_1^2\widetilde{p}_\kappa(t)$

$$\frac{\partial \widetilde{p}_\xi}{\partial t} = -\partial_\xi \widetilde{H} = \frac{\phi}{\psi}\widetilde{p}(t)\widetilde{p}_\xi(t) \tag{2.3.230}$$

$$\frac{\partial \widetilde{p}_\gamma}{\partial t} = -\partial_\gamma \widetilde{H} = \frac{\phi}{\psi}\widetilde{p}(t)\widetilde{p}_\gamma(t) \tag{2.3.231}$$

$$\frac{\partial \widetilde{p}_\kappa}{\partial t} = -\partial_\kappa \widetilde{H} = \frac{\phi}{\psi}\widetilde{p}(t)\widetilde{p}_\kappa(t) \tag{2.3.232}$$

Similarly, the characteristic ODEs for the scalars κ,γ,ξ are given by

$$\frac{\partial \xi}{\partial t} = \partial_{p_\xi}\widetilde{H} = -\frac{\phi}{\psi}\widetilde{p}(t)\xi \tag{2.3.233}$$

$$\frac{\partial \gamma}{\partial t} = \partial_{p_\gamma}\widetilde{H} = -b_2^2(1-\frac{\phi}{\psi}) - \frac{\phi}{\psi}\Big[2b_2^2\gamma\widetilde{p}_\gamma + (b_1^2\gamma + b_2^2\kappa)\widetilde{p}_\kappa + b_2^2\xi\,\widetilde{p}_\xi\Big] \tag{2.3.234}$$

$$\frac{\partial \kappa}{\partial t} = \partial_{p_\kappa}\widetilde{H} = -b_1^2(1-\frac{\phi}{\psi}) - \frac{\phi}{\psi}\Big[2b_1^2\kappa\widetilde{p}_\kappa + (b_1^2\gamma + b_2^2\kappa)\widetilde{p}_\gamma + b_1^2\xi\,\widetilde{p}_\xi\Big] \tag{2.3.235}$$

We can solve (2.3.231) and (2.3.232) in the same way as in the un-augmented case. This leads to

$$\widetilde{p}(t) = \left(\widetilde{p}(0)^{-1} - \frac{\phi}{\psi}t\right)^{-1} \tag{2.3.236}$$

$$\widetilde{p}_\gamma(t) = \frac{\widetilde{p}_\gamma(0)}{1 - \frac{\phi}{\psi}\widetilde{p}(0)t} \tag{2.3.237}$$

$$\widetilde{p}_\kappa(t) = \frac{\widetilde{p}_\kappa(0)}{1 - \frac{\phi}{\psi}\widetilde{p}(0)t} \tag{2.3.238}$$

Plugging (2.3.236) into (2.3.230), and solving the separable ODE we have

$$\widetilde{p}_\xi(t) = \frac{\widetilde{p}_\xi(0)}{1 - \frac{\phi}{\psi}\widetilde{p}(0)t} \tag{2.3.239}$$

This allows us to solve the ODE for $\xi(t)$, which gives

$$\xi(t) = \xi_0\left(1 - \frac{\phi}{\psi}\widetilde{p}(0)t\right). \tag{2.3.240}$$

The upside of (2.3.240) is that we can send $\xi(t) \to 0$ by beginning the $\xi-$characteristic curve at $\xi_0 = 0$.

Now let $\widetilde{\eta}(t) = 1 - \frac{\phi}{\psi}\widetilde{p}(0)t$, we will show $\widetilde{\eta}(t)|_{\xi=0} = \eta(t) = 1 - \frac{\phi}{\psi}p(0)t$ recovers the fixed point equation from the un-augmented H. First, by (2.3.207), the initial conditions for $\widetilde{p}_\xi(0), \widetilde{p}_\gamma(0), \widetilde{p}_\kappa(0)$ are all given by

$$\widetilde{p}_\gamma(0) = p_\gamma(0) + \frac{1}{n}\frac{\xi_0\partial_\gamma \boldsymbol{Q}_0}{1 + \xi_0\boldsymbol{Q}_0} \tag{2.3.241}$$

$$\widetilde{p}_\kappa(0) = p_\kappa(0) + \frac{1}{n}\frac{\xi_0\partial_\kappa \boldsymbol{Q}_0}{1 + \xi_0\boldsymbol{Q}_0} \tag{2.3.242}$$

$$\widetilde{p}_\xi(0) = \frac{1}{n}\frac{\boldsymbol{Q}_0}{1 + \xi_0\boldsymbol{Q}_0} \tag{2.3.243}$$

where $\boldsymbol{Q}_0 = \boldsymbol{Q}(0, \kappa_0, \gamma_0)$. Then, by taking $\xi_0 = 0$ and thereby sending $\xi(t) \to 0$ we have

$$\widetilde{p}_\gamma|_{\xi=0} = p_\gamma(0), \quad \widetilde{p}_\kappa(0)|_{\xi=0} = p_\kappa(0), \quad \widetilde{p}_\xi(0)|_{\xi=0} = \frac{1}{n}\boldsymbol{Q}_0. \tag{2.3.244}$$

Further, the characteristic curves for $(\kappa(t), \gamma(t))$, (2.3.235) and (2.3.234), return the original un-augmented solutions on the $\xi = 0$ slice. Therefore, on the $\xi = 0$ slice the augmented system $(\widetilde{H}, \kappa, \gamma, \widetilde{p}_\kappa, \widetilde{p}_\gamma)$ fully reduces to the original Hamiltonian $(H, \kappa, \gamma, p_\kappa, p_\gamma)$. Combined with matching initial conditions (2.3.244), uniqueness of ODEs gives $\widetilde{p}_\kappa(t)|_{\xi=0} = p_\kappa(t)$, $\widetilde{p}_\gamma(t)|_{\xi=0} = p_\gamma(t)$ for all $t \in [0,1]$. As a consequence, we must also have $\widetilde{p}(t)|_{\xi=0} = p(t)$, and $\widetilde{\eta}(t)|_{\xi=0} = \eta(t)$.

Finally using the deterministic equivalent for $p_\xi(0) = \boldsymbol{Q}_0$ (2.3.219) , this gives us

$$m_\beta(z) \asymp \frac{1}{\phi\,\kappa_0\eta^*} F_\beta(s), \qquad F_\beta(s) = \sum_{j=1}^{p} \beta_j^2 \frac{s\sigma_j}{s\sigma_j + 1} \tag{2.3.245}$$

which concludes the proof. □

As an immediate corollary under the dual-power law assumption (**A0**), we have

$$m_\beta(z) \asymp \frac{1}{\phi\,\kappa_0\eta^*} F_r(s), \qquad F_r(s) = \sum_{j=1}^{p} \frac{s j^{-(\alpha+2r)}}{s j^{-\alpha} + 1} \tag{2.3.246}$$

Proof of NTK Scaling Law

Proof. The proof has three steps. First we estimate the effective dimension $\mathscr{N}_p(\mu)$; then we estimate the source-bias profile $\mathscr{B}_p(\mu)$; finally we optimize the deterministic equivalent (2.2.60).

Step 1: effective dimension.

Since $\alpha > 1$,

$$T_p = \sum_{j=1}^{p} j^{-\alpha} \longrightarrow \zeta(\alpha) \in (0, \infty),$$

so $\bar{\sigma}_{p,j} = T_p^{-1} j^{-\alpha} \asymp j^{-\alpha}$ uniformly in $j \le p$. Let

$$J_\mu := \left\lfloor \mu^{-1/\alpha} \right\rfloor.$$

The constants c_{NTK} and T_p only change J_μ by a multiplicative constant, so they are suppressed in the comparison estimates. Assume $J_\mu = o(p)$. For $j \le J_\mu$, one has $j^{-\alpha} \gtrsim \mu$, and therefore

$$\frac{c_{\mathrm{NTK}} \bar{\sigma}_{p,j}}{c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu} \asymp 1.$$

For $j > J_\mu$, one has $j^{-\alpha} \lesssim \mu$, and therefore

$$\frac{c_{\mathrm{NTK}} \bar{\sigma}_{p,j}}{c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu} \asymp \frac{j^{-\alpha}}{\mu}.$$

Consequently,

$$\begin{aligned} \mathscr{N}_p(\mu) &\asymp \sum_{j \le J_\mu} 1 + \mu^{-1} \sum_{J_\mu < j \le p} j^{-\alpha} \\ &\asymp J_\mu + \mu^{-1} J_\mu^{1-\alpha} \asymp \mu^{-1/\alpha}. \end{aligned}$$

Thus

$$\mathscr{N}_p(\mu) \asymp \mu^{-1/\alpha} \tag{2.3.247}$$

whenever $\mu \downarrow 0$ and $J_\mu = o(p)$.

Step 2: source-bias profile.

By Assumption 1,

$$\sigma_j \boldsymbol{\beta}_j^2 = c_{p,\alpha,r}^2 j^{-(\alpha+2r)}.$$

Since $\alpha + 2r > 1$, the normalizing constants $c_{p,\alpha,r}$ converge to a positive finite limit. Hence $\sigma_j \boldsymbol{\beta}_j^2 \asymp j^{-(\alpha+2r)}$. Splitting the bias at J_μ gives

$$\mathscr{B}_p(\mu) = \sum_{j \le J_\mu} \sigma_j \boldsymbol{\beta}_j^2 \left(\frac{\mu}{c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu} \right)^2 + \sum_{j > J_\mu} \sigma_j \boldsymbol{\beta}_j^2 \left(\frac{\mu}{c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu} \right)^2 =: B_{\le}(\mu) + B_{>}(\mu).$$

For $j \le J_\mu$, $\mu/(c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu) \asymp \mu j^{\alpha}$, and therefore

$$B_{\le}(\mu) \asymp \mu^2 \sum_{j \le J_\mu} j^{\alpha - 2r}. \tag{2.3.248}$$

For $j > J_\mu$, $\mu/(c_{\mathrm{NTK}} \bar{\sigma}_{p,j} + \mu) \asymp 1$, and therefore

$$B_{>}(\mu) \asymp \sum_{j > J_\mu} j^{-(\alpha+2r)}. \tag{2.3.249}$$

If $\alpha - 2r > -1$, equivalently $\alpha + 2r < 2\alpha + 1$, then

$$\begin{aligned} B_{\le}(\mu) &\asymp \mu^2 J_\mu^{\alpha - 2r + 1} = \mu^{(\alpha+2r-1)/\alpha}, \\ B_{>}(\mu) &\asymp J_\mu^{1-\alpha-2r} = \mu^{(\alpha+2r-1)/\alpha}. \end{aligned}$$

Thus

$$\mathscr{B}_p(\mu) \asymp \mu^{(\alpha+2r-1)/\alpha} \qquad \text{if } \alpha + 2r < 2\alpha + 1. \tag{2.3.250}$$

If $\alpha - 2r < -1$, equivalently $\alpha + 2r > 2\alpha + 1$, then $\sum_{j \le J_\mu} j^{\alpha-2r} = O(1)$, so $B_{\le}(\mu) \asymp \mu^2$, while $B_{>}(\mu) \asymp \mu^{(\alpha+2r-1)/\alpha} = o(\mu^2)$. Hence

$$\mathscr{B}_p(\mu) \asymp \mu^2 \qquad \text{if } \alpha + 2r > 2\alpha + 1. \tag{2.3.251}$$

Combining (2.3.250) and (2.3.251), and excluding the logarithmic boundary $\alpha + 2r = 2\alpha + 1$, we have

$$\mathscr{B}_p(\mu) \asymp \mu^{q_{\alpha,r}}. \tag{2.3.252}$$

Step 3: optimization.

Substituting (2.3.247) and (2.3.252) into the local deterministic equivalent (2.2.60) gives, uniformly in the scaling window,

$$R_{n,N,p}(\lambda) \asymp \mu_\lambda^{q_{\alpha,r}} + \delta_{n,N}\mu_\lambda^{-1/\alpha}. \tag{2.3.253}$$

The two terms in (2.3.253) balance when

$$\mu_\lambda^{q_{\alpha,r}} \asymp \delta_{n,N}\mu_\lambda^{-1/\alpha},$$

that is,

$$\mu_{\lambda_\star} \asymp \delta_{n,N}^{1/(q_{\alpha,r}+1/\alpha)} = \delta_{n,N}^{\alpha/(\alpha q_{\alpha,r}+1)}.$$

At this value,

$$\mu_{\lambda_\star}^{q_{\alpha,r}} \asymp \delta_{n,N}\mu_{\lambda_\star}^{-1/\alpha} \asymp \delta_{n,N}^{\alpha q_{\alpha,r}/(\alpha q_{\alpha,r}+1)}.$$

This proves the upper bound in (2.2.62). For the matching lower bound, let $\mu_\star := \delta_{n,N}^{\alpha/(\alpha q_{\alpha,r}+1)}$. If $\mu \geq \mu_\star$, then the bias term in (2.3.253) is at least a constant multiple of $\mu_\star^{q_{\alpha,r}}$. If $\mu \leq \mu_\star$, then the variance term in (2.3.253) is at least a constant multiple of $\delta_{n,N}\mu_\star^{-1/\alpha}$. Both quantities are of order $\delta_{n,N}^{\alpha q_{\alpha,r}/(\alpha q_{\alpha,r}+1)}$. The condition (2.2.61) is exactly $\mu_\star^{-1/\alpha} = o(p)$, so the estimates above apply at the optimizer and in the neighboring scaling window. This proves (2.2.62) and (2.2.63). The special cases (2.2.64) and the saturated-source formula follow by substituting the two possible values of $q_{\alpha,r}$. Finally, if $N \asymp n^\ell$, then $\delta_{n,N} = n^{-1} + n^{-\ell} \asymp n^{-\min\{1,\ell\}}$, which gives (2.2.66). □

Remark. At the boundary $\alpha + 2r = 2\alpha + 1$, the proof gives

$$\mathscr{B}_p(\mu) \asymp \mu^2 \log(1/\mu), \qquad \mathscr{N}_p(\mu) \asymp \mu^{-1/\alpha}.$$

Consequently,

$$\mu_{\lambda_\star} \asymp \left(\frac{\delta_{n,N}}{\log(1/\delta_{n,N})} \right)^{\alpha/(2\alpha+1)},$$

and

$$\inf_{\lambda \geq 0} R_{n,N,p}(\lambda) \asymp \delta_{n,N}^{2\alpha/(2\alpha+1)} \left(\log \frac{1}{\delta_{n,N}} \right)^{1/(2\alpha+1)}.$$

2.4 Numerical Experiments

In this section, we present figures and experiments validating the main results presented in Section 2.2.7.

2.4.1 Numerical root solver

The fixed-point equation (2.2.34) can be solved using a standard root solving algorithm, e.g brentq. Here we present details on the algorithm implementation. Recall the fixed point equation defining $\eta(s)$ to be

$$\Psi(\eta^*) = F_{\mathbf{\Sigma}}\big(s(\eta^*)\big) + \zeta(\eta^*) s(\eta^*) - \phi = 0, \qquad F_{\mathbf{\Sigma}}(s) = \frac{1}{p} \sum_{j=1}^{p} \frac{s\sigma_j}{s\sigma_j + 1} \tag{2.4.1}$$

where $D_z = z + \tau^2 \sigma_W^2 (c_1 - b_1^2)$ and

$$s(\eta) = \frac{b_1^2(\eta + \sigma_W^2)}{\eta\big(D_z - b_2^2 \sigma_W^2\big)} \big(\phi\eta - \psi(1-\eta)(\eta + \sigma_W^2)\big), \qquad \zeta(\eta) = \frac{b_2^2 \eta + D_z}{b_1^2\big(\eta + \sigma_W^2\big)}. \tag{2.4.2}$$

The fixed point equation (2.4.1) admits a unique physical root $\eta^* > 0$ corresponding to $s = s(\eta^*) > 0$. Define the quadratic term in $s(\eta)$ as $N(\eta) := \psi\eta^2 + \big(\phi - \psi + \psi\sigma_W^2\big)\eta - \psi\sigma_W^2$ and

the denominator $D := D_z - b_2^2\sigma_W^2$. Then we can write

$$s(\eta) = \underbrace{\frac{b_1^2(\eta+\sigma_W^2)}{\eta}}_{\kappa_0(\eta)>0} \frac{N(\eta)}{D} \tag{2.4.3}$$

The first term in (2.4.3) is positive for $\eta > 0$, therefore the sign of s is controlled by the sign of $N(\eta)/D$. We handle each in separate cases. First the quadratic $N(\eta)$ has two real roots of opposite sign $\eta_0 < 0 < \eta_1 < 1$. Indeed, the product of the two roots is negative $\eta_0\eta_1 = -\sigma_W^2 < 0$. Further $N(0) = -\psi\sigma_W^2 < 0$ and $N(1) = \phi > 0$, forcing $0 < \eta_1 < 1$. Therefore $N(\eta) < 0$ on $(0,\eta_1)$ and $N(\eta) > 0$ on (η_1,∞).

Case one: $D < 0$

In this case $s(\eta) > 0$ for $\eta \in (0,\eta_1)$. It is straightforward to verify that Ψ is strictly decreasing on $(0,\eta_1)$, with

$$\Psi(\eta_1^-) \to -\phi < 0, \qquad \Psi(0^+) \to +\infty > 0$$

Hence there is exactly one root of (2.4.1) in $(0,\eta_1)$ corresponding to $s(\eta) > 0$.

Case two: $D > 0$ Now the physical root lies in (η_1,∞), where $s(\eta)$ is positive. Ψ is strictly increasing in this interval, with

$$\Psi(\eta_1^+) \to -\phi < 0, \qquad \Psi(\infty) \to \infty$$

As in the first case, there is a unique root of (2.4.1) in (η_1,∞) corresponding to $s(\eta) > 0$.

Scaling law of NTK regression

Figure 2.1 evaluates the unnormalized trace-class scaling of a finite-width Gaussian-equivalent NTK surrogate. The simulation uses a data spectral decay of $\alpha = 1.5$ and a teacher complexity of $r = 0.5$, with both the network width N and input dimension d scaling proportionally to the sample size. By testing sample sizes n ranging from 384 to 1500 across 10

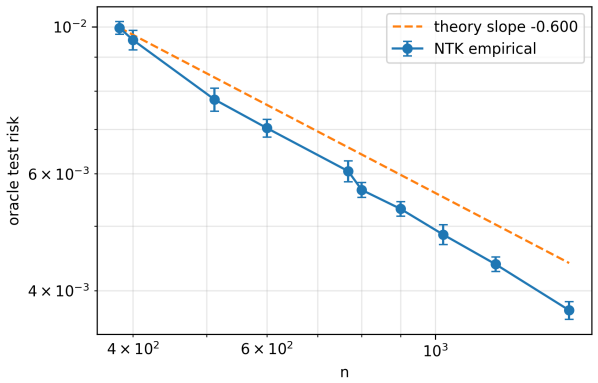


Figure 2.1. Test risk scaling of a finite-width Gaussian-equivalent NTK on unnormalized power-law data ($\alpha = 1.5, r = 0.5$).

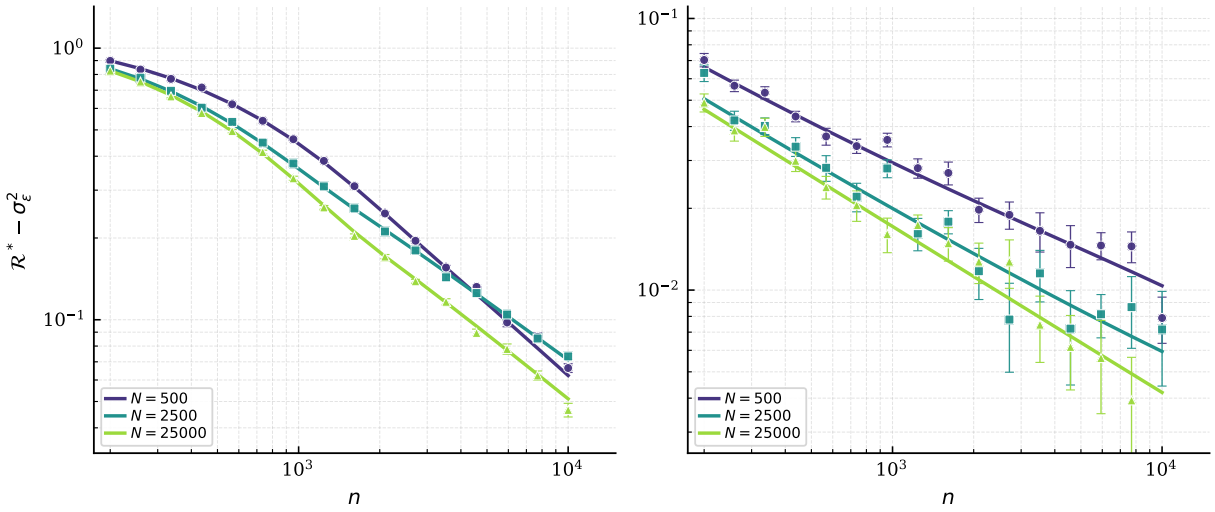


Figure 2.2. NTK risk at optimally-tuned ridge λ^* versus sample size n with log-log axis. Empirical test loss reported as markers $\pm$SE over 20 trials. Theoretical curves from (2.2.51) plotted as solid lines. **(Left)** slowly-decaying spectrum $\alpha = 0.2$, with visible curvature on log-log plot. **(Right)** rapidly-decaying spectrum $\alpha = 0.8$, exhibiting a power-law decay in n. Plots generated with $r = 1.5, p = 1200, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0$

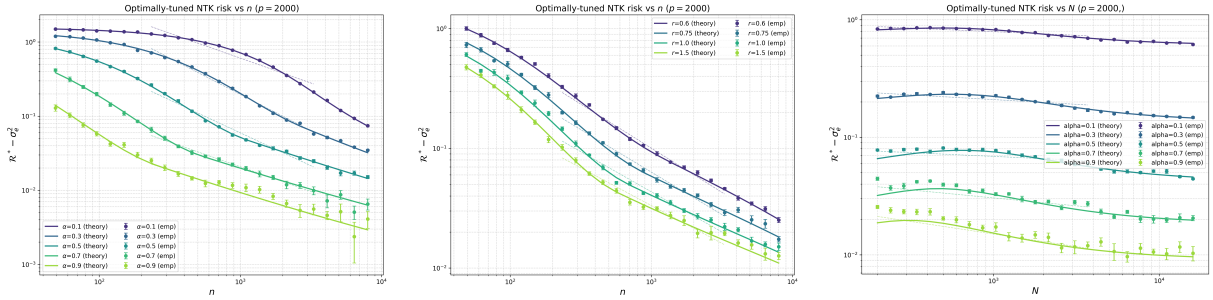


Figure 2.3. Optimally-tuned Risk: Excess NTK risk of optimally-tuned ridge $\lambda = \lambda^*$ (markers with $\pm$SE over 10 trials) on log-log axes. Theoretical curves from (2.2.51) plotted in solid lines, showing close alignment across all regimes.**Left**: excess risk vs training samples n with varying capacity exponent α and fixed $N = 5000, r = 1.0$. As α increases, curves straighten out to follow power-law decay. **Middle**: Varying values of source exponent r. **Right:** risk against width N for sweeping α, and fixed $n = 2000, r = 0.7$. Throughout $\sigma_W = 1, \sigma_\varepsilon^2 = 0.01$, $p = 2000$.

random seeds, the experiment successfully demonstrates a power-law decay in test risk. The resulting empirical risk exponent of $\theta_{\text{NTK}} \approx 0.687$ aligns closely with the theoretical prediction of 0.6, confirming that the Gaussian-equivalent model accurately captures the scaling behavior of finite-width neural networks in this regime.

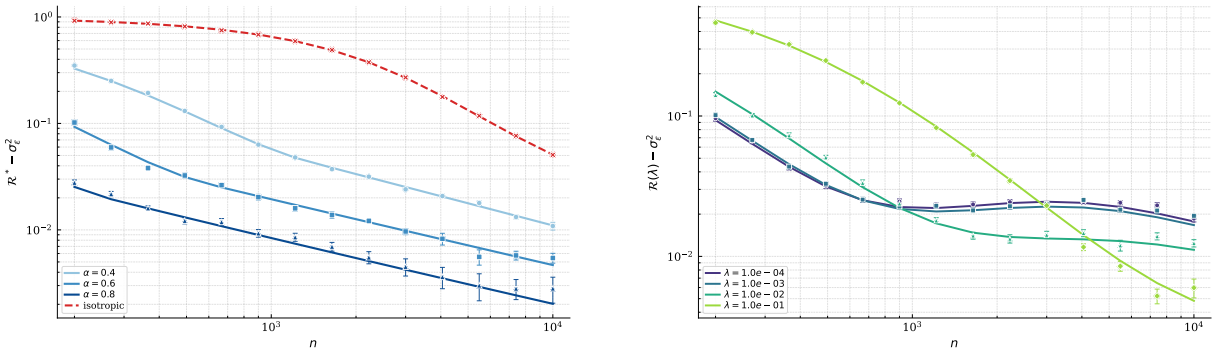


Figure 2.4. Isotropic vs power-law: **(Left)** Optimal-tuned excess NTK risk plotted for varying α against isotropic baseline. Isotropic case exhibits curve in log-log axes plot, while power law curve straightens as α increases. **(Right)** excess risk against n for various ridge λ with fixed $\alpha = 0.6$. Both plots generated with $p = 2000, \sigma_\varepsilon^2 = 0.3, r = 1.5, N = 2500, \sigma_W^2 = 1.0$. Error bars are plotted with marker $\pm$ SE over 20 trials.

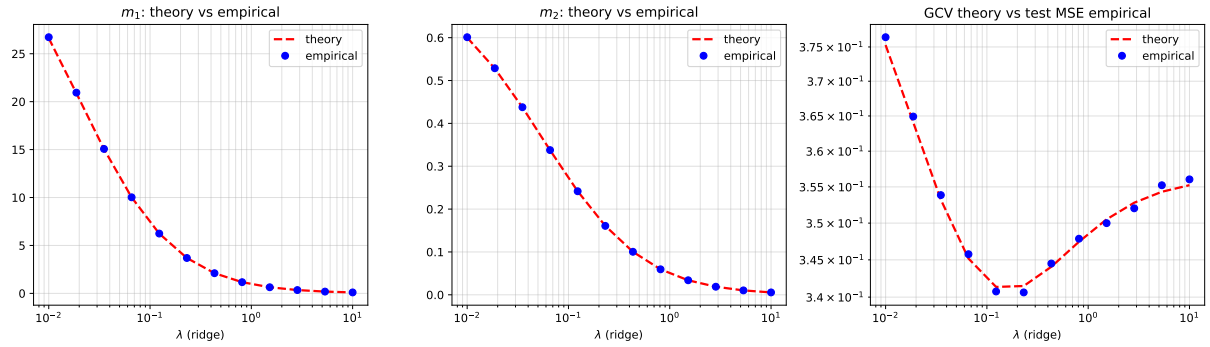


Figure 2.5. Theory vs empirical $\alpha = 0.5$: plots for m_1 (left), m_2 (middle), and GCV_λ (right) against ridge λ. All three show strong agreement across. Plots generated with $n = 2400, N = 1800, p = 1200, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0$ and isotropic labels.

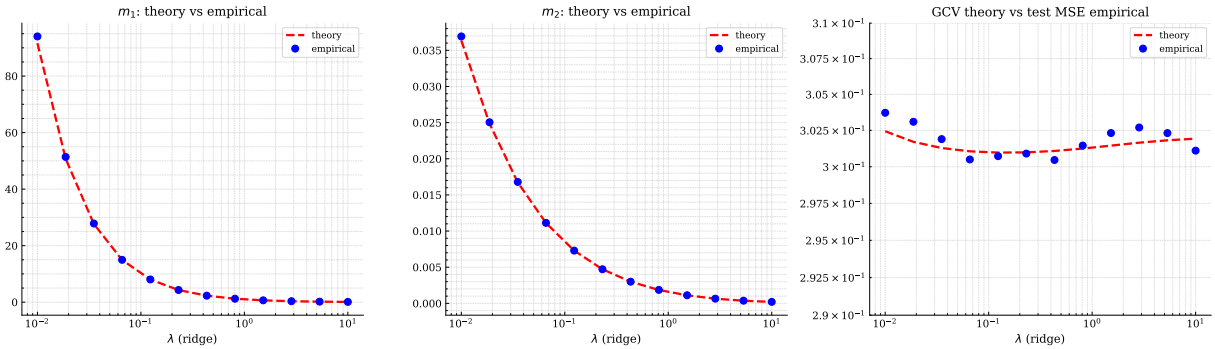


Figure 2.6. Theory vs empirical $\alpha = 1.5$: Same setup as Figure 2.5, with larger capacity exponent. Test loss (right) is noisier, but still tracked by theoretical curve.

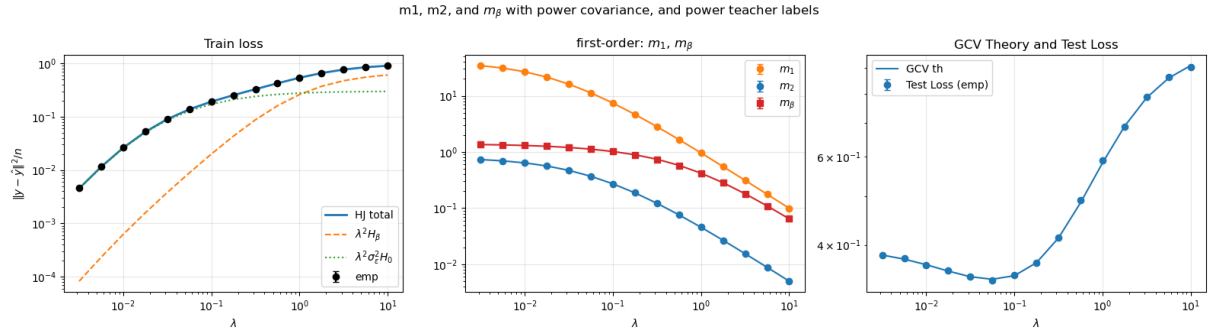


Figure 2.7. Signal resolvent: All three trace resolvents m_1, m_2, m_β plotted against ridge λ (middle). m_β shows divergence from m_2. Theoretical curves for all three trace resolvents track empirical plots well. Right shows GCV estimator against test MSE. Plots generated with $n = 2250, N = 3000, p = 1500, \sigma_\varepsilon^2 = 0.3, \sigma_W^2 = 1.0, \alpha = 0.5, r = 0.8$ and error bars over 20 trials

In conclusion, we study finite-width NTK regression under anisotropic dual power-law data, where both the input covariance spectrum and the teacher vector exhibit polynomial decay. By combining Gaussian-equivalent reductions with a Brownian embedding of the linearized NTK Gram matrix, we derive a Hamilton–Jacobi PDE for the limiting log-determinant potential and use its characteristic curves to obtain deterministic equivalents for the resolvent traces governing generalization. Under the stated Gaussian-equivalence and GCV-consistency assumptions, our formulas yield source–capacity scaling laws for optimally tuned NTK ridge regression and recover the infinite-width KRR behavior as a limiting case. Numerical experiments validate the predicted resolvent traces, GCV risk curves, and scaling exponents, supporting the proposed Hamilton–Jacobi approach as a flexible tool for studying neural scaling laws in anisotropic kernel models.

Acknowledgments

Chapter 2 is extracted paper from "Collin Cranston, Zhichao Wang, Todd Kemp. NTK Regression under dual power-law model: Deterministic Equivalents via SDE and PDE Methods." Currently under review. The dissertation author is the co-author of this paper.

Chapter 3

Eigen-spike emergence for Nonlinear Data

3.1 Introduction

Spectral information—from weight matrices, kernel matrices, and Hessians—provides a quantitative window into representation learning and efficient optimization in neural networks (NNs): e.g., eigenvalues in weight matrices of NNs capture the strength of learned directions, while corresponding eigenvectors indicate which features in the data are amplified or extracted during training [MM21, MM19a, MM20, MPM21, YTH$^+$22, LM21, WES$^+$23, BES$^+$22, SBL16, GKX19]. A major direction in deep learning theory is to use spectral tools to understand when a network can *propagate, preserve, transform, and amplify* task-relevant structure that may be hidden in high-dimensional datasets.

A concrete and widely used spectral mechanism is the emergence of *spikes* (outlier eigen/singular values) and the *alignment* of their eigen/singular vectors with signal directions in the dataset. Random matrix theory (RMT) provides a powerful tool to analyze such eigen-spike emergence. In high-dimensional spiked models where a low-rank signal is embedded in Gaussian noise, BBP (Baik–Ben Arous–Péché) transitions characterize when an informative outlier separates from the bulk spectrum and when its associated eigenvector becomes correlated with the underlying signal [BAP05, Nad08]. Recent work has developed nonlinear counterparts for random-feature and kernel-type matrices, providing sharp predictions for outlier locations and eigenvector alignment [WZ21, Fel23, BES$^+$23, WWF24]. Such nonlinear RMT can be

applied to study deep learning theory, e.g., the infinite-width *Neural Tangent Kernel* (NTK) and the *Conjugate Kernel* (CK) [JGH18, PW17, AP20] for multi-layer NNs. Their high-dimensional spectra have been studied extensively [LM25, WZ21, LCM20, LCM20]. In this work, we focus on the CK, and we study *emergent spikes* induced by *nonlinear* structure in the data. Concretely, we consider the following spectral question for the CK: *When can NNs transfer a nonlinear pattern in the data into a linearly accessible direction (an aligned eigen-spike) that enables linear classification?*

In this chapter, the CK is defined by a one-hidden-layer random NN: $\boldsymbol{Y} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{X}) \in \mathbb{R}^{N\times n}$, with nonlinear activation σ, random weights $\boldsymbol{W} \in \mathbb{R}^{N\times d}$, and data $\boldsymbol{X} \in \mathbb{R}^{d\times n}$. The CK matrix is the Gram matrix of hidden activations, $\boldsymbol{K} = \boldsymbol{Y}^\top\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ for n data points. The spectral distribution of $\boldsymbol{K}$ is particularly useful for generalization error of random feature regression, as demonstrated in [MM19b, HL22b] for isotropic Gaussian data $\boldsymbol{X}$. However, real datasets often contain low-rank or nonlinear structure that may be hard to analyze using PCA [Jol02, SSM98, TdSL00, RS00, BN03, CL06, JL09]. To highlight genuinely nonlinear structure, we instead take $\boldsymbol{X}$ to be the high-dimensional *XOR Gaussian mixture* dataset (formally defined in Section 3.2.1). This dataset is a balanced four-component Gaussian mixture whose two classes are unions of opposite components [RGKZ21]. The key property is that XOR is *not linearly separable* in first-order statistics: the class means vanish, and any method that only "sees" linear correlations cannot recover labels. However, XOR is *quadratically separable*, so a second-order feature can potentially make the task linearly solvable. Heuristically, the eigenvalues of $\boldsymbol{K}$ describe the spectrum of learned similarities, while the alignment of leading eigenvectors with task structure (e.g., labels or cluster indicators) may lead to spectral methods solving the task. Therefore, this chapter aims to answer the following spectral question for $\boldsymbol{K}$: *When do outlier eigenvalues emerge in the CK spectrum, and when do their spiked eigenvectors align with the XOR labels and enable linear classification?*

In practice, NN performance is affected by hyperparameters, dataset structure, and architecture; and, as a practical matter, machine learning (ML) practitioners typically fiddle

(quite aggressively) with various "knobs" of the ML training process in order to improve model performance. Motivated by this, we consider properties of the CK matrix for the XOR problem, as various "knobs" for $\boldsymbol{K}$ are varied; and, depending on their values, we observe very different spectral behavior. Concretely, we vary the following knobs (which are among the standard parameters that are tuned in ML practice): sampling size scaling n with input features d and NN width N; signal-to-noise ratio (SNR) in the data $\boldsymbol{X}$; the pretrained weight matrix $\boldsymbol{W}$; and the choice of nonlinear activation σ. We show that, depending on the values of these "knobs," one can obtain qualitatively different spectral properties, various phase transitions, and the emergence of *quadratic informative spikes* in the CK matrix. Importantly, the emergence of these quadratically informative spikes makes the nonlinear XOR linearly classifiable.

3.1.1 Related works

Simplicity bias and low-rank structure.

Recent work suggests that deep networks and transformers exhibit an implicit *distributional simplicity bias*, learning simpler (lower-order) statistical structure of the data earlier in training [BPQ+24, RGLG24]. Complementarily, [HMZ+23] report a *low-rank simplicity bias* in deep networks, arguing that depth/over-parameterization can implicitly favor low effective-rank feature representations. From a broader perspective, [Wil25] argues that many generalization phenomena often viewed as "mysterious" in deep learning can be reconciled with classical theory via soft inductive biases.

Spectrum of kernel random matrices.

The study of kernel random matrices and their spectral distributions has become a major topic in RMT. Global convergence of the empirical spectral distribution (ESD) is known for broad classes of nonlinear kernels [EK10, CS13, DV13]. In the proportional regime, [CS13] and [DV13] characterize limiting spectra of inner-product kernels. The CK and NTK can be viewed as random kernels, whose spectra were further studied [FW20, AP20, Péc19, PW17, DM24, GP26] and connected to generalization in random feature regression [MM19b, AP20] and memorization

of spurious features [BM24] .

Nonlinear spiked covariance matrices.

Spiked models exhibit phase transitions where informative eigenvalues separate from the bulk [BAP05, Péc06]. Recent work extends this to *nonlinear* random matrices, i.e., matrices whose entries are nonlinear functions of high-dimensional random projections (such as $\sigma(\boldsymbol{WX})$), leading to nonlinear analogues of BBP transitions and alignment formulas [WZ21, Fel23, BES+23, WWF24, GKK+23, BP22]. Outlier eigen-structure of CK and NTK have been studied [LHY25], using the deterministic equivalence from [GDZ+22]. Although they can deal with random deep networks and general Gaussian mixture datasets, their CKs and NTKs are expected kernels (namely first take width $N \to \infty$) and only equivalent to linear kernels.

Polynomial sample-size regimes.

In the proportional regime ($n \asymp d$), LE/Gaussian-equivalence phenomena often imply linear-kernel behavior for broad families of kernels. Moving to polynomial regimes ($n \asymp d^k$), several works show that many kernel matrices behave like polynomial kernels of degree k (with rigorous deterministic equivalents and sharp asymptotics) [XHM+22, LY22, CM24, MS24b]. In the quadratic regime ($k = 2$), nonlinear kernels can reduce in operator norm to quadratic polynomial kernels [GMMM19, PWZ24], enabling classification of quadratically separable data such as XOR [RGKZ21]. [WHL+25] recently proved a conditional Gaussian equivalence for random feature model at quadratic regime to show the asymptotic generalization error in this case.

Gaussian mixtures and XOR problem.

Gaussian mixture classification, specifically the XOR mixture has become an important testbed for statistics and ML to evaluate various algorithms. [RGKZ21] documented a sharp contrast between trained two-layer networks and kernel methods on such mixtures, motivating a spectral analysis of when (and why) kernels can succeed. Related RMT analyses for mixture classification with kernels appear in [LC19, Cou19]. [Gla23] showed how training two layer

NNs with SGD can learn this XOR problem even with a sample complexity $\tilde{O}(d)$. Recently, [TGM25] analyzed continual learning by studying one-hidden-layer quadratic NNs trained with gradient descent on an XOR dataset. [DD25] analyzed two-layer networks after one gradient step on $\boldsymbol{W}$ with Gaussian-mixture data using higher-order polynomial equivalence. This motivates our spiked feature case in Section 3.2.5, while our result focuses on label-aligned outliers for linear classification on XOR, unlike trained feature regression tasks in [DD25].

3.1.2 Contributions

We develop a RMT framework for the CK that is based on a *quadratic equivalent* (QE) that captures precisely when nonlinear learnability emerges for XOR, in regimes where the *linear equivalence* (LE) behavior is insufficient. We are particularly interested in this emergence as a function of the following four "knobs" that are widely-used by ML practitioners.

1. **Finite-SNR proportional limit ($n \asymp d$): two negative results.** In the proportional regime with $\mathrm{SNR} = \Theta(1)$, CK can exhibit outlier eigenvalues, but their eigenvectors do *not* align with XOR labels; hence spectral methods with linear readouts on CK features fail to classify XOR (Theorem 5; Figure 3.1). We also prove failure for Euclidean distance kernel spectral clustering.

2. **Large-SNR proportional limit: emergence of quadratic informative spike.** We show that with increasing SNR scaling and with proper activation choice, a pair of informative outliers separates and aligns with the XOR class direction, enabling linear spectral clustering/classification (Theorem 7).

3. **Pretrained/spiked features: test-time BBP transition.** A low-rank perturbation of a random $\boldsymbol{W}$ can yield the emergence of an outlier whose eigenvector aligns with the induced task direction (Theorem 8) even for finite-SNR and proportional limit. Then, empirically, we show that linear spectral clustering occurs for CK matrix with weights from a pretrained NN on CIFAR-2.

4. **Quadratic sample-size regime ($n \asymp d^2$): a quadratic kernel.** We prove an operator-norm reduction of the CK to a quadratic polynomial kernel in this regime for ultra-wide NNs; with proper activation choice, the reduced model has an informative outlier direction to recover XOR labels (Theorem 9).

We summarize these results in the following table.

Table 3.1. Summary of different knobs. From Case 2–4, $c_\sigma \neq 0$ is **necessary** in all these cases. If $c_\sigma = 0$ (e.g., tanh after centering/normalization), linear classification via CK is impossible.

Regime	SNR	Sample	Weights	Linear Label Align?
Case 1 (Baseline)	$\Theta(1)$	$n \asymp d$	Random	No
Case 2 (Large SNR)	$\Theta(n^{1/4})$	$n \asymp d$	Random	Yes
Case 3 (Trained weight $\mathbf{W}_1$)	$\Theta(1)$	$n \asymp d$	Spiked	Yes
Case 4 (Quad. sample size)	$\Theta(1)$	$n \asymp d^2$	Random	Yes

3.1.3 Approach

Much of the existing RMT literature on the CK matrix analysis lies in a proportional regime ($n \asymp d \asymp N$), and it often involves the assumption of isotropic Gaussian data $\boldsymbol{Z} \in \mathbb{R}^{d \times n}$. In this setting, a standard tool is a *linear equivalent* (LE) for the *feature matrix* $\sigma(\boldsymbol{WZ})$. This approach replaces the nonlinear features with an affine Gaussian surrogate and an independent Gaussian noise term $\boldsymbol{N}$:

$$\sigma(\boldsymbol{WZ}) \approx b_\sigma \boldsymbol{WZ} + a_\sigma \boldsymbol{N}, \qquad \boldsymbol{N}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{3.1.1}$$

for constants a_σ, b_σ determined by σ. This approximation is also called *Gaussian equivalence* in prior work [MM19b, HL22b, GLR+21, BPH23]. When applying the LE, one can show that the spectra of nonlinear CK matrices are *asymptotically equivalent* to a *linear* random matrix model, thus making classical RMT directly applicable. However, the LE also suggests that

the induced kernel behaves essentially like a *linear* kernel plus isotropic noise; this can hide important nonlinear task structure; see Figure 3.1.

For the XOR problem, the nonlinearity σ is essential: it can transform a nonlinear classification task into one that is linearly accessible. To understand this, we vary the ML "knobs" listed above to identify regimes in which the LE approximation works, versus when it breaks down. To go beyond the classical LE in (3.1.1), we develop a refined (quadratic) equivalent model that captures emergent spiked eigenvalues and eigenvectors carrying nonlinear information about CK and the dataset, thereby enabling linear classification on the nonlinear XOR problem. We refer to such spikes as *quadratic informative spikes*. Addressing these questions will clarify the regimes in which NNs, and in particular various choices for training knobs of the ML training process, can recover and even amplify beneficial nonlinear features concealed in the dataset.

Quadratic Equivalents (QE) .

We model the XOR data as a low-rank perturbation $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{M}$, where $\boldsymbol{Z}$ is Gaussian noise and $\boldsymbol{M}$ is a rank-two signal that encodes the XOR structure. Our goal is to understand how this signal shows up in the CK spectrum. We go beyond the LE in (3.1.1), which captures only the bulk spectrum, and derive a QE that precisely tracks all spikes via a second-order Taylor approximation of σ. At a high level, this yields the decomposition

$$\sigma(\boldsymbol{W}\boldsymbol{X}) \approx \underbrace{\sigma(\boldsymbol{W}\boldsymbol{Z})}_{\text{bulk+uninformative spikes}} + \underbrace{\boldsymbol{T}_1}_{\text{linear spikes}} + \underbrace{\boldsymbol{T}_2}_{\text{quadratic spikes}} \tag{3.1.2}$$

where $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are finite-rank contributions induced by the signal $\boldsymbol{M}$ and σ (see (3.3.78) and Proposition 35). In this QE, there are three terms. The first term, $\sigma(\boldsymbol{W}\boldsymbol{Z})$, determines the bulk spectrum; and it may also possibly induce *uninformative outliers*, solely due to the architecture [BP22]. The second term, $\boldsymbol{T}_1$, corresponds to a linear transformation of the signal $\boldsymbol{M}$; while it can induce outliers, their eigenvectors turn out *not* to be aligned with the XOR labels, meaning that they are not useful for nonlinear XOR classification. The third term, $\boldsymbol{T}_2$, captures quadratic

properties: depending on the correct tuning of the ML training "knobs," this quadratic term $\boldsymbol{T}_2$ can dominate and create *quadratic informative spikes* (see (3.3.78)). This effectively provides a "quadratic feature channel" that makes the nonlinear XOR problem analyzable with RMT and linearly separable.

3.2 Main Results

3.2.1 Notations and Preliminaries

Settings of One-layer NN model.

We study the output of one-layer NNs when fed a dataset $\boldsymbol{X} \in \mathbb{R}^{d\times n}$. Here, d is the input dimension and n is the sample size. We consider different scalings between n and d to determine the sample complexity of the model. Let the weight matrix at random initialization be $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ with $[\boldsymbol{W}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i \in [N], j \in [d]$, where N is the width of NN. Denote the output of our one-layer NN and the CK matrix as

$$\boldsymbol{Y} := \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{X}), \qquad \boldsymbol{K} := \boldsymbol{Y}^\top \boldsymbol{Y}, \tag{3.2.1}$$

where σ is a nonlinear activation. Our target is the eigenstructure of this CK when $\boldsymbol{X}$ has a nonlinear separable pattern. We now introduce the assumptions of the activation function σ.

Assumption 3. *Assume σ is three-times differentiable with $\sup_{x\in\mathbb{R}} |\sigma'(x)|, |\sigma''(x)|, |\sigma'''(x)| \le \lambda_\sigma$, for some $\lambda_\sigma \in (0,\infty)$, and is centered and normalized with respect to $\xi \sim \mathcal{N}(0,1)$:*

$$\mathbb{E}[\sigma(\xi)] = 0, \qquad \mathbb{E}[\sigma(\xi)^2] = 1. \tag{3.2.2}$$

Define the Hermite coefficients of σ as

$$b_\sigma := \mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi\,\sigma(\xi)] \in \mathbb{R}, \tag{3.2.3}$$

$$c_\sigma := \mathbb{E}[\sigma''(\xi)] = \mathbb{E}[(\xi^2-1)\sigma(\xi)] \in \mathbb{R}. \tag{3.2.4}$$

Many smooth activations satisfy Assumption 3, after centering and normalization (e.g., tanh, erf, smoothed-ReLU, GELU). The condition b_σ can be zero, which is needed for the pure-quadratic-channel regime in Theorem 7. While ReLU is not C^3, it can be handled by smoothing or by approximation arguments; we keep Assumption 3 to avoid technical distractions.

XOR Dataset.

The XOR dataset is a binary class Gaussian Mixture dataset with four balanced clusters. We denote the collection of n data points by $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$, where $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \frac{1}{4}\sum_{a=1}^{4} \mathcal{N}(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_a, \frac{1}{d}\mathbf{I}_d)$. Without loss of generality, we reorder the data points so that we can define

$$\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{Z}, \tag{3.2.5}$$

where $\boldsymbol{Z} \in \mathbb{R}^{d\times n}$ has i.i.d. entries with distribution $\mathcal{N}(0, \frac{1}{d})$ and

$$\boldsymbol{M} = r\sqrt{\frac{n}{2d}}\big(\boldsymbol{u}_1 \cdot \boldsymbol{v}_1^\top + \boldsymbol{u}_2 \cdot \boldsymbol{v}_2^\top\big), \tag{3.2.6}$$

for some $r \geq 0$, with unit-norm vectors

$$\boldsymbol{u}_1 := \frac{1}{\sqrt{d}}\begin{bmatrix}\mathbf{1}_{d/2}\\ -\mathbf{1}_{d/2}\end{bmatrix}, \quad \boldsymbol{u}_2 := \frac{1}{\sqrt{d}}\begin{bmatrix}\mathbf{1}_{d/2}\\ \mathbf{1}_{d/2}\end{bmatrix}, \quad \boldsymbol{v}_1 := \sqrt{\frac{2}{n}}\begin{bmatrix}\mathbf{1}_{n/4}\\ -\mathbf{1}_{n/4}\\ \mathbf{0}_{n/2}\end{bmatrix}, \quad \boldsymbol{v}_2 := \sqrt{\frac{2}{n}}\begin{bmatrix}\mathbf{0}_{n/2}\\ \mathbf{1}_{n/4}\\ -\mathbf{1}_{n/4}\end{bmatrix}. \tag{3.2.7}$$

Here, the signal-to-noise ratio (SNR) is defined as $\text{SNR} := r^2$. In this case, the mean of each data point is defined by

$$\boldsymbol{\mu}_1 = r\cdot\boldsymbol{u}_1, \quad \boldsymbol{\mu}_2 = r\cdot\boldsymbol{u}_2, \quad \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_1, \quad \boldsymbol{\mu}_4 = -\boldsymbol{\mu}_2,$$

Following the above definition of $\boldsymbol{X}$, we consider a binary classification dataset

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n, \qquad y_i \in \{\pm 1\}$$

We assume the dataset is not linearly separable in the input space, i.e., there do not exist $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \quad \forall i \in [n].$$

For simplicity, in this case, the labels of data $\boldsymbol{X}$ are defined by

$$\boldsymbol{y}^\top = \{y_i\}_{i=1}^n = [\mathbf{1}_{n/2}^\top, -\mathbf{1}_{n/2}^\top]. \tag{3.2.8}$$

Our goal is to study the emergence of outlier eigenvalues in the spectrum of $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ defined by (3.2.1) and (3.2.5), and verify whether the corresponding eigenvectors (spikes) are aligned to this label vector $\boldsymbol{y}$, in order to (partially) recover the classes of $\boldsymbol{X}$.

Marchenko–Pastur law.

Given sample covariance $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z} \in \mathbb{R}^{n \times n}$ defined in (3.2.5), assume $d, n \to \infty$ with $\frac{n}{d} \to \psi \in (0, \infty)$. Then the ESD of $\boldsymbol{S}$ converges weakly almost surely to the Marchenko–Pastur (MP) law [MP67], denoted ρ_ψ^{MP}. Let $a = (1 - \sqrt{\psi})^2$, and $b = (1 + \sqrt{\psi})^2$. When $\psi \leq 1$, the MP law has an absolutely continuous part on $[a, b]$ with density

$$\rho_\psi^{\mathrm{MP}}(x) = \frac{1}{2\pi\psi x} \sqrt{(b-x)(x-a)}\, \mathbf{1}_{[a,b]}(x),$$

and, when $\psi > 1$, an additional atom at zero $\rho_\psi^{\mathrm{MP}} = \left(1 - \frac{1}{\psi}\right)\delta_0 + \rho_\psi^{\mathrm{MP}}(x)\, dx$.

Stieltjes transform.

The Stieltjes transform provides an analytical tool for studying the ESD in RMT. For $z \in \mathbb{C} \setminus [0, \infty)$, define the Stieltjes transform of ρ_ψ^{MP} by $m_\psi(z) = \int_{\mathbb{R}} \frac{1}{x - z} \rho_\psi^{\mathrm{MP}}(dx)$. This Stieltjes

transform fully captures the spectral distribution ρ_ψ^{MP}, as we can recover the density from an inversion formula: $\rho_\phi^{\mathrm{MP}}(x) = \frac{1}{\pi}\lim_{\eta\downarrow 0}\mathrm{Im}\, m_\phi(x+i\eta)$.

Limiting spectral distribution of CK.

From [LLC18, BP21, FW20], we know the limiting ESD of $\boldsymbol{K}$ is a deformed MP law, under certain conditions of $\boldsymbol{X}$ and in the proportional regime. In general, we denote the deformed MP law as

$$\mu := \rho_\phi^{\mathrm{MP}} \boxtimes \nu, \quad \nu := (1-b_\sigma^2) + b_\sigma^2 \rho_\psi^{\mathrm{MP}}, \tag{3.2.9}$$

where ρ_ϕ^{MP} is a standard MP law with parameter $\phi \in (0,\infty)$ and ν is a shift of MP law ρ_ψ^{MP} with aspect ratio $\psi \in (0,\infty)$ which is compactly supported in $[0,\infty)$. Here $\boxtimes$ denotes free multiplicative convolution. We can fully characterize this deformed MP law μ by its Stieltjes transform $m(z) = \int \frac{1}{x-z} d\mu(x)$. Notice that $\boldsymbol{Y}\boldsymbol{Y}^T$ and $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ share the same non-zero eigenvalues. Hence, we can also define a limit measure for the ESD of $\boldsymbol{Y}\boldsymbol{Y}^T$ by $\tilde{\mu}$ whose the Stieltjes transform is denoted by $\tilde{m}(z)$, which we will refer to as the companion Stieltjes transform of $m(z)$. Moreover, we have that $\widetilde{m}(z) = \frac{\phi-1}{z} + \phi m(z)$. We refer to [BS10, FW20, WWF24] for more details of free multiplicative convolution and these Stieltjes transforms. We denote the following transforms:

$$z(s) = -\frac{1}{s} + \phi \int \frac{x}{1+xs}\nu(dx), \qquad \varphi(s) = -\frac{sz'(s)}{z(s)}, \tag{3.2.10}$$

$$T(s) = \frac{z(s)s^2 - (\phi-1)s}{\phi}, \qquad T^{-1}(t) = -\frac{t\left(1-\psi b_\sigma^2 t\right)}{1+\left(1-\psi b_\sigma^2\right)t - \psi b_\sigma^2(1-b_\sigma^2)t^2}, \tag{3.2.11}$$

which will determine the outliers of $\boldsymbol{K}$. More properties of these transforms are introduced in Section 3.3.4.

3.2.2 Statement of Main Results

In this section, we present our main results on the outlier eigenvalues and eigenvectors of the CK $\mathbf{K}$ across several regimes. In the first three regimes, we consider proportional limit, where

$d, n, N \to \infty$ with fixed aspect ratios, while varying other "knobs" discussed in the introduction of this chapter. In particular: Section 3.2.3 treats the finite SNR $= \Theta(1)$ regime; Section 3.2.4 analyzes large SNR $= \Theta(n^{1/4})$; and Section 3.2.5 replaces the random weight matrix $\mathbf{W}$ with a "pretrained" weight matrix modeled as a rank-one perturbation, still with SNR $= \Theta(1)$. Then, in Section 3.2.6, we consider the quadratic sample-size regime $n \asymp d^2$, with finite SNR and random weights. In each of these cases, we predict the emergence of outlier eigenvalues and whether their associated eigenvectors enable *linear* classification.

3.2.3 Proportional Limit with Finite SNR Case

Here, we consider a baseline regime in which all dimensions grow proportionally and the SNR is finite. We show that the outlier eigenvectors cannot be used to linearly classify the dataset, even partially. When SNR $= \Theta(1)$, the structured XOR signal is too weak for quadratic informative spikes to emerge in the CK spectrum under proportional scaling.

Assumption 4 (Proportional limit)**.** *Let $n, d, N \to \infty$ such that $n/N \to \phi \in (0, +\infty)$ and $n/d \to \psi \in (0, +\infty)$ for some fixed constants ϕ and ψ.*

We denote some constants used to define the BBP thresholds in the following theorem:

$$\tau := \frac{c_\sigma^2}{2}\psi, \quad \Lambda_\tau := (1 - b_\sigma^2) + \tau + \frac{b_\sigma^2 \tau}{\tau - b_\sigma^2 \psi}, \quad \beta_{\text{mean}} := \frac{c_\sigma^2}{2}\psi, \quad \beta_{\text{lin}} := \frac{r^2 \psi}{2} b_\sigma^2. \tag{3.2.12}$$

Define $s_{\text{cov}} := -1/\Lambda_\tau$, $s_{\text{mean}} := T^{-1}(1/\beta_{\text{mean}})$ and $s_{\text{lin}} := T^{-1}(1/\beta_{\text{lin}})$ where $T^{-1}(\cdot)$ is given by (3.2.11). Denote $\tau_{\text{crit}} := b_\sigma^2 \sqrt{\psi}(1 + \sqrt{\psi})$.

Theorem 5. *Under Assumptions 3 and 4, we further assume that r is a constant. Consider the $\boldsymbol{K}$ with XOR data $\boldsymbol{X}$ defined in Section 3.2.1. Recall the probability measure μ with transforms $z(\cdot), \varphi(\cdot)$ and $T(\cdot)$ defined by* (3.2.10) *and* (3.2.11)*. Denote $\boldsymbol{u} = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $\boldsymbol{m} := \mathbb{E}[\sigma(\boldsymbol{w}^\top \boldsymbol{Z}) \mid \boldsymbol{Z}] \in \mathbb{R}^n$, where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ independent of $\boldsymbol{Z}$. Then, the CK matrix $\boldsymbol{K}$ satisfies the following:*

(i) ***(Bulk.)*** *The ESD of $\boldsymbol{K}$ converges weakly to μ defined in* (3.2.9)*.*

(ii) (BBP for covariance spike.) *If $\tau > \tau_{\text{crit}}$ and $z'(s_{\text{cov}}) > 0$, then $\boldsymbol{K}$ has an outlier eigenvalue $\widehat{\lambda}_\tau = z(s_{\text{cov}}) + o_{\mathbb{P}}(1)$ and the corresponding unit-norm eigenvector $\widehat{\boldsymbol{v}}_\tau$ satisfies $|\widehat{\boldsymbol{v}}_\tau^\top \boldsymbol{u}|^2 \xrightarrow{\mathbb{P}} \varphi(s_{\text{cov}})$. If $\tau \le \tau_{\text{crit}}$ or $z'(s_{\text{cov}}) \le 0$, then $\boldsymbol{K}$ has no such type of outlier eigenvalue.*

(iii) (BBP for mean spike.) *If $z'(s_{\text{mean}}) > 0$, then $\boldsymbol{K}$ has an outlier eigenvalue given by $\widehat{\lambda}_{\text{mean}} = z(s_{\text{mean}}) + o_{\mathbb{P}}(1)$, and its unit-norm eigenvector $\widehat{\boldsymbol{v}}_{\text{mean}}$ satisfies*

$$|\langle \widehat{\boldsymbol{v}}_{\text{mean}}, \boldsymbol{m}/\|\boldsymbol{m}\| \rangle|^2 = -\frac{\varphi(s_{\text{mean}})}{\beta_{\text{mean}}^2 s_{\text{mean}}^2 T'(s_{\text{mean}})} + o_{\mathbb{P}}(1)$$

when $s_{\text{mean}} \neq s_{\text{lin}}$. If $z'(s_{\text{mean}}) \le 0$, $\boldsymbol{K}$ has no such type of outlier eigenvalue.

(iv) (BBP linear spikes from XOR.) *If $z'(s_{\text{lin}}) > 0$, $\boldsymbol{K}$ has two outlier eigenvalues both satisfying $\widehat{\lambda}_{\text{lin},1}, \widehat{\lambda}_{\text{lin},2} = z(s_{\text{lin}}) + o_{\mathbb{P}}(1)$, and the projector of the corresponding two-dimension eigenspace $\widehat{\boldsymbol{V}}_{\text{lin}}$ satisfies*

$$\left\| \text{Proj}_{\text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}} \widehat{\boldsymbol{V}}_{\text{lin}} \right\|_F^2 = \; 2 \cdot -\frac{\varphi(s_{\text{lin}})}{\beta_{\text{lin}}^2 s_{\text{lin}}^2 T'(s_{\text{lin}})} + o_{\mathbb{P}}(1), \tag{3.2.13}$$

when $s_{\text{mean}} \neq s_{\text{lin}}$, where $\text{Proj}_{\text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}} = \boldsymbol{v}_1 \boldsymbol{v}_1^\top + \boldsymbol{v}_2 \boldsymbol{v}_2^\top$. If $z'(s_{\text{lin}}) \le 0$, $\boldsymbol{K}$ has no such type of outlier eigenvalues.

(v) (No label alignment.) *For the XOR labels $\boldsymbol{y}$, every outlier eigenvector of $\boldsymbol{K}$ is asymptotically orthogonal to $\boldsymbol{y}$, i.e., the orthogonal projector $\widehat{\boldsymbol{V}}_{\text{out}}$ onto the direct sum of all separated outlier eigenspaces of $\boldsymbol{K}$ described above satisfies:*

$$\frac{1}{n} \|\widehat{\boldsymbol{V}}_{\text{out}} \boldsymbol{y}\|^2 = o_{\mathbb{P}}(1). \tag{3.2.14}$$

Theorem 5 exhibits a BBP-type phase transition for the CK matrix with XOR data: e.g., $z'(-1/\Lambda_\tau) > 0$ is the threshold showing when a spike separates from the deformed MP bulk μ; in the separated case, the associated eigenvectors have nontrivial asymptotic overlap with the

spike directions.

There are up to four candidate spikes outside $\mathrm{supp}(\mu)$: two outliers (Theorem 5(iv)) induced by the linear Hermite component of σ and spike strength s_{lin} depending on r, and two potential "architectural" outliers (Theorem 5(ii)&(iii)) corresponding to $c_\sigma \neq 0$. The latter is "uninformative" in the sense that it persists even in the null model $r = 0$ (pure noise data). See Section 3.4 for further discussion. This is why some earlier analyses impose $c_\sigma = 0$ [WWF24]; see also [BP22] for related phenomena. The proof of Theorem 5 is given in Section 3.3.8.

Remark. The uninformative spike in Theorem 5(ii) can be removed if $\tau \leq \tau_{\mathrm{crit}}$ holds, namely $\sqrt{\psi}\big(\frac{c_\sigma^2}{2} - b_\sigma^2\big) \leq b_\sigma^2$. Hence, this provides a necessary condition for the absence of an uninformative spike, e.g., the second Hermite coefficient $c_\sigma/\sqrt{2}$ is smaller than the first b_σ in absolute value (for instance, a centered and normalized GELU function in Section 3.3.5).

Remark. From the final statement of Theorem 5 in (3.2.14), the separated outlier eigenvectors are asymptotically orthogonal to the label vector $\boldsymbol{y}$. Thus, these outliers cannot be used for linear classification. Although the spikes induced by λ_{lin} carry information about the within-class structure (alignment with $\boldsymbol{v}_1, \boldsymbol{v}_2$ in (3.2.13)), they do not recover the XOR labels without an additional nonlinear transform.

Kernel Clustering also fails

We also consider classical kernel spectral clustering with an Euclidean distance kernel. We focus in particular on the kernel spectral clustering algorithm introduced by [NJW01], using the leading (nontrivial) eigenvectors of the normalized Laplacian for clustering. [CBG16] studied this kernel clustering for general Gaussian mixture data. The conclusion is similar as above. Results in this section on kernel spectral clustering are applications of the main results of [CBG16].

Recall the four-component XOR Gaussian mixture data $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ defined in (3.2.5),

with cluster means

$$\boldsymbol{\mu}_1 = \frac{r}{\sqrt{d}} \begin{pmatrix} \mathbf{1}_{d/2} \\ -\mathbf{1}_{d/2} \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \frac{r}{\sqrt{d}} \begin{pmatrix} \mathbf{1}_{d/2} \\ \mathbf{1}_{d/2} \end{pmatrix}, \quad \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_1, \quad \boldsymbol{\mu}_4 = -\boldsymbol{\mu}_2,$$

and common covariance $\boldsymbol{C} = \mathbf{I}_d$. Define the matrix of cluster means $\boldsymbol{M} := [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4] \in \mathbb{R}^{d \times 4}$, and cluster-indicator matrix $\boldsymbol{J}$

$$\boldsymbol{J} \in \{0,1\}^{n \times 4}, \qquad J_{i,a} = 1 \Leftrightarrow \boldsymbol{x}_i \text{ belongs to cluster } a,$$

so that each cluster $a \in [4]$ has size $n_a = n/4$. For an affinity function $f : \mathbb{R} \to \mathbb{R}$, define the Kernel matrix $\boldsymbol{K}^{(f)} \in \mathbb{R}^{n \times n}$ as

$$(\boldsymbol{K}^{(f)})_{ij} = f\Big(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\Big), \qquad i, j \in [n] \tag{3.2.15}$$

We consider the normalized Laplacian, which is defined as

$$\boldsymbol{L} := n\boldsymbol{D}^{-1/2}\boldsymbol{K}^{(f)}\boldsymbol{D}^{-1/2}, \quad \boldsymbol{D} = \operatorname{diag}(\boldsymbol{K}^{(f)}\mathbf{1}_n)$$

where $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{K}^{(f)}\mathbf{1}_n)$ is the degree matrix. Similar to the results on the spectra of the CK in Section 3.2.2, the analysis of the eigenvalues on $\boldsymbol{L}$ relies on a LE for $\boldsymbol{L}$. We refer to Theorem 1 of [CBG16] for more details.

It is not difficult to prove that under Assumption 4, the pairwise distances between distinct samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ of the XOR dataset concentrate around $\tau = 2$ (see Lemma 24 for details). The core of the analysis of $\boldsymbol{L}$ then relies upon a careful entry-wise Taylor expansion of the Kernel $\boldsymbol{K}^{(f)}$ around the limiting value τ. We therefore make the following assumption on the kernel function $f : \mathbb{R} \to \mathbb{R}$.

Assumption 5. *Let $\tau = 2$. The kernel function $f \in C^3$ is three-times differentiable in some*

neighborhood of τ. Further, $f(\tau) > 0$ and $f'(\tau) \neq 0$.

Theorem 6. *Suppose* SNR $r^2 = \Theta(1)$ *and let Assumptions 4 and 5 hold.*

(i) (BBP for two linear informative outliers.) *If $r^2 > 2\sqrt{\psi^{-1}}$, then there exists two isolated eigenvalues of $\boldsymbol{L}$, which are asymptotically approximated by*

$$-\frac{2f'(\tau)}{f(\tau)}\rho + \frac{f(0) - f(\tau) + \tau f'(\tau)}{f(\tau)} \tag{3.2.16}$$

where

$$\rho = \psi\big(1 + \frac{r^2}{2}\big) + \frac{2 + r^2}{r^2} \tag{3.2.17}$$

(ii) (BBP for additional non-informative outlier.) *Let $\ell_+ = \frac{5f'(\tau)}{4f(\tau)} - \frac{f''(\tau)}{f'(\tau)}$. If, in addition, $(1-\ell_+)^2 > \psi^{-1}$, then there is an additional corresponding isolated eigenvalue of $\boldsymbol{L}$, which is approximated by*

$$\lambda_+ := -\frac{2f'(\tau)}{f(\tau)}\rho_+ + \frac{f(0) - f(\tau) + \tau f'(\tau)}{f(\tau)},$$

in probability, where $\rho_+ = \ell_+\big(\psi - \frac{1}{1-\ell_+}\big)$. These and the trivial eigenvalue n represent all isolated eigenvalues of $\boldsymbol{L}$.

(iii) (Eigenvector alignment.) *Moreover, if ρ is such an isolated eigenvalue of $\boldsymbol{L}$ corresponding to* (3.2.16)*, let $\boldsymbol{\Pi}_\rho$ denote the orthogonal projection onto the two-dimensional eigenspace corresponding to ρ. Then we have*

$$\frac{1}{d}\boldsymbol{J}^\top \boldsymbol{\Pi}_\rho \boldsymbol{J} \to \frac{1}{1 + \frac{r^2}{2}}\Big(\psi - \frac{4}{r^4}\Big)\frac{r^2}{16}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2 \tag{3.2.18}$$

where $\otimes$ denotes the Kronecker product. In particular, in probability

$$\frac{1}{d}\boldsymbol{y}^\top \boldsymbol{\Pi}_\rho \boldsymbol{y} \to 0.$$

where $\boldsymbol{y}$ denotes the XOR-labels.

Remark. The trivial eigenvalue n, and the possible additional isolated eigenvalue of $\boldsymbol{L}$ corresponding to λ_+ carry no relevant clustering information. The other non-trivial isolating eigenvalues of $\boldsymbol{L}$ characterized in Theorem 6 (i) turn out to be the only eigenvalues that carry relevant clustering information. Similar to the eigenvectors of the CK matrix, the last statement of Theorem 6 implies that the outlier eigenvectors of the Laplacian $\boldsymbol{L}$ are asymptotically orthogonal to the XOR-labels $\boldsymbol{y}$.

3.2.4 Large-SNR Proportional Limit Case

Theorem 5 shows that spectral methods based on outlier eigenvectors fail to linearly classify XOR in the finite-SNR proportional regime. Linear informative spikes may still emerge, but a purely linear approximation of σ cannot capture the quadratic feature channel needed for XOR. To make XOR linearly learnable, we can increase the SNR so that the quadratic Hermite component of σ becomes non-vanishing. Here, we confirm this: in the large-SNR regime, the additional two outliers created by the quadratic component become label-informative, enabling linear classification.

Theorem 7. *Under Assumptions 3 and 4, $r = r_0 n^{1/4}$ for constant $r_0 \in (0,\infty)$, recall τ in* (3.2.12)*, and let*

$$\boldsymbol{u} := \mathbf{1}_n/\sqrt{n}, \qquad \kappa_0 = \frac{r_0^4 \psi^2}{4}, \qquad \beta_{\text{quad}} = c_\sigma^2 \kappa_0, \qquad \tilde{\beta}_{\text{quad}} = \beta_{\text{quad}} + \tau.$$

Let $\boldsymbol{P}_{\text{lin}} = \boldsymbol{v}_1\boldsymbol{v}_1^\top + \boldsymbol{v}_2\boldsymbol{v}_2^\top$, $\boldsymbol{P}_\perp = \boldsymbol{I} - \boldsymbol{P}_{\text{lin}}$, and $\boldsymbol{K}_\perp := \boldsymbol{P}_\perp \boldsymbol{K} \boldsymbol{P}_\perp$. Then, we have the following:

(i) (Bulk.) *The ESD of $\boldsymbol{K}$ converges weakly to μ.*

(ii) (Two diverging linear outliers.) *If $b_\sigma \neq 0$, then the top two eigenvalues satisfy*

$$\widehat{\lambda}_1(\boldsymbol{K}) = b_\sigma^2 r_0^2 \psi \cdot n^{1/2}\,(1 + o_{\mathbb{P}}(1)), \qquad \widehat{\lambda}_2(\boldsymbol{K}) = b_\sigma^2 r_0^2 \psi \cdot n^{1/2}\,(1 + o_{\mathbb{P}}(1)),$$

and the associated 2D eigenspace projector $\widehat{\boldsymbol{P}}_{\text{lin}}$ *satisfies* $\|\widehat{\boldsymbol{P}}_{\text{lin}} - \boldsymbol{P}_{\text{lin}}\|_{\text{F}} \xrightarrow{\mathbb{P}} 0$.

(iii) (BBP for mean spike.) *Same BBP result of Theorem 5 (iii) holds for* $\boldsymbol{K}_{\perp}$.

(iv) (Quadratic u-spike.) *Denote* $\tilde{s}_{\text{quad}} = T^{-1}(1/\tilde{\beta}_{\text{quad}})$. *If* $z'(\tilde{s}_{\text{quad}}) > 0$, *then* $\boldsymbol{K}_{\perp}$ *has an outlier* $\widehat{\lambda}_u^{\perp} = z(\tilde{s}_{\text{quad}}) + o_{\mathbb{P}}(1)$ *outside* $\text{supp}(\mu)$ *and the corresponding unit eigenvector* $\widehat{\boldsymbol{v}}_u$ *satisfies* $|\widehat{\boldsymbol{v}}_u^{\top}\boldsymbol{u}|^2 \to \frac{\tilde{s}_{\text{quad}} z'(\tilde{s}_{\text{quad}})}{\tilde{\beta}_{\text{quad}} T'(\tilde{s}_{\text{quad}})}$ *in probability.*

(v) (Quadratic y-spike; label alignment.) *Denote* $s_{\text{quad}} = T^{-1}(1/\beta_{\text{quad}})$. *If* $z'(s_{\text{quad}}) > 0$, *then* $\boldsymbol{K}_{\perp}$ *has an outlier* $\widehat{\lambda}_y^{\perp} = z(s_{\text{quad}}) + o_{\mathbb{P}}(1)$ *outside* $\text{supp}(\mu)$ *and the corresponding unit eigenvector* $\widehat{\boldsymbol{v}}_y$ *satisfies*

$$\frac{1}{n}|\widehat{\boldsymbol{v}}_y^{\top}\boldsymbol{y}|^2 \xrightarrow{\mathbb{P}} \frac{s_{\text{quad}} z'(s_{\text{quad}})}{\beta_{\text{quad}} T'(s_{\text{quad}})} > 0. \tag{3.2.19}$$

In this case (and in contrast to Theorem 5), when $r = \Theta(n^{1/4})$ the linear outliers characterized by Theorem 5 become diverging. Although Theorem 7 yields two diverging linear outliers, their eigenvectors remain asymptotically orthogonal to the label vector $\boldsymbol{y}$. Meanwhile, the quadratic component of σ is no longer negligible, and instead it produces two additional $O(1)$ outliers (Theorem 7(iv)&(v)). Here, (3.2.19) gains a positive alignment with labels $\boldsymbol{y}$ because the quadratic features $\boldsymbol{v}_1^{\odot 2}$ and $\boldsymbol{v}_2^{\odot 2}$ are preserved. Hence, a linear readout on the $\boldsymbol{y}$-spike eigenspace in (3.2.19) can classify XOR. The proof of Theorem 7 is in Section 3.3.15.

When $b_\sigma = 0$, the diverging linear outliers vanish, simplifying spectral classification. In contrast, if $c_\sigma = 0$, the quadratic component is absent, and linear classification is no longer possible, thus highlighting the role of the activation knob in preserving nonlinear structure. For further discussion about the role of non-linearity choice, we refer the reader to Section 3.4.

3.2.5 Pretrained / Spiked Weight Matrix Case

Empirically, pretrained weight matrices in well-trained NNs often exhibit spectral spikes (and even heavy-tailed) [MM21, MM19a, MM20, MPM21, YTH+22]. Such spectral behavior suggests that training induces nontrivial feature structure in the weights [BES+22, BES+23,

DD25]. However, the above theorems consider only random features, and therefore they do not capture task-dependent structure embedded in the weight $\boldsymbol{W}$. As a toy (but principled) model for feature learning, we consider a simple but nontrivial spiked model in which the weight is a rank-one perturbation of the random initialization:

$$\boldsymbol{W}_1 = \boldsymbol{W} + \theta\, \boldsymbol{a}\boldsymbol{b}^\top, \tag{3.2.20}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent of the *test* dataset $\boldsymbol{X}$. A related rank-one structure arises, for example, after one-step gradient descent in certain teacher–student settings [BES$^+$22, DD25], and it is also connected to LoRA [HSW$^+$21]. Below, we analyze a *test-time* BBP transition, i.e., we evaluate the CK built from $\boldsymbol{W}_1$ on independent XOR test data $\boldsymbol{X}$, and we characterize the emergence and alignment of the induced spike. Empirical simulations of test-time BBP are presented in Section 3.4.

Theorem 8. *Assume that $\theta = \theta_0 N^{1/4}$ for some constant $\theta_0 > 0$, $\boldsymbol{b} = \mathbf{1}_d$, and $\boldsymbol{a} \sim \mathcal{N}(0, \frac{1}{N}\boldsymbol{I})$ in* (3.2.20)*. Under Assumptions 3 and 4, consider CK computed from $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ and $\boldsymbol{Y} = \sigma(\boldsymbol{W}_1 \boldsymbol{X})/\sqrt{N}$ on test data $\boldsymbol{X}$ independent of $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$. Suppose that r is a fixed constant. Define*

$$\kappa = 1 + \frac{r^2}{2}, \qquad \beta_y = \frac{3c_\sigma^2}{4}\theta_0^4\, \phi\, \kappa_\perp, \qquad \kappa_\perp = 2 + 2r^2 + \frac{r^4}{4}, \qquad s_y = T^{-1}(1/\beta_y).$$

Then we have the following:

- ***(i) (Bulk.)*** *The ESD of $\boldsymbol{K}$ converges to μ.*
- ***(ii) (One giant outlier.)*** *The top eigenvalue satisfies $\widehat{\lambda}_1(\boldsymbol{K}) = b_\sigma^2 \theta_0^2\, \phi\, \kappa \sqrt{N}\,(1 + o_{\mathbb{P}}(1))$ and its eigenvector $\widehat{\boldsymbol{v}}_1$ asymptotically orthogonal to $\boldsymbol{y}$.*
- ***(iii) (Order-one spikes orthogonal to y.)*** *All outliers captured by Theorem 5 still hold here.*
- ***(iv) (Additional order-one spike.)*** *If $z'(s_y) > 0$, $\boldsymbol{K}$ exhibits one additional outlier eigenvalue $\widehat{\lambda}_y = z(s_y) + o_{\mathbb{P}}(1)$ separating from the bulk. If $z'(s_y) \le 0$, no such separation occurs.*

(v) ***(Label alignments.)*** *In the supercritical case $z'(s_y) > 0$, let $\widehat{\boldsymbol{v}}_y$ be the unit eigenvector of $\boldsymbol{K}$ associated to $\widehat{\lambda}_y$. Then $\frac{1}{n}\left|\widehat{\boldsymbol{v}}_y^\top \boldsymbol{y}\right|^2 \xrightarrow{\mathbb{P}} \frac{s_y z'(s_y)}{\beta_y T'(s_y)} \cdot \frac{r^4/4}{\kappa_\perp} > 0$.*

This result shows that the large spike in the pretrained matrix $\boldsymbol{W}_1$ amplifies the signals in XOR and alters the CK approximation: the usual LE is no longer sufficient, and a QE appears to capture the induced quadratic feature channel. Theorem 8(v) shows this quadratic eigenvector aligns with the labels. However, if $c_\sigma = 0$, then this quadratic channel is absent and we do not obtain label alignment.

3.2.6 Quadratic Sample-size Case

In many ML settings, sample complexity dictates how many data points are sufficient for an algorithm to succeed in high dimensions. In Section 3.2.3, sample complexity $n = \Theta(d)$ implies that LE washes out higher-order structure, so the kernel behaves like a "linear" model. This makes the XOR problem difficult to learn, because it is not linearly separable in first-order statistics.

Here, we retain the same data model as in Section 3.2.3, but we change into the quadratic sample-size regime $n = \Theta(d^2)$. Under this regime, broad kernel classes admit operator-norm reductions to quadratic polynomial kernels [XHM+22, WHL+25]. Below we show that the CK behaves like a quadratic kernel and thus becomes label-informative for XOR even at finite SNR.

Theorem 9. *Assume $2n/d^2 \to \gamma \in (0,\infty)$, $n^2/N \to 0$, and Assumption 3 holds with $c_\sigma \neq 0$. Let $\ell = r^4/4$. We further assume that there exists some fixed $\delta' > 0$ such that $\mathbb{E}[\xi\,\sigma(s\xi)] = \mathbb{E}[\sigma(s\xi)] = 0$ for all $s \in [1-\delta', 1+\delta']$ where $\xi \sim \mathcal{N}(0,1)$. Recall $\boldsymbol{P} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, $\boldsymbol{H}_c = \boldsymbol{P}(\boldsymbol{X}^\top\boldsymbol{X})^{\odot 2}\boldsymbol{P}$, and $\boldsymbol{K}_c = \boldsymbol{P}\boldsymbol{K}\boldsymbol{P}$. Then we have*

(A) Bulk.

The ESD of $\boldsymbol{K}$ converges to MP law $\rho_\gamma^{\mathrm{MP}}$ whose upper edge is $\lambda_+(\gamma) := (1+\sqrt{\gamma})^2$. And

(B) BBP for $\boldsymbol{H}_c$.

1. *(*Subcritical*) If $\ell \le 1/\sqrt{\gamma}$ (equivalently $\theta \le \gamma^{1/4}$), then*

$$\lambda_{\max}(\boldsymbol{H}_c) \xrightarrow{\mathbb{P}} \lambda_+(\gamma), \qquad \left|\langle \boldsymbol{v}_{\max}(\boldsymbol{H}_c), \boldsymbol{y}/\sqrt{n}\rangle\right|^2 \xrightarrow{\mathbb{P}} 0.$$

2. *(*Supercritical*) If $\ell > 1/\sqrt{\gamma}$ (equivalently $\theta > \gamma^{1/4}$), then $\boldsymbol{H}_c$ has a unique outlier:*

$$\lambda_{\max}(\boldsymbol{H}_c) \xrightarrow{\mathbb{P}} \lambda_{\text{out}}(\gamma,\ell) := (1+\theta^2)\Big(1+\frac{\gamma}{\theta^2}\Big) = 1+\gamma+\gamma\ell+\frac{1}{\ell},$$

and the top eigenvector aligns with the label:

$$\left|\langle \boldsymbol{v}_{\max}(\boldsymbol{H}_c), \boldsymbol{y}/\sqrt{n}\rangle\right|^2 \xrightarrow{\mathbb{P}} \text{Align}(\gamma,\ell) := \frac{1-\gamma/\theta^4}{1+\gamma/\theta^2} = \frac{\gamma\ell^2-1}{\gamma\ell(\ell+1)}.$$

(C) Transfer to centered CK.

With $\alpha_2 = c_\sigma^2/2$ and $\alpha_0 = 1-\alpha_2$ (since $b_\sigma = 0$),

$$\lambda_{\max}(\boldsymbol{K}_c) = \alpha_0 + \alpha_2\,\lambda_{\max}(\boldsymbol{H}_c) + o_{\mathbb{P}}(1), \quad \left|\langle \boldsymbol{v}_{\max}(\boldsymbol{K}_c), \boldsymbol{y}/\sqrt{n}\rangle\right|^2 = \left|\langle \boldsymbol{v}_{\max}(\boldsymbol{H}_c), \boldsymbol{y}/\sqrt{n}\rangle\right|^2 + o_{\mathbb{P}}(1).$$

Hence $\boldsymbol{K}_c$ exhibits the same BBP threshold and the same label-alignment curve.

For the full proof, see Section 3.3.19. In addition to quadratic sample complexity, Theorem 9 also assumes an ultra-wide network, leading to the concentration of the random kernel $\boldsymbol{K}$ around a fixed kernel conditional on the dataset. This simplifies the analysis. We leave the case $N \asymp n$ (linear width) for future work [WHL$^+$25], as this would involve additional spikes. Moreover, we assume $b_\sigma = 0$ to remove the linear term $\boldsymbol{X}^\top\boldsymbol{X}$, which contributes rank-d spikes.

3.3 Proofs of Main Results

3.3.1 Preliminaries

Throughout this chapter, we utilize the stochastic domination notation introduced by [EKY13].

Definition 22 (Stochastic domination)**.** *Consider any two families of random variables*

$$X = \{X_n(u) : n \in \mathbb{N}, u \in U_n\},\ Y = \{Y_n(u) : n \in \mathbb{N}, u \in U_n\},$$

with parameter $u \in U_n$ where U_n is a possibly n-dependent deterministic set. Let $\mathscr{E} = \{E_n\}_{n\in\mathbb{N}}$ be a sequence of measure subsets, where each E_n is related to $X_n(u)$ and $Y_n(u)$.

1. *We denote $|X| \prec Y$ or $X = O_{\prec}(Y)$ as the stochastic domination of X by Y uniformly in $u \in U_n$: if for all small $\varepsilon > 0$ and large $D > 0$, there exists some $N_0(\varepsilon, D)$, such that for all $n \geq N_0(\varepsilon, D)$,*

$$\sup_{u\in U_n} \mathbb{P}(|X_n(u)| \geq n^{\varepsilon} Y_n(u)) \leq n^{-D}.$$

2. *For simplicity, for a family of random matrices $\boldsymbol{A}$ and a family of non-negative random variables ζ, $\boldsymbol{A} = O_{\prec}(\zeta)$ represents $|\langle \boldsymbol{v}, \boldsymbol{A}\boldsymbol{u}\rangle| \prec \zeta \|\boldsymbol{v}\|\|\boldsymbol{u}\|$ uniformly for all deterministic vectors $\boldsymbol{v}$ and $\boldsymbol{u}$.*

If $\{\boldsymbol{A}_n\}$ are random matrices, we interpret

$$\|\boldsymbol{A}_n\|_{\text{op}} = o_{\mathbb{P}}(1) \quad \Longleftrightarrow \quad \|\boldsymbol{A}_n\|_{\text{op}} \xrightarrow{\mathbb{P}} 0.$$

For simplicity, in the proofs, we use $\|\boldsymbol{A}_n\|$ to represent operator norm $\|\boldsymbol{A}_n\|_{\text{op}}$.

3.3.2 Properties of XOR Data

We introduce the following notion of $(\tau_n, B)-$**orthonormality** for our feature matrix [FW20].

Definition 23. *A matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ is $(\tau_n, B)-$**orthonormal** if the columns $\{\boldsymbol{x}_\alpha\}_{\alpha=1}^n$ of $\boldsymbol{X}$ satisfy*

$$\big|\|\boldsymbol{x}_\alpha\| - 1\big| \le \tau_n, \quad \big|\|\boldsymbol{x}_\beta\| - 1\big| \le \tau_n, \quad \big|\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta\big| \le \tau_n \tag{3.3.1}$$

for all pairs $\alpha \neq \beta \in [n]$, and also

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

Lemma 24. *Let $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ be the XOR data matrix:*

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{M}, \qquad (\boldsymbol{Z})_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(0, \frac{1}{d}\Big), \qquad \boldsymbol{M} = \theta\, \boldsymbol{u}_1 \boldsymbol{v}_1^\top + \theta\, \boldsymbol{u}_2 \boldsymbol{v}_2^\top, \qquad \theta_{\text{SNR}} = r\sqrt{\frac{n}{2d}},$$

where $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$ are orthonormal vectors defined by (3.2.7)*. Assume either*

(A1) *$n \asymp d$ and $r = O(n^{1/4})$, or*

(A2) *$n = O(d^2)$ and $r = O(1)$.*

Then there exist constants $c, C > 0$ such that with probability at least $1 - n^{-c}$, the matrix $\boldsymbol{X}$ satisfies (3.3.1) *with*

$$\tau_n \;=\; C\Big(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d}\Big). \tag{3.3.2}$$

Moreover, on the same event,

$$\sum_{\alpha=1}^n \big(\|\boldsymbol{x}_\alpha\| - 1\big)^2 \;\le\; \sum_{\alpha=1}^n \big(\|\boldsymbol{x}_\alpha\|^2 - 1\big)^2 \;\le\; C\Big(\frac{n}{d} + \frac{nr^2}{d^2} + \frac{nr^4}{d^2}\Big), \tag{3.3.3}$$

and also

$$\|\boldsymbol{X}\| \le C\Big(1+(1+r)\sqrt{\frac{n}{d}}\Big). \tag{3.3.4}$$

In particular, in regime (A1) one may take $\tau_n = C\sqrt{\frac{\log n}{n}}$ *and* $B = C(1+r)$*, and in regime (A2) one may take* $\tau_n = C\sqrt{\frac{\log n}{d}}$ *and* $B = C\sqrt{\frac{n}{d}}$ *such that* $\boldsymbol{X}$ *is* $(\tau_n, B)-$***orthonormal*** *with probability at least* $1-n^{-c}$*.*

Proof. Throughout the proof, $c, C > 0$ denote constants that may change from line to line. This proof is basically following the proof of Proposition 3.3 of [FW20].

Write the columns as $\boldsymbol{x}_\alpha = \boldsymbol{z}_\alpha + \boldsymbol{m}_\alpha$, where $\boldsymbol{z}_\alpha$ is the α-th column of $\boldsymbol{Z}$ and $\boldsymbol{m}_\alpha$ is the α-th column of $\boldsymbol{M}$. Since $\boldsymbol{M} = \theta_{\mathrm{SNR}}\,\boldsymbol{u}_1\boldsymbol{v}_1^\top + \theta_{\mathrm{SNR}}\,\boldsymbol{u}_2\boldsymbol{v}_2^\top$ and $\mathrm{supp}(\boldsymbol{v}_1)\cap\mathrm{supp}(\boldsymbol{v}_2) = \emptyset$, for each α exactly one of $\boldsymbol{v}_1[\alpha], \boldsymbol{v}_2[\alpha]$ is nonzero, and we have

$$\boldsymbol{m}_\alpha = \theta_{\mathrm{SNR}}\,\boldsymbol{u}_1\boldsymbol{v}_1[\alpha] \;\text{ or }\; \theta_{\mathrm{SNR}}\,\boldsymbol{u}_2\boldsymbol{v}_2[\alpha].$$

Using $|\boldsymbol{v}_1[\alpha]| = |\boldsymbol{v}_2[\alpha]| = \sqrt{2/n}$ on the support, we obtain for all α

$$\|\boldsymbol{m}_\alpha\| = \theta_{\mathrm{SNR}}\sqrt{\frac{2}{n}} = r\sqrt{\frac{n}{2d}}\sqrt{\frac{2}{n}} = \frac{r}{\sqrt{d}}, \qquad \|\boldsymbol{m}_\alpha\|^2 = \frac{r^2}{d}, \qquad \|\boldsymbol{m}_\alpha\|^4 = \frac{r^4}{d^2}. \tag{3.3.5}$$

Now we verify the four requirements in the (τ_n, B)-orthonormal definition: (i) $\max_\alpha |\|\boldsymbol{x}_\alpha\| - 1| \le \tau_n$, (ii) $\max_{\alpha\neq\beta} |\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta| \le \tau_n$, (iii) $\sum_\alpha (\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \le B^2$, (iv) $\|\boldsymbol{X}\|_{\mathrm{op}} \le B$.

Step 1: Uniform control of column norms.

For each $\alpha \in [n]$,

$$\|\boldsymbol{x}_\alpha\|^2 = \|\boldsymbol{z}_\alpha\|^2 + 2\,\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha + \|\boldsymbol{m}_\alpha\|^2. \tag{3.3.6}$$

(1a) Noise norm term. Since $\boldsymbol{z}_\alpha \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$, we have $d\|\boldsymbol{z}_\alpha\|^2 \sim \chi^2_d$. By Lemma 1 of [LM00]:

for all $t > 0$,

$$\mathbb{P}\Big(\big|\|\boldsymbol{z}_\alpha\|^2 - 1\big| \geq 2\sqrt{\tfrac{t}{d}} + 2\tfrac{t}{d}\Big) \leq 2e^{-t}. \tag{3.3.7}$$

Taking $t = C_0 \log n$ and union bounding over $\alpha \in [n]$ yields: with probability at least $1 - 2n \cdot n^{-C_0}$, i.e. $1 - n^{-c}$ for C_0 large enough,

$$\max_{\alpha\in[n]} \big|\|\boldsymbol{z}_\alpha\|^2 - 1\big| \;\leq\; C\sqrt{\frac{\log n}{d}}. \tag{3.3.8}$$

In particular, on the same event, $\max_\alpha \|\boldsymbol{z}_\alpha\|^2 \leq 2$ for all large n.

(1b) Cross term. Conditioning on $\boldsymbol{m}_\alpha$, the scalar $\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha$ is Gaussian with

$$\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha \,\big|\, \boldsymbol{m}_\alpha \sim \mathcal{N}\Big(0, \frac{\|\boldsymbol{m}_\alpha\|^2}{d}\Big) = \mathcal{N}\Big(0, \frac{r^2}{d^2}\Big),$$

using (3.3.5). Hence for $t = C_0 \log n$,

$$\mathbb{P}\Big(|\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha| \geq \frac{r}{d}\sqrt{2t}\Big) \leq 2e^{-t}.$$

A union bound over $\alpha \in [n]$ yields: with probability at least $1 - n^{-c}$,

$$\max_{\alpha\in[n]} |\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha| \;\leq\; C\,\frac{r\sqrt{\log n}}{d}. \tag{3.3.9}$$

(1c) Combine. Plugging (3.3.8), (3.3.9), and $\|\boldsymbol{m}_\alpha\|^2 = r^2/d$ into (3.3.6) gives, with probability at least $1 - n^{-c}$,

$$\max_{\alpha\in[n]} \big|\|\boldsymbol{x}_\alpha\|^2 - 1\big| \;\leq\; C\Big(\sqrt{\frac{\log n}{d}} + \frac{r\sqrt{\log n}}{d} + \frac{r^2}{d}\Big). \tag{3.3.10}$$

We now simplify the RHS under regimes (A1) and (A2). In both regimes, for large n,

$$\frac{r\sqrt{\log n}}{d} \;\leq\; \sqrt{\frac{\log n}{d}} + \frac{r^2}{d}.$$

Indeed:

- If $r^2 \geq \sqrt{\log n}$, then $\frac{r\sqrt{\log n}}{d} \leq \frac{r^2}{d}$.
- If $r^2 \leq \sqrt{\log n}$, then $r \leq (\log n)^{1/4}$ so $\frac{r\sqrt{\log n}}{d} \leq \frac{(\log n)^{3/4}}{d} \leq \sqrt{\frac{\log n}{d}}$ for all large d.

Therefore (3.3.10) implies

$$\max_{\alpha\in[n]} \big|\|\boldsymbol{x}_\alpha\|^2 - 1\big| \;\leq\; C\Big(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d}\Big). \tag{3.3.11}$$

Finally, convert squared-norm deviation to norm deviation: since $\|\boldsymbol{x}_\alpha\| + 1 \geq 1$,

$$\big|\|\boldsymbol{x}_\alpha\| - 1\big| = \frac{\big|\|\boldsymbol{x}_\alpha\|^2 - 1\big|}{\|\boldsymbol{x}_\alpha\| + 1} \leq \big|\|\boldsymbol{x}_\alpha\|^2 - 1\big|,$$

so (3.3.11) yields

$$\max_{\alpha\in[n]} \big|\|\boldsymbol{x}_\alpha\| - 1\big| \;\leq\; C\Big(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d}\Big). \tag{3.3.12}$$

This verifies the column-norm part of (τ_n, B)-orthonormality with τ_n as in (3.3.2).

Step 2: Uniform control of pairwise inner products.

Fix $\alpha \neq \beta \in [n]$. Expand

$$\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta = \boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta + \boldsymbol{z}_\alpha^\top \boldsymbol{m}_\beta + \boldsymbol{m}_\alpha^\top \boldsymbol{z}_\beta + \boldsymbol{m}_\alpha^\top \boldsymbol{m}_\beta. \tag{3.3.13}$$

(2a) Noise–noise term. The scalar $\boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta$ is not exactly Gaussian, but it is conditionally Gaussian: conditioning on $\boldsymbol{z}_\alpha$,

$$\boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta \,\big|\, \boldsymbol{z}_\alpha \sim \mathcal{N}\Big(0, \frac{\|\boldsymbol{z}_\alpha\|^2}{d}\Big).$$

On the event $\max_\alpha \|\boldsymbol{z}_\alpha\|^2 \leq 2$ from Step 1, we have conditional variance $\leq 2/d$ and hence

$$\mathbb{P}\Big(|\boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta| \geq t \;\Big|\; \boldsymbol{z}_\alpha\Big) \leq 2\exp\Big(-\frac{dt^2}{4\|\boldsymbol{z}_\alpha\|^2}\Big) \leq 2\exp\Big(-c\,dt^2\Big).$$

Taking $t = C\sqrt{\frac{\log n}{d}}$ and union bounding over all $\alpha \neq \beta \in [n]$ (at most n^2 pairs) yields with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta| \;\leq\; C\sqrt{\frac{\log n}{d}}. \tag{3.3.14}$$

(2b) Noise–signal terms. For $\alpha \neq \beta$, conditional on $\boldsymbol{m}_\beta$,

$$\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\beta \sim \mathcal{N}\Big(0, \frac{\|\boldsymbol{m}_\beta\|^2}{d}\Big) = \mathcal{N}\Big(0, \frac{r^2}{d^2}\Big),$$

and similarly $\boldsymbol{m}_\alpha^\top \boldsymbol{z}_\beta \sim \mathcal{N}(0, r^2/d^2)$. A union bound over all $\alpha \neq \beta$ gives, with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\beta| \;\leq\; C\frac{r\sqrt{\log n}}{d}, \qquad \max_{\alpha \neq \beta} |\boldsymbol{m}_\alpha^\top \boldsymbol{z}_\beta| \;\leq\; C\frac{r\sqrt{\log n}}{d}. \tag{3.3.15}$$

(2c) Signal–signal term. Each $\boldsymbol{m}_\alpha$ equals $\pm\frac{r}{\sqrt{d}}\boldsymbol{u}_1$ or $\pm\frac{r}{\sqrt{d}}\boldsymbol{u}_2$. By Cauchy–Schwarz inequality,

$$|\boldsymbol{m}_\alpha^\top \boldsymbol{m}_\beta| \leq \|\boldsymbol{m}_\alpha\|\|\boldsymbol{m}_\beta\| = \frac{r^2}{d}.$$

Hence

$$\max_{\alpha \neq \beta} |\boldsymbol{m}_\alpha^\top \boldsymbol{m}_\beta| \;\leq\; \frac{r^2}{d}. \tag{3.3.16}$$

(2d) Combine. Combining (3.3.13)–(3.3.16) and using the same simplification as in Step 1, we obtain with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\boldsymbol{x}_\alpha^\top \boldsymbol{x}_\beta| \;\leq\; C\Big(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d}\Big). \tag{3.3.17}$$

This verifies the pairwise-inner-product part of (τ_n, B)-orthonormality with τ_n as in (3.3.2).

Step 3: Sum of squared squared-norm deviations.

From (3.3.6),

$$\|\boldsymbol{x}_\alpha\|^2 - 1 = (\|\boldsymbol{z}_\alpha\|^2 - 1) + 2\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha + \|\boldsymbol{m}_\alpha\|^2.$$

Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we get

$$(\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \le 3(\|\boldsymbol{z}_\alpha\|^2 - 1)^2 + 12(\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha)^2 + 3\|\boldsymbol{m}_\alpha\|^4.$$

Summing over α and using (3.3.5) gives

$$\sum_{\alpha=1}^{n} (\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \le 3\sum_{\alpha=1}^{n} (\|\boldsymbol{z}_\alpha\|^2 - 1)^2 + 12\sum_{\alpha=1}^{n} (\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha)^2 + 3n\frac{r^4}{d^2}. \tag{3.3.18}$$

(3a) Control of $\sum_{\alpha=1}^n (\|\boldsymbol{z}_\alpha\|^2 - 1)^2$. Set $g_\alpha := \sqrt{d}\,\boldsymbol{z}_\alpha \sim \mathcal{N}(0, I_d)$ so that $\|\boldsymbol{z}_\alpha\|^2 = \|g_\alpha\|^2/d$ and $\|g_\alpha\|^2 \sim \chi^2_d$. Define

$$U_\alpha := \|g_\alpha\|^2 - d, \qquad X_\alpha := (\|\boldsymbol{z}_\alpha\|^2 - 1)^2 = \frac{U_\alpha^2}{d^2}.$$

Then $\mathbb{E}[X_\alpha] = \mathrm{Var}(\|\boldsymbol{z}_\alpha\|^2) = 2/d$, hence $\mathbb{E}\sum_{\alpha=1}^n X_\alpha = 2n/d$.

Moment bound. Write $U_\alpha = \sum_{i=1}^d (\xi_{i\alpha}^2 - 1)$ with $\xi_{i\alpha} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. For any integer $p \ge 2$, Rosenthal's inequality (Theorem 3 of [Ros70]) yields

$$\mathbb{E}|U_\alpha|^{2p} \;\le\; C_p\Big(d\,\mathbb{E}|\xi_{11}^2 - 1|^{2p} + (d\,\mathrm{Var}(\xi_{11}^2 - 1))^p\Big) \;\le\; C_p' d^p, \tag{3.3.19}$$

where C_p, C_p' depend only on p (not on d, n). Consequently,

$$\mathbb{E}|X_\alpha|^p = \frac{\mathbb{E}|U_\alpha|^{2p}}{d^{2p}} \;\le\; \frac{C_p'}{d^p}, \qquad \mathbb{E}|X_\alpha - \mathbb{E}X_\alpha|^p \;\le\; \frac{C_p''}{d^p}. \tag{3.3.20}$$

Concentration of the sum. Let $\widetilde{X}_\alpha := X_\alpha - \mathbb{E}X_\alpha$ so that $\mathbb{E}\widetilde{X}_\alpha = 0$. Applying Rosenthal's inequality again (now over $\alpha = 1,\dots,n$) gives, for any integer $p \geq 2$,

$$\mathbb{E}\Big|\sum_{\alpha=1}^{n} \widetilde{X}_\alpha\Big|^p \leq C_p\Big(n\,\mathbb{E}|\widetilde{X}_1|^p + (n\,\mathbb{E}\widetilde{X}_1^2)^{p/2}\Big) \leq C_p'\Big(\frac{n}{d^2}\Big)^{p/2},$$

using (3.3.20) and $\mathbb{E}\widetilde{X}_1^2 \lesssim d^{-2}$. Therefore, by Markov's inequality, for any fixed $t > 0$,

$$\mathbb{P}\Big(\sum_{\alpha=1}^{n} X_\alpha \geq \frac{2n}{d} + t\frac{n}{d}\Big) = \mathbb{P}\Big(\sum_{\alpha=1}^{n} \widetilde{X}_\alpha \geq t\frac{n}{d}\Big) \leq \frac{\mathbb{E}|\sum_{\alpha=1}^{n}\widetilde{X}_\alpha|^p}{(tn/d)^p} \leq C_p'' t^{-p} n^{-p/2}.$$

Taking $t = 1$ and p large enough (so that $p/2 \geq c$) yields, for some constants $c, C > 0$,

$$\sum_{\alpha=1}^{n} (\|\boldsymbol{z}_\alpha\|^2 - 1)^2 \leq C\frac{n}{d} \qquad \text{with probability at least } 1 - n^{-c}. \tag{3.3.21}$$

This bound is uniform in (n,d) and in particular covers both regimes $n \asymp d$ and $n = O(d^2)$.

(3b) Control of $\sum(\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha)^2$. Conditioning on $\boldsymbol{m}_\alpha$, the random variable $\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha$ is Gaussian $\mathcal{N}(0, r^2/d^2)$, hence $(\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha)^2$ is sub-exponential with mean r^2/d^2. By the Bernstein inequality,

$$\sum_{\alpha=1}^{n} (\boldsymbol{z}_\alpha^\top \boldsymbol{m}_\alpha)^2 \leq C\frac{nr^2}{d^2} \qquad \text{with probability at least } 1 - n^{-c}. \tag{3.3.22}$$

(3c) Combine. Plugging (3.3.21), (3.3.22) into (3.3.18) gives, w.h.p.,

$$\sum_{\alpha=1}^{n} (\|\boldsymbol{x}_\alpha\|^2 - 1)^2 \leq C\Big(\frac{n}{d} + \frac{nr^2}{d^2} + \frac{nr^4}{d^2}\Big).$$

This proves (3.3.3) since

$$(\|\boldsymbol{x}_\alpha\| - 1)^2 = \frac{(\|\boldsymbol{x}_\alpha\|^2 - 1)^2}{(\|\boldsymbol{x}_\alpha\| + 1)^2} \leq (\|\boldsymbol{x}_\alpha\|^2 - 1)^2.$$

Step 4: Operator norm bound.

We bound

$$\|\boldsymbol{X}\| \le \|\boldsymbol{Z}\| + \|\boldsymbol{M}\|.$$

A standard Gaussian operator norm bound states that for all $t \ge 0$,

$$\mathbb{P}\Big(\|\boldsymbol{Z}\| \ge 1 + \sqrt{\frac{n}{d}} + t\Big) \le 2\exp(-cdt^2).$$

Taking $t = \sqrt{\frac{\log n}{d}}$ yields $\|\boldsymbol{Z}\| \le C(1+\sqrt{n/d})$ with probability at least $1 - n^{-c}$. Since $\boldsymbol{u}_1, \boldsymbol{u}_2$ are orthonormal and $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthonormal, $\|\boldsymbol{M}\| = \theta_{\text{SNR}} = r\sqrt{\frac{n}{2d}} \le Cr\sqrt{\frac{n}{d}}$. Therefore, with probability at least $1 - n^{-c}$,

$$\|\boldsymbol{X}\| \le C\Big(1 + \sqrt{\frac{n}{d}}\Big) + Cr\sqrt{\frac{n}{d}} \le C\Big(1 + (1+r)\sqrt{\frac{n}{d}}\Big).$$

This completes the proof. □

3.3.3 Properties of CK Matrix

The following intermediary results are necessary to construct the QE in Section 3.3.6.

Proposition 25. *Let χ_d denote the chi distribution with d degrees of freedom. Then, as $d \to \infty$,*

$$\frac{\mathbb{E}[\chi_d]}{\sqrt{d}} = 1 - \frac{1}{4d} + \frac{1}{32d^2} + \frac{5}{128d^3} + O(d^{-4}).$$

Proof. It is classical that

$$\mathbb{E}[\chi_d] = \sqrt{2}\,\frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)}.$$

(See, e.g., [AS64, Ch. 6], [DLM, Sec. 5.11].) Set $m := d/2$. Then

$$\frac{\mathbb{E}[\chi_d]}{\sqrt{d}} = \sqrt{\frac{2}{2m}}\,\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)} = \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)\sqrt{m}} =: Q_m.$$

We use the Stirling series for $\log\Gamma$ (with Bernoulli numbers), valid as $z \to \infty$ in $|\arg z| < \pi$:

$$\log\Gamma(z) = (z-\tfrac{1}{2})\log z - z + \tfrac{1}{2}\log(2\pi) + \frac{1}{12z} - \frac{1}{360z^3} + O(z^{-5}), \tag{3.3.23}$$

see [DLM, (5.11.2)] or [AS64, §6.1, §6.3]. Write

$$\log Q_m = \big(\log\Gamma(m+\tfrac{1}{2}) - \log\Gamma(m)\big) - \tfrac{1}{2}\log m.$$

Using (3.3.23) with $z = m+\frac{1}{2}$ and $z = m$, we obtain

$$\begin{aligned}\log Q_m = {} & m\log(m+\tfrac{1}{2}) - (m+\tfrac{1}{2}) - (m-\tfrac{1}{2})\log m + m \\ & + \frac{1}{12(m+\frac{1}{2})} - \frac{1}{12m} - \frac{1}{360(m+\frac{1}{2})^3} + \frac{1}{360m^3} - \tfrac{1}{2}\log m + O(m^{-5}).\end{aligned}$$

The logarithmic terms simplify to

$$m\log\left(1+\frac{1}{2m}\right) - \frac{1}{2}.$$

Expanding with the binomial and Taylor series $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + O(x^4)$ and $(1+x)^{-1} = 1 - x + x^2 - x^3 + O(x^4)$, we get

$$m\log\left(1+\frac{1}{2m}\right) = \frac{1}{2} - \frac{1}{8m} + \frac{1}{24m^2} - \frac{1}{64m^3} + O(m^{-4}),$$

and

$$\frac{1}{12(m+\frac{1}{2})} - \frac{1}{12m} = -\frac{1}{24m^2} + \frac{1}{48m^3} + O(m^{-4}).$$

Further, the difference $-\frac{1}{360(m+\frac{1}{2})^3} + \frac{1}{360m^3}$ is $O(m^{-4})$ (since the leading term cancels and the remainder begins at order m^{-4}). Collecting terms, we find

$$\log Q_m = -\frac{1}{8m} + \frac{1}{192m^3} + O(m^{-4}).$$

Exponentiating and using $\exp(u) = 1 + u + \frac{u^2}{2} + \frac{u^3}{6} + O(u^4)$ with $u = -\frac{1}{8m} + \frac{1}{192m^3}$, we obtain

$$Q_m = 1 - \frac{1}{8m} + \frac{1}{128m^2} + \frac{5}{1024m^3} + O(m^{-4}).$$

Finally, substituting $m = d/2$ yields

$$\frac{\mathbb{E}[\chi_d]}{\sqrt{d}} = 1 - \frac{1}{4d} + \frac{1}{32d^2} + \frac{5}{128d^3} + O(d^{-4}),$$

as claimed. □

Remark. The same expansion follows immediately from the general ratio formula

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} \sim z^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2z} + \cdots\right)$$

for $z \to \infty$ (take $a = \frac{1}{2}, b = 0$); see [TE+51, DLM].

Let us define the null model as

$$\boldsymbol{Y}_0 := \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z}), \tag{3.3.24}$$

where in this case, the dataset is pure standard noise.

Lemma 26. *Under the Assumption 3, we know that*

$$\|\mathbb{E}[\boldsymbol{Y}_0]\| \lesssim \frac{\sqrt{n}}{d},$$

as $n, d \to \infty$.

Proof. Notice that $\mathbb{E}[\boldsymbol{Y}_0] = \frac{C_d}{\sqrt{N}}\mathbf{1}_N\mathbf{1}_n^\top$ where $C_d := \mathbb{E}[\sigma(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top\boldsymbol{z})]$ where $\boldsymbol{z}, \boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$.

Hence,

$$\|\mathbb{E}[\boldsymbol{Y}_0]\| = \sqrt{n}|C_d|. \tag{3.3.25}$$

Denote $\xi \sim \mathcal{N}(0,1)$ which is independent with $\boldsymbol{w}$. Applying $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma'(\xi)\xi] = \mathbb{E}[\sigma''(\xi)] = c_\sigma$, and a Taylor approximation of σ at ξ, we can get

$$\begin{aligned} C_d &= \mathbb{E}[\sigma(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|\xi)] - \mathbb{E}[\sigma(\xi)] \\ &= \mathbb{E}[\sigma'(\xi)\xi(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)] + \mathbb{E}[\sigma''(\eta)\xi^2(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)^2] \\ &= c_\sigma(\mathbb{E}[\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|] - 1) + \mathbb{E}[\sigma''(\eta)\xi^2(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)^2], \end{aligned}$$

for some η between ξ and $\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|\xi$. Notice that $\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| \overset{\mathscr{D}}{\sim} \frac{1}{\sqrt{d}}\chi_d$ where χ_d is chi distribution with degree d. Thus,

$$\mathbb{E}[\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|] = \frac{\sqrt{2}\Gamma(\frac{1}{2}(d+1))}{\sqrt{d}\Gamma(\frac{1}{2}d)} = \sqrt{1 - \frac{1}{2d}} + o(d^{-1}) = 1 - \frac{1}{4d} + o(d^{-1}). \tag{3.3.26}$$

Then, we can get

$$|C_d| \lesssim \frac{1}{d} + \mathbb{E}[(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)^2] \lesssim \frac{1}{d},$$

where we apply the Assumption 3 and the fact $\mathbb{E}[(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)^2] = 2 - 2\mathbb{E}[\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|] = O(\frac{1}{d})$. □

Lemma 27. *Under the Assumption 3, we know that*

$$|\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})] - b_\sigma| \lesssim \frac{1}{d}$$

and

$$|\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})] - c_\sigma| \lesssim \frac{1}{\sqrt{d}}$$

where $\boldsymbol{z}, \boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$.

Proof. Denote $\xi \sim \mathcal{N}(0,1)$ which is independent with $\boldsymbol{w}$. Recall that $\mathbb{E}[\sigma'(\xi)] = b_\sigma$ and $\mathbb{E}[\sigma'(\xi)\xi] = \mathbb{E}[\sigma''(\xi)] = c_\sigma$. Similarly to the proof of Lemma 26, we can apply a Taylor approximation of σ' at ξ:

$$|\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})] - b_\sigma| = |\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|\xi)] - \mathbb{E}[\sigma'(\xi)]| \tag{3.3.27}$$

$$\leq |\mathbb{E}[\sigma''(\xi)\xi]| \cdot |\frac{1}{\sqrt{d}}\mathbb{E}[\|\boldsymbol{w}\|] - 1| + |\mathbb{E}[\sigma'''(\eta)\xi^2]|\mathbb{E}(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\| - 1)^2 \tag{3.3.28}$$

$$\lesssim |\frac{1}{\sqrt{d}}\mathbb{E}[\chi_d] - 1| + \mathbb{E}(\frac{1}{\sqrt{d}}\chi_d - 1)^2 \lesssim \frac{1}{d}, \tag{3.3.29}$$

for some η between ξ and $\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|\xi$. Similarly, for σ'', we have

$$|\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})] - c_\sigma| = |\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}}\|\boldsymbol{w}\|\xi)] - \mathbb{E}[\sigma''(\xi)]| \tag{3.3.30}$$

$$\leq |\mathbb{E}[\sigma''(\eta)\xi(\frac{1}{\sqrt{d}}\mathbb{E}[\|\boldsymbol{w}\|] - 1)]| \tag{3.3.31}$$

$$\lesssim \mathbb{E}[\sigma''(\eta)^2\xi^2]^{1/2}\mathbb{E}[(\frac{1}{\sqrt{d}}\chi_d - 1)^2]^{1/2} \lesssim \frac{1}{\sqrt{d}}. \tag{3.3.32}$$

Then, we complete the proof. □

Lemma 28. *Consider $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_N)$. Then we have $\|\boldsymbol{g}\|_\infty \prec \sqrt{2\log N}$.*

Proof. Apply Proposition 2.7.6 of [Ver18] and use the fact that $\mathbb{E}[\|\boldsymbol{g}\|_\infty] = \mathbb{E}[\max_{i\in[N]}|g_i|]$. □

Lemma 29. *Consider a matrix $\boldsymbol{M} \in \mathbb{R}^{N\times n}$ and two vectors $\boldsymbol{a} \in \mathbb{R}^N$, $\boldsymbol{b} \in \mathbb{R}^n$. We have the following identity*

$$\boldsymbol{M} \odot (\boldsymbol{a}\boldsymbol{b}^\top) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{M}\,\mathrm{diag}(\boldsymbol{b}).$$

This is a classical linear algebra result, so we will omit the proof here.

Lemma 30. *Assume σ is globally L-Lipschitz: $|\sigma(a) - \sigma(b)| \leq L|a-b|$, $\forall a, b \in \mathbb{R}$. Then there*

exist absolute constants $C, c > 0$ such that for every $t \geq 0$, with probability at least $1 - 6e^{-ct^2}$,

$$\left\| \frac{1}{\sqrt{N}} \Big(\sigma(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}\left[\sigma(\boldsymbol{W}\boldsymbol{Z})\right] \Big) \right\| \leq CL \Big(1 + \sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}} \Big) \Big(1 + \sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}} \Big).$$

Proof. Let $\boldsymbol{z}_j \in \mathbb{R}^d$ denote the jth column of $\boldsymbol{Z}$ and define

$$\boldsymbol{Y}_0 = \frac{1}{\sqrt{N}} \sigma(\boldsymbol{W}\boldsymbol{Z}) \in \mathbb{R}^{N \times n}, \qquad m := \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d),\, \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)} \left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}) \right].$$

By exchangeability, $\mathbb{E}\boldsymbol{Y}_0 = \frac{m}{\sqrt{N}} \mathbf{1}_N \mathbf{1}_n^\top$.

Firstly, condition on $\boldsymbol{Z}$. Define $\boldsymbol{\mu}(\boldsymbol{Z}) \in \mathbb{R}^n$ whose entries are

$$\mu_j(\boldsymbol{Z}) := \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)} \left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}_j) \right], \qquad j = 1, \ldots, n.$$

Then each row of $\boldsymbol{Y}_0$ has conditional mean $\boldsymbol{\mu}(\boldsymbol{Z})^\top$, hence

$$\mathbb{E}_{\boldsymbol{W}}[\boldsymbol{Y}_0 \mid \boldsymbol{Z}] = \frac{1}{\sqrt{N}} \mathbf{1}_N \boldsymbol{\mu}(\boldsymbol{Z})^\top.$$

Write

$$\|\boldsymbol{Y}_0 - \mathbb{E}\boldsymbol{Y}_0\| = \underbrace{\left\| \left(\boldsymbol{Y}_0 - \mathbb{E}_{\boldsymbol{W}}[\boldsymbol{Y}_0 \mid \boldsymbol{Z}] \right) \right\|}_{T_1} + \underbrace{\left\| \left(\mathbb{E}_{\boldsymbol{W}}[\boldsymbol{Y}_0 \mid \boldsymbol{Z}] - \mathbb{E}\boldsymbol{Y}_0 \right) \right\|}_{T_2}. \tag{3.3.33}$$

We then control T_1 and T_2 separately.

Step 1: Bound T_2 (fluctuation of the conditional mean). We have

$$T_2 = \left\| \frac{1}{\sqrt{N}} \mathbf{1}_N (\boldsymbol{\mu}(\boldsymbol{Z}) - m\mathbf{1}_n)^\top \right\| = \|\boldsymbol{\mu}(\boldsymbol{Z}) - m\mathbf{1}_n\|.$$

Define $\mu(\boldsymbol{z}) := \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{z})]$. For any $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$,

$$|\mu(\boldsymbol{z}) - \mu(\boldsymbol{z}')| \leq \mathbb{E}_{\boldsymbol{w}} \left| \sigma(\boldsymbol{w}^\top \boldsymbol{z}) - \sigma(\boldsymbol{w}^\top \boldsymbol{z}') \right| \leq L \mathbb{E}_{\boldsymbol{w}} |\boldsymbol{w}^\top (\boldsymbol{z} - \boldsymbol{z}')| = L \sqrt{\frac{2}{\pi}} \|\boldsymbol{z} - \boldsymbol{z}'\|_2.$$

Thus $\mu(\cdot)$ is $(L\sqrt{2/\pi})$-Lipschitz. Write $\boldsymbol{z}_j = \boldsymbol{g}_j/\sqrt{d}$ with $\boldsymbol{g}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then $f(\boldsymbol{g}) := \mu(\boldsymbol{g}/\sqrt{d})$ is Lipschitz with constant $O(L/\sqrt{d})$. By Gaussian concentration for Lipschitz functions (e.g., Chapter 5 of [Ver18]), each coordinate $\mu(\boldsymbol{z}_j) - \mathbb{E}\mu(\boldsymbol{z}_j) = \mu(\boldsymbol{z}_j) - m$ is subgaussian with subgaussian norm

$$\|\mu(\boldsymbol{z}_j) - m\|_{\psi_2} \le C\frac{L}{\sqrt{d}}.$$

Since $(\boldsymbol{z}_j)$ are independent across j, the coordinates of $\boldsymbol{\mu}(\boldsymbol{Z}) - m\boldsymbol{1}_n$ are independent subgaussian. A standard concentration bound for the Euclidean norm of an i.i.d. subgaussian vector (e.g., see Sec. 3.1 of [Ver18]) yields: for every $t \ge 0$, with probability at least $1 - 2e^{-ct^2}$,

$$T_2 = \|\boldsymbol{\mu}(\boldsymbol{Z}) - m\boldsymbol{1}_n\|_2 \le C\frac{L}{\sqrt{d}}(\sqrt{n} + t) = CL\Big(\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\Big). \tag{3.3.34}$$

Step 2: Bound T_1 (fluctuation around the conditional mean). Let

$$\boldsymbol{A} := \sigma(\boldsymbol{W}\boldsymbol{Z}) - \boldsymbol{1}_N\boldsymbol{\mu}(\boldsymbol{Z})^\top.$$

Then $T_1 = \|\boldsymbol{A}/\sqrt{N}\|$. Conditional on $\boldsymbol{Z}$, the rows $\boldsymbol{A}_i \in \mathbb{R}^n$ are i.i.d. (over rows $\boldsymbol{w}_i$ of $\boldsymbol{W}$) and mean-zero. Fix any $\boldsymbol{x} \in \mathbb{S}^{n-1}$ and define the scalar function of $\boldsymbol{w} \in \mathbb{R}^d$

$$F_{\boldsymbol{x}}(\boldsymbol{w}) := \sum_{j=1}^{n} x_j\,\sigma(\boldsymbol{w}^\top \boldsymbol{z}_j) = \boldsymbol{x}^\top \sigma(\boldsymbol{w}^\top \boldsymbol{Z}) \quad \in \mathbb{R}.$$

Then $\boldsymbol{A}_i^\top \boldsymbol{x} = F_{\boldsymbol{x}}(\boldsymbol{w}_i) - \mathbb{E}[F_{\boldsymbol{x}}(\boldsymbol{w}_i) \mid \boldsymbol{Z}]$. Moreover, $F_{\boldsymbol{x}}$ is Lipschitz with constant at most $L\|\boldsymbol{Z}\|$: for any $\boldsymbol{w}, \boldsymbol{w}'$,

$$\begin{aligned}|F_{\boldsymbol{x}}(\boldsymbol{w}) - F_{\boldsymbol{x}}(\boldsymbol{w}')| &\le \sum_{j=1}^{n} |x_j|\,|\sigma(\boldsymbol{w}^\top \boldsymbol{z}_j) - \sigma(\boldsymbol{w}'^\top \boldsymbol{z}_j)| \le L\sum_{j=1}^{n} |x_j|\,|(\boldsymbol{w} - \boldsymbol{w}')^\top \boldsymbol{z}_j| \\ &\le L\|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{Z}^\top(\boldsymbol{w} - \boldsymbol{w}')\|_2 \le L\|\boldsymbol{Z}\|\,\|\boldsymbol{w} - \boldsymbol{w}'\|_2,\end{aligned}$$

since $\|\boldsymbol{x}\|_2 = 1$. Therefore, by Gaussian concentration for Lipschitz functions, conditional on $\boldsymbol{Z}$

we have for all $u \geq 0$,

$$\mathbb{P}\Big(|\boldsymbol{A}_i^\top \boldsymbol{x}| \geq u \,\Big|\, \boldsymbol{Z}\Big) \leq 2\exp\Big(-c\,\frac{u^2}{L^2\|\boldsymbol{Z}\|^2}\Big). \tag{3.3.35}$$

Now fix $\boldsymbol{x} \in \mathbb{S}^{n-1}$ and $\boldsymbol{y} \in \mathbb{S}^{N-1}$. Then

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^{N} y_i\,(\boldsymbol{A}_i^\top \boldsymbol{x})$$

is a sum of independent (over i) centered subgaussian random variables. Using (3.3.35) and $\|\boldsymbol{y}\|_2 = 1$, we obtain the conditional tail bound

$$\mathbb{P}\Big(|\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle| \geq u \,\Big|\, \boldsymbol{Z}\Big) \leq 2\exp\Big(-c\,\frac{u^2}{L^2\|\boldsymbol{Z}\|^2}\Big). \tag{3.3.36}$$

Next, take $1/4$-nets $\mathscr{N}_n \subset \mathbb{S}^{n-1}$ and $\mathscr{N}_N \subset \mathbb{S}^{N-1}$ with $|\mathscr{N}_n| \leq 9^n$ and $|\mathscr{N}_N| \leq 9^N$. By the standard net reduction for operator norms, one has

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}\in\mathbb{S}^{n-1},\, \boldsymbol{y}\in\mathbb{S}^{N-1}} \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \;\leq\; 2 \max_{\boldsymbol{x}\in\mathscr{N}_n,\, \boldsymbol{y}\in\mathscr{N}_N} \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle. \tag{3.3.37}$$

Combining (3.3.36)–(3.3.37) with a union bound yields, for any $u \geq 0$,

$$\mathbb{P}\Big(\|\boldsymbol{A}\| \geq 2u \,\Big|\, \boldsymbol{Z}\Big) \leq 2\,|\mathscr{N}_n|\,|\mathscr{N}_N|\,\exp\Big(-c\,\frac{u^2}{L^2\|\boldsymbol{Z}\|^2}\Big) \leq 2\exp\Big((n+N)\log 9 - c\,\frac{u^2}{L^2\|\boldsymbol{Z}\|^2}\Big).$$

Choosing $u = CL\|\boldsymbol{Z}\|\big(\sqrt{n}+\sqrt{N}+t\big)$ and taking C large enough (absorbing $(n+N)\log 9$) gives: for every $t \geq 0$, with conditional probability at least $1-2e^{-ct^2}$,

$$\|\boldsymbol{A}\| \leq CL\|\boldsymbol{Z}\|\big(\sqrt{n}+\sqrt{N}+t\big), \quad \text{hence} \quad T_1 = \Big\|\frac{\boldsymbol{A}}{\sqrt{N}}\Big\| \leq CL\|\boldsymbol{Z}\|\Big(1+\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}\Big).$$

Since the bound holds conditionally on $\boldsymbol{Z}$ with failure probability $\leq 2e^{-ct^2}$, it also holds unconditionally.

Step 3: Bound $\|\boldsymbol{Z}\|$ and conclude. By the norm bound for Gaussian matrices, with probability at least $1-2e^{-ct^2}$,

$$\|\boldsymbol{Z}\| \le 1+\sqrt{\frac{n}{d}}+\frac{t}{\sqrt{d}}.$$

On the intersection of the three high-probability events (3.3.34), the T_1 bound above, and the $\|\boldsymbol{Z}\|$ bound, we get

$$T_1 \le CL\Big(1+\sqrt{\frac{n}{d}}+\frac{t}{\sqrt{d}}\Big)\Big(1+\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}\Big)$$
$$T_2 \le CL\Big(\sqrt{\frac{n}{d}}+\frac{t}{\sqrt{d}}\Big) \le CL\Big(1+\sqrt{\frac{n}{d}}+\frac{t}{\sqrt{d}}\Big)\Big(1+\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}\Big)$$

and then (3.3.33) yields the desired bound. A union bound over the three events gives total failure probability at most $6e^{-ct^2}$, which completes the proof. □

3.3.4 The z-transform and the T-transform

Let ν be a compactly supported probability measure on $[0,\infty)$ and $\phi \in (0,\infty)$. Recall for $s \in \mathbb{R}$ such that $1+\lambda s \neq 0$ for all $\lambda \in \operatorname{supp}(\nu)$,

$$z(s) = -\frac{1}{s}+\phi\int \frac{\lambda}{1+\lambda s}\,\nu(d\lambda), \qquad z'(s) = \frac{1}{s^2}-\phi\int \frac{\lambda^2}{(1+\lambda s)^2}\,\nu(d\lambda), \tag{3.3.38}$$

the alignment map

$$\varphi(s) = -\frac{s z'(s)}{z(s)}, \tag{3.3.39}$$

and the additive transfer function

$$T(s) = \frac{z(s)s^2-(\phi-1)s}{\phi}. \tag{3.3.40}$$

Lemma 31. *Let ν be a compactly supported probability measure on $[0,\infty)$ and let $\phi \in (0,\infty)$.*

Let $\mu := \rho_\phi^{\mathrm{MP}} \boxtimes \nu$ be the deformed MP law, and define its companion probability measure

$$\tilde{\mu} \;:=\; \phi\,\mu + (1-\phi)\,\delta_0.$$

Let $m_\mu(z) = \int (x-z)^{-1}\,\mu(dx)$ and $m_{\tilde{\mu}}(z) = \int (x-z)^{-1}\,\tilde{\mu}(dx)$ be the Stieltjes transforms, defined on $\mathbb{C}\setminus \mathrm{supp}(\mu)$ and $\mathbb{C}\setminus \mathrm{supp}(\tilde{\mu})$ respectively. Then for all $z \in \mathbb{C}\setminus(\mathrm{supp}(\mu)\cup\{0\})$,

$$m_{\tilde{\mu}}(z) \;=\; \tilde{m}_\mu(z) \;:=\; \frac{\phi-1}{z} + \phi\, m_\mu(z). \tag{3.3.41}$$

Define the set of singularities

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

For $s \in \mathbb{C}\setminus T$, $z(s)$ and $z'(s)$ in (3.3.38) *are well defined.*

Then, we have

(i) *For any real $\lambda \in \mathbb{R}\setminus \mathrm{supp}(\tilde{\mu})$, the quantity $s := m_{\tilde{\mu}}(\lambda)$ belongs to $\mathbb{R}\setminus\mathscr{T}$ and satisfies*

$$\lambda \;=\; z(s), \qquad \textit{and} \qquad z'(s) \;>\; 0. \tag{3.3.42}$$

Equivalently, since $\mathrm{supp}(\mu)$ *and* $\mathrm{supp}(\tilde{\mu})$ *may differ only at* $\{0\}$*, the same conclusion holds for any real $\lambda \in \mathbb{R}\setminus(\mathrm{supp}(\mu)\cup\{0\})$ with $s = \tilde{m}_\mu(\lambda)$.*

(ii) *Conversely, if $s \in \mathbb{R}\setminus\mathscr{T}$ satisfies $z'(s) > 0$, then $\lambda := z(s)$ belongs to $\mathbb{R}\setminus\mathrm{supp}(\tilde{\mu})$ and*

$$m_{\tilde{\mu}}(\lambda) = s.$$

Moreover, let I be a maximal *open interval contained in $\mathbb{R}\setminus\mathscr{T}$ on which $z'(s) > 0$. Then z is strictly increasing on I and maps I bijectively onto a connected component of $\mathbb{R}\setminus\mathrm{supp}(\tilde{\mu})$*

(an exterior component or an internal gap). Finally, every nonzero boundary point $\lambda_\star \neq 0$ of $\operatorname{supp}(\tilde{\mu})$ *can be written as $\lambda_\star = z(s_\star)$ for some $s_\star \in \mathbb{R} \setminus \mathscr{T}$ satisfying $z'(s_\star) = 0$.*

Proof. This is exactly the classical result of Silverstein–Choi (Theorems. 4.1–4.2 of [SC95]); see also Proposition 8 in Appendix A.2 of [WWF24]. □

Lemma 32. *With the same notation as Lemma 31, recall the T-transform defined by* (3.3.40). *For any real $\lambda \in \mathbb{R} \setminus (\operatorname{supp}(\mu) \cup \{0\})$, let $s = \tilde{m}_\mu(\lambda) = m_{\tilde{\mu}}(\lambda)$. Then $\lambda = z(s)$ and*

$$\lambda\, s\, m_\mu(\lambda) \;=\; T(s). \tag{3.3.43}$$

Proof. Since $\lambda \neq 0$ and $s = \tilde{m}_\mu(\lambda) = \frac{\phi-1}{\lambda} + \phi m_\mu(\lambda)$, we can solve for $m_\mu(\lambda)$:

$$m_\mu(\lambda) = \frac{1}{\phi}\Big(s - \frac{\phi-1}{\lambda}\Big).$$

Therefore

$$\lambda s m_\mu(\lambda) = \lambda s \cdot \frac{1}{\phi}\Big(s - \frac{\phi-1}{\lambda}\Big) = \frac{\lambda s^2 - (\phi-1)s}{\phi}.$$

By Lemma 31(i), $\lambda = z(s)$, so the right-hand side equals $T(s)$ by definition. This proves (3.3.43). □

Lemma 33 (Strict monotonicity of T on each analytic branch)**.** *On any real interval $I \subset \mathbb{R}$ that does not cross a pole $s = -1/\lambda$ with $\lambda \in \operatorname{supp}(\nu)$, the function T is C^1 and strictly decreasing:*

$$T'(s) < 0, \qquad s \in I. \tag{3.3.44}$$

Hence, T is invertible on this interval.

Proof. For $s \in I$, we define

$$I_1(s) := \int \frac{\lambda}{1+\lambda s}\, \nu(d\lambda), \qquad I_2(s) := \int \frac{\lambda^2}{(1+\lambda s)^2}\, \nu(d\lambda).$$

Then $I_1 \in C^1(I)$ and $I_1'(s) = -I_2(s)$. Using $T(s) = -s + s^2 I_1(s)$ (equivalent to (3.3.40) and (3.3.38)) we compute

$$T'(s) = -1 + 2sI_1(s) - s^2 I_2(s). \tag{3.3.45}$$

Let $F(\lambda) := \frac{\lambda}{1+\lambda s}$. Then $I_1(s) = \mathbb{E}_\nu[F]$ and $I_2(s) = \mathbb{E}_\nu[F^2]$. Hence

$$T'(s) = -1 + 2s\mathbb{E}[F] - s^2\mathbb{E}[F^2] = -(1 - s\mathbb{E}[F])^2 - s^2\mathrm{Var}(F).$$

If ν is not a single atom, then $\mathrm{Var}(F) > 0$ for all $s \in I$, so $T'(s) < 0$. In the degenerate atomic case, $T' < 0$ still holds since $(1 - s\mathbb{E}[F])^2 > 0$. This proves (3.3.44). □

Remark (Relation to the classical D-transform in [BGN12]). If one defines

$$D(\lambda) := \lambda m_\mu(\lambda)\tilde{m}_\mu(\lambda) \tag{3.3.46}$$

then by Lemma 32, $D(\lambda) = T(s)$ whenever $\lambda = z(s)$ and $s = \tilde{m}_\mu(\lambda)$. Thus all additive outlier equations $\beta D(\lambda) = 1$ are exactly $\beta T(s) = 1$ in z-coordinates. This D-transform and $\beta D(\lambda) = 1$ are used by [BGN12] to characterize the outlier singular values in additive spiked random matrix models.

Lemma 34 (Closed-form inverse of T for a shifted MP bulk)**.** *Fix $\psi \in (0,\infty)$ and let $b_\sigma^2 \in [0,1]$. Define the shifted MP law*

$$\nu \;:=\; (1 - b_\sigma^2) + b_\sigma^2\, \rho_\psi^{\mathrm{MP}},$$

and define, for any $s \in \mathbb{C} \setminus \{0\}$ such that $1 + \lambda s \neq 0$ for all $\lambda \in \mathrm{supp}(\nu)$, T transform defined by (3.3.40) *can be written as*

$$T(s) \;=\; -s + s^2 \int \frac{\lambda}{1+\lambda s}\, \nu(d\lambda). \tag{3.3.47}$$

Let $m_{\mathrm{MP},\psi}(\cdot)$ denote the Stieltjes transform of ρ_ψ^{MP}.

(i) *(**Closed form of T in terms of $m_{\mathrm{MP},\psi}$**) Let*

$$\xi(s) \;:=\; -\frac{1+(1-b_\sigma^2)s}{b_\sigma^2\, s}. \tag{3.3.48}$$

Then

$$T(s) \;=\; -\frac{1}{b_\sigma^2}\, m_{\mathrm{MP},\psi}\big(\xi(s)\big). \tag{3.3.49}$$

If $b_\sigma = 0$, then $T(s) = -s/(1+s)$.

(ii) *(**Exact formula for T^{-1}**) Let $t \in \mathbb{C}$ be such that the denominator in* (3.3.50) *below is nonzero. Define*

$$T^{-1}(t) \;:=\; -\frac{t\big(1-\psi b_\sigma^2\, t\big)}{1+\big(1-\psi b_\sigma^2\big)t - \psi b_\sigma^2(1-b_\sigma^2)t^2}. \tag{3.3.50}$$

Then $T\big(T^{-1}(t)\big) = t$. In particular, on any real monotone branch of T (e.g. $s < 0$), (3.3.50) *is the functional inverse of T.*

Proof. Write $\alpha := 1 - b_\sigma^2$ and $\beta := b_\sigma^2$ for short, so that $\nu = \alpha + \beta\rho_\psi^{\mathrm{MP}}$. Let $X \sim \rho_\psi^{\mathrm{MP}}$ and set $\lambda = \alpha + \beta X$. Then

$$\int \frac{\lambda}{1+\lambda s}\, \nu(d\lambda) \;=\; \mathbb{E}\Big[\frac{\alpha+\beta X}{1+s(\alpha+\beta X)}\Big].$$

Now note the following *exact* algebraic identity:

$$\frac{\alpha+\beta X}{1+s(\alpha+\beta X)} = \frac{1}{s}\Big(1 - \frac{1}{1+s(\alpha+\beta X)}\Big). \tag{3.3.51}$$

Taking expectations in (3.3.51) gives

$$\int \frac{\lambda}{1+\lambda s}\, \nu(d\lambda) = \frac{1}{s}\Big(1 - \mathbb{E}\Big[\frac{1}{1+s(\alpha+\beta X)}\Big]\Big). \tag{3.3.52}$$

Insert (3.3.52) into the definition (3.3.47):

$$T(s) = -s + s^2 \cdot \frac{1}{s}\Big(1 - \mathbb{E}\Big[\frac{1}{1+s(\alpha+\beta X)}\Big]\Big) = -s + s - s\mathbb{E}\Big[\frac{1}{1+s(\alpha+\beta X)}\Big].$$

Hence

$$T(s) = -s\,\mathbb{E}\Big[\frac{1}{1+s\alpha+s\beta X}\Big]. \tag{3.3.53}$$

Now rewrite the expectation in Stieltjes form. Factor out $s\beta$:

$$1 + s\alpha + s\beta X = s\beta\Big(X + \frac{1+s\alpha}{s\beta}\Big) = s\beta\Big(X - \Big(-\frac{1+s\alpha}{s\beta}\Big)\Big).$$

Therefore

$$\mathbb{E}\Big[\frac{1}{1+s\alpha+s\beta X}\Big] = \frac{1}{s\beta}\,\mathbb{E}\Big[\frac{1}{X-\xi(s)}\Big], \qquad \xi(s) = -\frac{1+s\alpha}{s\beta}.$$

By definition of the MP Stieltjes transform, $\mathbb{E}\big[\frac{1}{X-\xi}\big] = m_{\mathrm{MP},\psi}(\xi)$, so

$$\mathbb{E}\Big[\frac{1}{1+s\alpha+s\beta X}\Big] = \frac{1}{s\beta}\,m_{\mathrm{MP},\psi}\big(\xi(s)\big).$$

Plug this into (3.3.53):

$$T(s) = -s \cdot \frac{1}{s\beta}\,m_{\mathrm{MP},\psi}\big(\xi(s)\big) = -\frac{1}{\beta}\,m_{\mathrm{MP},\psi}\big(\xi(s)\big).$$

Recalling $\beta = b_\sigma^2$ and $\xi(s) = -(1+(1-b_\sigma^2)s)/(b_\sigma^2 s)$ gives (3.3.49).

Lemma 33 shows T is invertible on certain intervals. Let $t = T(s)$. By (3.3.49), this is equivalent to

$$m_{\mathrm{MP},\psi}\big(\xi(s)\big) = -\beta t. \tag{3.3.54}$$

Set $m := -\beta t$. For MP, $m = m_{\mathrm{MP},\psi}(\xi)$ satisfies the exact quadratic identity

$$\psi\,\xi\,m^2 + (\xi + \psi - 1)m + 1 = 0. \tag{3.3.55}$$

Solve (3.3.55) for ξ in terms of m:

$$\big(\psi m^2 + m\big)\xi + (\psi - 1)m + 1 = 0 \quad \Longrightarrow \quad \xi = -\frac{1 + (\psi - 1)m}{m(1 + \psi m)}.$$

Thus, with $m = -\beta t$,

$$\xi = -\frac{1 + (\psi - 1)(-\beta t)}{(-\beta t)\big(1 + \psi(-\beta t)\big)} = \frac{1 - (\psi - 1)\beta t}{\beta t(1 - \psi\beta t)}. \tag{3.3.56}$$

By definition of $\xi(s)$,

$$\xi(s) = -\frac{1 + \alpha s}{\beta s}.$$

Hence, we have

$$s = -\frac{1}{\alpha + \beta\xi}.$$

Substituting ξ from (3.3.56) yields the claimed closed form

$$s = -\frac{t(1 - \psi\beta t)}{1 + (1 - \psi\beta)t - \psi\beta(1 - \beta)t^2}.$$

Recalling $\beta = b_\sigma^2$ proves (3.3.50). Let $s = T^{-1}(t)$ given by (3.3.50). The above computation is reversible: from t we obtain ξ via (3.3.56), then s via $s = -1/(\alpha + \beta\xi)$, and finally $m_{\mathrm{MP},\psi}(\xi) = -\beta t$ by construction. Plugging into (3.3.49) gives $T(s) = -(1/\beta)m_{\mathrm{MP},\psi}(\xi) = t$. This shows $T(T^{-1}(t)) = t$ whenever the algebra is well-defined. □

3.3.5 Centered and Normalized GELU

Let $\xi \sim \mathcal{N}(0,1)$ and denote the standard normal PDF and CDF by

$$\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) := \int_{-\infty}^{x} \varphi(t)\, dt.$$

The GELU activation function [HG16] is

$$g(x) := x\Phi(x).$$

We construct a centered and normalized version of GELU (w.r.t. ξ) by

$$\mu_g := \mathbb{E}[g(\xi)], \qquad v_g^2 := \mathrm{Var}(g(\xi)), \qquad \sigma(x) := \frac{g(x)-\mu_g}{v_g}. \tag{3.3.57}$$

Then $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma(\xi)^2] = 1$ by construction.

Assumption 3 is satisfied for $\sigma(\xi)$.

A direct calculation gives

$$g'(x) = \Phi(x) + x\varphi(x), \qquad g''(x) = (2-x^2)\varphi(x), \qquad g^{(3)}(x) = (x^3-4x)\varphi(x).$$

Since $0 \le \Phi(x) \le 1$ and $|x|^k\varphi(x)$ is bounded for any fixed k, we have

$$\sup_x\{|g'(x)|, |g''(x)|, |g^{(3)}(x)|\} < \infty$$

Hence the same holds for σ (with an extra factor $1/v_g$). Therefore σ satisfies Assumption 3.

Compute the centering constant $\mu_g = \mathbb{E}[g(\xi)]$.

We have

$$\mu_g = \mathbb{E}[\xi\Phi(\xi)] = \int_{\mathbb{R}} x\Phi(x)\,\varphi(x)\,dx.$$

Using $\varphi'(x) = -x\varphi(x)$, we write $x\varphi(x)\,dx = -d\varphi(x)$ and integrate by parts:

$$\begin{aligned}\mu_g &= \int_{\mathbb{R}} \Phi(x)\,x\varphi(x)\,dx = -\int_{\mathbb{R}} \Phi(x)\,d\varphi(x) \\ &= -\Big[\Phi(x)\varphi(x)\Big]_{-\infty}^{\infty} + \int_{\mathbb{R}} \varphi(x)\,d\Phi(x) = \int_{\mathbb{R}} \varphi(x)^2\,dx.\end{aligned}$$

Since $\varphi(x)^2 = \frac{1}{2\pi} e^{-x^2}$, we obtain

$$\mu_g = \int_{\mathbb{R}} \frac{1}{2\pi} e^{-x^2}\, dx = \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}}. \tag{3.3.58}$$

Compute the variance $v_g^2 = \mathrm{Var}(g(\xi))$.

Let

$$m_{2,g} := \mathbb{E}[g(\xi)^2] = \mathbb{E}[\xi^2 \Phi(\xi)^2] = \int_{\mathbb{R}} x^2 \Phi(x)^2 \varphi(x)\, dx.$$

Define the antiderivative

$$V(x) := \int_{-\infty}^{x} t^2 \varphi(t)\, dt.$$

A standard identity gives $V(x) = -x\varphi(x) + \Phi(x)$ (differentiate to check $V'(x) = x^2\varphi(x)$). Integrating by parts with $u = \Phi(x)^2$ and $dv = x^2\varphi(x)\,dx$ yields

$$\begin{aligned} m_{2,g} &= \Big[\Phi(x)^2 V(x)\Big]_{-\infty}^{\infty} - \int_{\mathbb{R}} V(x)\, d(\Phi(x)^2) \\ &= 1 - 2\int_{\mathbb{R}} V(x)\Phi(x)\varphi(x)\, dx = 1 - 2\int_{\mathbb{R}} (-x\varphi(x) + \Phi(x))\Phi(x)\varphi(x)\, dx \\ &= 1 + 2\int_{\mathbb{R}} x\Phi(x)\varphi(x)^2\, dx - 2\int_{\mathbb{R}} \Phi(x)^2\varphi(x)\, dx. \end{aligned}$$

The last term equals $\mathbb{E}[\Phi(\xi)^2]$. Since $U := \Phi(\xi) \sim \mathrm{Unif}(0,1)$, we have

$$\int_{\mathbb{R}} \Phi(x)^2 \varphi(x)\, dx = \mathbb{E}[U^2] = \frac{1}{3}. \tag{3.3.59}$$

For the mixed term, note $\varphi(x)^2 = \frac{1}{2\pi} e^{-x^2}$, hence

$$\int_{\mathbb{R}} x\Phi(x)\varphi(x)^2\, dx = \frac{1}{2\pi}\int_{\mathbb{R}} x\Phi(x) e^{-x^2}\, dx.$$

Integrate by parts with $u = \Phi(x)$ and $dv = xe^{-x^2}\,dx$ (so $v = -\frac{1}{2}e^{-x^2}$ and $du = \varphi(x)\,dx$):

$$\begin{aligned}\int_{\mathbb{R}} x\Phi(x)e^{-x^2}\,dx &= \Big[-\frac{1}{2}\Phi(x)e^{-x^2}\Big]_{-\infty}^{\infty} + \frac{1}{2}\int_{\mathbb{R}} e^{-x^2}\varphi(x)\,dx \\ &= \frac{1}{2}\int_{\mathbb{R}} e^{-x^2}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = \frac{1}{2\sqrt{2\pi}}\int_{\mathbb{R}} e^{-3x^2/2}\,dx \\ &= \frac{1}{2\sqrt{2\pi}}\sqrt{\frac{2\pi}{3}} = \frac{1}{2\sqrt{3}}.\end{aligned}$$

Therefore

$$\int_{\mathbb{R}} x\Phi(x)\varphi(x)^2\,dx = \frac{1}{2\pi}\cdot\frac{1}{2\sqrt{3}} = \frac{1}{4\pi\sqrt{3}}. \tag{3.3.60}$$

Plugging (3.3.59)–(3.3.60) into the expression for $m_{2,g}$ gives

$$m_{2,g} = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}}. \tag{3.3.61}$$

Finally, by (3.3.58) and (3.3.61),

$$v_g^2 = \mathrm{Var}(g(\xi)) = m_{2,g} - \mu_g^2 = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}} - \frac{1}{4\pi}. \tag{3.3.62}$$

Numerically, $v_g^2 \approx 0.3456440110$ and $v_g \approx 0.5879149692$.

Compute b_σ and c_σ.

Recall your definitions $b_\sigma = \mathbb{E}[\sigma'(\xi)]$ and $c_\sigma = \mathbb{E}[\sigma''(\xi)]$. Since $\sigma(x) = (g(x) - \mu_g)/v_g$, centering does not change derivatives and we have

$$\sigma'(x) = \frac{g'(x)}{v_g} = \frac{\Phi(x) + x\varphi(x)}{v_g}, \qquad \sigma''(x) = \frac{g''(x)}{v_g} = \frac{(2-x^2)\varphi(x)}{v_g}.$$

Hence

$$b_\sigma = \mathbb{E}[\sigma'(\xi)] = \frac{1}{v_g}\mathbb{E}[\Phi(\xi) + \xi\,\varphi(\xi)] = \frac{1}{v_g}\Big(\mathbb{E}[\Phi(\xi)] + \mathbb{E}[\xi\,\varphi(\xi)]\Big) = \frac{1}{v_g}\cdot\frac{1}{2} = \frac{1}{2v_g}, \quad (3.3.63)$$

$$c_\sigma = \mathbb{E}[\sigma''(\xi)] = \frac{1}{v_g}\mathbb{E}[(2-\xi^2)\varphi(\xi)] = \frac{1}{v_g}\int_{\mathbb{R}}(2-x^2)\varphi(x)^2\,dx. \quad (3.3.64)$$

To evaluate the last integral, use $\varphi(x)^2 = \frac{1}{2\pi}e^{-x^2}$ and the standard Gaussian integrals $\int_{\mathbb{R}} e^{-x^2}\,dx = \sqrt{\pi}$ and $\int_{\mathbb{R}} x^2 e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2}$:

$$\int_{\mathbb{R}}(2-x^2)\varphi(x)^2\,dx = \frac{1}{2\pi}\int_{\mathbb{R}}(2-x^2)e^{-x^2}\,dx = \frac{1}{2\pi}\Big(2\sqrt{\pi} - \frac{\sqrt{\pi}}{2}\Big) = \frac{3}{4\sqrt{\pi}}.$$

Plugging back into (3.3.64) yields

$$c_\sigma = \frac{1}{v_g}\cdot\frac{3}{4\sqrt{\pi}} = \frac{3}{4\sqrt{\pi}\,v_g}. \quad (3.3.65)$$

Numerically,

$$b_\sigma \approx 0.8504631217, \qquad c_\sigma \approx 0.7197336517.$$

Therefore, GELU does not spike in Theorem 5(ii) since $\tau \le \tau_{\text{crit}}$ and $\sqrt{\psi}\big(\frac{c_\sigma^2}{2} - b_\sigma^2\big) \le b_\sigma^2$.

3.3.6 Quadratic Equivalence for Proportional Limit

Let us recall the null model:

$$\boldsymbol{Y}_0 = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z}).$$

and QE model of the random feature matrix on XOR dataset:

$$\boldsymbol{Y}_{\text{QE}} = \boldsymbol{Y}_0 + \frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}}\Big(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top\Big) + \frac{\theta_{\text{SNR}}^2 c_\sigma}{\sqrt{N}}\Big(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\Big),$$

where $\boldsymbol{g}_1 = \boldsymbol{W}\boldsymbol{u}_1$, $\boldsymbol{g}_2 = \boldsymbol{W}\boldsymbol{u}_2$, and $\theta_{\mathrm{SNR}} := r\sqrt{\frac{n}{2d}}$.

Proof of Proposition 35. In this proof, we aim to prove a more precise statement:

$$\|\boldsymbol{Y} - \boldsymbol{Y}_{\mathrm{QE}}\| \prec n^{-1/4}. \tag{3.3.66}$$

Let us define

$$\boldsymbol{Y}_{\mathrm{QE}} := \boldsymbol{Y}_0 + \frac{\theta_{\mathrm{SNR}} b_\sigma}{\sqrt{N}}(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top) + \frac{\theta_{\mathrm{SNR}}^2 c_\sigma}{\sqrt{N}}(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}) \tag{3.3.67}$$

where $\boldsymbol{g}_1 := \boldsymbol{W}\boldsymbol{u}_1$ and $\boldsymbol{g}_2 := \boldsymbol{W}\boldsymbol{u}_2$. Assume that $r = O(n^{1/4})$. Then, under Assumptions 3 and 4, we can show that

$$\|\boldsymbol{Y} - \boldsymbol{Y}_{\mathrm{QE}}\| \prec \frac{1}{n^{1/4}}. \tag{3.3.68}$$

Recall that $\boldsymbol{Y} = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{M})$ where $\boldsymbol{Z}$ and $\boldsymbol{M}$ are defined by (3.2.6). Then, we can take Taylor approximations for each entry of $\boldsymbol{Y}$ around $\boldsymbol{Y}_0$:

$$\boldsymbol{Y} = \boldsymbol{Y}_0 + \frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{W}\boldsymbol{M}) + \frac{1}{2\sqrt{N}}\sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{W}\boldsymbol{M})^{\odot 2} + \frac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\boldsymbol{W}\boldsymbol{M})^{\odot 3}, \tag{3.3.69}$$

for some random matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N\times n}$ such that $\Theta_{i,i}$ is between $[\boldsymbol{W}\boldsymbol{Z}]_{i,j}$ and $[\boldsymbol{W}\boldsymbol{X}]_{i,j}$ for all $i \in [N]$ and $j \in [n]$. Notice that

$$\boldsymbol{W}\boldsymbol{M} = \theta_{\mathrm{SNR}}(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top), \tag{3.3.70}$$

and $\boldsymbol{g}_1, \boldsymbol{g}_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_N)$. By the definitions of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2 \in \mathbb{R}^n$, we have

$$(\boldsymbol{W}\boldsymbol{M})^{\odot 2} = \theta_{\mathrm{SNR}}^2(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top} + (\boldsymbol{g}_1 \odot \boldsymbol{g}_2)(\boldsymbol{v}_1 \odot \boldsymbol{v}_2)^\top) = \theta_{\mathrm{SNR}}^2(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}). \tag{3.3.71}$$

Thus, from (3.3.69), we have that

$$\boldsymbol{Y} = \boldsymbol{Y}_0 + \frac{\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})]\theta_{\text{SNR}}}{\sqrt{N}}(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top) + \frac{\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}}\boldsymbol{w}^\top \boldsymbol{z})]\theta_{\text{SNR}}^2}{2\sqrt{N}}(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}) \tag{3.3.72}$$

$$+ \mathbb{E}[\boldsymbol{Y}_0] + \frac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\boldsymbol{W}\boldsymbol{M})^{\odot 3} + \frac{1}{\sqrt{N}}(\sigma'(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}[\sigma'(\boldsymbol{W}\boldsymbol{Z})]) \odot (\boldsymbol{W}\boldsymbol{M}) \tag{3.3.73}$$

$$+ \frac{1}{2\sqrt{N}}(\sigma''(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}[\sigma''(\boldsymbol{W}\boldsymbol{Z})]) \odot (\boldsymbol{W}\boldsymbol{M})^{\odot 2}, \tag{3.3.74}$$

where $\boldsymbol{z}, \boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Next, we address the terms on the right-hand side of (3.3.74) individually.

(a) Since $\|\sigma'''\|_\infty = O(1)$, we have that

$$\begin{aligned}
\left\| \tfrac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\boldsymbol{W}\boldsymbol{M})^{\odot 3} \right\| &\lesssim \theta^3 \left\| \tfrac{1}{\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \right\| \sum_{i,j,k=1}^{2} \|\boldsymbol{g}_i\|_\infty \|\boldsymbol{g}_j\|_\infty \|\boldsymbol{g}_k\|_\infty \\
&\qquad \times \|\boldsymbol{v}_i\|_\infty \|\boldsymbol{v}_j\|_\infty \|\boldsymbol{v}_k\|_\infty \\
&\lesssim n^{-3/4} \left\| \tfrac{1}{\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \right\|_F \|\boldsymbol{g}_1\|_\infty^3 \\
&\prec \frac{1}{n^{1/4}},
\end{aligned} \tag{3.3.75}$$

where we apply Lemmas 28 and 29.

(b) Next, notice that

$$\begin{aligned}
&\|\frac{1}{\sqrt{N}}(\sigma'(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}[\sigma'(\boldsymbol{W}\boldsymbol{Z})]) \odot (\boldsymbol{W}\boldsymbol{M})\| \\
&\leq \|\frac{1}{\sqrt{N}}(\sigma'(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}[\sigma'(\boldsymbol{W}\boldsymbol{Z})])\| \cdot \theta_{\text{SNR}} \cdot (\|\boldsymbol{g}_1\|_\infty \|\boldsymbol{v}_1\|_\infty + \|\boldsymbol{g}_2\|_\infty \|\boldsymbol{v}_2\|_\infty) \\
&\lesssim \frac{\sqrt{\log n}}{n^{1/4}} \|\frac{1}{\sqrt{N}}(\sigma'(\boldsymbol{W}\boldsymbol{Z}) - \mathbb{E}[\sigma'(\boldsymbol{W}\boldsymbol{Z})])\| \prec \frac{1}{n^{1/4}}
\end{aligned} \tag{3.3.76}$$

where we utilize Lemma 30 in the last line since σ' is Lipschitz.

(c) Analogously, we have that

$$\frac{1}{2\sqrt{N}}\|(\sigma''(\boldsymbol{W}\boldsymbol{Z})-\mathbb{E}[\sigma''(\boldsymbol{W}\boldsymbol{Z})])\odot(\boldsymbol{W}\boldsymbol{M})^{\odot 2}\| \prec \frac{1}{\sqrt{n}}.$$

(d) Lastly, we apply Lemma 27 to obtain that

$$\boldsymbol{Y}=\boldsymbol{Y}_0+\frac{\theta_{\mathrm{SNR}}b_\sigma}{\sqrt{N}}(\boldsymbol{g}_1\boldsymbol{v}_1^\top+\boldsymbol{g}_2\boldsymbol{v}_2^\top)+\frac{\theta_{\mathrm{SNR}}^2c_\sigma}{2\sqrt{N}}(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top}+\boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top})+O_\prec(n^{-1/4}). \tag{3.3.77}$$

Combining all these together, we proved (3.3.66). Then, if $\|\boldsymbol{Y}\|, \|\boldsymbol{Y}_{\mathrm{QE}}\| \lesssim 1$, we can conclude that

$$\|\boldsymbol{K}-\boldsymbol{K}_{\mathrm{QE}}\| \prec n^{-1/4}$$

and they have asymptotically same outlier eigenvalues and eigenvector alignments. □

Remark. The adove proof can be extended to the case when $\sigma(x)$ is a piece-wise linear function defined by

$$\sigma(x)=\begin{cases} ax+b, & x>0, \\ cx+b, & x\le 0, \end{cases}$$

for some constants $a,b,c\in\mathbb{R}$ such that (3.2.2) holds. In this case, instead of taking Taylor approximation, we need the Hermite decomposition of σ.

3.3.7 Proof Strategy for Main Results

Before proceeding to the main proofs, we briefly outline the strategy. Recall $\boldsymbol{X}=\boldsymbol{Z}+\boldsymbol{M}$, where $\boldsymbol{M}$ is a rank-2 XOR signal and $\boldsymbol{Z}$ is isotropic noise. The main idea is to decompose the CK into a bulk component (driven by $\boldsymbol{Z}$) and low-rank signal components (driven by $\boldsymbol{M}$). We first identify any uninformative spikes, then analyze the signal-induced spikes by improving standard

tools from random matrix theory [BGN12].

In the proportional limit, for Theorems 5, 7, and 8, we expand $\sigma(\boldsymbol{W}(\boldsymbol{Z}+\boldsymbol{M}))$ around $\boldsymbol{WZ}$ and retain Hermite components of σ up to second order. This yields a bulk term $\sigma(\boldsymbol{WZ})$ and a low-rank spike decomposition for $\boldsymbol{Y}$. Let $\boldsymbol{Y}_0 = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{WZ})$ and $\theta_{\text{snr}} = r\sqrt{\frac{n}{2d}}$. Consider Gaussian random vectors $\boldsymbol{g}_1 = \boldsymbol{W}\boldsymbol{u}_1,\ \boldsymbol{g}_2 = \boldsymbol{W}\boldsymbol{u}_2$. Theorems 5, 7, and 8 repeatedly use the following QE:

$$\boldsymbol{Y}_{\text{QE}} := \boldsymbol{Y}_0 + \underbrace{\frac{\theta_{\text{snr}} b_\sigma}{\sqrt{N}}\left(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top\right)}_{\boldsymbol{T}_1 \text{ (linear spikes)}} + \underbrace{\frac{\theta_{\text{snr}}^2 c_\sigma}{\sqrt{N}}\left(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\right)}_{\boldsymbol{T}_2 \text{ (quadratic spikes)}}. \tag{3.3.78}$$

Proposition 35 (QE for XOR in Proportional Regime)**.** *Assume $r = O(n^{1/4})$ and Assumptions 3 and 4 hold. Let $\boldsymbol{Y}_{\text{QE}}$ be* (3.3.78)*. Then, in probability, $\|\boldsymbol{Y} - \boldsymbol{Y}_{\text{QE}}\| \to 0$. Consequently, if $\|\boldsymbol{Y}\|, \|\boldsymbol{Y}_{\text{QE}}\| \lesssim 1$, then $\boldsymbol{K} = \boldsymbol{Y}^\top\boldsymbol{Y}$ and $\boldsymbol{K}_{\text{QE}} := \boldsymbol{Y}_{\text{QE}}^\top\boldsymbol{Y}_{\text{QE}}$ have asymptotically matching outlier eigenvalues and eigenvector alignments.*

After this proposition, we can analyze informative spikes from $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, although we have to carefully consider the uninformative spikes from the null model $\boldsymbol{Y}_0$, which gives more technical difficulty. The magnitude of $\boldsymbol{T}_1$ scales like $\theta_{\text{snr}} b_\sigma$, while $\boldsymbol{T}_2$ scales like $\theta_{\text{snr}}^2 c_\sigma/\sqrt{n}$. For finite SNR (or when $c_\sigma = 0$), the quadratic spikes $\boldsymbol{T}_2$ are negligible relative to the bulk. In this case, only the linear spikes in $\boldsymbol{T}_1$ contribute; but these are uninformative for linear clustering. In the large-SNR regime (with $c_\sigma \neq 0$), $\boldsymbol{T}_2$ becomes comparable to the bulk. Then since $\boldsymbol{y} \in \text{span}\{\boldsymbol{v}_1^{\odot 2}, \boldsymbol{v}_2^{\odot 2}\}$, the corresponding outlier eigenvectors become label-informative.

In Section 3.2.6, due to the additional assumption of ultra-wide scaling $N \gg n$, the random kernel concentrates around a deterministic quadratic polynomial kernel, which depends solely on the first two Hermite coefficients of σ. In particular, we have the following concentration result:

Proposition 36 (Concentration of CK for Quadratic Regime; informal)**.** *Let $a_\sigma^2 = 1 - b_\sigma^2 - c_\sigma^2/2$. Assume that $2n/d^2 \to \gamma \in (0,\infty)$ and $n^2/N \to 0$. Under Assumption 3, we have that*

$$\left\|\boldsymbol{K} - \left(a_\sigma^2\boldsymbol{P} + b_\sigma^2\boldsymbol{X}^\top\boldsymbol{X} + \frac{c_\sigma^2}{2}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) \odot \left(\boldsymbol{X}^\top\boldsymbol{X}\right)\right)\right\| \xrightarrow{\mathbb{P}} 0. \tag{3.3.79}$$

3.3.8 Proofs for Finite-SNR and Proportional Limit Regime

In this section, we present the proof for Theorem 5 when we consider proportional limit and finite SNR regime. In this case, the rank-two XOR signal satisfies

$$\boldsymbol{M} = \theta_{\mathrm{SNR}}\,(\boldsymbol{u}_1\boldsymbol{v}_1^\top + \boldsymbol{u}_2\boldsymbol{v}_2^\top), \qquad \theta_{\mathrm{SNR}} := r\sqrt{\frac{n}{2d}} \;\to\; r\sqrt{\frac{\psi}{2}}.$$

Furthermore, we know that both $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$ are deterministic orthonormal vectors satisfying

$$\boldsymbol{v}_1^\top \boldsymbol{v}_2 = 0, \qquad \|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = 1, \qquad \boldsymbol{u}^\top \boldsymbol{v}_1 = \boldsymbol{u}^\top \boldsymbol{v}_2 = 0, \qquad \boldsymbol{u} := \frac{\mathbf{1}_n}{\sqrt{n}}. \tag{3.3.80}$$

Recall that $\boldsymbol{Z} \in \mathbb{R}^{d\times n}$ have i.i.d. $\mathcal{N}(0, 1/d)$ entries, and $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, independent of $\boldsymbol{Z}$. Recall the null feature matrix

$$\boldsymbol{Y}_0 = \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z}),$$

and its columnwise conditional mean

$$\boldsymbol{m} := \mathbb{E}[\sigma(\boldsymbol{w}^\top \boldsymbol{Z}) \mid \boldsymbol{Z}] \in \mathbb{R}^n,$$

whose entries are

$$m_j = F(\|\boldsymbol{z}_j\|^2), \quad F(s) := \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma(\sqrt{s}\,\xi)], \quad j \in [n].$$

Here $\boldsymbol{z}_j$ denotes the j-th column of $\boldsymbol{Z}$, and $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ independent of $\boldsymbol{Z}$. Notice that conditioning on $\boldsymbol{Z}$, $\boldsymbol{Y}_0$ has independent and identical distributed rows. However, when $c_\sigma \neq 0$, the mean of rows in $\boldsymbol{Y}_0$, $\boldsymbol{m}$, is not vanishing which will potentially give us some uninformative spike, see [BP22, WWF24]. Hence, we fist need to subtract the mean vector $\boldsymbol{m}$ to make all rows

centered and to apply theorems in [WWF24]. Then we consider the effect the mean vector $\boldsymbol{m}$. Define the centered null feature matrix by

$$\boldsymbol{G} := \boldsymbol{Y}_0 - \frac{1}{\sqrt{N}} \mathbf{1}_N \boldsymbol{m}^\top \tag{3.3.81}$$

and the null model of CK matrix by

$$\boldsymbol{K}_0 := \boldsymbol{G}^\top \boldsymbol{G}.$$

3.3.9 Analysis of the Population Covariance Matrix

Conditional on $\boldsymbol{Z}$, the rows of $\boldsymbol{G}$ are i.i.d. centered vectors in $\mathbb{R}^n$. Let $\bar{\boldsymbol{g}}^\top := \sigma(\boldsymbol{w}^\top \boldsymbol{Z}) - \boldsymbol{m}^\top$ for $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ independent of $\boldsymbol{Z}$, and define the conditional population covariance

$$\boldsymbol{\Sigma} := \mathbb{E}[\bar{\boldsymbol{g}}\, \bar{\boldsymbol{g}}^\top \mid \boldsymbol{Z}] \in \mathbb{R}^{n \times n}. \tag{3.3.82}$$

Lemma 37 (Approximation of $\boldsymbol{\Sigma}$)**.** *Let $\boldsymbol{S} := \boldsymbol{Z}^\top \boldsymbol{Z}$. Under Assumptions 3 and 4, we have*

$$\boldsymbol{\Sigma} = (1 - b_\sigma^2) \boldsymbol{I}_n + b_\sigma^2 \boldsymbol{S} + \frac{c_\sigma^2}{2} \left(\boldsymbol{S}^{\odot 2} - \boldsymbol{I}_n \right) + \boldsymbol{R}_\Sigma, \tag{3.3.83}$$

such that $\|\boldsymbol{R}_\Sigma\| = \widetilde{O}_{\mathbb{P}}(n^{-1/2})$.

Proof. With Lemma 5.2 of [WZ21] and Lemma 24, we have

$$\boldsymbol{\Sigma} = (1 - b_\sigma^2) \boldsymbol{I}_n + b_\sigma^2 \boldsymbol{S} + \frac{c_\sigma^2}{2} \left(\boldsymbol{S}^{\odot 2} - \boldsymbol{I}_n \right) + d_\sigma^2 (\boldsymbol{S}^{\odot 3} - \boldsymbol{I}_n) + O_{\mathbb{P}}(n^{-1/2}),$$

where d_σ is the third-order Hermite coefficient of σ. Note that under event of Lemma 24,

$$\|(\boldsymbol{Z}^\top \boldsymbol{Z})^{\odot 3} - \mathbf{I}_n\| \le \|\mathrm{diag}\big((\boldsymbol{Z}^\top \boldsymbol{Z})^{\odot 3} - \mathbf{I}_n\big)\| + \|\mathrm{offdiag}\big(\boldsymbol{Z}^\top \boldsymbol{Z}\big)^{\odot 3}\|_F$$

$$\le \max_{\alpha \in [n]} \big| |\boldsymbol{z}_\alpha|^6 - 1 \big| + n \cdot \max_{\alpha \neq \beta \in [n]} |\boldsymbol{z}_\alpha^\top \boldsymbol{z}_\beta|^3 \le C\big(\tau_n + n\tau_n^3\big) \le 2C\sqrt{(\log^3 n)/n} \tag{3.3.84}$$

for some constant $C \geq 0$. This completes the proof. □

Lemma 38. *Let $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z}$ and $\boldsymbol{u} = \mathbf{1}_n/\sqrt{n}$. Then*

$$\boldsymbol{S}^{\odot 2} - \boldsymbol{I}_n = \frac{n}{d}\boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{E}_2, \qquad \|\boldsymbol{E}_2\| = o_{\mathbb{P}}(1). \tag{3.3.85}$$

Proof. The proof is directly derived from the proof of Theorem 2.1 of [EK10], where $\boldsymbol{S}^{\odot 2}$ is the quadratic term in the Taylor expansion of the inner-product kernel of $\boldsymbol{Z}$. □

Corollary 39. *Define*

$$\boldsymbol{\Sigma}_0 := (1-b_\sigma^2)\boldsymbol{I}_n + b_\sigma^2 \boldsymbol{S}, \qquad \tau := \frac{c_\sigma^2}{2}\frac{n}{d} \to \frac{c_\sigma^2}{2}\psi, \qquad \boldsymbol{u} := \frac{\mathbf{1}_n}{\sqrt{n}}. \tag{3.3.86}$$

Then

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \tau \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{R}, \qquad \|\boldsymbol{R}\| = o_{\mathbb{P}}(1). \tag{3.3.87}$$

Moreover the limiting ESD of $\boldsymbol{\Sigma}_0$ (and of $\boldsymbol{\Sigma}$) equals

$$\nu = (1-b_\sigma^2) + b_\sigma^2 \rho_\psi^{\mathrm{MP}}. \tag{3.3.88}$$

Thus, the right edge of ν is $\lambda_+^\nu = (1-b_\sigma^2) + b_\sigma^2(1+\sqrt{\psi})^2$.

Proof. Combine Lemma 37 with Lemma 38 and absorb all $o_{\mathbb{P}}(1)$ terms into $\boldsymbol{R}$. The rank-one term $\tau\boldsymbol{u}\boldsymbol{u}^\top$ does not change the limiting ESD. The limiting ESD of $\boldsymbol{\Sigma}_0 = (1-b_\sigma^2)\boldsymbol{I} + b_\sigma^2 \boldsymbol{Z}^\top\boldsymbol{Z}$ is the shifted MP law ν in (3.3.88). □

3.3.10 Analysis of Spike in Population Covariance

Lemma 40 (Population outlier from $\boldsymbol{\Sigma}_0 + \tau\boldsymbol{u}\boldsymbol{u}^\top$)**.** *Let $\boldsymbol{A} := \boldsymbol{\Sigma}_0$ and consider $\boldsymbol{A} + \tau\boldsymbol{u}\boldsymbol{u}^\top$. For $\tau \geq 0$, any deterministic limit point $\Lambda_\tau > \lambda_+^\nu$ of an isolated eigenvalue of $\boldsymbol{A} + \tau\boldsymbol{u}\boldsymbol{u}^\top$ must satisfy*

$$1 + \tau m_\nu(\Lambda_\tau) = 0, \tag{3.3.89}$$

where m_ν is the Stieltjes transform of ν. Moreover, such a solution exists iff

$$\tau > \tau_{\text{crit}} := b_\sigma^2 \sqrt{\psi}(1+\sqrt{\psi}). \tag{3.3.90}$$

Proof. For any $\lambda \notin \operatorname{spec}(\boldsymbol{A})$, the rank-one perturbation identity gives

$$\lambda \in \operatorname{spec}(\boldsymbol{A} + \tau \boldsymbol{u}\boldsymbol{u}^\top) \setminus \operatorname{spec}(\boldsymbol{A}) \quad \Longleftrightarrow \quad 1 + \tau\, \boldsymbol{u}^\top (\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \boldsymbol{u} = 0. \tag{3.3.91}$$

Write $\boldsymbol{A} = (1-b_\sigma^2)\boldsymbol{I} + b_\sigma^2 \boldsymbol{S}$ with $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z}$. For $\lambda > \lambda_+^\nu + \kappa$ (fixed $\kappa > 0$), define

$$\zeta := \frac{\lambda - (1-b_\sigma^2)}{b_\sigma^2} > (1+\sqrt{\psi})^2 + \kappa',$$

so that

$$(\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} = \frac{1}{b_\sigma^2} (\boldsymbol{S} - \zeta \boldsymbol{I})^{-1}.$$

The isotropic MP local law [BEK+14] yields, uniformly for such ζ,

$$\boldsymbol{u}^\top (\boldsymbol{S} - \zeta \boldsymbol{I})^{-1} \boldsymbol{u} = m_{\text{MP},\psi}(\zeta) + o_{\mathbb{P}}(1),$$

hence

$$\boldsymbol{u}^\top (\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \boldsymbol{u} = \frac{1}{b_\sigma^2} m_{\text{MP},\psi}(\zeta) + o_{\mathbb{P}}(1) = m_\nu(\lambda) + o_{\mathbb{P}}(1). \tag{3.3.92}$$

Define $F(\lambda) := 1 + \tau m_\nu(\lambda)$ for $\lambda > \lambda_+^\nu$. Since $m_\nu(\lambda) < 0$ and is strictly increasing on (λ_+^ν, ∞), F is strictly increasing and $\lim_{\lambda \to \infty} F(\lambda) = 1$. Thus a solution exists iff $F(\lambda_+^\nu) < 0$, i.e. $\tau > -1/m_\nu(\lambda_+^\nu)$. At the MP edge, $m_{\text{MP},\psi}((1+\sqrt{\psi})^2) = -\frac{1}{\sqrt{\psi}(1+\sqrt{\psi})}$. Using $m_\nu(\lambda) = \frac{1}{b_\sigma^2} m_{\text{MP},\psi}(\zeta)$, at $\lambda = \lambda_+^\nu$ we get

$$m_\nu(\lambda_+^\nu) = -\frac{1}{b_\sigma^2 \sqrt{\psi}(1+\sqrt{\psi})}.$$

Hence the critical τ is $\tau_{\text{crit}} = -1/m_\nu(\lambda_+^\nu) = b_\sigma^2 \sqrt{\psi}(1+\sqrt{\psi})$, proving (3.3.90). □

Lemma 41 (Closed form for Λ_τ when ν is shifted MP)**.** *Let* $\nu = (1-b_\sigma^2) + b_\sigma^2 \rho_\psi^{\mathrm{MP}}$ *and assume* $\tau > \tau_{\mathrm{crit}}$*. Then the unique solution* $\Lambda_\tau > \lambda_+^\nu$ *of* $1 + \tau m_\nu(\Lambda_\tau) = 0$ *equals*

$$\Lambda_\tau = (1-b_\sigma^2) + \tau + \frac{b_\sigma^2\,\tau}{\tau - b_\sigma^2\,\psi}. \tag{3.3.93}$$

Proof. Write $\Lambda = (1-b_\sigma^2) + b_\sigma^2\,\zeta$. Then $m_\nu(\Lambda) = \frac{1}{b_\sigma^2} m_{\mathrm{MP},\psi}(\zeta)$, and

$$m_{\mathrm{MP},\psi}(\zeta) = -\frac{b_\sigma^2}{\tau}. \tag{3.3.94}$$

The Stieltjes transform $m = m_{\mathrm{MP},\psi}(\zeta)$ satisfies the quadratic equation

$$\psi\,\zeta\,m^2 + (\zeta + \psi - 1)m + 1 = 0. \tag{3.3.95}$$

Solving (3.3.95) for ζ in terms of m yields

$$\zeta = -\frac{1 + (\psi - 1)m}{m(1+\psi m)} = -\frac{1}{m} + \frac{1}{1+\psi m}. \tag{3.3.96}$$

Substitute $m = -b_\sigma^2/\tau$ from (3.3.94) into (3.3.96):

$$\zeta = \frac{\tau}{b_\sigma^2} + \frac{1}{1 - \psi b_\sigma^2/\tau} = \frac{\tau}{b_\sigma^2} + \frac{\tau}{\tau - \psi b_\sigma^2}.$$

Therefore, we have

$$\Lambda_\tau = (1-b_\sigma^2) + b_\sigma^2\,\zeta = (1-b_\sigma^2) + \tau + \frac{b_\sigma^2\,\tau}{\tau - b_\sigma^2\,\psi},$$

which is (3.3.93). Uniqueness follows from strict monotonicity of $m_\nu(\lambda)$ on (λ_+^ν, ∞). □

Corollary 42 (Subcritical case: no population outlier)**.** *If* $\tau \le \tau_{\mathrm{crit}}$*, then* $\mathbf{\Sigma}$ *has no isolated eigenvalue outside* $\mathrm{supp}(\nu)$ *and* $\lambda_{\max}(\mathbf{\Sigma}) = \lambda_+^\nu + o_{\mathbb{P}}(1)$.

Proof. If $\tau \le \tau_{\text{crit}}$, then Lemma 40 shows the limiting secular equation has no root in $(\lambda_+^{\nu}, \infty)$. Using (3.3.91)–(3.3.92) and [BS98], no separated eigenvalue exists and $\lambda_{\max}(\boldsymbol{\Sigma}) = \lambda_+^{\nu} + o_{\mathbb{P}}(1)$. □

3.3.11 BBP Transition for Null Model

Let z, φ be defined from (ν, ϕ) by (3.3.38)–(3.3.39) where ν is defiend by (3.3.88). In this section, we analyze the uninformative spike of $\boldsymbol{K}_0$ derived from the potential spike in the population covariance matrix $\boldsymbol{\Sigma}$.

Lemma 43 (Bulk of $\boldsymbol{K}_0$ and multiplicative BBP outlier from Λ_τ)**.** *Let $\boldsymbol{K}_0 = \boldsymbol{G}^\top \boldsymbol{G}$.*

(i) (Bulk.) *The ESD of $\boldsymbol{K}_0$ converges weakly in probability to $\mu = \rho_\phi^{\text{MP}} \boxtimes \nu$.*

(ii) (BBP for covariance outlier.) *If $\tau > \tau_{\text{crit}}$, let Λ_τ be the population outlier of $\boldsymbol{\Sigma}$ in* (3.3.93)*. Then:*

$$\text{If additionally } z'\Big(-\frac{1}{\Lambda_\tau}\Big) > 0, \quad \boldsymbol{K}_0 \text{ has a separated outlier at } \widehat{\lambda}_\tau = z\Big(-\frac{1}{\Lambda_\tau}\Big) + o_{\mathbb{P}}(1),$$

where $\widehat{\lambda}_\tau$ lies in a connected component of $\mathbb{R} \setminus \text{supp}(\mu)$. If $z'(-1/\Lambda_\tau) \le 0$ or $\tau \le \tau_{\text{crit}}$, the corresponding eigenvalue sticks to $\text{supp}(\mu)$ and $\boldsymbol{K}_0$ has no outlier.

(iii) (Alignment.) *In the separated case, if $\widehat{\boldsymbol{v}}_\tau$ is a unit eigenvector of $\boldsymbol{K}_0$ associated to $\widehat{\lambda}_\tau$, then*

$$|\widehat{\boldsymbol{v}}_\tau^\top \boldsymbol{u}|^2 \xrightarrow{\mathbb{P}} \varphi\Big(-\frac{1}{\Lambda_\tau}\Big). \tag{3.3.97}$$

In general, any fixed unit vectors $\boldsymbol{v} \in \mathbb{R}^n$ (potentially $\boldsymbol{Z}$ dependent),

$$|\widehat{\boldsymbol{v}}_\tau^\top \boldsymbol{v}|^2 - \varphi\Big(-\frac{1}{\Lambda_\tau}\Big)|\boldsymbol{u}^\top \boldsymbol{v}|^2 \xrightarrow{\mathbb{P}} 0. \tag{3.3.98}$$

Proof. This lemma is an application of Theorem 12 of [WWF24]. It suffices to verify Assumptions 5 and 6 of [WWF24]. Assumption 6 is verified by Lemmas 40 and 41, and Corollary 42:

$\boldsymbol{\Sigma}$ has bulk law ν and at most one population spike Λ_τ (in the $\boldsymbol{u}$ direction), present iff $\tau > \tau_{\text{crit}}$. Notice that from Corollary 3.5 of [WZ21], we have

$$\bar{\boldsymbol{g}}^\top \boldsymbol{A}\bar{\boldsymbol{g}} - \operatorname{Tr}\boldsymbol{A}\boldsymbol{\Sigma} \prec \|\boldsymbol{A}\|_F,$$

for any deterministic matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then, Assumption 5 can be verified based on Lemmas 32, 33, and 34 in [WWF24]. We ignore the details here for simplicity. □

3.3.12 BBP Transition for Rank-three Additive Deformation

First, we prove that, in the finite-SNR regime $r = \Theta(1)$, the explicit quadratic term in Proposition 35 is negligible.

Lemma 44 (Finite-SNR: the explicit quadratic spike is negligible)**.** *Assume $r = \Theta(1)$ and the assumptions of Proposition 35. Recall*

$$\boldsymbol{T}_2 = \frac{\theta_{\text{SNR}}^2 c_\sigma}{\sqrt{N}}\big(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\big).$$

Then $\|\boldsymbol{T}_2\| = O_\prec(n^{-1/2})$. Furthermore, $\|\boldsymbol{Y}_{\text{QE}}\| \prec 1$ and

$$\big\|(\boldsymbol{Y}_{\text{QE}} - \boldsymbol{T}_2)^\top(\boldsymbol{Y}_{\text{QE}} - \boldsymbol{T}_2) - \boldsymbol{Y}_{\text{QE}}^\top\boldsymbol{Y}_{\text{QE}}\big\| \prec n^{-1/2}.$$

Proof. We bound

$$\|\boldsymbol{T}_2\| \le \frac{|\theta_{\text{SNR}}^2 c_\sigma|}{\sqrt{N}} \sum_{\ell=1}^{2} \|\boldsymbol{g}_\ell^{\odot 2}\|\,\|\boldsymbol{v}_\ell^{\odot 2}\|.$$

Since $\boldsymbol{g}_\ell \sim \mathcal{N}(0, \boldsymbol{I}_N)$, $\|\boldsymbol{g}_\ell^{\odot 2}\|^2 = \sum_{i=1}^N g_{\ell,i}^4 = O_\prec(N)$, hence $\|\boldsymbol{g}_\ell^{\odot 2}\| = O_\prec(\sqrt{N})$. Also $\|\boldsymbol{v}_\ell^{\odot 2}\|^2 = \sum_{j=1}^n v_{\ell,j}^4 = 2/n$ for XOR sign vectors, so $\|\boldsymbol{v}_\ell^{\odot 2}\| = \sqrt{2/n}$. Therefore $\|\boldsymbol{T}_2\| = O_\prec(n^{-1/2})$.

By Lemmas 26 and 30, we have $\sigma(\boldsymbol{W}\boldsymbol{Z})/\sqrt{N} \prec 1$. And since $\boldsymbol{g}_1, \boldsymbol{g}_2 \prec \sqrt{N}$ and $\|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = 1$, we have that $\boldsymbol{Y}_{\text{QE}} \prec 1$ by the definition of $\boldsymbol{Y}_{\text{QE}}$ in (3.3.78). The Gram difference bound

follows by expansion:

$$(\boldsymbol{Y}_{\mathrm{QE}}-\boldsymbol{T}_2)^\top(\boldsymbol{Y}_{\mathrm{QE}}-\boldsymbol{T}_2)-\boldsymbol{Y}_{\mathrm{QE}}^\top\boldsymbol{Y}_{\mathrm{QE}}=-\boldsymbol{Y}_{\mathrm{QE}}^\top\boldsymbol{T}_2-\boldsymbol{T}_2^\top\boldsymbol{Y}_{\mathrm{QE}}+\boldsymbol{T}_2^\top\boldsymbol{T}_2,$$

hence, applying $\|\boldsymbol{Y}_{\mathrm{QE}}\| \prec 1$ and $\|\boldsymbol{T}_2\| = O_\prec(n^{-1/2})$ we can get the final result. □

In the next proposition, we demonstrate that, in addition to the potential spike in $\boldsymbol{G}$ characterized in Lemma 43, the matrix $\boldsymbol{Y}$ may exhibit further potential spikes arising from a rank-three additive perturbation of $\boldsymbol{G}$.

Proposition 45 (Rank-three additive deformation of $\boldsymbol{G}$)**.** *Assume $r = \Theta(1)$. Define*

$$\boldsymbol{Y}^\sharp := \boldsymbol{G} + \underbrace{\frac{1}{\sqrt{N}}\mathbf{1}_N\boldsymbol{m}^\top}_{\textit{mean spike}} + \underbrace{\frac{\theta_{\mathrm{SNR}}b_\sigma}{\sqrt{N}}\left(\boldsymbol{g}_1\boldsymbol{v}_1^\top+\boldsymbol{g}_2\boldsymbol{v}_2^\top\right)}_{\textit{two XOR linear spikes}}. \tag{3.3.99}$$

Denote a new kernel matrix by $\boldsymbol{K}^\sharp := \boldsymbol{Y}^{\sharp\top}\boldsymbol{Y}^\sharp$. Then

$$\|\boldsymbol{Y}-\boldsymbol{Y}^\sharp\| \prec n^{-1/4}, \qquad \textit{and hence} \qquad \|\boldsymbol{K}-\boldsymbol{K}^\sharp\| \prec n^{-1/4}. \tag{3.3.100}$$

In particular, all the isolated eigenvalues and their eigenvectors of $\boldsymbol{K}$ and $\boldsymbol{K}^\sharp$ are asymptotically the same.

Proof. By the proof of Proposition 35 and Lemma 44, in finite SNR, we can further drop $\boldsymbol{T}_2$ to get $\|\boldsymbol{Y}-\boldsymbol{Y}^\sharp\| \prec n^{-1/4}$. From the proof of Lemma 44, we also know that $\|\boldsymbol{Y}\|, \|\boldsymbol{Y}_{\mathrm{QE}}\|, \|\boldsymbol{Y}^\sharp\| \prec 1$. Hence we have

$$\|\boldsymbol{K}-\boldsymbol{K}^\sharp\| = \|\boldsymbol{Y}^\top\boldsymbol{Y}-\boldsymbol{Y}^{\sharp\top}\boldsymbol{Y}^\sharp\| \le \|\boldsymbol{Y}-\boldsymbol{Y}^\sharp\|(\|\boldsymbol{Y}\|+\|\boldsymbol{Y}^\sharp\|) \prec n^{-1/4},$$

which implies asumptotic matching of isolated eigenvalues/eigenspaces by Weyl's inequality and Davis–Kahan Theorem [DK70, Wed72]. □

Orthogonal Decomposition of Weight

Next, to analyze two XOR linear spikes in (3.3.99), we need to make bulk matrix $\boldsymbol{G}$ independent with the vectors $\boldsymbol{g}_1, \boldsymbol{g}_2$ by the following decomposition.

Lemma 46 (Orthogonal decomposition and independence)**.** *Let* $\boldsymbol{P}_\perp := \boldsymbol{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top$ *and define*

$$\boldsymbol{W}_\perp := \boldsymbol{W}\boldsymbol{P}_\perp, \qquad \boldsymbol{g}_1 = \boldsymbol{W}\boldsymbol{u}_1, \qquad \boldsymbol{g}_2 = \boldsymbol{W}\boldsymbol{u}_2.$$

Then:

(i) $\boldsymbol{W} = \boldsymbol{W}_\perp + \boldsymbol{g}_1\boldsymbol{u}_1^\top + \boldsymbol{g}_2\boldsymbol{u}_2^\top$.

(ii) $\boldsymbol{g}_1, \boldsymbol{g}_2$ *are independent* $\mathcal{N}(0, \boldsymbol{I}_N)$ *vectors and are independent of* $\boldsymbol{W}_\perp$.

Proof. For each row $\boldsymbol{w}_i^\top$ of $\boldsymbol{W}$, decompose in $\mathbb{R}^d$:

$$\boldsymbol{w}_i = (\boldsymbol{w}_i^\top \boldsymbol{u}_1)\boldsymbol{u}_1 + (\boldsymbol{w}_i^\top \boldsymbol{u}_2)\boldsymbol{u}_2 + \boldsymbol{P}_\perp \boldsymbol{w}_i.$$

Stacking over i yields $\boldsymbol{W} = \boldsymbol{W}_\perp + \boldsymbol{g}_1\boldsymbol{u}_1^\top + \boldsymbol{g}_2\boldsymbol{u}_2^\top$. Since $\boldsymbol{w}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$, the scalar components $\boldsymbol{w}_i^\top \boldsymbol{u}_1$, $\boldsymbol{w}_i^\top \boldsymbol{u}_2$, and the vector $\boldsymbol{P}_\perp \boldsymbol{w}_i$ are independent Gaussians because they are orthogonal linear functionals of a Gaussian vector. Stacking over i gives independence of $\boldsymbol{g}_1, \boldsymbol{g}_2$ and $\boldsymbol{W}_\perp$. □

Lemma 47 (The null bulk is asymptotically independent of $\boldsymbol{g}_1, \boldsymbol{g}_2$)**.** *Define*

$$\widetilde{\boldsymbol{Y}}_0 := \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}_\perp \boldsymbol{Z}), \qquad \widetilde{\boldsymbol{m}} := \mathbb{E}[\sigma(\widetilde{\boldsymbol{w}}^\top \boldsymbol{Z}) \mid \boldsymbol{Z}], \;\; \widetilde{\boldsymbol{w}} \sim \mathcal{N}(0, \boldsymbol{P}_\perp), \qquad \widetilde{\boldsymbol{G}} := \widetilde{\boldsymbol{Y}}_0 - \frac{1}{\sqrt{N}}\boldsymbol{1}_N \widetilde{\boldsymbol{m}}^\top.$$

Then

$$\|\boldsymbol{m} - \widetilde{\boldsymbol{m}}\|, \|\boldsymbol{G} - \widetilde{\boldsymbol{G}}\| \prec 1/\sqrt{N}, \tag{3.3.101}$$

and $\widetilde{\boldsymbol{G}}$ *is independent of* $(\boldsymbol{g}_1, \boldsymbol{g}_2)$.

Proof. By Lemma 46,

$$\boldsymbol{W}\boldsymbol{Z} = \boldsymbol{W}_\perp \boldsymbol{Z} + \boldsymbol{g}_1(\boldsymbol{u}_1^\top \boldsymbol{Z}) + \boldsymbol{g}_2(\boldsymbol{u}_2^\top \boldsymbol{Z}).$$

Let $\boldsymbol{t}_1^\top := \boldsymbol{u}_1^\top \boldsymbol{Z}$ and $\boldsymbol{t}_2^\top := \boldsymbol{u}_2^\top \boldsymbol{Z}$. Each $\boldsymbol{t}_k \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0, 1/d)$ entries and is independent of $\boldsymbol{W}$. Set $\boldsymbol{A} := \boldsymbol{W}_\perp \boldsymbol{Z}$ and $\boldsymbol{\Delta} := \boldsymbol{g}_1 \boldsymbol{t}_1^\top + \boldsymbol{g}_2 \boldsymbol{t}_2^\top$, so $\sigma(\boldsymbol{WZ}) = \sigma(\boldsymbol{A} + \boldsymbol{\Delta})$ and $\sigma(\boldsymbol{W}_\perp \boldsymbol{Z}) = \sigma(\boldsymbol{A})$.

Following the same proof of Proposition 35, we can entrywisely a apply the following Taylor approximation, with $a = A_{ij}$, $\delta = \Delta_{ij}$, bounded $\sigma^{(3)}$ gives

$$\sigma(a+\delta) = \sigma(a) + \sigma'(a)\delta + \frac{1}{2}\sigma''(\xi)\delta^2$$

for some ξ between a and $a+\delta$. Therefore

$$\boldsymbol{Y}_0 - \widetilde{\boldsymbol{Y}}_0 = \frac{1}{\sqrt{N}}\big(\sigma(\boldsymbol{A}+\boldsymbol{\Delta}) - \sigma(\boldsymbol{A})\big) = \widetilde{\boldsymbol{T}}_1 + \widetilde{\boldsymbol{T}}_2,$$

where

$$\widetilde{\boldsymbol{T}}_1 := \frac{1}{\sqrt{N}}\big(\sigma'(\boldsymbol{A}) \odot \boldsymbol{\Delta}\big), \quad \widetilde{\boldsymbol{T}}_2 := \frac{1}{2\sqrt{N}}\big(\sigma''(\boldsymbol{\Xi}) \odot \boldsymbol{\Delta}^{\odot 2}\big).$$

First, we have

$$\|\widetilde{\boldsymbol{T}}_1\| \le \frac{1}{\sqrt{N}} \sum_{i=1}^{2} \|\operatorname{diag}(\boldsymbol{g}_i)\| \, \|\sigma'(\boldsymbol{A})\| \, \|\operatorname{diag}(\boldsymbol{t}_i)\| \prec 1/\sqrt{n},$$

where we apply Lemmas 28 and 29, and $\|\sigma'(\boldsymbol{A})\| = O_\prec(\sqrt{N})$ because σ' is bounded and $\boldsymbol{A}$ has i.i.d. bounded sub-Gaussian norm of rows. Similarly we have $\|\widetilde{\boldsymbol{T}}_2\| \prec 1/\sqrt{n}$ since $\|\sigma'; (\boldsymbol{\Xi})\| = O_\prec(1)$. A parallel argument applies to the conditional mean vectors, giving $\|\boldsymbol{m} - \widetilde{\boldsymbol{m}}\| \prec 1/\sqrt{n}$ since $\|\boldsymbol{m} - \widetilde{\boldsymbol{m}}\| \le \mathbb{E}[\|\boldsymbol{Y}_0 - \widetilde{\boldsymbol{Y}}_0\| | \boldsymbol{Z}]$. Therefore we complete the proof. □

Mean Spike Strength and Orthogonality

Lemma 48 (Mean spike strength and orthogonality)**.** *Under the Assumptions 3 and 4, we have*

$$\|\boldsymbol{m}\|^2 \xrightarrow{\mathbb{P}} \beta_{\text{mean}} := \frac{c_\sigma^2}{2}\psi, \tag{3.3.102}$$

and

$$\boldsymbol{u}^\top \boldsymbol{m} = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}_1^\top \boldsymbol{m} = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}_2^\top \boldsymbol{m} = o_{\mathbb{P}}(1). \tag{3.3.103}$$

Proof. Fix $j \in [n]$ and condition on $\boldsymbol{z}_j$. Since $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ is independent of $\boldsymbol{z}_j$, $\boldsymbol{w}^\top \boldsymbol{z}_j \sim \mathcal{N}(0, \|\boldsymbol{z}_j\|^2)$. Define for $s > 0$

$$F(s) := \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}\big[\sigma(\sqrt{s}\,\xi)\big],$$

so that $m_j = F(\|\boldsymbol{z}_j\|^2)$. By $\mathbb{E}[\sigma(\xi)] = 0$ we have $F(1) = 0$.

Under Assumption 3, F is C^2 on a neighborhood of 1 and we may differentiate under the expectation:

$$F'(s) = \frac{1}{2\sqrt{s}} \mathbb{E}\big[\sigma'(\sqrt{s}\xi)\,\xi\big].$$

Hence, using Stein's identity $\mathbb{E}[\xi f(\xi)] = \mathbb{E}[f'(\xi)]$ with $f = \sigma'$,

$$F'(1) = \frac{1}{2}\mathbb{E}[\sigma'(\xi)\xi] = \frac{1}{2}\mathbb{E}[\sigma''(\xi)] = \frac{c_\sigma}{2}.$$

Therefore, for s in a fixed neighborhood of 1, Taylor expansion of F at $s = 1$ gives us

$$F(s) = \frac{c_\sigma}{2}(s-1) + R(s), \qquad |R(s)| \le C(s-1)^2, \tag{3.3.104}$$

for some constant C depending only on $\sup_{t \in [1-\delta, 1+\delta]} |F''(t)|$.

Then let $\Delta_j := \|\boldsymbol{z}_j\|^2 - 1$. By (3.3.104),

$$m_j = \frac{c_\sigma}{2}\Delta_j + R_j, \qquad |R_j| \le C\Delta_j^2$$

on the event $\|\boldsymbol{z}_j\|^2 \in [1-\delta, 1+\delta]$, which holds w.h.p. since $\|\boldsymbol{z}_j\|^2 = (1/d)\chi^2_d$. Then

$$m_j^2 = \frac{c_\sigma^2}{4}\Delta_j^2 + \underbrace{\frac{c_\sigma}{1}\Delta_j R_j + R_j^2}_{=:e_j}, \qquad |e_j| \le C\big(|\Delta_j|^3 + \Delta_j^4\big).$$

Summing over j gives

$$\|\boldsymbol{m}\|^2 = \frac{c_\sigma^2}{4}\sum_{j=1}^{n}\Delta_j^2 + \sum_{j=1}^{n} e_j. \tag{3.3.105}$$

Since $\Delta_j = (\chi_d^2 - d)/d$ has moments $\mathbb{E}|\Delta_j|^3 = O(d^{-3/2})$ and $\mathbb{E}[\Delta_j^4] = O(d^{-2})$, and $n/d \to \psi$, we have

$$\mathbb{E}\Big[\sum_{j=1}^{n}|e_j|\Big] \le Cn\Big(\mathbb{E}|\Delta_j|^3 + \mathbb{E}[\Delta_j^4]\Big) = O\Big(\frac{n}{d^{3/2}} + \frac{n}{d^2}\Big) = o(1),$$

hence $\sum_{j=1}^{n} e_j = o_{\mathbb{P}}(1)$ by Markov inequality. Thus (3.3.105) reduces to

$$\|\boldsymbol{m}\|^2 = \frac{c_\sigma^2}{4}\sum_{j=1}^{n}\Delta_j^2 + o_{\mathbb{P}}(1). \tag{3.3.106}$$

Lastly, we give the concentration of $\sum\Delta_j^2$. Because $\boldsymbol{z}_j \sim \mathcal{N}(0, \boldsymbol{I}_d/d)$, we have $\Delta_j = (\chi_{d,j}^2 - d)/d$ with i.i.d. $\chi_{d,j}^2$. Then

$$\mathbb{E}[\Delta_j^2] = \mathrm{Var}(\|\boldsymbol{z}_j\|^2) = \frac{2}{d}, \qquad \mathrm{Var}(\Delta_j^2) = O(d^{-2}).$$

Therefore

$$\mathrm{Var}\Big(\sum_{j=1}^{n}\Delta_j^2\Big) = n\,\mathrm{Var}(\Delta_j^2) = O\Big(\frac{n}{d^2}\Big) = O\Big(\frac{1}{d}\Big) \to 0,$$

so the Chebyshev inequality implies

$$\sum_{j=1}^{n}\Delta_j^2 = \frac{2n}{d} + o_{\mathbb{P}}(1).$$

Plugging into (3.3.106) yields

$$\|\boldsymbol{m}\|^2 = \frac{c_\sigma^2}{4}\cdot\frac{2n}{d} + o_{\mathbb{P}}(1) \to \frac{c_\sigma^2}{2}\psi,$$

proving (3.3.102).

Let $\boldsymbol{v} \in \mathbb{R}^n$ be any deterministic unit vector (in particular $\boldsymbol{u}, \boldsymbol{v}_1, \boldsymbol{v}_2$). Since m_j are indepen-

dent and $\mathrm{Var}(m_j) = O(1/d)$ (from $m_j = (c_\sigma/2)\Delta_j + O(\Delta_j^2)$ and $\mathrm{Var}(\Delta_j) = 2/d$),

$$\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{m}) = \sum_{j=1}^{n} v_j^2 \mathrm{Var}(m_j) = O(1/d) \to 0.$$

Also $\mathbb{E}[m_j] = O(1/d)$ (because $\mathbb{E}[\Delta_j] = 0$ and the remainder is $O(\Delta_j^2)$), hence

$$|\mathbb{E}[\boldsymbol{v}^\top \boldsymbol{m}]| = \Big| \sum_{j=1}^{n} v_j \mathbb{E}[m_j] \Big| \le \|\boldsymbol{v}\|_1 \, O(1/d) \le \sqrt{n}\, O(1/d) = O(d^{-1/2}) \to 0.$$

Thus $\boldsymbol{v}^\top \boldsymbol{m} \to 0$ in L^2 and hence in probability. Taking $\boldsymbol{v} = \boldsymbol{u}, \boldsymbol{v}_1, \boldsymbol{v}_2$ yields (3.3.103). □

Lemma 49. *Let $\boldsymbol{g}_k = \boldsymbol{W}\boldsymbol{u}_k$ for fixed orthonormal $\boldsymbol{u}_1, \boldsymbol{u}_2$. Then*

$$\frac{\|\boldsymbol{g}_k\|^2}{N} - 1 \prec \frac{1}{\sqrt{N}}, \qquad \frac{\boldsymbol{g}_1^\top \boldsymbol{g}_2}{N} \prec \frac{1}{\sqrt{N}}.$$

Consequently, the two XOR additive spikes in (3.3.99) *have asymptotic strength*

$$\beta_{\text{linear}} := \theta_{\text{SNR}}^2 b_\sigma^2 \;\to\; \frac{r^2 \psi}{2} b_\sigma^2. \tag{3.3.107}$$

Proof. Each $\boldsymbol{g}_k \sim \mathcal{N}(0, \boldsymbol{I}_N)$ and $(\boldsymbol{g}_1, \boldsymbol{g}_2)$ is jointly Gaussian with $\mathrm{Cov}(\boldsymbol{g}_1, \boldsymbol{g}_2) = \boldsymbol{0}$. Therefore, we can directly apply Proposition 2.6.6 and Theorem 3.1.1 of [Ver18] to get conclusion. □

Resolvent Quadratic Forms for the Spike Analysis

Fix a compact set $\mathscr{D} \subset \mathbb{R}$ such that $\mathscr{D} \cap \mathrm{supp}(\mu) = \emptyset$. For $\lambda \in \mathscr{D}$, define $s(\lambda) \in \mathbb{R}$ by the unique relation

$$\lambda = z(s(\lambda)), \qquad z'(s(\lambda)) > 0, \tag{3.3.108}$$

which is well-defined by Lemma 31.

Define the left and right resolvents

$$\boldsymbol{Q}_L(\lambda) := (\boldsymbol{G}\boldsymbol{G}^\top - \lambda \boldsymbol{I}_N)^{-1}, \qquad \boldsymbol{Q}_R(\lambda) := (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1}.$$

Lemma 50 (Left resolvent limits for quadratic forms)**.** *Uniformly for* $\lambda \in \mathscr{D} \subset \mathbb{R} \setminus \{0\}$,

$$\frac{1}{N}\mathbf{1}_N^\top \boldsymbol{Q}_L(\lambda)\mathbf{1}_N = s(\lambda) + o_{\mathbb{P}}(1), \tag{3.3.109}$$

$$\frac{1}{N}\boldsymbol{g}_k^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{g}_\ell = \delta_{k\ell}\, s(\lambda) + o_{\mathbb{P}}(1), \tag{3.3.110}$$

$$\frac{1}{N}\mathbf{1}_N^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{g}_k = o_{\mathbb{P}}(1). \tag{3.3.111}$$

Proof. First, for (3.3.109), we can employ rowwise Lindeberg replacement for $\boldsymbol{G}$ to transform it into a Gaussian Wishart equivalent model and subsequently apply the classical anisotropic local law. Nevertheless, for the sake of simplicity, we can directly utilize Theorem 2.10 of [FMPW26] to derive the anisotropic local law of $\boldsymbol{G}\boldsymbol{G}^\top$:

$$\frac{1}{N}\mathbf{1}_N^\top \boldsymbol{Q}_L(\lambda)\mathbf{1}_N - s(\lambda) \prec \frac{1}{\sqrt{N}}$$

uniformly for $\lambda \in \mathscr{D} \subset \mathbb{R} \setminus \{0\}$ where $\mathscr{D} \subset \mathbb{R}$ is a fixed compact set such that $\mathscr{D} \cap (\operatorname{supp}(\mu) \cup \{0\}) = \emptyset$.

Second, we can replace $\boldsymbol{G}$ by an independent copy of the bulk. By Lemma 47, $\|\boldsymbol{G} - \widetilde{\boldsymbol{G}}\| \prec 1/\sqrt{n}$, hence $\|\boldsymbol{G}\boldsymbol{G}^\top - \widetilde{\boldsymbol{G}}\widetilde{\boldsymbol{G}}^\top\| \prec 1/\sqrt{n}$. Resolvent stability gives, uniformly on $\mathscr{D}$,

$$\|\boldsymbol{Q}_L(\lambda) - \widetilde{\boldsymbol{Q}}_L(\lambda)\| \prec 1/\sqrt{n}, \qquad \widetilde{\boldsymbol{Q}}_L(\lambda) := (\widetilde{\boldsymbol{G}}\widetilde{\boldsymbol{G}}^\top - \lambda \boldsymbol{I})^{-1}.$$

Thus it suffices to prove (3.3.110)–(3.3.111) with $\boldsymbol{Q}_L$ replaced by $\widetilde{\boldsymbol{Q}}_L$.

By Lemma 46, $\boldsymbol{g}_1, \boldsymbol{g}_2$ are independent standard Gaussian random vectors in $\mathbb{R}^N$ and independent of $\widetilde{\boldsymbol{G}}$, hence independent of $\widetilde{\boldsymbol{Q}}_L(\lambda)$. Conditional on $\widetilde{\boldsymbol{Q}}_L(\lambda)$, for $k \neq \ell$, $\boldsymbol{g}_k^\top \widetilde{\boldsymbol{Q}}_L(\lambda)\boldsymbol{g}_\ell$ is centered with variance $O(N)$, so $N^{-1}\boldsymbol{g}_k^\top \widetilde{\boldsymbol{Q}}_L(\lambda)\boldsymbol{g}_\ell = o_{\mathbb{P}}(1)$. For $k = \ell$, conditional expectation gives

$$\mathbb{E}\Big[\frac{1}{N}\boldsymbol{g}_k^\top \widetilde{\boldsymbol{Q}}_L(\lambda)\boldsymbol{g}_k \;\Big|\; \widetilde{\boldsymbol{Q}}_L(\lambda)\Big] = \frac{1}{N}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_L(\lambda),$$

and Hanson–Wright inequality yields concentration around this trace at $O_{\mathbb{P}}(N^{-1/2})$, uniformly

on $\mathscr{D}$ since $\|\widetilde{\boldsymbol{Q}}_L(\lambda)\| \le 1/\mathrm{dist}(\lambda, \mathrm{spec}(\widetilde{\boldsymbol{G}}\widetilde{\boldsymbol{G}}^\top)) \le C(\mathscr{D})$. Similarly, $N^{-1}\mathbf{1}^\top \widetilde{\boldsymbol{Q}}_L(\lambda)\boldsymbol{g}_k = o_{\mathbb{P}}(1)$.

□

Lemma 51 (Wishart resolvent quadratic form with a column-dependent vector)**.** *Let $\boldsymbol{Z} \in \mathbb{R}^{d\times n}$ have i.i.d. entries $Z_{\ell j} \sim \mathscr{N}(0, 1/d)$, and set*

$$\boldsymbol{S} := \boldsymbol{Z}^\top \boldsymbol{Z} \in \mathbb{R}^{n\times n}, \qquad \boldsymbol{G}(\zeta) := (\boldsymbol{S} - \zeta \boldsymbol{I}_n)^{-1}, \qquad \zeta \in \mathbb{C} \setminus \mathbb{R}.$$

Assume $n/d \to \psi \in (0,\infty)$ and define MP edges

$$a := (1-\sqrt{\psi})^2, \qquad b := (1+\sqrt{\psi})^2.$$

Fix $\kappa > 0$ and $\omega \in (0,1)$, and consider the outside-spectrum domain

$$\mathscr{S}_{\mathrm{out}} := \Big\{ \zeta = E + i\eta : \ \mathrm{dist}(E,[a,b]) \ge \kappa, \ \ n^{-1+\omega} \le \eta \le 1 \Big\}.$$

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be Lipschitz with $f(1) = 0$ and polynomial growth, and define the column-separable random vector

$$\boldsymbol{m} \in \mathbb{R}^n, \qquad m_j := f(\|\boldsymbol{z}_j\|^2), \quad j \in [n].$$

Let $m_{\mathrm{MP},\psi}(\zeta)$ be the Stieltjes transform of ρ_ψ^{MP}. Then uniformly for $\zeta \in \mathscr{S}_{\mathrm{out}}$,

$$\boldsymbol{m}^\top \boldsymbol{G}(\zeta)\boldsymbol{m} \ = \ m_{\mathrm{MP},\psi}(\zeta)\,\|\boldsymbol{m}\|^2 \ + \ O_\prec(d^{-1/2}). \tag{3.3.112}$$

In particular, for any fixed real E with $\mathrm{dist}(E,[a,b]) \ge \kappa$,

$$\boldsymbol{m}^\top (\boldsymbol{S} - E\boldsymbol{I})^{-1}\boldsymbol{m} = m_{\mathrm{MP},\psi}(E)\,\|\boldsymbol{m}\|^2 + o_{\mathbb{P}}(1).$$

Proof. Since $\|\boldsymbol{z}_j\|^2 = 1 + O_\prec(d^{-1/2})$ and f is Lipschitz with $f(1) = 0$,

$$m_j = O_\prec(d^{-1/2}) \qquad \text{uniformly in } j. \tag{3.3.113}$$

Hence $\|\boldsymbol{m}\|^2 = \sum_{j=1}^n m_j^2 = O_\prec(n/d) = O_\prec(1)$, and moreover

$$\|\boldsymbol{m}\| = O_\prec(1). \tag{3.3.114}$$

We first recall isotropic local law of [BEK+14] for $\boldsymbol{S}$. By the entrywise local law for sample covariance matrices and its minor-stability [BEK+14, KY17], we have

$$\max_{i,j\in[n]} \left|\boldsymbol{G}_{ij}(\zeta) - \delta_{ij} m_{\mathrm{MP},\psi}(\zeta)\right| \prec \Psi(\zeta) := n^{-1/2}, \tag{3.3.115}$$

and the same holds for all resolvent minors $\boldsymbol{G}^{(j)}(\zeta) := (\boldsymbol{S}^{(j)} - \zeta \boldsymbol{I})^{-1}$, uniformly in $j \in [n]$), where $\boldsymbol{S}^{(j)}$ denote the $(n-1)\times(n-1)$ minor of $\boldsymbol{S}$ obtained by removing j-th row and column.

We will also use the standard resolvent comparison identity between $\boldsymbol{G}$ and $\boldsymbol{G}^{(j)}$: for any $i,k \neq j$,

$$\boldsymbol{G}_{ik}(\zeta) = \boldsymbol{G}^{(j)}_{ik}(\zeta) + \frac{\boldsymbol{G}_{ij}(\zeta)\boldsymbol{G}_{jk}(\zeta)}{\boldsymbol{G}_{jj}(\zeta)}. \tag{3.3.116}$$

Since $\boldsymbol{G}_{ij} = O_\prec(n^{-1/2})$ for $i \neq j$ and $\boldsymbol{G}_{jj} = O_\prec(1)$ on $\mathscr{S}_{\text{out}}$,

$$\max_{i,k\neq j} \left|\boldsymbol{G}_{ik}(\zeta) - \boldsymbol{G}^{(j)}_{ik}(\zeta)\right| \prec \Psi(\zeta)^2 = n^{-1}. \tag{3.3.117}$$

Write the block form of $\boldsymbol{S} - \zeta \boldsymbol{I}$ w.r.t. index j:

$$\boldsymbol{S} - \zeta \boldsymbol{I} = \begin{pmatrix} q_j - \zeta & \boldsymbol{r}_j^\top \\ \boldsymbol{r}_j & \boldsymbol{S}^{(j)} - \zeta \boldsymbol{I} \end{pmatrix}, \qquad q_j := \|\boldsymbol{z}_j\|^2, \quad \boldsymbol{r}_j := (\boldsymbol{z}_k^\top \boldsymbol{z}_j)_{k\neq j} \in \mathbb{R}^{n-1}.$$

Denote vector $\boldsymbol{G}_{j,-j}(\zeta) = (\boldsymbol{G}_{jk}(\zeta))_{k\neq j} \in \mathbb{R}^{1\times(n-1)}$. The Schur complement identities give

$$\boldsymbol{G}_{jj}(\zeta) = \frac{1}{q_j - \zeta - \boldsymbol{r}_j^\top \boldsymbol{G}^{(j)}(\zeta)\boldsymbol{r}_j}, \tag{3.3.118}$$

$$\boldsymbol{G}_{j,-j}(\zeta) = -\boldsymbol{G}_{jj}(\zeta)\boldsymbol{r}_j^\top \boldsymbol{G}^{(j)}(\zeta). \tag{3.3.119}$$

Let $\boldsymbol{m}^{(j)}$ be $\boldsymbol{m}$ with the j-th component removed, and define

$$a_j(\zeta) := \boldsymbol{r}_j^\top \boldsymbol{G}^{(j)}(\zeta)\boldsymbol{m}^{(j)}. \tag{3.3.120}$$

Then $(\boldsymbol{G}\boldsymbol{m})_j = \boldsymbol{G}_{jj}(m_j - a_j(\zeta))$, so

$$\boldsymbol{m}^\top \boldsymbol{G}(\zeta)\boldsymbol{m} = \sum_{j=1}^n m_j^2 \boldsymbol{G}_{jj}(\zeta) - \sum_{j=1}^n m_j \boldsymbol{G}_{jj}(\zeta) a_j(\zeta) =: T_1(\zeta) - T_2(\zeta). \tag{3.3.121}$$

Now we control $T_1(\zeta)$ and $T_2(\zeta)$ separately.

For $T_1(\zeta)$ term, by (3.3.115) with $i = j$, $\boldsymbol{G}_{jj}(\zeta) = m_{\mathrm{MP},\psi}(\zeta) + O_\prec(n^{-1/2})$. Using $\|\boldsymbol{m}\|^2 = O_\prec(1)$,

$$T_1(\zeta) = m_{\mathrm{MP},\psi}(\zeta)\,\|\boldsymbol{m}\|^2 + O_\prec(n^{-1/2}). \tag{3.3.122}$$

Hence, it suffices to control the dependence term $T_2(\zeta)$ in the rest of proof. Write $X_j(\zeta) := m_j \boldsymbol{G}_{jj}(\zeta) a_j(\zeta)$, so $T_2(\zeta) = \sum_{j=1}^n X_j(\zeta)$. We follow the fluctuation average idea from [WWF24] to show that $T_2(\zeta)$ is vanishing.

Firstly, we compute the mean of this $T_2(\zeta)$. Let $\mathbb{E}_j[\cdot]$ be conditional expectation over the single column $\boldsymbol{z}_j$ (i.e. conditioning on $\boldsymbol{Z}^{(j)}$). Then: (i) $m_j = f(\|\boldsymbol{z}_j\|^2)$ is even in $\boldsymbol{z}_j$; (ii) $\boldsymbol{G}_{jj}(\zeta)$ is even in $\boldsymbol{z}_j$ by (3.3.118); (iii) $a_j(\zeta)$ is odd in $\boldsymbol{z}_j$ because $\boldsymbol{r}_j = \boldsymbol{Z}^{(j)\top}\boldsymbol{z}_j$ is linear in $\boldsymbol{z}_j$, while $\boldsymbol{G}^{(j)}$ and $\boldsymbol{m}^{(j)}$ are $\boldsymbol{Z}^{(j)}$-measurable, independent with $\boldsymbol{z}_j$. Hence $X_j(\zeta)$ is odd in $\boldsymbol{z}_j$, so

$$\mathbb{E}_j[X_j(\zeta)] = 0. \tag{3.3.123}$$

Next, we control the size of a_j and X_j. Conditional on $\boldsymbol{Z}^{(j)}$, we have $\boldsymbol{r}_j = \boldsymbol{Z}^{(j)\top}\boldsymbol{z}_j$, with $\boldsymbol{z}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$ independent of $\boldsymbol{Z}^{(j)}$. Thus $\boldsymbol{r}_j$ is centered Gaussian with $\mathrm{Cov}_j(\boldsymbol{r}_j) = (1/d)\,\boldsymbol{S}^{(j)}$. Therefore $a_j(\zeta) = \boldsymbol{m}^{(j)\top}\boldsymbol{G}^{(j)}\boldsymbol{r}_j$ is centered Gaussian with conditional variance

$$\mathrm{Var}_j(a_j(\zeta)) = \frac{1}{d}\,\boldsymbol{m}^{(j)\top}\boldsymbol{G}^{(j)}(\zeta)\,\boldsymbol{S}^{(j)}\,\boldsymbol{G}^{(j)}(\zeta)\,\boldsymbol{m}^{(j)} \;\le\; \frac{C}{d}\,\|\boldsymbol{m}\|^2,$$

because on $\mathscr{S}_{\mathrm{out}}$, $\|\boldsymbol{G}^{(j)}(\zeta)\| \le C(\kappa)$ and $\|\boldsymbol{S}^{(j)}\| = O_{\prec}(1)$. Using $\|\boldsymbol{m}\| = O_{\prec}(1)$, we get

$$a_j(\zeta) = O_{\prec}(d^{-1/2}). \tag{3.3.124}$$

Combining (3.3.113), $\boldsymbol{G}_{jj}(\zeta) = O_{\prec}(1)$, and (3.3.124) yields

$$X_j(\zeta) = O_{\prec}(d^{-1}). \tag{3.3.125}$$

Finally, we take high-moment fluctuation averaging for T_2. Fix an integer $p \ge 1$. We claim

$$\mathbb{E}\big|T_2(\zeta)\big|^{2p} \;\le\; C_p\Big(\frac{n}{d^2}\Big)^p \;\asymp\; C_p\,d^{-p}, \qquad \text{uniformly for } \zeta \in \mathscr{S}_{\mathrm{out}}. \tag{3.3.126}$$

This implies $T_2(\zeta) = O_{\prec}(\sqrt{n}/d) = O_{\prec}(d^{-1/2})$.

We can expand moments and classify by the number of distinct indices. Notice that

$$\mathbb{E}|T_2|^{2p} = \sum_{j_1,\dots,j_{2p}\in[n]} \mathbb{E}\big[X_{j_1}\cdots X_{j_{2p}}\big].$$

For any fixed multi-index $\boldsymbol{j} = (j_1,\dots,j_{2p})$, let $\mathscr{I}(\boldsymbol{j})$ be the set of distinct values in $\boldsymbol{j}$, and write $r := |\mathscr{I}(\boldsymbol{j})|$. For each $t \in \mathscr{I}(\boldsymbol{j})$, let ν_t be its multiplicity in the list $\boldsymbol{j}$, so $\sum_{t\in\mathscr{I}}\nu_t = 2p$. Let

$\mathscr{S}(\boldsymbol{j}) := \{t \in \mathscr{I}(\boldsymbol{j}) : \nu_t = 1\}$ be the set of singleton indices, and write $q := |\mathscr{S}(\boldsymbol{j})|$. Notice that

$$2p = \sum_{t \in \mathscr{I}} \nu_t \geq \sum_{t \in \mathscr{S}} 1 + \sum_{t \notin \mathscr{S}} 2 = q + 2(r-q) = 2r - q,$$

hence $q \geq 2r - 2p$ for any fixed multi-index $\boldsymbol{j}$.

Fix a monomial $X_{j_1} \cdots X_{j_{2p}}$ and pick a singleton index $t \in \mathscr{S}(\boldsymbol{j})$. Write the product as $X_t \cdot F_t$, where F_t is the product of the remaining $2p-1$ factors. Define $F_t^{(t)}$ by replacing in F_t every resolvent entry of $\boldsymbol{G}$ (and of any minors) by the corresponding entry of a resolvent/minor with index t removed whenever the indices allow. Formally, $F_t^{(t)}$ is constructed so that it is $\boldsymbol{Z}^{(t)}$-measurable, hence independent of $\boldsymbol{z}_t$, and such that

$$F_t = F_t^{(t)} + \Delta_t, \qquad |\Delta_t| \prec \Psi(\zeta)^2 \prod_{\alpha \neq t} |X_{j_\alpha}|. \tag{3.3.127}$$

The estimate (3.3.127) follows by repeated use of the resolvent identity (3.3.116) and the replacement bound (3.3.117): each replacement introduces at least one factor $\boldsymbol{G}_{it}\boldsymbol{G}_{tk}/\boldsymbol{G}_{tt}$, hence a factor $O_\prec(\Psi^2)$, and the total number of replacements is $O_p(1)$ since p is fixed. Hence, (3.3.127) decouples each singleton index and gains a factor $\Psi(\zeta)^2$ per singleton.

Now take $\mathbb{E}_t[\cdot]$. Since $F_t^{(t)}$ is $\boldsymbol{Z}^{(t)}$-measurable, by (3.3.123), we have

$$\mathbb{E}_t[X_t F_t^{(t)}] = F_t^{(t)}\, \mathbb{E}_t[X_t] = 0.$$

Hence

$$\mathbb{E}[X_t F_t] = \mathbb{E}[X_t \Delta_t]. \tag{3.3.128}$$

Using (3.3.125) and (3.3.127), we obtain

$$\big|\mathbb{E}[X_t \Delta_t]\big| \prec \Psi(\zeta)^2\, \mathbb{E}\Big[|X_t| \prod_{\alpha \neq t} |X_{j_\alpha}|\Big] \prec \Psi(\zeta)^2 d^{-2p}.$$

Then, we can repeat this decoupling *for each singleton index* in $\mathscr{S}(\boldsymbol{j})$, and each repetition contributes an additional factor $\Psi(\zeta)^2$. Therefore, for the monomial with q singleton indices,

$$\left|\mathbb{E}[X_{j_1}\cdots X_{j_{2p}}]\right| \;\prec\; d^{-2p}\,\Psi(\zeta)^{2q}. \tag{3.3.129}$$

For fixed $r = |\mathscr{I}|$, the number of multi-indices with r distinct values is $O(n^r)$. Using (3.3.129) and $q \geq 2r-2p$, we get

$$\sum_{\boldsymbol{j}:\,|\mathscr{I}(\boldsymbol{j})|=r} \left|\mathbb{E}[X_{j_1}\cdots X_{j_{2p}}]\right| \;\prec\; n^r\, d^{-2p}\,\Psi(\zeta)^{2(2r-2p)}.$$

On $\mathscr{S}_{\text{out}}$, $\Psi(\zeta) = n^{-1/2}$, so $n^r d^{-2p}\Psi^{4r-4p} = n^{2p-r}d^{-2p} \;\asymp\; d^{-r}$. Summing over $r = 1,\ldots,2p$ gives

$$\mathbb{E}|T_2(\zeta)|^{2p} \;\prec\; \sum_{r=1}^{2p} d^{-r} \;\prec\; d^{-1}.$$

This is already $o(1)$, but we need the sharper d^{-p} scaling. For that, we treat separately the case $r \leq p$, where there are no (or few) singletons and pairing dominates:

- If $r \leq p$, then trivially $q \geq 0$ and we use the crude size bound $|X_j| \prec d^{-1}$: there are $O(n^r) \leq O(n^p)$ such monomials, each of size $O_\prec(d^{-2p})$, so their total contribution is $O_\prec(\, n^p d^{-2p}) = O_\prec(d^{-p})$.
- If $r \geq p+1$, then $q \geq 2r-2p \geq 2$, and the singleton-decoupling bound above yields a total contribution $O_\prec(\sum_{r\geq p+1} d^{-r}) \prec d^{-(p+1)}$, which is negligible compared to d^{-p}.

Combining both regimes yields (3.3.126), proving $T_2(\zeta) = O_\prec(d^{-1/2})$. By (3.3.121), (3.3.122), and $T_2(\zeta) = O_\prec(d^{-1/2})$, we can conclude that

$$\boldsymbol{m}^\top \boldsymbol{G}(\zeta)\,\boldsymbol{m} = m_{\text{MP},\psi}(\zeta)\,\|\boldsymbol{m}\|^2 + O_\prec(n^{-1/2}) + O_\prec(d^{-1/2}).$$

Since $n \asymp d$, this is exactly (3.3.112). □

Lemma 52 (Right resolvent limits for quadratic forms)**.** *Uniformly for* $\lambda \in \mathscr{D}$*,*

$$\boldsymbol{v}_k^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{v}_\ell = \delta_{k\ell} m_\mu(\lambda) + o_{\mathbb{P}}(1), \tag{3.3.130}$$

$$\boldsymbol{m}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{m} = \|\boldsymbol{m}\|^2 m_\mu(\lambda) + o_{\mathbb{P}}(1), \tag{3.3.131}$$

$$\boldsymbol{v}_k^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{m} = o_{\mathbb{P}}(1). \tag{3.3.132}$$

Proof. Conditional on $\boldsymbol{Z}$, $\boldsymbol{K}_0 = \boldsymbol{G}^\top \boldsymbol{G}$ is a general sample covariance matrix with population covariance $\boldsymbol{\Sigma}$ and aspect ratio $\phi = n/N$. Similar as the proof of Lemma 43, we can verify that Assumption 5 of [WWF24] is satisfied for $\boldsymbol{G}$. Hence, we can apply Theorem 10 of [WWF24] to obtain a deterministic equivalence of quadratic form of the resolvent $\boldsymbol{Q}_R(\lambda)$. Precisely for any deterministic unit vectors $\boldsymbol{b}, \boldsymbol{b}'$ (possibly $\boldsymbol{Z}$-measurable),

$$\boldsymbol{b}^\top (\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} \boldsymbol{b}' = \boldsymbol{b}^\top \Big(-\lambda \tilde{m}_\mu(\lambda)\boldsymbol{\Sigma} - \lambda \boldsymbol{I} \Big)^{-1} \boldsymbol{b}' + O_\prec(N^{-1/2}), \tag{3.3.133}$$

uniformly for $\lambda \in \mathscr{D}$.

By Corollary 39, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \tau \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{R}$ with $\|\boldsymbol{R}\| = o_{\mathbb{P}}(1)$. Let $\boldsymbol{A} := -\lambda s(\lambda)\boldsymbol{\Sigma}_0 - \lambda \boldsymbol{I}$. Then

$$-\lambda s(\lambda)\boldsymbol{\Sigma} - \lambda \boldsymbol{I} = \boldsymbol{A} - \lambda s(\lambda)\tau \boldsymbol{u}\boldsymbol{u}^\top + o_{\mathbb{P}}(1).$$

By Woodbury identity, we have

$$(\boldsymbol{A} - \lambda s \tau \boldsymbol{u}\boldsymbol{u}^\top)^{-1} = \boldsymbol{A}^{-1} + \frac{\lambda s \tau}{1 - \lambda s \tau \boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}} \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{A}^{-1}.$$

If $\boldsymbol{b}^\top \boldsymbol{u} = 0$ and $\boldsymbol{b}'^\top \boldsymbol{u} = 0$, then $\boldsymbol{b}^\top (\boldsymbol{A}^{-1}\boldsymbol{u}) = O(1) \cdot (\boldsymbol{b}^\top \boldsymbol{u}) = 0$, hence the rank-one correction vanishes:

$$\boldsymbol{b}^\top (\boldsymbol{A} - \lambda s \tau \boldsymbol{u}\boldsymbol{u}^\top)^{-1} \boldsymbol{b}' = \boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b}'.$$

Applying this with $\boldsymbol{b} = \boldsymbol{v}_k$, $\boldsymbol{b}' = \boldsymbol{v}_\ell$ (note $\boldsymbol{v}_k^\top \boldsymbol{u} = 0$ by (3.3.80)) yields that the τ-spike does not

affect (3.3.130). Also, Lemma 48 implies that $\boldsymbol{b}^\top(\boldsymbol{A}-\lambda s\tau\boldsymbol{u}\boldsymbol{u}^\top)^{-1}\boldsymbol{b}' = \boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{b}' + o_{\mathbb{P}}(1)$ for $\boldsymbol{b}=\boldsymbol{b}'=\boldsymbol{m}/\|\boldsymbol{m}\|$.

For $\boldsymbol{v}_k^\top\boldsymbol{A}^{-1}\boldsymbol{v}_\ell$, since $\boldsymbol{\Sigma}_0 = (1-b_\sigma^2)\boldsymbol{I}+b_\sigma^2\boldsymbol{S}$ is orthogonally invariant (shifted Wishart), the anisotropic local law from [KY17] implies $\boldsymbol{v}_k^\top\boldsymbol{A}^{-1}\boldsymbol{v}_\ell = \delta_{k\ell}m_\mu(\lambda)+o_{\mathbb{P}}(1)$, giving (3.3.130). Here, we have removed the rank-one $\tau\boldsymbol{u}\boldsymbol{u}^\top$ term in $\boldsymbol{\Sigma}$ when applying (3.3.133).

For the quadratic from with respect to $\boldsymbol{m}$, we apply (3.3.133) with $\boldsymbol{b}=\boldsymbol{b}'=\boldsymbol{m}/\|\boldsymbol{m}\|$ and then show that $\boldsymbol{m}^\top\boldsymbol{A}^{-1}\boldsymbol{m} = \|\boldsymbol{m}\|^2 m_\mu(\lambda)+o_{\mathbb{P}}(1)$ even though $\boldsymbol{m}$ depends on $\boldsymbol{Z}$. Write

$$\boldsymbol{A}^{-1} = \Big(-\lambda s(\lambda)\big((1-b_\sigma^2)\boldsymbol{I}+b_\sigma^2\boldsymbol{S}\big)-\lambda\boldsymbol{I}\Big)^{-1} = \alpha(\lambda)\,(\boldsymbol{S}-\zeta(\lambda)\boldsymbol{I})^{-1},$$

for explicit scalars $\alpha(\lambda)$ and $\zeta(\lambda)$ depending on λ (and bounded away from $\operatorname{supp}(\rho_\psi^{\mathrm{MP}})$ for $\lambda\in\mathscr{D}$). From Lemma 51, we can derive that

$$\boldsymbol{m}^\top(\boldsymbol{S}-\zeta\boldsymbol{I})^{-1}\boldsymbol{m} = m_{\mathrm{MP},\psi}(\zeta)\,\|\boldsymbol{m}\|^2+o_{\mathbb{P}}(1), \tag{3.3.134}$$

uniformly for $\zeta=\zeta(\lambda)$ in the relevant domain.

For the cross term $\boldsymbol{v}_k^\top\boldsymbol{Q}_R(\lambda)\boldsymbol{m}$, we use (3.3.133) with $\boldsymbol{b}=\boldsymbol{v}_k$, $\boldsymbol{b}'=\boldsymbol{m}$. The deterministic equivalent reduces to $\boldsymbol{v}_k^\top\boldsymbol{A}^{-1}\boldsymbol{m}$ since $\boldsymbol{v}_k\perp\boldsymbol{u}$. Because $\boldsymbol{v}_k^\top\boldsymbol{m}=o_{\mathbb{P}}(1)$ (Lemma 48) and $\|\boldsymbol{A}^{-1}\|=O(1)$ on $\mathscr{D}$, we get $\boldsymbol{v}_k^\top\boldsymbol{A}^{-1}\boldsymbol{m}=o_{\mathbb{P}}(1)$, yielding (3.3.132). □

Analyses of Additive Spikes and Alignments

Define the rank-three spike matrix

$$\boldsymbol{P} := \frac{1}{\sqrt{N}}\mathbf{1}_N\boldsymbol{m}^\top + \frac{\theta_{\mathrm{SNR}}b_\sigma}{\sqrt{N}}\big(\boldsymbol{g}_1\boldsymbol{v}_1^\top+\boldsymbol{g}_2\boldsymbol{v}_2^\top\big) = \sum_{k=0}^{2}\sqrt{\beta_k}\,\boldsymbol{a}_k\boldsymbol{b}_k^\top, \tag{3.3.135}$$

where

$$\boldsymbol{a}_0 := \frac{\mathbf{1}_N}{\sqrt{N}},\quad \boldsymbol{b}_0 := \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|},\quad \boldsymbol{a}_1 := \frac{\boldsymbol{g}_1}{\|\boldsymbol{g}_1\|},\quad \boldsymbol{b}_1 := \boldsymbol{v}_1,\quad \boldsymbol{a}_2 := \frac{\boldsymbol{g}_2}{\|\boldsymbol{g}_2\|},\quad \boldsymbol{b}_2 := \boldsymbol{v}_2,$$

and (by Lemmas 48 and 49), asymptotically,

$$\beta_0 = \|\boldsymbol{m}\|^2 \to \beta_{\text{mean}} = \frac{c_\sigma^2}{2}\psi, \qquad \beta_1 = \beta_2 = \theta_{\text{SNR}}^2 b_\sigma^2 \to \beta_{\text{linear}} = \frac{r^2\psi}{2} b_\sigma^2. \tag{3.3.136}$$

Then, we recall $\boldsymbol{Y}^\sharp = \boldsymbol{G} + \boldsymbol{P}$ and $\boldsymbol{K}^\sharp = \boldsymbol{Y}^{\sharp\top}\boldsymbol{Y}^\sharp$ defined by Proposition 45.

Lemma 53. *Let $\boldsymbol{G} \in \mathbb{R}^{N\times n}$ and $\boldsymbol{K}_0 := \boldsymbol{G}^\top\boldsymbol{G}$. Assume $\boldsymbol{K}_0$ has an* isolated *outlier eigenvalue $\widehat{\lambda}_\tau^{(0)}$ with unit-norm eigenvector $\widehat{\boldsymbol{u}}_\tau \in \mathbb{R}^n$ so that for any $\varepsilon > 0$,*

$$\boldsymbol{K}_0\widehat{\boldsymbol{u}}_\tau = \widehat{\lambda}_\tau^{(0)}\widehat{\boldsymbol{u}}_\tau, \qquad \widehat{\lambda}_\tau^{(0)} \notin \operatorname{supp}(\mu) + (-\varepsilon, \varepsilon)$$

for all large n. Define the associated unit left singular vector of $\boldsymbol{G}$ as

$$\widehat{\boldsymbol{w}}_\tau := \frac{\boldsymbol{G}\widehat{\boldsymbol{u}}_\tau}{\sqrt{\widehat{\lambda}_\tau^{(0)}}} \in \mathbb{R}^N, \qquad \|\widehat{\boldsymbol{w}}_\tau\| = 1.$$

Consider the rank–3 additive spike matrix $\boldsymbol{P} = \sum_{k=0}^2 \sqrt{\beta_k}\boldsymbol{a}_k\boldsymbol{b}_k^\top$ with $\|\boldsymbol{a}_k\| = \|\boldsymbol{b}_k\| = 1$ and $\beta_k = O(1)$, $\boldsymbol{Y}^\sharp = \boldsymbol{G} + \boldsymbol{P}$, and $\boldsymbol{K}^\sharp = \boldsymbol{Y}^{\sharp\top}\boldsymbol{Y}^\sharp$. Assume the decoupling conditions*:*

$$\max_{k\in\{0,1,2\}} |\boldsymbol{b}_k^\top\widehat{\boldsymbol{u}}_\tau| = o_{\mathbb{P}}(1), \qquad \max_{k\in\{0,1,2\}} |\boldsymbol{a}_k^\top\widehat{\boldsymbol{w}}_\tau| = o_{\mathbb{P}}(1). \tag{3.3.137}$$

Then:

(i) (Eigenvalue stability.) *There exists an eigenvalue $\widehat{\lambda}_\tau^\sharp \in \operatorname{spec}(\boldsymbol{K}^\sharp)$ such that*

$$\left|\widehat{\lambda}_\tau^\sharp - \widehat{\lambda}_\tau^{(0)}\right| = o_{\mathbb{P}}(1).$$

Hence, $\widehat{\lambda}_\tau^\sharp$ is an outlier eigenvalue of $\boldsymbol{K}^\sharp$.

(ii) (Eigenvector-alignment stability.) *Fix any constant $\delta > 0$. Let $\boldsymbol{\Pi}_\delta$ be the spectral projector*

of $\boldsymbol{K}^\sharp$ *onto* $I_\delta := (\widehat{\lambda}_\tau^{(0)} - \delta, \widehat{\lambda}_\tau^{(0)} + \delta)$. *Then*

$$\|(\boldsymbol{I} - \boldsymbol{\Pi}_\delta)\widehat{\boldsymbol{u}}_\tau\| = o_{\mathbb{P}}(1).$$

In particular, the normalized vector

$$\widehat{\boldsymbol{u}}_\tau^\sharp := \frac{\boldsymbol{\Pi}_\delta \widehat{\boldsymbol{u}}_\tau}{\|\boldsymbol{\Pi}_\delta \widehat{\boldsymbol{u}}_\tau\|} \in \mathrm{Range}(\boldsymbol{\Pi}_\delta)$$

satisfies $|\langle \widehat{\boldsymbol{u}}_\tau^\sharp, \widehat{\boldsymbol{u}}_\tau \rangle|^2 = 1 - o_{\mathbb{P}}(1)$. *Thus* $\widehat{\boldsymbol{u}}_\tau^\sharp$ *inherits the same eigenvector alignment limit as* $\widehat{\boldsymbol{u}}_\tau$ *(e.g. alignment with* $\boldsymbol{u} = \mathbf{1}_n/\sqrt{n}$ *in Lemma 43).*

(iii) ***(Projected resolvents remove the singularity at*** $\widehat{\lambda}_\tau^{(0)}$***.)*** *Define the orthogonal projectors*

$$\boldsymbol{P}_R^\perp := \boldsymbol{I}_n - \widehat{\boldsymbol{u}}_\tau \widehat{\boldsymbol{u}}_\tau^\top, \qquad \boldsymbol{P}_L^\perp := \boldsymbol{I}_N - \widehat{\boldsymbol{w}}_\tau \widehat{\boldsymbol{w}}_\tau^\top,$$

and the spike-removed resolvents

$$\begin{aligned} \boldsymbol{Q}_R^\perp(\lambda) &:= \boldsymbol{P}_R^\perp (\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} \boldsymbol{P}_R^\perp, \\ \boldsymbol{Q}_L^\perp(\lambda) &:= \boldsymbol{P}_L^\perp (\boldsymbol{G}\boldsymbol{G}^\top - \lambda \boldsymbol{I})^{-1} \boldsymbol{P}_L^\perp, \end{aligned}$$

which admit a continuous extension to $\lambda = \widehat{\lambda}_\tau^{(0)}$. *For each fixed* $k, \ell \in \{0,1,2\}$,

$$\boldsymbol{b}_k^\top \boldsymbol{Q}_R^\perp(\widehat{\lambda}_\tau^{(0)})\, \boldsymbol{b}_\ell \quad \textit{and} \quad \boldsymbol{a}_k^\top \boldsymbol{Q}_L^\perp(\widehat{\lambda}_\tau^{(0)})\, \boldsymbol{a}_\ell$$

are well-defined and remain $O_{\mathbb{P}}(1)$. *Hence, applying* (3.3.137)*, we have*

$$\boldsymbol{b}_k^\top \boldsymbol{Q}_R(\widehat{\lambda}_\tau^{(0)})\, \boldsymbol{b}_\ell = \boldsymbol{b}_k^\top \boldsymbol{Q}_R^\perp(\widehat{\lambda}_\tau^{(0)})\, \boldsymbol{b}_\ell + o_{\mathbb{P}}(1) \tag{3.3.138}$$

and

$$\boldsymbol{a}_k^\top \boldsymbol{Q}_L(\widehat{\lambda}_\tau^{(0)})\,\boldsymbol{a}_\ell = \boldsymbol{a}_k^\top \boldsymbol{Q}_L^\perp(\widehat{\lambda}_\tau^{(0)})\,\boldsymbol{a}_\ell + o_{\mathbb{P}}(1). \tag{3.3.139}$$

Proof. We can expand $\boldsymbol{K}^\sharp$ as a low-rank perturbation of $\boldsymbol{K}_0$. By definition,

$$\boldsymbol{K}^\sharp = (\boldsymbol{G}+\boldsymbol{P})^\top(\boldsymbol{G}+\boldsymbol{P}) = \boldsymbol{K}_0 + \underbrace{\boldsymbol{G}^\top\boldsymbol{P}+\boldsymbol{P}^\top\boldsymbol{G}+\boldsymbol{P}^\top\boldsymbol{P}}_{=:\,\boldsymbol{\Delta}}.$$

Thus $\boldsymbol{\Delta}$ has rank at most 9 and $\|\boldsymbol{\Delta}\| = O_{\mathbb{P}}(1)$ (since $\|\boldsymbol{G}\| = O_{\mathbb{P}}(1)$ and $\|\boldsymbol{P}\| = O_{\mathbb{P}}(1)$).

Part (i). Let $\boldsymbol{A}$ be symmetric, $\lambda \in \mathbb{R}$, and $\boldsymbol{v}$ a unit vector. Then

$$\operatorname{dist}\big(\lambda, \operatorname{spec}(\boldsymbol{A})\big) \;\le\; \|(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{v}\|. \tag{3.3.140}$$

Indeed, considering eigen-decomposition $\boldsymbol{A} = \sum_i \lambda_i \boldsymbol{q}_i\boldsymbol{q}_i^\top$ and $\boldsymbol{v} = \sum_i c_i\boldsymbol{q}_i$, we have

$$\|(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{v}\|^2 = \sum_i (\lambda_i-\lambda)^2|c_i|^2 \ge \min_i(\lambda_i-\lambda)^2\sum_i|c_i|^2 = \operatorname{dist}(\lambda,\operatorname{spec}(\boldsymbol{A}))^2$$

Since $\boldsymbol{K}_0\widehat{\boldsymbol{u}}_\tau = \widehat{\lambda}_\tau^{(0)}\widehat{\boldsymbol{u}}_\tau$, we have

$$(\boldsymbol{K}^\sharp - \widehat{\lambda}_\tau^{(0)}\boldsymbol{I})\widehat{\boldsymbol{u}}_\tau = (\boldsymbol{K}_0 - \widehat{\lambda}_\tau^{(0)}\boldsymbol{I})\widehat{\boldsymbol{u}}_\tau + \boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau = \boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau.$$

Therefore,

$$\operatorname{dist}\big(\widehat{\lambda}_\tau^{(0)}, \operatorname{spec}(\boldsymbol{K}^\sharp)\big) \;\le\; \|\boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau\|.$$

It remains to show $\|\boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau\| = o_{\mathbb{P}}(1)$.

Write $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{\Theta}\boldsymbol{B}^\top$ with $\boldsymbol{A} = [\boldsymbol{a}_0,\boldsymbol{a}_1,\boldsymbol{a}_2]$, $\boldsymbol{B} = [\boldsymbol{b}_0,\boldsymbol{b}_1,\boldsymbol{b}_2]$ and $\boldsymbol{\Theta} = \operatorname{diag}(\sqrt{\beta_0},\sqrt{\beta_1},\sqrt{\beta_2})$.

Then

$$\begin{aligned}
\boldsymbol{G}^\top \boldsymbol{P}\widehat{\boldsymbol{u}}_\tau &= \boldsymbol{G}^\top \boldsymbol{A}\boldsymbol{\Theta}\,(\boldsymbol{B}^\top \widehat{\boldsymbol{u}}_\tau),\\
\boldsymbol{P}^\top \boldsymbol{P}\widehat{\boldsymbol{u}}_\tau &= \boldsymbol{B}\boldsymbol{\Theta}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\Theta}\,(\boldsymbol{B}^\top \widehat{\boldsymbol{u}}_\tau),\\
\boldsymbol{P}^\top \boldsymbol{G}\widehat{\boldsymbol{u}}_\tau &= \boldsymbol{B}\boldsymbol{\Theta}\boldsymbol{A}^\top(\boldsymbol{G}\widehat{\boldsymbol{u}}_\tau) = \sqrt{\widehat{\lambda}_\tau^{(0)}}\,\boldsymbol{B}\boldsymbol{\Theta}\,(\boldsymbol{A}^\top \widehat{\boldsymbol{w}}_\tau),
\end{aligned}$$

where we used $\boldsymbol{G}\widehat{\boldsymbol{u}}_\tau = \sqrt{\widehat{\lambda}_\tau^{(0)}}\,\widehat{\boldsymbol{w}}_\tau$. Hence,

$$\|\boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau\| \;\le\; \|\boldsymbol{G}^\top \boldsymbol{A}\boldsymbol{\Theta}\| \cdot \|\boldsymbol{B}^\top \widehat{\boldsymbol{u}}_\tau\| + \sqrt{\widehat{\lambda}_\tau^{(0)}}\,\|\boldsymbol{B}\boldsymbol{\Theta}\| \cdot \|\boldsymbol{A}^\top \widehat{\boldsymbol{w}}_\tau\| + \|\boldsymbol{B}\boldsymbol{\Theta}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\Theta}\| \cdot \|\boldsymbol{B}^\top \widehat{\boldsymbol{u}}_\tau\|.$$

Now $\|\boldsymbol{G}^\top \boldsymbol{A}\boldsymbol{\Theta}\| \le \|\boldsymbol{G}\|\,\|\boldsymbol{A}\|\,\|\boldsymbol{\Theta}\| = O_{\mathbb{P}}(1)$, $\|\boldsymbol{B}\boldsymbol{\Theta}\| = O(1)$, and $\|\boldsymbol{B}\boldsymbol{\Theta}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{\Theta}\| = O(1)$. By the decoupling assumptions (3.3.137),

$$\|\boldsymbol{B}^\top \widehat{\boldsymbol{u}}_\tau\| = o_{\mathbb{P}}(1), \qquad \|\boldsymbol{A}^\top \widehat{\boldsymbol{w}}_\tau\| = o_{\mathbb{P}}(1).$$

Therefore $\|\boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau\| = o_{\mathbb{P}}(1)$, and combining with (3.3.140) yields

$$\operatorname{dist}\big(\widehat{\lambda}_\tau^{(0)}, \operatorname{spec}(\boldsymbol{K}^\sharp)\big) = o_{\mathbb{P}}(1),$$

which proves **(i)**.

Part (ii). Let $\boldsymbol{K}^\sharp = \sum_i \widehat{\lambda}_i^\sharp \widehat{\boldsymbol{v}}_i \widehat{\boldsymbol{v}}_i^\top$ be an eigen-decomposition, and write $\widehat{\boldsymbol{u}}_\tau = \sum_i c_i \widehat{\boldsymbol{v}}_i$. Then

$$\|(\boldsymbol{K}^\sharp - \widehat{\lambda}_\tau^{(0)} \boldsymbol{I})\widehat{\boldsymbol{u}}_\tau\|^2 = \sum_i (\widehat{\lambda}_i^\sharp - \widehat{\lambda}_\tau^{(0)})^2 |c_i|^2 = \|\boldsymbol{\Delta}\widehat{\boldsymbol{u}}_\tau\|^2 = o_{\mathbb{P}}(1).$$

Split the sum into indices with $\widehat{\lambda}_i^\sharp \in I_\delta$ and those with $\widehat{\lambda}_i^\sharp \notin I_\delta$. For $\widehat{\lambda}_i^\sharp \notin I_\delta$ we have $|\widehat{\lambda}_i^\sharp - \widehat{\lambda}_\tau^{(0)}| \ge \delta$, hence

$$\delta^2 \sum_{\widehat{\lambda}_i^\sharp \notin I_\delta} |c_i|^2 \;\le\; \sum_{\widehat{\lambda}_i^\sharp \notin I_\delta} (\widehat{\lambda}_i^\sharp - \widehat{\lambda}_\tau^{(0)})^2 |c_i|^2 \;\le\; \|(\boldsymbol{K}^\sharp - \widehat{\lambda}_\tau^{(0)} \boldsymbol{I})\widehat{\boldsymbol{u}}_\tau\|^2 = o_{\mathbb{P}}(1).$$

Therefore $\sum_{\widehat{\lambda}_i^\sharp \notin I_\delta} |c_i|^2 = o_{\mathbb{P}}(1)$, i.e. $\|(\boldsymbol{I} - \boldsymbol{\Pi}_\delta)\widehat{\boldsymbol{u}}_\tau\| = o_{\mathbb{P}}(1)$, proving **(ii)**. This argument makes *no* assumption about how many eigenvalues lie inside I_δ, so it remains valid under collisions when there are more than one eigenvalues of $\boldsymbol{K}^\sharp$ in I_δ.

Part (iii). Let $\{\widehat{\lambda}_i^{(0)}, \widehat{\boldsymbol{u}}_i^{(0)}\}$ be an eigen-decomposition of $\boldsymbol{K}_0$ with $\widehat{\lambda}_\tau^{(0)}$ corresponding to $\widehat{\boldsymbol{u}}_\tau$. Then for $\lambda \neq \widehat{\lambda}_\tau^{(0)}$,

$$(\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} = \frac{1}{\widehat{\lambda}_\tau^{(0)} - \lambda} \widehat{\boldsymbol{u}}_\tau \widehat{\boldsymbol{u}}_\tau^\top + \sum_{i \neq \tau} \frac{1}{\widehat{\lambda}_i^{(0)} - \lambda} \widehat{\boldsymbol{u}}_i^{(0)} \widehat{\boldsymbol{u}}_i^{(0)\top}.$$

Multiplying by $\boldsymbol{P}_R^\perp$ on both sides removes the pole term, giving

$$\boldsymbol{Q}_R^\perp(\lambda) = \sum_{i \neq \tau} \frac{1}{\widehat{\lambda}_i^{(0)} - \lambda} \widehat{\boldsymbol{u}}_i^{(0)} \widehat{\boldsymbol{u}}_i^{(0)\top},$$

which is continuous at $\lambda = \widehat{\lambda}_\tau^{(0)}$ (since the sum is over $i \neq \tau$). The same argument applies to $\boldsymbol{Q}_L^\perp(\lambda)$ using the eigen-decomposition of $\boldsymbol{G}\boldsymbol{G}^\top$. Hence the quadratic forms in **(iii)** are well-defined and $O_{\mathbb{P}}(1)$. This is exactly what is needed to write the additive outlier determinant equation in a way that still makes sense at the collision point $\lambda = \widehat{\lambda}_\tau^{(0)}$. □

Lemma 54 (Verification of the decoupling conditions (3.3.137) for $\boldsymbol{K}_0$)**.** *Under the assumptions of Lemma 43, additionally suppose the separated regime, that is, $\tau > \tau_{\text{crit}}$ and $z'(-1/\Lambda_\tau) > 0$. Under these conditions, the matrix $\boldsymbol{K}_0 = \boldsymbol{G}^\top \boldsymbol{G}$ possesses a unique isolated outlier eigenvalue, denoted by λ_τ, associated with a unit-norm eigenvector $\boldsymbol{u}_\tau$. Define the associated left singular vector $\boldsymbol{w}_\tau := \boldsymbol{G}\boldsymbol{u}_\tau / \sqrt{\lambda_\tau}$. Recall the spike directions*

$$\boldsymbol{b}_0 = \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}, \quad \boldsymbol{b}_1 = \boldsymbol{v}_1, \quad \boldsymbol{b}_2 = \boldsymbol{v}_2, \qquad \boldsymbol{a}_0 = \frac{\mathbf{1}_N}{\sqrt{N}}, \quad \boldsymbol{a}_1 = \frac{\boldsymbol{g}_1}{\|\boldsymbol{g}_1\|}, \quad \boldsymbol{a}_2 = \frac{\boldsymbol{g}_2}{\|\boldsymbol{g}_2\|}.$$

Then the decoupling conditions (3.3.137) *hold:*

$$\max_{k \in \{0,1,2\}} |\boldsymbol{b}_k^\top \boldsymbol{u}_\tau| = o_{\mathbb{P}}(1), \qquad \max_{k \in \{0,1,2\}} |\boldsymbol{a}_k^\top \boldsymbol{w}_\tau| = o_{\mathbb{P}}(1).$$

Proof. We prove right and left decoupling conditions separately.

Part I: right decoupling condition $\max_k |\boldsymbol{b}_k^\top \boldsymbol{u}_\tau| = o_{\mathbb{P}}(1)$**.**

By the XOR construction, $\boldsymbol{v}_1^\top \boldsymbol{u} = \boldsymbol{v}_2^\top \boldsymbol{u} = 0$ where $\boldsymbol{u} = \frac{1}{\sqrt{n}}\mathbf{1}_n$. Applying (3.3.98) in Lemma 43 with $\boldsymbol{v} = \boldsymbol{v}_k$ gives

$$\boldsymbol{v}_k^\top \boldsymbol{u}_\tau = o_{\mathbb{P}}(1), \qquad k \in \{1,2\}.$$

Moreover, Lemma 48 gives $\boldsymbol{u}^\top \boldsymbol{m} = o_{\mathbb{P}}(1)$ and $\|\boldsymbol{m}\| \to \sqrt{\beta_{\text{mean}}} > 0$. Therefore, we can obtain

$$\boldsymbol{b}_0^\top \boldsymbol{u} = \frac{\boldsymbol{m}^\top \boldsymbol{u}}{\|\boldsymbol{m}\|} = o_{\mathbb{P}}(1).$$

Applying Lemma 43 again with $\boldsymbol{v} = \boldsymbol{b}_0$ in (3.3.98) yields

$$\boldsymbol{b}_0^\top \boldsymbol{u}_\tau = o_{\mathbb{P}}(1).$$

Taking the maximum over $k \in \{0,1,2\}$ completes Part I.

Part II: left decoupling condition $\max_k |\boldsymbol{a}_k^\top \boldsymbol{w}_\tau| = o_{\mathbb{P}}(1)$**.**

Let $\boldsymbol{G}^e$ be $\widetilde{\boldsymbol{G}}$ defined in Lemma 47, so Lemma 47 indicates that $\|\boldsymbol{G} - \boldsymbol{G}^e\| \prec N^{-1/2}$ and $\boldsymbol{G}^e$ is independent of $(\boldsymbol{g}_1, \boldsymbol{g}_2)$. Define $\boldsymbol{K}_0^e := (\boldsymbol{G}^e)^\top \boldsymbol{G}^e$ and let $(\lambda_\tau^e, \boldsymbol{u}_\tau^e)$ denote the isolated eigenpair of $\boldsymbol{K}_0^e$ corresponding to the outlier of $\boldsymbol{K}_0$. In fact, Lemma 43 holds for $\boldsymbol{K}_0^e$ as well. Set $\boldsymbol{w}_\tau^e := \boldsymbol{G}^e \boldsymbol{u}_\tau^e / \sqrt{\lambda_\tau^e}$. Since

$$\|\boldsymbol{K}_0 - \boldsymbol{K}_0^e\| = \|\boldsymbol{G}^\top \boldsymbol{G} - (\boldsymbol{G}^e)^\top \boldsymbol{G}^e\| \le (\|\boldsymbol{G}\| + \|\boldsymbol{G}^e\|)\, \|\boldsymbol{G} - \boldsymbol{G}^e\| = O_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(N^{-1/2}) = o_{\mathbb{P}}(1),$$

we can claim that

$$\|\boldsymbol{w}_\tau - \boldsymbol{w}_\tau^e\| = o_{\mathbb{P}}(1). \tag{3.3.141}$$

by Davis–Kahan Theorem. Also, we have $\|\boldsymbol{u}_\tau - \boldsymbol{u}_\tau^e\| = o_{\mathbb{P}}(1)$ and $|\lambda_\tau - \lambda_\tau^e| = o_{\mathbb{P}}(1)$.

Fix $k \in \{1,2\}$. Since $\boldsymbol{a}_k$ is unit,

$$|\boldsymbol{a}_k^\top \boldsymbol{w}_\tau| \le |\boldsymbol{a}_k^\top \boldsymbol{w}_\tau^e| + \|\boldsymbol{w}_\tau - \boldsymbol{w}_\tau^e\| = |\boldsymbol{a}_k^\top \boldsymbol{w}_\tau^e| + o_{\mathbb{P}}(1)$$

by (3.3.141). Condition on $\boldsymbol{w}_\tau^e$ (which is measurable w.r.t. $\boldsymbol{G}^e$). By Lemmas 46 and 47, $\boldsymbol{g}_k \sim \mathcal{N}(0, \boldsymbol{I}_N)$ is independent of $\boldsymbol{G}^e$, hence independent of $\boldsymbol{w}_\tau^e$. Therefore $\boldsymbol{g}_k^\top \boldsymbol{w}_\tau^e \sim \mathcal{N}(0,1)$ and $\|\boldsymbol{g}_k\| = \sqrt{N} + o_{\mathbb{P}}(\sqrt{N})$, so

$$|\boldsymbol{a}_k^\top \boldsymbol{w}_\tau^e| = \frac{|\boldsymbol{g}_k^\top \boldsymbol{w}_\tau^e|}{\|\boldsymbol{g}_k\|} = O_{\mathbb{P}}(1) \cdot \frac{1}{\sqrt{N} + o_{\mathbb{P}}(\sqrt{N})} = o_{\mathbb{P}}(1).$$

Hence $|\boldsymbol{a}_k^\top \boldsymbol{w}_\tau| = o_{\mathbb{P}}(1)$ for $k \in \{1,2\}$.

Lastly, we consider $\boldsymbol{a}_0 = \mathbf{1}_N/\sqrt{N}$. Let $\boldsymbol{\Pi}_\delta$ be the spectral projector of $\boldsymbol{G}\boldsymbol{G}^\top$ onto interval $(\lambda_\tau - \delta, \lambda_\tau + \delta)$ for some fixed and small $\delta > 0$. On the event that the outlier is simple, $\boldsymbol{\Pi}_\delta = \boldsymbol{w}_\tau \boldsymbol{w}_\tau^\top$, so

$$|\boldsymbol{a}_0^\top \boldsymbol{w}_\tau|^2 = \boldsymbol{a}_0^\top \boldsymbol{\Pi}_\delta \boldsymbol{a}_0 = -\frac{1}{2\pi i} \oint_\Gamma \boldsymbol{a}_0^\top \boldsymbol{Q}_L(z) \boldsymbol{a}_0 \, dz, \qquad \boldsymbol{Q}_L(z) := \left(\boldsymbol{G}\boldsymbol{G}^\top - z\boldsymbol{I}_N\right)^{-1},$$

where $\Gamma := \{z \in \mathbb{C} : |z - \lambda_\tau| = \delta/2\}$ and $\boldsymbol{a}_0^\top \boldsymbol{Q}_L(z) \boldsymbol{a}_0$ is well-defined on Γ. Following the proof of Lemma 50 we can apply Theorem 2.10 in [FMPW26] to get: uniformly for $z \in \Gamma$,

$$\boldsymbol{a}_0^\top \boldsymbol{Q}_L(z) \boldsymbol{a}_0 = s(z) + o_{\mathbb{P}}(1),$$

where $s(z) = \tilde{m}_\mu(z)$ is analytic in a neighborhood of λ_τ containing Γ. Therefore $\oint_\Gamma s(z)\, dz = 0$ and we obtain $|\boldsymbol{a}_0^\top \boldsymbol{w}_\tau|^2 = o_{\mathbb{P}}(1)$. □

Lemma 55 (Master equation for *additive* outliers of $\boldsymbol{K}^\sharp$)**.** *Fix $\lambda \in \mathbb{C}$ such that*

$$\lambda \notin \operatorname{supp}(\mu) \cup \{0\}. \tag{3.3.142}$$

Define $\boldsymbol{Q}_L(\lambda) = (\boldsymbol{G}\boldsymbol{G}^\top - \lambda \boldsymbol{I}_N)^{-1}$ *and* $\boldsymbol{Q}_R(\lambda) = (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1}$. *Let us define three* 3×3 *matrices as*

$$\boldsymbol{A}(\lambda) := \left(\boldsymbol{a}_k^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{a}_\ell\right)_{k,\ell=0}^2, \qquad \boldsymbol{B}(\lambda) := \left(\boldsymbol{b}_k^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b}_\ell\right)_{k,\ell=0}^2, \tag{3.3.143}$$

and $\boldsymbol{\Theta} := \operatorname{diag}(\sqrt{\beta_0}, \sqrt{\beta_1}, \sqrt{\beta_2})$. *Then* λ *is an outlier eigenvalue of* $\boldsymbol{K}^\sharp$ *if*

$$\det\Big(\boldsymbol{I}_3 - \lambda\, \boldsymbol{\Theta}\boldsymbol{A}(\lambda)\,\boldsymbol{\Theta}\cdot \boldsymbol{B}(\lambda)\Big) = 0. \tag{3.3.144}$$

Moreover, (3.3.144) *characterizes* all *outlier eigenvalues of* $\boldsymbol{K}^\sharp$*, induced by* $\{\boldsymbol{a}_k, \boldsymbol{b}_k\}_{k=0}^2$*, that lie in any compact set* $\mathscr{D} \subset \mathbb{C}$ *satisfying* (3.3.142).

Proof. Notice that from the last part of Lemma 53, we know all the entries of $\boldsymbol{A}(\lambda)$ and $\boldsymbol{B}(\lambda)$ are well-defined and finite, even if $\lambda = \widehat{\lambda}_\tau^{(0)} \notin \operatorname{supp}(\mu)\cup\{0\}$ where $\widehat{\lambda}_\tau^{(0)}$ is characterized by Lemma 43.

Define

$$\boldsymbol{H}^\sharp := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{Y}^\sharp \\ \boldsymbol{Y}^{\sharp\top} & \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{H}_0 := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{G} \\ \boldsymbol{G}^\top & \boldsymbol{0} \end{pmatrix}.$$

If $\rho \neq 0$, then $\rho \in \operatorname{spec}(\boldsymbol{H}^\sharp)$ iff $\rho^2 \in \operatorname{spec}(\boldsymbol{K}^\sharp)$. Set $\lambda := \rho^2$. We now consider the low rank matrix $\boldsymbol{P} = \boldsymbol{A}_0\boldsymbol{\Theta}\boldsymbol{B}_0^\top$ where $\boldsymbol{A}_0 := [\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2] \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{B}_0 := [\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2] \in \mathbb{R}^{n\times 3}$. Then

$$\boldsymbol{H}^\sharp = \boldsymbol{H}_0 + \boldsymbol{\Delta}, \qquad \boldsymbol{\Delta} := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{P} \\ \boldsymbol{P}^\top & \boldsymbol{0} \end{pmatrix}.$$

Factor $\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{C}\boldsymbol{U}^\top$ with

$$\boldsymbol{U} := \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_0 \end{pmatrix} \in \mathbb{R}^{(N+n)\times 6}, \qquad \boldsymbol{C} := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{6\times 6}.$$

Under (3.3.142), $\rho \notin \operatorname{spec}(\boldsymbol{H}_0)$ so $(\boldsymbol{H}_0 - \rho \boldsymbol{I})$ is invertible and

$$\det(\boldsymbol{H}^\sharp - \rho \boldsymbol{I}) = \det(\boldsymbol{H}_0 - \rho \boldsymbol{I}) \det\Big(\boldsymbol{I}_6 + \boldsymbol{C}\boldsymbol{U}^\top (\boldsymbol{H}_0 - \rho \boldsymbol{I})^{-1} \boldsymbol{U}\Big).$$

Hence $\rho \in \operatorname{spec}(\boldsymbol{H}^\sharp) \setminus \operatorname{spec}(\boldsymbol{H}_0)$ iff the 6×6 determinant above vanishes. For $\lambda = \rho^2 \notin \operatorname{spec}(\boldsymbol{K}_0) \cup \operatorname{spec}(\boldsymbol{G}\boldsymbol{G}^\top)$ we have

$$(\boldsymbol{H}_0 - \rho \boldsymbol{I})^{-1} = \begin{pmatrix} -\rho(\boldsymbol{G}\boldsymbol{G}^\top - \lambda \boldsymbol{I})^{-1} & \boldsymbol{G}(\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} \\ (\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} \boldsymbol{G}^\top & -\rho(\boldsymbol{K}_0 - \lambda \boldsymbol{I})^{-1} \end{pmatrix} = \begin{pmatrix} -\rho\, \boldsymbol{Q}_L(\lambda) & \boldsymbol{G}\boldsymbol{Q}_R(\lambda) \\ \boldsymbol{Q}_R(\lambda)\boldsymbol{G}^\top & -\rho\, \boldsymbol{Q}_R(\lambda) \end{pmatrix}.$$

Using the identity $\boldsymbol{Q}_L(\lambda)\boldsymbol{G} = \boldsymbol{G}\boldsymbol{Q}_R(\lambda)$, the 6×6 determinant reduces (by an explicit Schur complement) to the equivalent 3×3 condition

$$\det\Big(\boldsymbol{I}_3 - \lambda\, \boldsymbol{\Theta}\boldsymbol{A}_0^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{A}_0 \boldsymbol{\Theta} \cdot \boldsymbol{B}_0^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{B}_0\Big) = 0,$$

which is exactly (3.3.144). The final uniform statement over $\mathscr{D}$ follows because all steps are analytic in λ on the domain (3.3.142). □

Lemma 56 (Scalar additive outlier equations)**.** *Let $\lambda \in \mathscr{D}$ and $s = s(\lambda)$ satisfy* (3.3.108)*. Then the determinant equation* (3.3.144) *reduces to three scalar equations:*

$$1 - \beta_k\, T(s) = o_{\mathbb{P}}(1), \qquad k \in \{0,1,2\}. \tag{3.3.145}$$

Equivalently, each additive spike of strength $\beta \in \{\beta_0, \beta_1, \beta_2\}$ produces a separated outlier eigenvalue at

$$\lambda = z(s) + o_{\mathbb{P}}(1) \quad \textit{for any real } s \textit{ with } z'(s) > 0 \textit{ satisfying } \beta\, T(s) = 1.$$

Proof. Fix $\lambda \in \mathscr{D}$ and let $s = s(\lambda)$. By Lemma 50, using $\boldsymbol{a}_0 = \mathbf{1}/\sqrt{N}$, $\boldsymbol{a}_1 = \boldsymbol{g}_1/\|\boldsymbol{g}_1\|$, $\boldsymbol{a}_2 =$

$\boldsymbol{g}_2/\|\boldsymbol{g}_2\|$, we have

$$\boldsymbol{A}_{k\ell}(\lambda) = \boldsymbol{a}_k^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{a}_\ell = \delta_{k\ell}\, s + o_{\mathbb{P}}(1), \qquad k,\ell \in \{0,1,2\}.$$

Thus $\boldsymbol{A}(\lambda) = s\boldsymbol{I}_3 + o_{\mathbb{P}}(1)$.

By Lemma 52 and orthogonality $\boldsymbol{v}_k^\top \boldsymbol{u} = 0$, $\boldsymbol{v}_k^\top \boldsymbol{m} = o_{\mathbb{P}}(1)$, we obtain

$$\boldsymbol{B}_{k\ell}(\lambda) = \boldsymbol{b}_k^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b}_\ell = \delta_{k\ell}\, m_\mu(\lambda) + o_{\mathbb{P}}(1), \qquad k,\ell \in \{0,1,2\}.$$

Thus $\boldsymbol{B}(\lambda) = m_\mu(\lambda)\boldsymbol{I}_3 + o_{\mathbb{P}}(1)$.

Plugging the above approximations of $\boldsymbol{A}(\lambda)$ and $\boldsymbol{B}(\lambda)$ into (3.3.144) gives

$$\det\Big(\boldsymbol{I}_3 - \lambda\, s m_\mu(\lambda)\, \mathrm{diag}(\beta_0,\beta_1,\beta_2)\Big) = o_{\mathbb{P}}(1),$$

so λ satisfies at least one of the following equations

$$1 - \beta_k \lambda\, s m_\mu(\lambda) = o_{\mathbb{P}}(1),$$

for $k = 0,1,2$. By Lemma 32, $\lambda s m_\mu(\lambda) = T(s)$, proving (3.3.145). □

Lemma 57 (Eigenvector alignment for additive outliers)**.** *Fix a distinct spike strength value* $\beta_\star \in \{\beta_0,\beta_1,\beta_2\}$*, and define*

$$J_\star := \{k \in \{0,1,2\} : \beta_k = \beta_\star\}, \qquad r_\star := |J_\star|.$$

Let

$$\boldsymbol{\Pi}_{B,\star} := \sum_{k\in J_\star} \boldsymbol{b}_k \boldsymbol{b}_k^\top, \qquad \boldsymbol{\Pi}_{A,\star} := \sum_{k\in J_\star} \boldsymbol{a}_k \boldsymbol{a}_k^\top.$$

Assume that $s_\star \in \mathbb{R}$ satisfies

$$\beta_\star T(s_\star) = 1, \qquad z'(s_\star) > 0, \qquad \lambda_\star := z(s_\star) \in \mathbb{R} \setminus (\mathrm{supp}(\mu) \cup \{0\}).$$

Then, there exists a sufficiently small $\delta > 0$ with interval $I_\delta = (\lambda_\star - \delta, \lambda_\star + \delta) \subset \mathbb{R} \setminus (\mathrm{supp}(\mu) \cup \{0\})$ such that the spectral projector $\widehat{\boldsymbol{P}}_\star$ of $\boldsymbol{K}^\sharp$ onto interval I_δ and the corresponding spectral projector $\widehat{\boldsymbol{P}}_\star^{(L)}$ of $\boldsymbol{Y}^\sharp \boldsymbol{Y}^{\sharp\top}$ onto I_δ satisfy:

$$\mathrm{Tr}\big(\widehat{\boldsymbol{P}}_\star \boldsymbol{\Pi}_{B,\star}\big) = r_\star \cdot \gamma_R(s_\star) + o_{\mathbb{P}}(1), \tag{3.3.146}$$

and

$$\mathrm{Tr}\big(\widehat{\boldsymbol{P}}_\star^{(L)} \boldsymbol{\Pi}_{A,\star}\big) = r_\star \cdot \gamma_L(s_\star) + o_{\mathbb{P}}(1), \tag{3.3.147}$$

where

$$\gamma_R(s) := \frac{m_\mu(z(s))\, z'(s)}{\beta_\star T'(s)} = -\frac{T(s)}{\beta_\star s^2 T'(s)}\, \varphi(s) = -\frac{\varphi(s)}{\beta_\star^2 s^2 T'(s)}. \tag{3.3.148}$$

and

$$\gamma_L(s) := \frac{s\, z'(s)}{\beta_\star T'(s)} = -\frac{z(s)}{\beta_\star T'(s)}\, \varphi(s). \tag{3.3.149}$$

In particular, if $r_\star = 1$ and $J_\star = \{k\}$, then the unique outlier eigenvalue inside I_δ is simple, and if $\widehat{\boldsymbol{v}}_k$ denotes its unit right eigenvector, then

$$|\langle \widehat{\boldsymbol{v}}_k, \boldsymbol{b}_k \rangle|^2 = \gamma_R(s_\star) + o_{\mathbb{P}}(1). \tag{3.3.150}$$

Proof. By Lemma 56, every additive outlier of $\boldsymbol{K}^\sharp$ in $\mathscr{D}$ is asymptotically described by a real solution s of

$$\beta\, T(s) = 1, \qquad z'(s) > 0, \qquad \lambda = z(s),$$

for some $\beta \in \{\beta_0, \beta_1, \beta_2\}$. Since $z'(s_\star) > 0$, Lemma 31 implies that $\lambda_\star = z(s_\star)$ lies in a connected component of $\mathbb{R} \setminus \mathrm{supp}(\mu)$, and $s(\lambda)$ is real-analytic near $\lambda_\star$.

Define

$$f_\star(\lambda) := 1 - \beta_\star T(s(\lambda)).$$

Then $f_\star(\lambda_\star) = 0$, and

$$f'_\star(\lambda_\star) = -\beta_\star T'(s_\star)\,\frac{ds}{d\lambda}(\lambda_\star) = -\beta_\star \frac{T'(s_\star)}{z'(s_\star)}. \tag{3.3.151}$$

Hence $f'_\star(\lambda_\star) \neq 0$ whenever $T'(s_\star) \neq 0$, so $f_\star$ has a simple zero at $\lambda_\star$. Now choose a sufficient small constant $\delta > 0$ such that: $I_\delta \subset \mathscr{D}$; $\lambda_\star$ is the only zero of $f_\star$ inside I_δ; and for every $\beta \neq \beta_\star$, the function $1 - \beta T(s(\lambda))$ does not vanish on I_δ. By Lemma 56, the eigenvalues of $\boldsymbol{K}^\sharp$ inside I_δ are precisely the outlier cluster associated with the repeated block $J_\star$; its total multiplicity is $r_\star$.

Let

$$\boldsymbol{R}^\sharp(\lambda) := (\boldsymbol{K}^\sharp - \lambda \boldsymbol{I}_n)^{-1}, \qquad \boldsymbol{Q}_R(\lambda) := (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1}, \qquad \boldsymbol{Q}_L(\lambda) := (\boldsymbol{G}\boldsymbol{G}^\top - \lambda \boldsymbol{I}_N)^{-1}.$$

Recall from Lemma 55 that

$$\boldsymbol{Y}^\sharp = \boldsymbol{G} + \boldsymbol{P}, \qquad \boldsymbol{P} = \boldsymbol{A}_0 \boldsymbol{\Theta} \boldsymbol{B}_0^\top, \qquad \boldsymbol{A}_0 = [\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2], \quad \boldsymbol{B}_0 = [\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2].$$

Introduce the linearizations

$$\boldsymbol{H}^\sharp := \begin{pmatrix} 0 & \boldsymbol{Y}^\sharp \\ \boldsymbol{Y}^{\sharp\top} & 0 \end{pmatrix}, \qquad \boldsymbol{H}_0 := \begin{pmatrix} 0 & \boldsymbol{G} \\ \boldsymbol{G}^\top & 0 \end{pmatrix},$$

and

$$\boldsymbol{U} := \begin{pmatrix} \boldsymbol{A}_0 & 0 \\ 0 & \boldsymbol{B}_0 \end{pmatrix} \in \mathbb{R}^{(N+n)\times 6}, \qquad \boldsymbol{C} := \begin{pmatrix} 0 & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & 0 \end{pmatrix} \in \mathbb{R}^{6\times 6},$$

so that $\boldsymbol{H}^\sharp = \boldsymbol{H}_0 + \boldsymbol{U}\boldsymbol{C}\boldsymbol{U}^\top$.

Fix ρ with $\rho^2 = \lambda$, and set

$$\boldsymbol{R}_0(\rho) := (\boldsymbol{H}_0 - \rho \boldsymbol{I}_{N+n})^{-1}, \qquad \boldsymbol{S}(\rho) := \boldsymbol{U}^\top \boldsymbol{R}_0(\rho) \boldsymbol{U}.$$

By the Woodbury identity,

$$(\boldsymbol{H}^\sharp - \rho \boldsymbol{I})^{-1} = \boldsymbol{R}_0(\rho) - \boldsymbol{R}_0(\rho) \boldsymbol{U} \boldsymbol{C} \big(\boldsymbol{I}_6 + \boldsymbol{S}(\rho) \boldsymbol{C}\big)^{-1} \boldsymbol{U}^\top \boldsymbol{R}_0(\rho). \tag{3.3.152}$$

Multiplying (3.3.152) on the left by $\boldsymbol{U}^\top$ and on the right by $\boldsymbol{U}$, we obtain the exact identity

$$\boldsymbol{U}^\top (\boldsymbol{H}^\sharp - \rho \boldsymbol{I})^{-1} \boldsymbol{U} = \big(\boldsymbol{I}_6 + \boldsymbol{S}(\rho) \boldsymbol{C}\big)^{-1} \boldsymbol{S}(\rho), \tag{3.3.153}$$

since

$$\boldsymbol{S} - \boldsymbol{S}\boldsymbol{C}(\boldsymbol{I}_6 + \boldsymbol{S}\boldsymbol{C})^{-1} \boldsymbol{S} = (\boldsymbol{I}_6 + \boldsymbol{S}\boldsymbol{C})^{-1} \boldsymbol{S}.$$

Now write $\lambda = \rho^2$. Using the block formula for $(\boldsymbol{H}_0 - \rho \boldsymbol{I})^{-1}$ from Lemma 55, we have

$$\boldsymbol{S}(\rho) = \begin{pmatrix} -\rho \boldsymbol{A}(\lambda) & \boldsymbol{X}(\lambda) \\ \boldsymbol{X}(\lambda)^\top & -\rho \boldsymbol{B}(\lambda) \end{pmatrix}, \qquad \boldsymbol{X}(\lambda) := \boldsymbol{A}_0^\top \boldsymbol{G} \boldsymbol{Q}_R(\lambda) \boldsymbol{B}_0 = \boldsymbol{A}_0^\top \boldsymbol{Q}_L(\lambda) \boldsymbol{G} \boldsymbol{B}_0.$$

Here $\boldsymbol{A}(\lambda)$ and $\boldsymbol{B}(\lambda)$ are defined by (3.3.143). Also, the lower-right block of $(\boldsymbol{H}^\sharp - \rho \boldsymbol{I})^{-1}$ equals

$$-\rho \, (\boldsymbol{K}^\sharp - \lambda \boldsymbol{I}_n)^{-1} = -\rho \, \boldsymbol{R}^\sharp(\lambda).$$

Therefore, taking the lower-right 3×3 block in (3.3.153), there exists a 3×3 matrix-valued function $\boldsymbol{H}_B(\lambda)$, analytic on a neighborhood of $\mathscr{D}$, such that

$$\boldsymbol{B}_0^\top \boldsymbol{R}^\sharp(\lambda) \boldsymbol{B}_0 = \boldsymbol{B}(\lambda) \boldsymbol{M}(\lambda)^{-1} + \boldsymbol{H}_B(\lambda), \qquad \boldsymbol{M}(\lambda) := \boldsymbol{I}_3 - \lambda \boldsymbol{\Theta} \boldsymbol{A}(\lambda) \boldsymbol{\Theta} \boldsymbol{B}(\lambda). \tag{3.3.154}$$

Hence every pole of $\boldsymbol{B}_0^\top \boldsymbol{R}^\sharp(\lambda)\boldsymbol{B}_0$ inside $\mathscr{D}$ comes from a zero of $\det \boldsymbol{M}(\lambda)$, that is, from an additive outlier of $\boldsymbol{K}^\sharp$.

By Lemmas 50 and 52, uniformly for $\lambda \in \mathscr{D} \subset \mathbb{C}$,

$$\boldsymbol{A}(\lambda) = s(\lambda)\boldsymbol{I}_3 + o_{\mathbb{P}}(1), \qquad \boldsymbol{B}(\lambda) = m_\mu(\lambda)\boldsymbol{I}_3 + o_{\mathbb{P}}(1).$$

Therefore, uniformly on $\mathscr{D}$, we have

$$\boldsymbol{M}(\lambda) = \operatorname{diag}\big(1-\beta_0 T(s(\lambda)),\, 1-\beta_1 T(s(\lambda)),\, 1-\beta_2 T(s(\lambda))\big) + o_{\mathbb{P}}(1), \tag{3.3.155}$$

because of $\lambda s(\lambda) m_\mu(\lambda) = T(s(\lambda))$.

Restrict now to the block $J_\star$. and let

$$E_\star := \sum_{k \in J_\star} \mathbf{e}_k \mathbf{e}_k^\top \in \mathbb{R}^{3\times 3},$$

where $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2$ are the standard basis vectors of $\mathbb{R}^3$. Since $\beta_k = \beta_\star$ for all $k \in J_\star$,

$$E_\star \boldsymbol{M}(\lambda) E_\star = \big(1-\beta_\star T(s(\lambda))\big) E_\star + o_{\mathbb{P}}(1), \qquad \lambda \in \mathscr{D}.$$

Hence,

$$E_\star \boldsymbol{M}(\lambda)^{-1} E_\star = \frac{1}{1-\beta_\star T(s(\lambda))} E_\star + o_{\mathbb{P}}(1), \qquad \lambda \in \mathscr{D}. \tag{3.3.156}$$

Combining (3.3.154), $\boldsymbol{B}(\lambda) = m_\mu(\lambda)\boldsymbol{I}_3 + o_{\mathbb{P}}(1)$, and (3.3.156), we obtain

$$E_\star \boldsymbol{B}_0^\top \boldsymbol{R}^\sharp(\lambda)\boldsymbol{B}_0 E_\star = \frac{m_\mu(\lambda)}{1-\beta_\star T(s(\lambda))} E_\star + \boldsymbol{H}_\star(\lambda) + o_{\mathbb{P}}(1), \qquad \lambda \in \mathscr{D}, \tag{3.3.157}$$

where $\boldsymbol{H}_\star(\lambda) := E_\star \boldsymbol{H}_B(\lambda) E_\star$ is analytic on a neighborhood of $\mathscr{D}$.

By the contour formula for spectral projectors,

$$\widehat{\boldsymbol{P}}_\star = -\frac{1}{2\pi i}\oint_{\Gamma_\star} \boldsymbol{R}^\sharp(\lambda)\,d\lambda,$$

for some circle $\Gamma_\star \subset \mathscr{D}$ with $\Gamma_\star \cap \mathbb{R} = \{\lambda_\star - \delta, \lambda_\star + \delta\}$. Therefore

$$\begin{aligned}\operatorname{Tr}\big(\widehat{\boldsymbol{P}}_\star \boldsymbol{\Pi}_{B,\star}\big) &= -\frac{1}{2\pi i}\oint_{\Gamma_\star} \operatorname{Tr}\big(\boldsymbol{\Pi}_{B,\star}\boldsymbol{R}^\sharp(\lambda)\big)\,d\lambda \\ &= -\frac{1}{2\pi i}\oint_{\Gamma_\star} \operatorname{Tr}\big(E_\star \boldsymbol{B}_0^\top \boldsymbol{R}^\sharp(\lambda)\boldsymbol{B}_0 E_\star\big)\,d\lambda. \end{aligned} \tag{3.3.158}$$

Insert (3.3.157). The analytic term $\boldsymbol{H}_\star(\lambda)$ has zero contour integral, and $\operatorname{Tr}(E_\star) = r_\star$, so

$$\operatorname{Tr}\big(\widehat{\boldsymbol{P}}_\star \boldsymbol{\Pi}_{B,\star}\big) = -\frac{1}{2\pi i}\oint_{\Gamma_\star} \frac{r_\star m_\mu(\lambda)}{1-\beta_\star T(s(\lambda))}\,d\lambda + o_{\mathbb{P}}(1). \tag{3.3.159}$$

Now $1-\beta_\star T(s(\lambda))$ has a simple zero at $\lambda_\star$, and by (3.3.151)

$$\frac{d}{d\lambda}\Big(1-\beta_\star T(s(\lambda))\Big)\Big|_{\lambda=\lambda_\star} = -\beta_\star \frac{T'(s_\star)}{z'(s_\star)}.$$

Hence, $-\operatorname{Res}_{\lambda=\lambda_\star} \frac{m_\mu(\lambda)}{1-\beta_\star T(s(\lambda))} = \frac{m_\mu(\lambda_\star)z'(s_\star)}{\beta_\star T'(s_\star)}$ and Cauchy Residue Theorem implies that

$$\operatorname{Tr}\big(\widehat{\boldsymbol{P}}_\star \boldsymbol{\Pi}_{B,\star}\big) = r_\star \cdot \frac{m_\mu(\lambda_\star)z'(s_\star)}{\beta_\star T'(s_\star)} + o_{\mathbb{P}}(1),$$

which is (3.3.146). By Lemma 32, $T(s) = z(s)\,s\,m_\mu(z(s))$ and $m_\mu(z(s)) = \frac{T(s)}{z(s)s}$. Substituting into (3.3.146),

$$\gamma_R(s) = \frac{T(s)}{\beta_\star z(s) s}\cdot\frac{z'(s)}{T'(s)} = -\frac{\varphi(s_\star)}{\beta_\star^2 s_\star^2 T'(s_\star)}$$

since $\varphi(s) := -\frac{s z'(s)}{z(s)}$ and $\beta_\star T(s_\star) = 1$.

The proof on the left singular vectors is identical, replacing $\boldsymbol{K}^\sharp$ by $\boldsymbol{Y}^\sharp \boldsymbol{Y}^{\sharp\top}$, $\boldsymbol{B}_0$ by $\boldsymbol{A}_0$, and $\boldsymbol{B}(\lambda)$ by $\boldsymbol{A}(\lambda)$. Since $\boldsymbol{A}(\lambda) = s(\lambda)\boldsymbol{I}_3 + o_{\mathbb{P}}(1)$, the numerator becomes $s(\lambda)$ instead of $m_\mu(\lambda)$,

yielding

$$\mathrm{Tr}\big(\widehat{\boldsymbol{P}}_\star^{(L)}\boldsymbol{\Pi}_{A,\star}\big) = r_\star \cdot \frac{s_\star z'(s_\star)}{\beta_\star T'(s_\star)} + o_{\mathbb{P}}(1).$$

If $r_\star = 1$, then $\widehat{\boldsymbol{P}}_\star = \widehat{\boldsymbol{v}}_k \widehat{\boldsymbol{v}}_k^\top$, so

$$|\langle \widehat{\boldsymbol{v}}_k, \boldsymbol{b}_k \rangle|^2 = \mathrm{Tr}(\widehat{\boldsymbol{P}}_\star \Pi_{B,\star}) = \gamma_R(s_\star) + o_{\mathbb{P}}(1),$$

proving (3.3.150). □

Proposition 58 (No label alignment of outliers)**.** *Under the assumptions of Theorem 5, let* $\widehat{\boldsymbol{V}}_{\mathrm{out}}$ *denote the orthogonal projector onto the direct sum of all separated outlier eigenspaces of* $\boldsymbol{K}$*. Specifically, we can define* $\widehat{\boldsymbol{V}}_{\mathrm{out}}$ *as the spectral projector of* $\boldsymbol{K}$ *on* $(0,\infty) \setminus (\mathrm{supp}(\mu) + (-\varepsilon, \varepsilon))$ *for any small* $\varepsilon > 0$*. Then*

$$\frac{1}{n}\|\widehat{\boldsymbol{V}}_{\mathrm{out}}\boldsymbol{y}\|^2 = \frac{1}{n}\boldsymbol{y}^\top \widehat{\boldsymbol{V}}_{\mathrm{out}}\boldsymbol{y} \xrightarrow{\mathbb{P}} 0. \tag{3.3.160}$$

Proof. Recall from Proposition 45 that

$$\|\boldsymbol{K} - \boldsymbol{K}^\sharp\| \prec n^{-1/4},$$

where $\boldsymbol{K}^\sharp = \boldsymbol{Y}^{\sharp\top}\boldsymbol{Y}^\sharp$ is the rank-three additive deformation model. Since the number of separated outlier clusters is finite and each cluster stays at positive distance from the bulk with probability tending to one, Weyl's inequality and Davis–Kahan theorem imply that the spectral projector $\widehat{\boldsymbol{V}}_{\mathrm{out}}$ of $\boldsymbol{K}$ and the corresponding outlier projector $\widehat{\boldsymbol{V}}^\sharp_{\mathrm{out}}$ of $\boldsymbol{K}^\sharp$ satisfy

$$\|\widehat{\boldsymbol{V}}_{\mathrm{out}} - \widehat{\boldsymbol{V}}^\sharp_{\mathrm{out}}\| = o_{\mathbb{P}}(1).$$

Hence

$$\frac{1}{n}\Big|\boldsymbol{y}^\top(\widehat{\boldsymbol{V}}_{\mathrm{out}} - \widehat{\boldsymbol{V}}^\sharp_{\mathrm{out}})\boldsymbol{y}\Big| \le \|\widehat{\boldsymbol{V}}_{\mathrm{out}} - \widehat{\boldsymbol{V}}^\sharp_{\mathrm{out}}\| = o_{\mathbb{P}}(1),$$

because $\|\boldsymbol{y}\|^2 = n$. Therefore it suffices to prove

$$\frac{1}{n}\boldsymbol{y}^\top \widehat{\boldsymbol{V}}_{\text{out}}^{\sharp}\boldsymbol{y} \xrightarrow{\mathbb{P}} 0. \tag{3.3.161}$$

By Lemmas 43, 53, and 57, the only possible separated outliers of $\boldsymbol{K}^{\sharp}$ are:

- the covariance-spike outlier inherited from $\boldsymbol{K}_0$ (the τ-outlier),
- the additive mean-spike outlier corresponding to $\boldsymbol{b}_0 = \boldsymbol{m}/\|\boldsymbol{m}\|$,
- the two additive linear XOR outliers corresponding to $\boldsymbol{b}_1 = \boldsymbol{v}_1$, $\boldsymbol{b}_2 = \boldsymbol{v}_2$.

Set $\bar{\boldsymbol{y}} := \frac{\boldsymbol{y}}{\sqrt{n}}$ and $\|\bar{\boldsymbol{y}}\| = 1$. From the XOR construction, $\bar{\boldsymbol{y}}^\top \boldsymbol{u} = 0$, $\bar{\boldsymbol{y}}^\top \boldsymbol{v}_1 = 0$, $\bar{\boldsymbol{y}}^\top \boldsymbol{v}_2 = 0$. Here, $\boldsymbol{u} = \mathbf{1}_n/\sqrt{n}$. Next, Lemma 48 shows that

$$\boldsymbol{u}^\top \boldsymbol{m} = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}_1^\top \boldsymbol{m} = o_{\mathbb{P}}(1), \qquad \boldsymbol{v}_2^\top \boldsymbol{m} = o_{\mathbb{P}}(1), \qquad \|\boldsymbol{m}\|^2 \xrightarrow{\mathbb{P}} \beta_{\text{mean}} \in (0,\infty).$$

Moreover, the same variance computation as in the proof of Lemma 48 gives $\boldsymbol{y}^\top \boldsymbol{m} = O_{\mathbb{P}}(1)$. Hence

$$\bar{\boldsymbol{y}}^\top \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|} = \frac{\boldsymbol{y}^\top \boldsymbol{m}}{\sqrt{n}\,\|\boldsymbol{m}\|} = o_{\mathbb{P}}(1). \tag{3.3.162}$$

Thus the normalized label direction $\bar{\boldsymbol{y}}$ is asymptotically orthogonal to all three additive spike directions

$$\boldsymbol{b}_0 = \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}, \qquad \boldsymbol{b}_1 = \boldsymbol{v}_1, \qquad \boldsymbol{b}_2 = \boldsymbol{v}_2.$$

We can claim that uniformly for $\lambda \in \mathscr{D}$,

$$\bar{\boldsymbol{y}}^\top \boldsymbol{Q}_R(\lambda)\bar{\boldsymbol{y}} = m_\mu(\lambda) + o_{\mathbb{P}}(1), \qquad \bar{\boldsymbol{y}}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b}_k = o_{\mathbb{P}}(1), \quad k = 0,1,2, \tag{3.3.163}$$

where $\boldsymbol{Q}_R(\lambda) = (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1}$. The proof of (3.3.163) is exactly followed by the proof Lemma 52. Hence, we ignore the details here.

Recall that $s_{\text{cov}} = -1/\Lambda_\tau$, $s_{\text{mean}} = T^{-1}(1/\beta_{\text{mean}})$ and $s_{\text{lin}} = T^{-1}(1/\beta_{\text{lin}})$. Denote $\lambda_0 = z(s_{\text{mean}})$, $\lambda_1 = \lambda_2 = z(s_{\text{lin}})$, and $\lambda_3 = z(s_{\text{cov}})$. Fix any distinct spike value $\lambda \in \{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$. Then, by Lemmas 53 and 56, there exists a small constant $\delta > 0$ such that for sufficiently large n, $\text{dist}(I_\delta, \text{supp}(\mu) \cup \{0\}) > \delta$ and all eigenvalues of $\boldsymbol{K}^\sharp$ inside I_δ will converge to λ in probability, where $I_\delta := (\lambda - \delta, \lambda + \delta)$. Let $\widehat{\boldsymbol{V}}_\lambda^\sharp$ be the spectral projector of $\boldsymbol{K}^\sharp$ onto outlier cluster in I_δ, and let $\Gamma_\lambda \subset \mathbb{C}$ be a small contour enclosing that cluster and no other part of $\text{spec}(\boldsymbol{K}^\sharp)$. Then

$$\widehat{\boldsymbol{V}}_\lambda^\sharp = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} (\boldsymbol{K}^\sharp - z\boldsymbol{I})^{-1} \, dz.$$

Hence

$$\bar{\boldsymbol{y}}^\top \widehat{\boldsymbol{V}}_\lambda^\sharp \bar{\boldsymbol{y}} = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\boldsymbol{y}}^\top (\boldsymbol{K}^\sharp - z\boldsymbol{I})^{-1} \bar{\boldsymbol{y}} \, dz. \tag{3.3.164}$$

Now use the same linearization trick as in Lemma 57. Recall that $\boldsymbol{A}_0 = [\boldsymbol{a}_0, \boldsymbol{a}_1, \boldsymbol{a}_2]$, $\boldsymbol{B}_0 = [\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2]$, and $\boldsymbol{\Theta} = \text{diag}(\sqrt{\beta_0}, \sqrt{\beta_1}, \sqrt{\beta_2})$. Then one obtains the exact projected-resolvent representation

$$(\boldsymbol{K}^\sharp - z\boldsymbol{I})^{-1} = \boldsymbol{Q}_R(z) + \boldsymbol{Q}_R(z)\boldsymbol{B}_0 \boldsymbol{S}(z) \boldsymbol{B}_0^\top \boldsymbol{Q}_R(z), \tag{3.3.165}$$

where $\boldsymbol{S}(z) = z\boldsymbol{\Theta}\boldsymbol{A}(z)\boldsymbol{\Theta}\big(\boldsymbol{I}_3 - z\boldsymbol{\Theta}\boldsymbol{A}(z)\boldsymbol{\Theta}\boldsymbol{B}(z)\big)^{-1}$ is a 3×3 matrix-valued meromorphic function, and all poles inside $\mathscr{D}$ come from the additive outlier equation described by Lemma 56. Hence, $\boldsymbol{S}(z)$ is analytic and bounded on Γ_λ. Insert (3.3.165) into (3.3.164). The possible pole term is

$$\bar{\boldsymbol{y}}^\top \widehat{\boldsymbol{V}}_\lambda^\sharp \bar{\boldsymbol{y}} = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\boldsymbol{y}}^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} \, dz - \frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\boldsymbol{y}}^\top \boldsymbol{Q}_R(z) \boldsymbol{B}_0 \boldsymbol{S}(z) \boldsymbol{B}_0^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} dz.$$

By (3.3.163), $\bar{\boldsymbol{y}} \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} = m_\mu(z) + o_{\mathbb{P}}(1)$ and

$$\boldsymbol{B}_0^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{b}_0^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} \\ \boldsymbol{b}_1^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} \\ \boldsymbol{b}_2^\top \boldsymbol{Q}_R(z) \bar{\boldsymbol{y}} \end{bmatrix} = o_{\mathbb{P}}(1)$$

uniformly on the contour $z \in \Gamma_\lambda$. Since $m_\mu(z)$ is analytic on Γ_λ, its contour integral vanishes. Hence

$$\bar{\boldsymbol{y}}^\top \widehat{\boldsymbol{V}}_\lambda^\sharp \bar{\boldsymbol{y}} = o_{\mathbb{P}}(1).$$

This completes the proof of this proposition. □

3.3.13 Proof of Theorem 5

By Proposition 45,

$$\|\boldsymbol{K} - \boldsymbol{K}^\sharp\| = o_{\mathbb{P}}(1)$$

Therefore isolated eigenvalues and their eigenspaces coincide up to $o_{\mathbb{P}}(1)$. It suffices to analyze $\boldsymbol{K}^\sharp$.

Lemma 43 gives: (i) the bulk ESD of $\boldsymbol{K}_0$ is μ; (ii) the population spike Λ_τ exists if $\tau > \tau_{\text{crit}}$, and the associated outlier eigenvalue of $\boldsymbol{K}_0$ exists if $z'(-1/\Lambda_\tau) > 0$ and $\tau > \tau_{\text{crit}}$, at location $z(-1/\Lambda_\tau)$, with eigenvector alignment $\varphi(-1/\Lambda_\tau)$ for unit vector $\boldsymbol{u}$. Then by Lemmas 53 and 54, $\boldsymbol{K}^\sharp$ has a vanishing perturbation of the this outlier eigenvalue and eigenvector derived from $\boldsymbol{K}_0$, and they have the same limiting bulk distribution μ, which proves item (i) and (ii) of Theorem 5. Lemma 43 gives an exact outlier condition for $\boldsymbol{K}^\sharp$ outside the limiting bulk distribution μ.

Next, Lemmas 55 and 56 show that each additive spike of $\boldsymbol{K}^\sharp$ yields an outlier eigenvalue $\widehat{\lambda}_k = z(s_k) + o_{\mathbb{P}}(1)$ if the condition $z'(s_k) > 0$ holds, where s satisfies $\beta_k T(s) = 1$ for $k = 0, 1, 2$. Here $s_0 = s_{\text{mean}}$ and $s_1 = s_2 = s_{\text{lin}}$ defined by Theorem 5. This proves item (iii) and (iv), including multiplicities via $\beta_1 = \beta_2$. Lemma 57 gives the formulas for eigenvector alignments for all these outliers in item (iii) and (iv) of Theorem 5 if $s_{\text{mean}} \neq s_{\text{lin}}$.

The covariance spike contributes at most one separated outlier, and the additive deformation has rank three, so it contributes at most three separated outliers (counting multiplicities). Because the additive right directions $\boldsymbol{b}_0, \boldsymbol{v}_1, \boldsymbol{v}_2$ are orthogonal to $\boldsymbol{u}$ (exactly for $\boldsymbol{v}_1, \boldsymbol{v}_2$, and $o_{\mathbb{P}}(1)$ for $\boldsymbol{b}_0$), Lemmas 53 and 54 show that the covariance-spike eigenspace does not mix with the

additive spikes at leading order; thus the covariance outlier persists at the stated location. Then, by Theorem 9 of [WWF24], there is no other outlier eigenvalues except for what we discussed above.

Finally, Lemma 58 proves item (v) of Theorem 5.

3.3.14 Proof of Kernel Spectral Clustering

Proof of Theorem 6. The common covariance $\boldsymbol{C} = \mathbf{I}_d$ of the XOR dataset greatly simplifies the analysis of the spectrum of $\boldsymbol{L}$. We apply Corollary 1 from [CBG16], which gives the isolating eigenvalues of $\boldsymbol{L}$ in terms of the isolated eigenvalues of $\boldsymbol{C} + \boldsymbol{M}\mathscr{D}(c)\boldsymbol{M}^\top$, where $\boldsymbol{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4]$, $\boldsymbol{c} = (\frac{1}{4})_{\alpha=1}^4$ is the vector of class proportions, and $\mathscr{D}(\cdot)$ is the diagonal operator. For the XOR dataset, this matrix becomes

$$\boldsymbol{C} + \boldsymbol{M}\mathscr{D}(c)\boldsymbol{M}^\top = \mathbf{I}_d + \frac{1}{4}\sum_{k=1}^{4} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top = \mathbf{I}_d + \frac{r^2}{d}\begin{bmatrix} \mathbf{1}_{\frac{d}{2}}\mathbf{1}_{\frac{d}{2}}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\frac{d}{2}}\mathbf{1}_{\frac{d}{2}}^\top \end{bmatrix} \tag{3.3.166}$$

The eigenvalues of (3.3.166) are $\ell = 1 + \frac{r^2}{2}$, of multiplicity two, and the rest are equal to 1. The phase transition in Corollary 1 can therefore be restated as:

$$|\ell - 1| = \frac{r^2}{2} > \sqrt{\psi^{-1}}.$$

In this case, by Corollary 1 of [CBG16] the isolating eigenvalues of $\boldsymbol{L}$ are asymptotically well approximated via

$$-\frac{2f'(\tau)}{f(\tau)}\rho + \frac{f(0) - f(\tau) + \tau f'(\tau)}{f(\tau)},$$

in probability, where $\rho = \psi\big(1 + \frac{r^2}{2}\big) + \frac{2+r^2}{r^2}$. Similarly, define $\ell_+ = \frac{5f'(\tau)}{4f(\tau)} - \frac{f''(\tau)}{f'(\tau)}$. The additional isolating eigenvalue of $\boldsymbol{L}$ corresponding to $\rho_+ = \ell_+\big(\psi - \frac{1}{1-\ell_+}\big)$ emerges if

$$(1 - \ell_+)^2 > \psi^{-1}$$

in which case, the corresponding outlier eigenvalue of $\boldsymbol{L}$ converges to, in probability,

$$-2\frac{f'(\tau)}{f(\tau)}\rho_+ + \frac{f(0)-f(\tau)+\tau f'(\tau)}{f(\tau)}. \tag{3.3.167}$$

For eigenvector alignment of the eigenvalues relevant for clustering, we observe the two-dimensional eigenspace corresponding to $\ell = 1+\frac{r^2}{2}$ is spanned by the unit-norm eigenvectors $\{\Upsilon_1, \Upsilon_2\}$, where

$$\Upsilon_1 = \frac{1}{\sqrt{d/2}}\begin{pmatrix}\mathbf{1}_{d/2}\\ \mathbf{0}_{d/2}\end{pmatrix}, \qquad \Upsilon_2 = \frac{1}{\sqrt{d/2}}\begin{pmatrix}\mathbf{0}_{d/2}\\ \mathbf{1}_{d/2}\end{pmatrix}. \tag{3.3.168}$$

Let $\Upsilon_\rho = (\Upsilon_1, \Upsilon_2) \in \mathbb{R}^{d\times 2}$ be the column-concatenated eigenvectors, then we apply Corollary 2 of [CBG16] to obtain

$$\tfrac{1}{d}\boldsymbol{J}^\top \boldsymbol{\Pi}_\rho \boldsymbol{J} = \frac{1}{\ell}\Big(\psi - \frac{1}{(1-\ell)^2}\Big)\mathscr{D}(c)\boldsymbol{M}^\top \Upsilon_\rho \Upsilon_\rho^\top \boldsymbol{M}\mathscr{D}(c) + o_{\mathbb{P}}(1) \tag{3.3.169}$$

Or equivalently,

$$\frac{1}{d}\boldsymbol{J}^\top \boldsymbol{\Pi}_\rho \boldsymbol{J} \to \frac{r^2}{16+8r^2}\Big(\psi - \frac{4}{r^4}\Big)\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\otimes \mathbf{I}_2. \tag{3.3.170}$$

in probability. We conclude the proof by observing the XOR-labels $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{J}\begin{pmatrix}1\\ -1\\ 1\\ -1\end{pmatrix}$$

Applying this to (3.3.170), we have

$$\frac{1}{d}\boldsymbol{y}^\top \boldsymbol{\Pi}_\rho \boldsymbol{y} \to \frac{r^2}{16+8r^2}\big(\psi - \frac{4}{r^4}\big) \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}^\top \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2 \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = 0.$$

in probability. Hence, we complete the proof of this theorem. □

3.3.15 Proof for Large-SNR Regime

In this section, we prove Theorem 7 when SNR is increasing with n as $n,d,N \to \infty$ proportionally. Recall that Proposition 35 implies that

$$\|\boldsymbol{Y} - \boldsymbol{Y}_{\mathrm{QE}}\| = O_\prec(n^{-1/4}) \tag{3.3.171}$$

where

$$\boldsymbol{Y}_{\mathrm{QE}} = \boldsymbol{Y}_0 + \frac{\theta_{\mathrm{SNR}}\, b_\sigma}{\sqrt{N}}\big(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top\big) + \frac{\theta_{\mathrm{SNR}}^2\, c_\sigma}{2\sqrt{N}}\big(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\big)$$

However since $\theta_{\mathrm{SNR}} = \Theta(n^{1/4})$, we observe that $\boldsymbol{Y}_{\mathrm{QE}}$ is no longer $O_\prec(1)$. To study the order-one spikes, we require an orthogonal projection $\boldsymbol{P}_\perp$ from the following lemma.

Lemma 59 (QE for Projected Feature $\boldsymbol{Y}\boldsymbol{P}_\perp$)**.** *Let*

$$\boldsymbol{P}_{\mathrm{lin}} := \boldsymbol{v}_1\boldsymbol{v}_1^\top + \boldsymbol{v}_2\boldsymbol{v}_2^\top, \qquad \boldsymbol{P}_\perp := \boldsymbol{I}_n - \boldsymbol{P}_{\mathrm{lin}}.$$

Define $\boldsymbol{g}_k := \bar{\boldsymbol{W}}\boldsymbol{u}_k$ *and the centered squares* $\widetilde{\boldsymbol{g}}_k := \boldsymbol{g}_k^{\odot 2} - \mathbf{1}_N$*. Then* (3.3.78) *is equivalently*

$$\boldsymbol{Y}_{\mathrm{QE}} = \boldsymbol{Y}_0 + \frac{\theta_{\mathrm{SNR}}\, b_\sigma}{\sqrt{N}}\big(\boldsymbol{g}_1\boldsymbol{v}_1^\top + \boldsymbol{g}_2\boldsymbol{v}_2^\top\big) + \frac{\theta_{\mathrm{SNR}}^2\, c_\sigma}{2\sqrt{N}}\Big(\widetilde{\boldsymbol{g}}_1\boldsymbol{v}_1^{\odot 2\top} + \widetilde{\boldsymbol{g}}_2\boldsymbol{v}_2^{\odot 2\top} + \mathbf{1}_N(\boldsymbol{v}_1^{\odot 2} + \boldsymbol{v}_2^{\odot 2})^\top\Big). \tag{3.3.172}$$

Assume $r = \Theta(n^{1/4})$. Then

$$\|\boldsymbol{Y}\boldsymbol{P}_\perp - \boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp\| = O_\prec(n^{-1/4}).$$

Recall $\boldsymbol{K} = \boldsymbol{Y}^\top\boldsymbol{Y}$, and define

$$\boldsymbol{K}_\perp := \boldsymbol{P}_\perp\boldsymbol{K}\boldsymbol{P}_\perp = (\boldsymbol{Y}\boldsymbol{P}_\perp)^\top(\boldsymbol{Y}\boldsymbol{P}_\perp), \qquad \boldsymbol{K}_{\mathrm{QE},\perp} := (\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp)^\top(\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp).$$

We have $\|\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp\| = O_\prec(1)$ and

$$\|\boldsymbol{K}_\perp - \boldsymbol{K}_{\mathrm{QE},\perp}\| = O_\prec(n^{-1/4}).$$

Moreover, since $\boldsymbol{P}_\perp\boldsymbol{v}_k = \boldsymbol{0}$, we have

$$\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp = \boldsymbol{Y}_0\boldsymbol{P}_\perp + \frac{\theta_{\mathrm{SNR}}^2\, c_\sigma}{2\sqrt{N}}\Big(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\Big)\boldsymbol{P}_\perp. \tag{3.3.173}$$

In the XOR construction used below, one has $\boldsymbol{v}_k^{\odot 2} \perp \boldsymbol{v}_\ell$ for all $k,\ell \in \{1,2\}$, hence $\boldsymbol{P}_\perp\boldsymbol{v}_k^{\odot 2} = \boldsymbol{v}_k^{\odot 2}$ and the quadratic term in (3.3.173) *is unchanged by $\boldsymbol{P}_\perp$.*

Proof. The first claim follows Proposition 35 and $\|\boldsymbol{P}_\perp\| = 1$:

$$\|\boldsymbol{Y}\boldsymbol{P}_\perp - \boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp\| \le \|\boldsymbol{Y} - \boldsymbol{Y}_{\mathrm{QE}}\|\,\|\boldsymbol{P}_\perp\| = O_\prec(n^{-1/4}).$$

For the Gram matrices, write

$$\boldsymbol{K}_\perp - \boldsymbol{K}_{\mathrm{QE},\perp} = (\boldsymbol{Y}\boldsymbol{P}_\perp)^\top(\boldsymbol{Y}\boldsymbol{P}_\perp) - (\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp)^\top(\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp).$$

Hence

$$\|\boldsymbol{K}_\perp - \boldsymbol{K}_{\mathrm{QE},\perp}\| \le (\|\boldsymbol{Y}\boldsymbol{P}_\perp\| + \|\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp\|)\,\|\boldsymbol{Y}\boldsymbol{P}_\perp - \boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp\|.$$

One can easily check $\|\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_{\perp}\| = O_{\prec}(1)$ based on (3.3.173). Since $\boldsymbol{Y}\boldsymbol{P}_{\perp} = \boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_{\perp} + O_{\prec}(n^{-1/4})$ in operator norm, we also have $\|\boldsymbol{Y}\boldsymbol{P}_{\perp}\| = O_{\prec}(1)$. Therefore $\|\boldsymbol{K}_{\perp} - \boldsymbol{K}_{\mathrm{QE},\perp}\| = O_{\prec}(n^{-1/4})$. The identities (3.3.172) and (3.3.173) are algebraic consequences of $\boldsymbol{g}_k^{\odot 2} = \widetilde{\boldsymbol{g}}_k + \mathbf{1}_N$ and $\boldsymbol{P}_{\perp}\boldsymbol{v}_k = \mathbf{0}$. □

Lemma 60 (Quadratic spikes in the XOR model)**.** *Let $\boldsymbol{u} := \mathbf{1}_n/\sqrt{n}$, let $\boldsymbol{y} \in \{\pm 1\}^n$ with $\|\boldsymbol{y}\|^2 = n$, and define $\bar{\boldsymbol{y}} := \boldsymbol{y}/\sqrt{n}$. For the XOR choice of $\boldsymbol{v}_1, \boldsymbol{v}_2$, one has the identities*

$$\boldsymbol{v}_1^{\odot 2} = \frac{1}{n}(\mathbf{1}_n + \boldsymbol{y}), \qquad \boldsymbol{v}_2^{\odot 2} = \frac{1}{n}(\mathbf{1}_n - \boldsymbol{y}), \tag{3.3.174}$$

equivalently

$$\boldsymbol{v}_1^{\odot 2} = \frac{1}{\sqrt{n}}(\boldsymbol{u} + \bar{\boldsymbol{y}}), \qquad \boldsymbol{v}_2^{\odot 2} = \frac{1}{\sqrt{n}}(\boldsymbol{u} - \bar{\boldsymbol{y}}). \tag{3.3.175}$$

Consequently,

$$\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top} = \frac{1}{\sqrt{n}}\Big[(\boldsymbol{g}_1^{\odot 2} + \boldsymbol{g}_2^{\odot 2})\boldsymbol{u}^{\top} + (\boldsymbol{g}_1^{\odot 2} - \boldsymbol{g}_2^{\odot 2})\bar{\boldsymbol{y}}^{\top}\Big]. \tag{3.3.176}$$

If we center squares by $\widetilde{\boldsymbol{g}}_k := \boldsymbol{g}_k^{\odot 2} - \mathbf{1}_N$ and define

$$\boldsymbol{h}_u := \widetilde{\boldsymbol{g}}_1 + \widetilde{\boldsymbol{g}}_2 = \boldsymbol{g}_1^{\odot 2} + \boldsymbol{g}_2^{\odot 2} - 2\mathbf{1}_N, \qquad \boldsymbol{h}_y := \widetilde{\boldsymbol{g}}_1 - \widetilde{\boldsymbol{g}}_2 = \boldsymbol{g}_1^{\odot 2} - \boldsymbol{g}_2^{\odot 2},$$

then (3.3.176) *becomes*

$$\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top} = \frac{1}{\sqrt{n}}\Big[(\boldsymbol{h}_u + 2\mathbf{1}_N)\boldsymbol{u}^{\top} + \boldsymbol{h}_y\bar{\boldsymbol{y}}^{\top}\Big]. \tag{3.3.177}$$

Proof. The identities (3.3.174)–(3.3.175) are verified entrywise using the XOR definition. Plugging (3.3.175) into $\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}$ yields (3.3.176). Finally, (3.3.177) follows from $\boldsymbol{g}_k^{\odot 2} = \widetilde{\boldsymbol{g}}_k + \mathbf{1}_N$, i.e. $\boldsymbol{g}_1^{\odot 2} + \boldsymbol{g}_2^{\odot 2} = \boldsymbol{h}_u + 2\mathbf{1}_N$ and $\boldsymbol{g}_1^{\odot 2} - \boldsymbol{g}_2^{\odot 2} = \boldsymbol{h}_y$. □

Lemma 61 (Normalization of centered quadratic features)**.** *Let $\boldsymbol{g}_1, \boldsymbol{g}_2 \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ and define*

$\boldsymbol{h}_u, \boldsymbol{h}_y$ *as in Lemma 60. Then*

$$\frac{1}{N}\|\boldsymbol{h}_u\|^2 = 4 + O_{\prec}(N^{-1/2}), \qquad \frac{1}{N}\|\boldsymbol{h}_y\|^2 = 4 + O_{\prec}(N^{-1/2}), \qquad \frac{1}{N}\boldsymbol{h}_u^\top \boldsymbol{h}_y = O_{\prec}(N^{-1/2}). \tag{3.3.178}$$

Defining the unit vectors $\boldsymbol{a}_u := \boldsymbol{h}_u/\|\boldsymbol{h}_u\|$ *and* $\boldsymbol{a}_y := \boldsymbol{h}_y/\|\boldsymbol{h}_y\|$*, and assuming the large-SNR scaling of Theorem 7 so that* $\theta_{\mathrm{SNR}}^4/n \to \kappa_0$*, the random part of the quadratic spike satisfies*

$$\frac{\theta_{\mathrm{SNR}}^2 c_\sigma}{2\sqrt{N}} \cdot \frac{1}{\sqrt{n}} \Big[\boldsymbol{h}_u \boldsymbol{u}^\top + \boldsymbol{h}_y \bar{\boldsymbol{y}}^\top\Big] = \sqrt{\beta_{\mathrm{quad}}}\, \boldsymbol{a}_u \boldsymbol{u}^\top + \sqrt{\beta_{\mathrm{quad}}}\, \boldsymbol{a}_y \bar{\boldsymbol{y}}^\top + o_{\mathbb{P}}(1), \tag{3.3.179}$$

where $\beta_{\mathrm{quad}} := c_\sigma^2 \kappa_0$ *and the* $o_{\mathbb{P}}(1)$ *is in operator norm.*

Proof. Since $\boldsymbol{g}_1, \boldsymbol{g}_2$ have i.i.d. standard normal coordinates, each coordinate of $\boldsymbol{h}_u$ and $\boldsymbol{h}_y$ is i.i.d. with mean 0 and

$$\mathbb{E}[(g_{1,i}^2 + g_{2,i}^2 - 2)^2] = 4, \qquad \mathbb{E}[(g_{1,i}^2 - g_{2,i}^2)^2] = 4, \qquad \mathbb{E}[(g_{1,i}^2 + g_{2,i}^2 - 2)(g_{1,i}^2 - g_{2,i}^2)] = 0,$$

which yields (3.3.178) by the law of large numbers. For $\boldsymbol{h}_u$ and $\boldsymbol{h}_y$, all the entries are are i.i.d. mean 0 and variance 4 with sub-exponential tails; apply norm concentration (see [GSS21]), we can get the norm control of $\boldsymbol{h}_u$ and $\boldsymbol{h}_y$. Cross terms $\boldsymbol{h}_u^\top \boldsymbol{h}_y$ are sums of i.i.d. centered sub-exponential variables, hence we can apply Hanson–Wright inequality (see Proposition 1.1 of [GSS21]).

Hence the coefficient in front of $\boldsymbol{a}_u \boldsymbol{u}^\top$ has squared magnitude

$$\Big(\frac{\theta_{\mathrm{SNR}}^2 c_\sigma}{2\sqrt{N}\sqrt{n}}\Big)^2 \|\boldsymbol{h}_u\|^2 = \frac{\theta_{\mathrm{SNR}}^4 c_\sigma^2}{4Nn}\|\boldsymbol{h}_u\|^2 = c_\sigma^2 \frac{\theta_{\mathrm{SNR}}^4}{n} \cdot \frac{\|\boldsymbol{h}_u\|^2}{4N} \xrightarrow{\mathbb{P}} c_\sigma^2 \kappa_0 = \beta_{\mathrm{quad}},$$

and the same holds for the $\boldsymbol{h}_y$ term. This gives (3.3.179). □

Lemma 62 (Left-resolvent quadratic forms)**.** *Recall* $\boldsymbol{G} = \boldsymbol{Y}_0 - \frac{1}{\sqrt{N}} \mathbf{1}_N \boldsymbol{m}^\top$*. Fix a compact set*

$D \subset \mathbb{R}$ such that $D \cap \operatorname{supp}(\mu) = \varnothing$ and $0 \notin D$. For $\lambda \in D$, define

$$\boldsymbol{Q}_L(\lambda) := ((\boldsymbol{G}\boldsymbol{P}_\perp)(\boldsymbol{G}\boldsymbol{P}_\perp)^\top - \lambda \boldsymbol{I}_N)^{-1}, \qquad s(\lambda) := \widetilde{m}_\mu(\lambda).$$

Let $\boldsymbol{a}_0 := \mathbf{1}_N/\sqrt{N}$, and let $\boldsymbol{a}_u, \boldsymbol{a}_y$ be as in Lemma 61. Then uniformly for $\lambda \in D$,

$$\boldsymbol{a}_i^\top \boldsymbol{Q}_L(\lambda) \boldsymbol{a}_j = s(\lambda)\, \delta_{ij} + o_{\mathbb{P}}(1), \qquad i, j \in \{0, u, y\}. \tag{3.3.180}$$

Proof. As in the proof of Lemma 50, replace $\boldsymbol{G}\boldsymbol{P}_\perp$ by an independent copy $\widetilde{\boldsymbol{G}}$ with $\|\boldsymbol{G}\boldsymbol{P}_\perp - \widetilde{\boldsymbol{G}}\| = O_\prec(n^{-1/2})$, so resolvent stability gives $\|\boldsymbol{Q}_L - \widetilde{\boldsymbol{Q}}_L\| = O_\prec(n^{-1/2})$ uniformly on D. It therefore suffices to prove (3.3.180) with $\boldsymbol{Q}_L$ replaced by $\widetilde{\boldsymbol{Q}}_L$, which is independent of $\boldsymbol{g}_1, \boldsymbol{g}_2$, hence independent of $\boldsymbol{a}_u, \boldsymbol{a}_y$.

Conditional on $\widetilde{\boldsymbol{Q}}_L(\lambda)$, the coordinates of $\boldsymbol{h}_u, \boldsymbol{h}_y$ are independent, centered, and subexponential, so a Hanson–Wright inequality [GSS21] yields concentration of quadratic forms:

$$\frac{1}{N} \boldsymbol{h}_u^\top \widetilde{\boldsymbol{Q}}_L(\lambda) \boldsymbol{h}_u = \frac{1}{N} \operatorname{Tr} \widetilde{\boldsymbol{Q}}_L(\lambda) \cdot \frac{1}{N} \|\boldsymbol{h}_u\|^2 + o_{\mathbb{P}}(1),$$

uniformly for $\lambda \in D$, and similarly for $\boldsymbol{h}_y$. Using (3.3.178) and $\frac{1}{N} \operatorname{Tr} \widetilde{\boldsymbol{Q}}_L(\lambda) \to s(\lambda)$, we obtain

$$\boldsymbol{a}_u^\top \widetilde{\boldsymbol{Q}}_L(\lambda) \boldsymbol{a}_u = \frac{\boldsymbol{h}_u^\top \widetilde{\boldsymbol{Q}}_L(\lambda) \boldsymbol{h}_u}{\|\boldsymbol{h}_u\|^2} = s(\lambda) + o_{\mathbb{P}}(1),$$

and similarly for $\boldsymbol{a}_y$. Cross-terms $\boldsymbol{a}_u^\top \widetilde{\boldsymbol{Q}}_L(\lambda) \boldsymbol{a}_y$ vanish by the same argument (using $\frac{1}{N} \boldsymbol{h}_u^\top \boldsymbol{h}_y \to 0$ and concentration of bilinear forms), and $\boldsymbol{a}_0^\top \widetilde{\boldsymbol{Q}}_L(\lambda) \boldsymbol{a}_0 \to s(\lambda)$ is the deterministic-vector case already proved in Lemma 50. This yields (3.3.180). □

Lemma 63 (Right-resolvent quadratic forms and the τ-correction along $\boldsymbol{u}$)**.** *Fix D as in Lemma 62. For $\lambda \in D$, define the bulk Gram matrix $\boldsymbol{K}_0 := (\boldsymbol{G}\boldsymbol{P}_\perp)^\top (\boldsymbol{G}\boldsymbol{P}_\perp)$,*

$$\boldsymbol{Q}_R(\lambda) := (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1},$$

and $\tau = \frac{c_\sigma^2}{2}\psi$. *Then uniformly for* $\lambda \in D$,

$$\boldsymbol{u}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{u} = \frac{m_\mu(\lambda)}{1-\tau T(s(\lambda))} + o_{\mathbb{P}}(1), \tag{3.3.181}$$

and for any deterministic unit vector $\boldsymbol{b} \perp \boldsymbol{u}$,

$$\boldsymbol{b}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b} = m_\mu(\lambda) + o_{\mathbb{P}}(1), \qquad \boldsymbol{b}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{u} = o_{\mathbb{P}}(1). \tag{3.3.182}$$

Proof. We follow the same proof as Lemma 52. Set

$$\boldsymbol{A} := -\lambda s(\lambda)\boldsymbol{\Sigma}_0 - \lambda \boldsymbol{I}_n = -\lambda(\boldsymbol{I}_n + s(\lambda)\boldsymbol{\Sigma}_0).$$

By (3.3.87),

$$-\lambda s(\lambda)\boldsymbol{\Sigma} - \lambda \boldsymbol{I}_n = \boldsymbol{A} - \lambda s(\lambda)\tau \boldsymbol{u}\boldsymbol{u}^\top + o_{\mathbb{P}}(1).$$

Ignoring the $o_{\mathbb{P}}(1)$ term for the moment (it will be absorbed into the final $o_{\mathbb{P}}(1)$), Woodbury–Sherman–Morrison identity yields

$$\left(\boldsymbol{A} - \lambda s\tau \boldsymbol{u}\boldsymbol{u}^\top\right)^{-1} = \boldsymbol{A}^{-1} + \frac{\lambda s\tau}{1-\lambda s\tau \boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}} \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{A}^{-1}. \tag{3.3.183}$$

Multiplying left/right by $\boldsymbol{u}^\top$ and $\boldsymbol{u}$,

$$\boldsymbol{u}^\top \left(\boldsymbol{A} - \lambda s\tau \boldsymbol{u}\boldsymbol{u}^\top\right)^{-1}\boldsymbol{u} = \frac{\boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}}{1-\lambda s\tau \boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}}. \tag{3.3.184}$$

We now identify $\boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}$. Since $\boldsymbol{A}^{-1} = -(1/\lambda)(\boldsymbol{I}_n + s\boldsymbol{\Sigma}_0)^{-1}$, an anisotropic local law gives

$$\boldsymbol{u}^\top (\boldsymbol{I}_n + s\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{u} = \frac{1}{n}\operatorname{Tr}(\boldsymbol{I}_n + s\boldsymbol{\Sigma}_0)^{-1} + o_{\mathbb{P}}(1). \tag{3.3.185}$$

Let ν denote the limiting ESD of $\boldsymbol{\Sigma}_0$. Then

$$\frac{1}{n}\operatorname{Tr}(\boldsymbol{I}_n + s\boldsymbol{\Sigma}_0)^{-1} \;\longrightarrow\; \int \frac{1}{1+st}\,\nu(dt) = 1 - s\int \frac{t}{1+st}\,\nu(dt) = -\frac{T(s)}{s}.$$

Thus (3.3.185) implies

$$\boldsymbol{u}^\top(\boldsymbol{I}_n + s\boldsymbol{\Sigma}_0)^{-1}\boldsymbol{u} = -\frac{T(s)}{s} + o_{\mathbb{P}}(1),$$

and therefore

$$\boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u} = -\frac{1}{\lambda}\Big(-\frac{T(s)}{s}\Big) + o_{\mathbb{P}}(1) = \frac{T(s)}{\lambda s} + o_{\mathbb{P}}(1). \tag{3.3.186}$$

Using Lemma 34, we obtain $\boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u} = m_\mu(\lambda) + o_{\mathbb{P}}(1)$. Plugging into (3.3.184) gives

$$\boldsymbol{u}^\top\big(\boldsymbol{A} - \lambda s\tau\,\boldsymbol{u}\boldsymbol{u}^\top\big)^{-1}\boldsymbol{u} = \frac{m_\mu(\lambda)}{1 - \tau T(s(\lambda))} + o_{\mathbb{P}}(1),$$

and this yields (3.3.181).

If $\boldsymbol{b} \perp \boldsymbol{u}$, the rank-one correction term in (3.3.183) does not contribute in the limit for $\boldsymbol{b}^\top(\cdot)\boldsymbol{b}$ and $\boldsymbol{b}^\top(\cdot)\boldsymbol{u}$, and the same anisotropic law gives (3.3.182). □

Proof of Theorem 7. Notice that

$$\theta_{\mathrm{SNR}}^2 \;=\; r^2\,\frac{n}{2d} \;=\; \frac{r_0^2\psi}{2}\,n^{1/2}, \qquad \frac{\theta_{\mathrm{SNR}}^4}{n} \;\longrightarrow\; \kappa_0 := \frac{r_0^4\psi^2}{4}. \tag{3.3.187}$$

Define $\boldsymbol{u} := \mathbf{1}_n/\sqrt{n}$, the label vector $\boldsymbol{y} \in \{\pm 1\}^n$ with $\|\boldsymbol{y}\|^2 = n$, and the orthogonal projectors

$$\boldsymbol{P}_{\mathrm{lin}} := \boldsymbol{v}_1\boldsymbol{v}_1^\top + \boldsymbol{v}_2\boldsymbol{v}_2^\top, \qquad \boldsymbol{P}_\perp := \boldsymbol{I}_n - \boldsymbol{P}_{\mathrm{lin}}, \qquad \boldsymbol{K}_\perp := \boldsymbol{P}_\perp \boldsymbol{K} \boldsymbol{P}_\perp.$$

Recall also the uninformative covariance-spike parameter

$$\tau = \frac{c_\sigma^2}{2}\,\psi, \tag{3.3.188}$$

and define

$$\beta_{\mathrm{quad}} := c_\sigma^2 \kappa_0, \qquad \widetilde{\beta}_{\mathrm{quad}} := \beta_{\mathrm{quad}} + \tau. \tag{3.3.189}$$

Step 1: QE reduction on $\boldsymbol{P}_\perp$ and spike decomposition.

By Proposition 35 and Lemma 59,

$$\|\boldsymbol{K}_\perp - \boldsymbol{K}_{\perp,\mathrm{QE}}\| = O_\prec(n^{-1/4}), \qquad \boldsymbol{K}_{\perp,\mathrm{QE}} := (\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp)^\top(\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp). \tag{3.3.190}$$

Therefore, it suffices to analyze outliers of $\boldsymbol{K}_{\perp,\mathrm{QE}}$ and transfer them to $\boldsymbol{K}_\perp$ by perturbation stability.

Next, we split the null feature matrix into a centered bulk plus a mean spike. Recall that

$$\boldsymbol{m} = \mathbb{E}[\sigma(\boldsymbol{w}^\top \boldsymbol{Z}) \mid \boldsymbol{Z}] \in \mathbb{R}^n, \qquad \boldsymbol{G} = \boldsymbol{Y}_0 - \frac{1}{\sqrt{N}}\mathbf{1}_N \boldsymbol{m}^\top,$$

so $\boldsymbol{Y}_0 = \boldsymbol{G} + \frac{1}{\sqrt{N}}\mathbf{1}_N \boldsymbol{m}^\top$. Hence

$$\boldsymbol{Y}_0\boldsymbol{P}_\perp = \boldsymbol{G}\boldsymbol{P}_\perp + \frac{1}{\sqrt{N}}\mathbf{1}_N(\boldsymbol{P}_\perp \boldsymbol{m})^\top. \tag{3.3.191}$$

Combining (3.3.173) and (3.3.191),

$$\boldsymbol{Y}_{\mathrm{QE}}\boldsymbol{P}_\perp = \boldsymbol{G}\boldsymbol{P}_\perp + \frac{1}{\sqrt{N}}\mathbf{1}_N(\boldsymbol{P}_\perp \boldsymbol{m})^\top + \frac{\theta_{\mathrm{SNR}}^2\, c_\sigma}{2\sqrt{N}}\big(\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top}\big), \tag{3.3.192}$$

where in the last term we used the XOR property $\boldsymbol{P}_\perp \boldsymbol{v}_k^{\odot 2} = \boldsymbol{v}_k^{\odot 2}$.

Step 2: Reallocating quadratic spike mass into u- and y-channels.

By Lemma 60,

$$\boldsymbol{g}_1^{\odot 2}\boldsymbol{v}_1^{\odot 2\top} + \boldsymbol{g}_2^{\odot 2}\boldsymbol{v}_2^{\odot 2\top} = \frac{1}{\sqrt{n}}\Big[(\boldsymbol{h}_u + 2\mathbf{1}_N)\boldsymbol{u}^\top + \boldsymbol{h}_y \bar{\boldsymbol{y}}^\top\Big]. \tag{3.3.193}$$

Thus, the quadratic term in (3.3.192) decomposes into a deterministic rank-one part $\propto \mathbf{1}_N \boldsymbol{u}^\top$ and a random rank-two part. By Lemma 61, the random part satisfies (3.3.179).

Let $\boldsymbol{a}_0 := \mathbf{1}_N/\sqrt{N}$ and $\boldsymbol{b}_0 := (\boldsymbol{P}_\perp \boldsymbol{m})/\|\boldsymbol{P}_\perp \boldsymbol{m}\|$. Then

$$\frac{1}{\sqrt{N}} \mathbf{1}_N (\boldsymbol{P}_\perp \boldsymbol{m})^\top = \sqrt{\beta_{\text{mean}}}\, \boldsymbol{a}_0 \boldsymbol{b}_0^\top, \qquad \beta_{\text{mean}} = \|\boldsymbol{P}_\perp \boldsymbol{m}\|^2,$$

and (as in the finite-SNR analysis) $\beta_{\text{mean}} \to c_\sigma^2 \psi/2$. Collecting terms, on $\boldsymbol{P}_\perp$,

$$\boldsymbol{Y}_{\text{QE}} \boldsymbol{P}_\perp = \underbrace{\boldsymbol{G}\boldsymbol{P}_\perp}_{\text{bulk}} + \underbrace{\sqrt{\beta_{\text{mean}}}\, \boldsymbol{a}_0 \boldsymbol{b}_0^\top}_{\text{mean spike}} + \underbrace{\sqrt{\beta_{\text{quad}}}\, \boldsymbol{a}_u \boldsymbol{u}^\top + \sqrt{\beta_{\text{quad}}}\, \boldsymbol{a}_y \bar{\boldsymbol{y}}^\top}_{\text{random quadratic spikes}} + o_{\mathbb{P}}(1), \tag{3.3.194}$$

where the $o_{\mathbb{P}}(1)$ is in operator norm.

Denote the bulk Gram matrix $\boldsymbol{K}_0 := (\boldsymbol{G}\boldsymbol{P}_\perp)^\top (\boldsymbol{G}\boldsymbol{P}_\perp)$. Its ESD converges to μ and it admits the same deterministic resolvent limits as in the finite-SNR analysis, with the same transforms m_μ and $\widetilde{m}_\mu$.

Step 3: Bulk ESD of $\boldsymbol{K}$.

Part (i) of the theorem states that the ESD of $\boldsymbol{K}$ converges weakly to μ. This follows from the fact that $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ differs from the null kernel $\boldsymbol{K}_{\text{null}} := \boldsymbol{Y}_0^\top \boldsymbol{Y}_0$ by a finite-rank perturbation plus a bounded operator-norm error: indeed, $\|\boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}_{\text{QE}}^\top \boldsymbol{Y}_{\text{QE}}\| \le (\|\boldsymbol{Y}\| + \|\boldsymbol{Y}_{\text{QE}}\|)\|\boldsymbol{Y} - \boldsymbol{Y}_{\text{QE}}\|$, and $\|\boldsymbol{Y}\|, \|\boldsymbol{Y}_{\text{QE}}\| = O_\prec(n^{1/4})$ while $\|\boldsymbol{Y} - \boldsymbol{Y}_{\text{QE}}\| = O_\prec(n^{-1/4})$, hence the difference is $O_\prec(1)$. Moreover, $\boldsymbol{Y}_{\text{QE}} = \boldsymbol{Y}_0 +$ (rank-4 matrix), so $\boldsymbol{Y}_{\text{QE}}^\top \boldsymbol{Y}_{\text{QE}}$ differs from $\boldsymbol{Y}_0^\top \boldsymbol{Y}_0$ by rank $O(1)$. Therefore the ESD of $\boldsymbol{K}$ equals that of $\boldsymbol{K}_{\text{null}}$ up to $O(1/n)$ in Kolmogorov distance, and since the ESD of $\boldsymbol{K}_{\text{null}}$ converges to μ, the same is true for $\boldsymbol{K}$. This proves (i).

Step 4: Two diverging linear outliers and their eigenspace.

Assume $b_\sigma \neq 0$. The linear part of (3.3.78) is

$$\boldsymbol{T}_1 := \frac{\theta_{\text{SNR}}\, b_\sigma}{\sqrt{N}} (\boldsymbol{g}_1 \boldsymbol{v}_1^\top + \boldsymbol{g}_2 \boldsymbol{v}_2^\top).$$

Since $\|\boldsymbol{g}_k\|^2/N \to 1$ and $\boldsymbol{g}_1^\top \boldsymbol{g}_2/N \to 0$ in probability,

$$\boldsymbol{T}_1^\top \boldsymbol{T}_1 = \theta_{\mathrm{SNR}}^2 b_\sigma^2 \,(\boldsymbol{v}_1 \boldsymbol{v}_1^\top + \boldsymbol{v}_2 \boldsymbol{v}_2^\top) + o_P(\theta_{\mathrm{SNR}}^2).$$

In particular, $\boldsymbol{T}_1^\top \boldsymbol{T}_1$ has two nonzero eigenvalues $\theta_{\mathrm{SNR}}^2 b_\sigma^2 (1 + o_{\mathbb{P}}(1))$, diverging as $n^{1/2}$ by (3.3.187), and its top-2 eigenspace equals $\mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ up to $o_{\mathbb{P}}(1)$.

Now write $\boldsymbol{K} = \boldsymbol{T}_1^\top \boldsymbol{T}_1 + \boldsymbol{R}$ where $\boldsymbol{R}$ collects all remaining terms ($\boldsymbol{Y}_0$, the quadratic part, and the QE error). A direct operator-norm bound using (3.3.78) shows $\|\boldsymbol{R}\| = O_\prec(1)$, whereas the two nonzero eigenvalues of $\boldsymbol{T}_1^\top \boldsymbol{T}_1$ are of order $n^{1/2}$. Hence, by Weyl's inequality,

$$\widehat{\lambda}_1(\boldsymbol{K}) = \theta_{\mathrm{SNR}}^2 b_\sigma^2 (1 + o_{\mathbb{P}}(1)), \qquad \widehat{\lambda}_2(\boldsymbol{K}) = \theta_{\mathrm{SNR}}^2 b_\sigma^2 (1 + o_{\mathbb{P}}(1)),$$

which gives the claimed divergence rate after substituting (3.3.187). Moreover, since the spectral gap between the diverging eigenvalues and the rest is of order $n^{1/2}$, Davis–Kahan yields convergence of the corresponding spectral projector: if $\widehat{\boldsymbol{P}}_{\mathrm{lin}}$ denotes the top-2 eigenspace projector of $\boldsymbol{K}$, then

$$\|\widehat{\boldsymbol{P}}_{\mathrm{lin}} - \boldsymbol{P}_{\mathrm{lin}}\|_F \;\le\; \frac{2\|\boldsymbol{R}\|}{\mathrm{gap}(\boldsymbol{T}_1^\top \boldsymbol{T}_1)} \;=\; o_{\mathbb{P}}(1).$$

This proves (ii). (In particular, since $\boldsymbol{y} \perp \boldsymbol{v}_1, \boldsymbol{v}_2$, these diverging eigenvectors carry no label alignment.)

Step 5: Deterministic equivalents for resolvent quadratic forms.

Fix a compact set $D \subset \mathbb{R}$ such that $D \cap \mathrm{supp}(\mu) = \varnothing$. For $\lambda \in D$, define

$$\boldsymbol{Q}_R(\lambda) := (\boldsymbol{K}_0 - \lambda \boldsymbol{I}_n)^{-1}, \qquad \boldsymbol{Q}_L(\lambda) := ((\boldsymbol{G}\boldsymbol{P}_\perp)(\boldsymbol{G}\boldsymbol{P}_\perp)^\top - \lambda \boldsymbol{I}_N)^{-1}, \qquad s(\lambda) := \widetilde{m}_\mu(\lambda).$$

By Lemma 62 and Lemma 63, uniformly for $\lambda \in D$,

$$\boldsymbol{a}_i^\top \boldsymbol{Q}_L(\lambda) \boldsymbol{a}_j = s(\lambda)\, \delta_{ij} + o_{\mathbb{P}}(1), \qquad i, j \in \{0, u, y\},$$

$$\boldsymbol{u}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{u} = \frac{m_\mu(\lambda)}{1-\tau T(s(\lambda))} + o_{\mathbb{P}}(1), \qquad \boldsymbol{b}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b} = m_\mu(\lambda) + o_{\mathbb{P}}(1) \quad (\boldsymbol{b} \perp \boldsymbol{u}).$$

Step 6: Outlier equations for additive spikes (determinant reduction).

Consider the rank-r perturbation $\boldsymbol{P}$ in (3.3.194):

$$\boldsymbol{P} = \sqrt{\beta_{\text{mean}}}\,\boldsymbol{a}_0\boldsymbol{b}_0^\top + \sqrt{\beta_{\text{quad}}}\,\boldsymbol{a}_u\boldsymbol{u}^\top + \sqrt{\beta_{\text{quad}}}\,\boldsymbol{a}_y\bar{\boldsymbol{y}}^\top \qquad (r=3).$$

Define the spiked Gram matrix $\boldsymbol{K}_{\text{sp}} := (\boldsymbol{G}\boldsymbol{P}_\perp + \boldsymbol{P})^\top(\boldsymbol{G}\boldsymbol{P}_\perp + \boldsymbol{P})$. By (3.3.190) and (3.3.194), $\|\boldsymbol{K}_{\perp,\text{QE}} - \boldsymbol{K}_{\text{sp}}\| = o_{\mathbb{P}}(1)$, so it suffices to analyze outliers of $\boldsymbol{K}_{\text{sp}}$.

Let $\lambda \in D$. Standard linearization/determinant lemmas for additive finite-rank deformations (see also Remark 25 in the draft relating T to the classical D-transform) show that $\lambda \notin \text{supp}(\mu)$ is an eigenvalue of $\boldsymbol{K}_{\text{sp}}$ iff

$$\det\Big(\boldsymbol{I}_r - \lambda\,\boldsymbol{B}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{B}\,\boldsymbol{D}\,\boldsymbol{A}^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{A}\Big) = 0, \tag{3.3.195}$$

where $\boldsymbol{A} = [\boldsymbol{a}_0, \boldsymbol{a}_u, \boldsymbol{a}_y]$, $\boldsymbol{B} = [\boldsymbol{b}_0, \boldsymbol{u}, \bar{\boldsymbol{y}}]$, and $\boldsymbol{D} = \text{diag}(\beta_{\text{mean}}, \beta_{\text{quad}}, \beta_{\text{quad}})$.

By the resolvent limits in Step 5, the $r \times r$ matrix inside (3.3.195) is asymptotically diagonal. More precisely, uniformly for $\lambda \in D$,

$$\begin{aligned}
1 - \lambda\,(\boldsymbol{b}_0^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{b}_0)\,\beta_{\text{mean}}\,(\boldsymbol{a}_0^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{a}_0) &= 1 - \beta_{\text{mean}}\,\lambda\,m_\mu(\lambda)\,s(\lambda) + o_{\mathbb{P}}(1),\\
1 - \lambda\,(\boldsymbol{u}^\top \boldsymbol{Q}_R(\lambda)\boldsymbol{u})\,\beta_{\text{quad}}\,(\boldsymbol{a}_u^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{a}_u) &= 1 - \beta_{\text{quad}}\,\frac{\lambda\,m_\mu(\lambda)\,s(\lambda)}{1-\tau T(s(\lambda))} + o_{\mathbb{P}}(1),\\
1 - \lambda\,(\bar{\boldsymbol{y}}^\top \boldsymbol{Q}_R(\lambda)\bar{\boldsymbol{y}})\,\beta_{\text{quad}}\,(\boldsymbol{a}_y^\top \boldsymbol{Q}_L(\lambda)\boldsymbol{a}_y) &= 1 - \beta_{\text{quad}}\,\lambda\,m_\mu(\lambda)\,s(\lambda) + o_{\mathbb{P}}(1).
\end{aligned}$$

Using Lemma 33 and (3.3.43), i.e. $\lambda m_\mu(\lambda)s(\lambda) = T(s(\lambda))$, the outlier equations become:

$$\text{mean: } 1 - \beta_{\text{mean}}\, T(s) = 0, \tag{3.3.196}$$

$$\text{quadratic } \boldsymbol{y}\text{: } 1 - \beta_{\text{quad}}\, T(s) = 0, \tag{3.3.197}$$

$$\text{quadratic } \boldsymbol{u}\text{: } 1 - \beta_{\text{quad}}\, \frac{T(s)}{1 - \tau T(s)} = 0 \quad \Longleftrightarrow \quad 1 - (\beta_{\text{quad}} + \tau)\, T(s) = 0. \tag{3.3.198}$$

This is the key place where the deterministic equivalent (3.3.181) is used: it modifies the usual additive-spike equation by the factor $(1 - \tau T(s))^{-1}$, and therefore adds τ to the effective spike strength, yielding $\widetilde{\beta}_{\text{quad}} = \beta_{\text{quad}} + \tau$.

Let

$$s_{\text{quad}} := T^{-1}(\beta_{\text{quad}}), \qquad \widetilde{s}_{\text{quad}} := T^{-1}(\widetilde{\beta}_{\text{quad}}), \qquad s_{\text{mean}} := T^{-1}(\beta_{\text{mean}}).$$

By the monotonicity of T on the relevant real branch and the characterization of outliers in z-coordinates, each of the equations (3.3.196)–(3.3.198) yields an outlier at $\lambda = z(s)$ precisely when $z'(s) > 0$. Therefore:

- The mean spike yields the same BBP transition as in Theorem 3(iii) for $\boldsymbol{K}_\perp$, proving (iii).
- If $z'(\widetilde{s}_{\text{quad}}) > 0$, then $\boldsymbol{K}_\perp$ has an outlier $\widehat{\lambda}_u^\perp = z(\widetilde{s}_{\text{quad}}) + o_{\mathbb{P}}(1)$, proving the location part of (iv).
- If $z'(s_{\text{quad}}) > 0$, then $\boldsymbol{K}_\perp$ has an outlier $\widehat{\lambda}_y^\perp = z(s_{\text{quad}}) + o_{\mathbb{P}}(1)$, proving the location part of (v).

Step 7: Eigenvector alignments (residue/derivative computation).

We now compute overlaps with $\boldsymbol{u}$ and $\boldsymbol{y}$. Let $\widehat{\boldsymbol{v}}_u$ be the unit eigenvector of $\boldsymbol{K}_\perp$ associated with $\widehat{\lambda}_u^\perp$, and similarly $\widehat{\boldsymbol{v}}_y$ for $\widehat{\lambda}_y^\perp$.

For a rank-one additive spike with strength β and right-direction $\boldsymbol{b}$ (unit), the standard

Benaych–Georges–Nadakuditi outlier-eigenvector formula gives

$$|\widehat{\boldsymbol{v}}^\top \boldsymbol{b}|^2 \;\longrightarrow\; \frac{s\,z'(s)}{\beta\,T'(s)}, \qquad \text{where } s = T^{-1}(\beta) \text{ and } z'(s) > 0. \tag{3.3.199}$$

In our setting, the $\boldsymbol{y}$-outlier satisfies $\beta_{\text{quad}} T(s) = 1$, so (3.3.199) yields

$$|\widehat{\boldsymbol{v}}_y^\top \bar{\boldsymbol{y}}|^2 \;\xrightarrow{P}\; \frac{s_{\text{quad}}\, z'(s_{\text{quad}})}{\beta_{\text{quad}}\, T'(s_{\text{quad}})}.$$

Since $|\widehat{\boldsymbol{v}}_y^\top \bar{\boldsymbol{y}}|^2 = (1/n)|\widehat{\boldsymbol{v}}_y^\top \boldsymbol{y}|^2$, this is exactly the claimed label alignment in (v).

For the $\boldsymbol{u}$-outlier, the effective spike strength is $\widetilde{\beta}_{\text{quad}}$ because of (3.3.198); applying the same residue/derivative argument with the modified equation $\widetilde{\beta}_{\text{quad}} T(s) = 1$ gives

$$|\widehat{\boldsymbol{v}}_u^\top \boldsymbol{u}|^2 \;\xrightarrow{P}\; \frac{\widetilde{s}_{\text{quad}}\, z'(\widetilde{s}_{\text{quad}})}{\widetilde{\beta}_{\text{quad}}\, T'(\widetilde{s}_{\text{quad}})},$$

which proves the eigenvector statement in (iv).

Step 8: Conclusion and transfer from QE to $\boldsymbol{K}_\perp$.

All outlier locations and alignments above were derived for $\boldsymbol{K}_{\text{sp}}$. By (3.3.190) and standard outlier stability (continuity of isolated eigenvalues/eigenvectors under vanishing operator-norm perturbations with nonzero spectral gap), the same limits hold for $\boldsymbol{K}_\perp$, completing the proof of (iii)–(v). Together with Steps 3–4, this finishes the proof of Theorem 7. □

3.3.16 Proof for Spiked Weight Regime

In this section, we prove Theorem 8. At the beginning, we give a comprehensive version of Theorem 8 below and prove this theorem instead.

Theorem 10 (Theorem 8)**.** *Assume Assumptions 3 and 4, and the linear-width scaling $n/N \to \phi \in (0,\infty)$ and $n/d \to \psi \in (0,\infty)$. Let $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{Z} \in \mathbb{R}^{d\times n}$ be the XOR data matrix with fixed*

SNR $r = \Theta(1)$ as in Section 3.2.1, with label vector $\boldsymbol{y} \in \mathbb{R}^n$ (half $+1$, half -1) and $\boldsymbol{y}^\top \boldsymbol{u} = 0$ where $\boldsymbol{u} = \mathbf{1}_n/\sqrt{n}$. Let $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, and let the spiked ("pretrained") weight matrix be

$$\boldsymbol{W}_1 := \boldsymbol{W} + \theta\, \boldsymbol{a}\boldsymbol{b}^\top, \qquad \boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N/N), \;\; \boldsymbol{b} := \mathbf{1}_d, \qquad \theta = \theta_0 N^{1/4}, \tag{3.3.200}$$

independent of $(\boldsymbol{W}, \boldsymbol{X})$, where $\theta_0 > 0$ is fixed. Define the random feature matrix and (uncentered) conjugate kernel

$$\boldsymbol{Y} := \frac{1}{\sqrt{N}}\, \sigma(\boldsymbol{W}_1 \boldsymbol{X}) \in \mathbb{R}^{N\times n}, \qquad \boldsymbol{K} := \boldsymbol{Y}^\top \boldsymbol{Y} \in \mathbb{R}^{n\times n}. \tag{3.3.201}$$

Let

$$\boldsymbol{s} := \boldsymbol{X}^\top \boldsymbol{b} \in \mathbb{R}^n, \qquad \boldsymbol{q} := \boldsymbol{s}^{\odot 2} \in \mathbb{R}^n, \qquad \boldsymbol{q}_\perp := \boldsymbol{q} - (\boldsymbol{u}^\top \boldsymbol{q})\, \boldsymbol{u} \in \mathbb{R}^n, \tag{3.3.202}$$

and the constants

$$\kappa := 1 + \frac{r^2}{2}, \qquad \kappa_\perp := 2 + 2r^2 + \frac{r^4}{4}. \tag{3.3.203}$$

Recall the limiting bulk law μ of Theorem 5 and the functions $z(\cdot)$ and $T(\cdot)$ from Section 3.3.4 (with T strictly monotone on each branch of z outside $\operatorname{supp}(\mu)$). Define the spike strength

$$\beta_y := \frac{3c_\sigma^2}{4}\, \theta_0^4\, \phi\, \kappa_\perp. \tag{3.3.204}$$

Then the following hold:

(i) Bulk. *The empirical spectral distribution of $\boldsymbol{K}$ converges weakly in probability to μ.*

(ii) Giant outlier. *The top eigenvalue of $\boldsymbol{K}$ diverges at rate $\sqrt{N}$:*

$$\hat{\lambda}_1(\boldsymbol{K}) = b_\sigma^2\, \theta_0^2\, \phi\, \kappa \sqrt{N}\, \big(1 + o_{\mathbb{P}}(1)\big). \tag{3.3.205}$$

Moreover, if $\hat{\boldsymbol{v}}_{b,1}$ is the associated unit eigenvector, then

$$\left|\frac{\hat{\boldsymbol{v}}_{b,1}^{\top}\boldsymbol{s}}{\|\boldsymbol{s}\|}\right|^2 \to 1 \qquad \textit{and} \qquad \frac{1}{n}\left|\hat{\boldsymbol{v}}_{b,1}^{\top}\boldsymbol{y}\right|^2 \to 0, \qquad \textit{in probability.} \tag{3.3.206}$$

(iii) Order-one outliers orthogonal to $\boldsymbol{y}$. *All $O(1)$ outliers already present in the random-weight model of Theorem 5 (e.g. the linear outliers and the possible "$\boldsymbol{u}$-outlier") remain $O(1)$ and have eigenvectors asymptotically orthogonal to $\boldsymbol{y}$. (Their locations and overlaps are as in Theorem 5, up to $o_{\mathbb{P}}(1)$ perturbations.)*

(iv) Additional quadratic (label-aligned) outlier. *Let s_y be any real solution to the additive BBP equation*

$$\beta_y T(s_y) = 1, \tag{3.3.207}$$

on a branch such that $z(s_y) \notin \operatorname{supp}(\mu)$ and $z'(s_y) > 0$. Then $\boldsymbol{K}$ has an additional outlier eigenvalue at

$$\hat{\lambda}_y(\boldsymbol{K}) = z(s_y) + o_{\mathbb{P}}(1), \tag{3.3.208}$$

whose unit eigenvector $\hat{\boldsymbol{v}}_{b,y}$ satisfies the alignment limit

$$\left|\frac{\hat{\boldsymbol{v}}_{b,y}^{\top}\boldsymbol{q}_{\perp}}{\|\boldsymbol{q}_{\perp}\|}\right|^2 \to \frac{s_y z'(s_y)}{\beta_y T'(s_y)} \qquad \textit{in probability.} \tag{3.3.209}$$

Consequently, the label alignment of this outlier satisfies

$$\frac{1}{n}\left|\hat{\boldsymbol{v}}_{b,y}^{\top}\boldsymbol{y}\right|^2 \to \frac{s_y z'(s_y)}{\beta_y T'(s_y)} \cdot \frac{r^4/4}{\kappa_{\perp}}, \qquad \textit{in probability.} \tag{3.3.210}$$

3.3.17 Step 1: A QE for spiked weights.

We follow the proof strategy of Theorem 5 in Section 3.3.8, with the key difference that the spiked weight $\theta = \theta_0 N^{1/4}$ forces a quadratic Taylor term to remain $O(1)$ and creates an additional outlier aligned with $\boldsymbol{q}_{\perp}$.

We first show the concentration of $\boldsymbol{s}$ and $\boldsymbol{q}_\perp$, and correlations with labels. We record the limiting norms and the key correlation $\boldsymbol{q}_\perp^\top \boldsymbol{y}$.

Lemma 64 (Concentration of $\boldsymbol{s}$ and $\boldsymbol{q}_\perp$)**.** *Under the assumptions of Theorem 10, as $n \to \infty$,*

$$\frac{1}{n}\|\boldsymbol{s}\|^2 \to \kappa \qquad \textit{and} \qquad \frac{1}{n}\|\boldsymbol{q}_\perp\|^2 \to \kappa_\perp, \qquad \textit{in probability,} \tag{3.3.211}$$

with $\kappa, \kappa_\perp$ as in (3.3.203)*. Moreover,*

$$\frac{1}{n}\boldsymbol{q}_\perp^\top \boldsymbol{y} \to -\frac{r^2}{2} \qquad \textit{and} \qquad \frac{1}{n}\big(\boldsymbol{q}_\perp^\top \boldsymbol{y}\big)^2 \Big/ \|\boldsymbol{q}_\perp\|^2 \to \frac{r^4/4}{\kappa_\perp}, \qquad \textit{in probability.} \tag{3.3.212}$$

Finally, $\boldsymbol{y}^\top \boldsymbol{s} = O_{\mathbb{P}}(\sqrt{n})$.

Proof. Recall $\boldsymbol{s} = \boldsymbol{X}^\top \boldsymbol{b}$ with $\boldsymbol{b} = \mathbf{1}_d$. By the XOR construction, each coordinate s_j is (up to the deterministic class-dependent shift) a sum of d independent $\mathcal{N}(0, 1/d)$ entries, hence s_j has variance 1 in the first half and $1 + r^2$ in the second half, and the mixture gives $\mathbb{E}[s_j^2] = 1 + r^2/2 = \kappa$ and $\mathbb{E}[s_j^4] = 3 + 3r^2 + r^4/4$. A law of large numbers yields $\|\boldsymbol{s}\|^2/n \to \kappa$.

Next, $\boldsymbol{q} = \boldsymbol{s}^{\odot 2}$ has mean $\mathbb{E}[q_j] = \kappa$, so $\boldsymbol{u}^\top \boldsymbol{q} = \frac{1}{\sqrt{n}} \sum_j q_j = \sqrt{n}\kappa + o_{\mathbb{P}}(\sqrt{n})$. Thus $\boldsymbol{q}_\perp = \boldsymbol{q} - (\boldsymbol{u}^\top \boldsymbol{q})\boldsymbol{u}$ satisfies

$$\begin{aligned} \frac{1}{n}\|\boldsymbol{q}_\perp\|^2 &= \frac{1}{n}\sum_{j=1}^{n} (q_j - \bar{q})^2 \to \mathrm{Var}(s_j^2) = \mathbb{E}[s_j^4] - \mathbb{E}[s_j^2]^2 \\ &= \Big(3 + 3r^2 + \frac{r^4}{4}\Big) - \Big(1 + \frac{r^2}{2}\Big)^2 = 2 + 2r^2 + \frac{r^4}{4} = \kappa_\perp. \end{aligned}$$

This proves (3.3.211).

Since $\boldsymbol{y}^\top \boldsymbol{u} = 0$, we have $\boldsymbol{q}_\perp^\top \boldsymbol{y} = \boldsymbol{q}^\top \boldsymbol{y}$. But $\boldsymbol{q}^\top \boldsymbol{y} = \sum_{j \le n/2} s_j^2 - \sum_{j > n/2} s_j^2$, so by the law of large numbers

$$\frac{1}{n}\boldsymbol{q}_\perp^\top \boldsymbol{y} \to \frac{1}{2}\mathbb{E}[s_j^2 \mid \text{class } +] - \frac{1}{2}\mathbb{E}[s_j^2 \mid \text{class } -] = \frac{1}{2}\cdot 1 - \frac{1}{2}\cdot(1 + r^2) = -\frac{r^2}{2}.$$

Dividing by $\|\boldsymbol{q}_\perp\|$ and using $\|\boldsymbol{q}_\perp\|^2/n \to \kappa_\perp$ yields the second limit in (3.3.212).

Finally, $\boldsymbol{y}^\top \boldsymbol{s} = \sum_{j \le n/2} s_j - \sum_{j > n/2} s_j$ is a sum of n i.i.d. centered terms (by symmetry within each class), hence is $O_{\mathbb{P}}(\sqrt{n})$ by the central limit theorem. □

Write

$$\boldsymbol{W}_1 \boldsymbol{X} = \boldsymbol{W}\boldsymbol{Z} + \underbrace{\boldsymbol{W}\boldsymbol{M}}_{\text{(data signal)}} + \underbrace{\theta\, \boldsymbol{a}(\boldsymbol{b}^\top \boldsymbol{X})}_{\text{(weight spike)}} = \boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}}, \tag{3.3.213}$$

where $\boldsymbol{\Delta}_{\text{lin}} := \boldsymbol{W}\boldsymbol{M}$ and $\boldsymbol{\Delta}_{\text{sp}} := \theta\, \boldsymbol{a}\boldsymbol{s}^\top$ with $\boldsymbol{s} = \boldsymbol{X}^\top \boldsymbol{b}$. As in the proof of Proposition 35 and Section 3.3.12, we expand $\sigma(\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}})$ to second order around $\boldsymbol{W}\boldsymbol{Z}$:

$$\sigma(\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}}) \tag{3.3.214}$$

$$= \sigma(\boldsymbol{W}\boldsymbol{Z}) + \sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}}) + \frac{1}{2}\sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}})^{\odot 2} + \boldsymbol{R}, \tag{3.3.215}$$

where $\boldsymbol{R}$ is the third-order remainder.

The next lemma records the operator-norm sizes of the terms in (3.3.214) under the finite-SNR regime $r = \Theta(1)$ and the spiked-weight scaling $\theta = \theta_0 N^{1/4}$.

Lemma 65 (Finite-SNR QE under spiked weights)**.** *Under the assumptions of Theorem 10, define*

$$\boldsymbol{Y}_{\text{sp}} := \frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z}) + \frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{\Delta}_{\text{lin}} + \frac{b_\sigma \theta}{\sqrt{N}}\, \boldsymbol{a}\boldsymbol{s}^\top + \frac{c_\sigma \theta^2}{2\sqrt{N}}\, \boldsymbol{a}^{\odot 2}\boldsymbol{q}^\top, \tag{3.3.216}$$

where $\boldsymbol{q} = \boldsymbol{s}^{\odot 2}$. *Then*

$$\left\|\boldsymbol{Y} - \boldsymbol{Y}_{\text{sp}}\right\| \prec n^{-1/4}. \tag{3.3.217}$$

Proof. We bound each discarded contribution in (3.3.214) after division by $\sqrt{N}$.

(a) The $\boldsymbol{\Delta}_{\mathrm{sp}}$ linearization. By definition, $\boldsymbol{\Delta}_{\mathrm{sp}} = \theta\, \boldsymbol{a}\boldsymbol{s}^\top$, hence

$$\frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{\Delta}_{\mathrm{sp}} = \frac{\theta}{\sqrt{N}}\big(\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{a}\boldsymbol{s}^\top\big).$$

Since σ' is bounded (Assumption 1), $\|\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{a}\boldsymbol{s}^\top\| \le \|\sigma'\|_\infty \|\boldsymbol{a}\boldsymbol{s}^\top\| = \|\sigma'\|_\infty \|\boldsymbol{a}\|\,\|\boldsymbol{s}\|$. Using $\|\boldsymbol{a}\|^2 \to 1$ and $\|\boldsymbol{s}\|^2/n \to \kappa$ in probability (Lemma 64),

$$\left\| \frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{\Delta}_{\mathrm{sp}} \right\| = O_{\mathbb{P}}\!\left(\frac{\theta}{\sqrt{N}} \|\boldsymbol{s}\| \right) = O_{\mathbb{P}}\big(\theta\sqrt{n/N}\big) = O_{\mathbb{P}}\big(N^{1/4}\big),$$

which is precisely the "giant" spike term. To obtain the QE form in (3.3.216), we replace $\sigma'(\boldsymbol{W}\boldsymbol{Z})$ by its Hermite coefficient b_σ:

$$\frac{1}{\sqrt{N}}\sigma'(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{\Delta}_{\mathrm{sp}} = \frac{b_\sigma \theta}{\sqrt{N}}\, \boldsymbol{a}\boldsymbol{s}^\top + \boldsymbol{E}_{\mathrm{sp},1},$$

and standard concentration plus boundedness of σ'' yields $\|\boldsymbol{E}_{\mathrm{sp},1}\| \prec n^{-1/4}$ (the same argument as in Proposition 35, but with perturbation size $\|\boldsymbol{\Delta}_{\mathrm{sp}}\|_\infty = O_{\mathbb{P}}(N^{-1/4})$ and Taylor remainder order $O(\|\Delta\|_\infty^2)$).

(b) The quadratic term $(\boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2}$. We have $(\boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2} = \theta^2 \boldsymbol{a}^{\odot 2}\boldsymbol{q}^\top$ with $\boldsymbol{q} = \boldsymbol{s}^{\odot 2}$. Thus

$$\frac{1}{2\sqrt{N}}\sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2} = \frac{\theta^2}{2\sqrt{N}}\big(\sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot \boldsymbol{a}^{\odot 2}\boldsymbol{q}^\top\big).$$

Replacing $\sigma''(\boldsymbol{W}\boldsymbol{Z})$ by its Hermite coefficient c_σ yields the rank-one term $\frac{c_\sigma \theta^2}{2\sqrt{N}}\boldsymbol{a}^{\odot 2}\boldsymbol{q}^\top$ plus an error $\boldsymbol{E}_{\mathrm{sp},2}$ with $\|\boldsymbol{E}_{\mathrm{sp},2}\| \prec n^{-1/4}$ (again by boundedness of $\sigma^{(3)}$ and the entrywise size $\|(\boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2}\|_\infty = O_{\mathbb{P}}(N^{-1/2})$).

(c) Mixed and purely-data quadratic terms. Expand $(\boldsymbol{\Delta}_{\mathrm{lin}} + \boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2} = (\boldsymbol{\Delta}_{\mathrm{lin}})^{\odot 2} + 2(\boldsymbol{\Delta}_{\mathrm{lin}} \odot \boldsymbol{\Delta}_{\mathrm{sp}}) + (\boldsymbol{\Delta}_{\mathrm{sp}})^{\odot 2}$. In the finite-SNR regime, Section 3.3.8 (Lemma 44) shows that the contribution $\frac{1}{2\sqrt{N}}\sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{\Delta}_{\mathrm{lin}})^{\odot 2}$ has operator norm $\prec n^{-1/2}$ and is negligible. For the mixed term, using $\|\boldsymbol{\Delta}_{\mathrm{lin}}\|_\infty = O_{\mathbb{P}}(n^{-1/2})$ and $\|\boldsymbol{\Delta}_{\mathrm{sp}}\|_\infty = O_{\mathbb{P}}(N^{-1/4})$, we have $\|\boldsymbol{\Delta}_{\mathrm{lin}} \odot \boldsymbol{\Delta}_{\mathrm{sp}}\|_\infty = O_{\mathbb{P}}(n^{-1/2}N^{-1/4})$.

Since σ'' is bounded,

$$\left\| \frac{1}{\sqrt{N}} \sigma''(\boldsymbol{W}\boldsymbol{Z}) \odot (\boldsymbol{\Delta}_{\text{lin}} \odot \boldsymbol{\Delta}_{\text{sp}}) \right\| \;\le\; \frac{\|\sigma''\|_\infty}{\sqrt{N}} \cdot \sqrt{Nn} \cdot O_{\mathbb{P}}(n^{-1/2} N^{-1/4}) \;=\; O_{\mathbb{P}}(N^{-1/4}) \;\prec\; n^{-1/4}.$$

(d) The third-order remainder. Since $\sigma^{(3)}$ is bounded and $\|\boldsymbol{\Delta}_{\text{lin}} + \boldsymbol{\Delta}_{\text{sp}}\|_\infty = O_{\mathbb{P}}(N^{-1/4})$, the remainder satisfies $\|\boldsymbol{R}\|_\infty = O_{\mathbb{P}}(N^{-3/4})$. Therefore

$$\left\|\boldsymbol{R}/\sqrt{N}\right\| \;\le\; \sqrt{Nn} \cdot O_{\mathbb{P}}(N^{-3/4})/\sqrt{N} \;=\; O_{\mathbb{P}}(n^{1/2} N^{-3/4}) \;=\; O_{\mathbb{P}}(N^{-1/4}) \;\prec\; n^{-1/4}.$$

Combining (a)–(d) yields (3.3.217). □

3.3.18 Proof of Theorem 10

Define the "spiked-QE" Gram matrix

$$\boldsymbol{K}_{\text{sp}} := \boldsymbol{Y}_{\text{sp}}^\top \boldsymbol{Y}_{\text{sp}}. \tag{3.3.218}$$

By Lemma 65 and the basic inequality $\|\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{B}^\top \boldsymbol{B}\| \le (\|\boldsymbol{A}\| + \|\boldsymbol{B}\|)\|\boldsymbol{A} - \boldsymbol{B}\|$, we obtain

$$\left\|\boldsymbol{K} - \boldsymbol{K}_{\text{sp}}\right\| \;\prec\; (\|\boldsymbol{Y}\| + \|\boldsymbol{Y}_{\text{sp}}\|)\, n^{-1/4}. \tag{3.3.219}$$

The operator norms satisfy $\|\boldsymbol{Y}\| + \|\boldsymbol{Y}_{\text{sp}}\| = O_{\mathbb{P}}(N^{1/4})$ because of the giant spike term, hence $\|\boldsymbol{K} - \boldsymbol{K}_{\text{sp}}\| = o_{\mathbb{P}}(\sqrt{N})$. This will be sufficient to transfer the asymptotics of the diverging top eigenvalue and to keep all $O(1)$ outliers stable.

Proof of Bulk limit (part (i)).

The matrix $\boldsymbol{K}_{\text{sp}}$ differs from the random-weight kernel (Theorem 5) by a finite-rank perturbation: it contains (i) the same null bulk term $\frac{1}{\sqrt{N}}\sigma(\boldsymbol{W}\boldsymbol{Z})$ plus the same $O(1)$ finite-rank terms from $\boldsymbol{\Delta}_{\text{lin}}$, and additionally (ii) the rank-one spike terms $\frac{b_\sigma \theta}{\sqrt{N}} \boldsymbol{a}\boldsymbol{s}^\top$ and $\frac{c_\sigma \theta^2}{2\sqrt{N}} \boldsymbol{a}^{\odot 2} \boldsymbol{q}^\top$. Finite-rank perturbations do not affect the limiting ESD, so the ESD of $\boldsymbol{K}_{\text{sp}}$ converges to the same

μ as in Theorem 5. Since $\|\boldsymbol{K} - \boldsymbol{K}_{\text{sp}}\| = o_{\mathbb{P}}(1)$ on compact spectral scales (from (3.3.219) and $\|\boldsymbol{Y}\| = O_{\mathbb{P}}(N^{1/4})$), the ESD of $\boldsymbol{K}$ also converges to μ.

Proof of Giant outlier (part (ii)).

Let

$$\boldsymbol{P}_s := \frac{b_\sigma \theta}{\sqrt{N}} \boldsymbol{a}\boldsymbol{s}^\top = \frac{b_\sigma \theta_0}{N^{1/4}} \boldsymbol{a}\boldsymbol{s}^\top. \tag{3.3.220}$$

Then $\boldsymbol{P}_s$ has rank one and

$$\boldsymbol{P}_s^\top \boldsymbol{P}_s = \frac{b_\sigma^2 \theta^2}{N} (\boldsymbol{a}^\top \boldsymbol{a}) \boldsymbol{s}\boldsymbol{s}^\top.$$

Hence its unique nonzero eigenvalue equals

$$\lambda_{\max}(\boldsymbol{P}_s^\top \boldsymbol{P}_s) = \frac{b_\sigma^2 \theta^2}{N} (\boldsymbol{a}^\top \boldsymbol{a}) \, \|\boldsymbol{s}\|^2.$$

Using $\boldsymbol{a}^\top \boldsymbol{a} \to 1$ and $\|\boldsymbol{s}\|^2/n \to \kappa$ (Lemma 64), and $\theta^2 = \theta_0^2 N^{1/2}$ with $n/N \to \phi$, we obtain

$$\lambda_{\max}(\boldsymbol{P}_s^\top \boldsymbol{P}_s) = b_\sigma^2 \, \theta_0^2 \, \phi \, \kappa \sqrt{N} \, (1 + o_{\mathbb{P}}(1)).$$

All other terms in $\boldsymbol{K}_{\text{sp}}$ have operator norm $O_{\mathbb{P}}(1)$ (the random-weight CK and the $O(1)$ finite-rank linear terms), so by Weyl's inequality,

$$\hat{\lambda}_1(\boldsymbol{K}_{\text{sp}}) = b_\sigma^2 \, \theta_0^2 \, \phi \, \kappa \sqrt{N} \, (1 + o_{\mathbb{P}}(1)).$$

Since $\|\boldsymbol{K} - \boldsymbol{K}_{\text{sp}}\| = o_{\mathbb{P}}(\sqrt{N})$, the same asymptotic holds for $\hat{\lambda}_1(\boldsymbol{K})$, proving (3.3.205).

For the eigenvector, note that $\boldsymbol{P}_s^\top \boldsymbol{P}_s$ has eigenvector $\boldsymbol{s}/\|\boldsymbol{s}\|$ and an eigen-gap of order $\sqrt{N}$ to the rest of the spectrum (which is $O(1)$). Therefore, by Davis–Kahan, the top eigenvector $\hat{\boldsymbol{v}}_{b,1}$ of $\boldsymbol{K}_{\text{sp}}$ satisfies $\left|\hat{\boldsymbol{v}}_{b,1}^\top \boldsymbol{s}/\|\boldsymbol{s}\|\right|^2 \to 1$, and stability under the $o_{\mathbb{P}}(\sqrt{N})$ perturbation from $\boldsymbol{K}_{\text{sp}}$ to $\boldsymbol{K}$ yields the same for $\boldsymbol{K}$. Finally, $\boldsymbol{y}^\top \boldsymbol{s} = O_{\mathbb{P}}(\sqrt{n})$ by Lemma 64, while $\|\boldsymbol{s}\| = \Theta_{\mathbb{P}}(\sqrt{n})$, hence $(1/n)|\hat{\boldsymbol{v}}_{b,1}^\top \boldsymbol{y}|^2 \to 0$, proving (3.3.206).

Proof of the additional quadratic outlier (part (iv)).

Define the "quadratic" rank-one term

$$\boldsymbol{P}_q := \frac{c_\sigma \theta^2}{2\sqrt{N}} \boldsymbol{a}^{\odot 2} \boldsymbol{q}^\top = \frac{c_\sigma \theta_0^2}{2} \boldsymbol{a}^{\odot 2} \boldsymbol{q}^\top. \tag{3.3.221}$$

Since $\boldsymbol{y}^\top \boldsymbol{u} = 0$, only the component of $\boldsymbol{q}$ orthogonal to $\boldsymbol{u}$ can correlate with labels; therefore we decompose $\boldsymbol{q} = (\boldsymbol{u}^\top \boldsymbol{q})\boldsymbol{u} + \boldsymbol{q}_\perp$ as in (3.3.202) and retain the $\boldsymbol{q}_\perp$ channel.

We now compute the spike strength associated with $\boldsymbol{q}_\perp$. Let $\tilde{\boldsymbol{a}} := \boldsymbol{a}^{\odot 2}/\|\boldsymbol{a}^{\odot 2}\|$ and $\tilde{\boldsymbol{q}}_\perp := \boldsymbol{q}_\perp/\|\boldsymbol{q}_\perp\|$. Then

$$\frac{c_\sigma \theta_0^2}{2} \boldsymbol{a}^{\odot 2} \boldsymbol{q}_\perp^\top = \sqrt{\beta_y}\, \tilde{\boldsymbol{a}}\, \tilde{\boldsymbol{q}}_\perp^\top \qquad \text{with} \qquad \beta_y = \Big(\frac{c_\sigma \theta_0^2}{2}\Big)^2 \|\boldsymbol{a}^{\odot 2}\|^2 \|\boldsymbol{q}_\perp\|^2.$$

Since $\boldsymbol{a}_i \sim \mathcal{N}(0, 1/N)$ i.i.d., we have $\|\boldsymbol{a}^{\odot 2}\|^2 = \sum_i a_i^4 \to 3/N$ in probability, and by Lemma 64, $\|\boldsymbol{q}_\perp\|^2/n \to \kappa_\perp$. Thus

$$\beta_y = \Big(\frac{c_\sigma \theta_0^2}{2}\Big)^2 \cdot \frac{3}{N} \cdot n\kappa_\perp\,(1 + o_{\mathbb{P}}(1)) = \frac{3c_\sigma^2}{4}\, \theta_0^4\, \phi\, \kappa_\perp\,(1 + o_{\mathbb{P}}(1)),$$

which matches (3.3.204).

At this point we invoke the additive outlier theory for spiked sample covariance matrices (Lemmas 53, 55, 56, and 57 in Section 3.3.8): condition on $\boldsymbol{X}$ so that $\tilde{\boldsymbol{q}}_\perp$ is deterministic, and use independence of $\boldsymbol{a}$ from the null bulk. Then, on any branch where $z'(s) > 0$ and $z(s) \notin \operatorname{supp}(\mu)$, the existence of an outlier associated with the rank-one spike $\sqrt{\beta_y}\tilde{\boldsymbol{a}}\tilde{\boldsymbol{q}}_\perp^\top$ is governed by the scalar equation $\beta_y T(s) = 1$ (Lemma 55), i.e. (3.3.207). If s_y is a solution on such a branch, then the corresponding outlier eigenvalue converges to $z(s_y)$, giving (3.3.208).

Moreover, Lemma 46 yields the eigenvector overlap with the spike direction:

$$\big|\hat{\boldsymbol{v}}_{b,y}^\top \tilde{\boldsymbol{q}}_\perp\big|^2 \to \frac{s_y z'(s_y)}{\beta_y T'(s_y)},$$

which is (3.3.209).

Finally, since $\boldsymbol{y} \perp \boldsymbol{u}$ we have $\boldsymbol{q}_\perp^\top \boldsymbol{y} = \boldsymbol{q}^\top \boldsymbol{y}$ and

$$\frac{1}{n}\left|\hat{\boldsymbol{v}}_{b,y}^\top \boldsymbol{y}\right|^2 = \frac{1}{n}\left|\hat{\boldsymbol{v}}_{b,y}^\top \tilde{\boldsymbol{q}}_\perp\right|^2 \cdot \frac{\left(\boldsymbol{q}_\perp^\top \boldsymbol{y}\right)^2}{\|\boldsymbol{q}_\perp\|^2} + o_{\mathbb{P}}(1),$$

and Lemma 64 gives $\frac{1}{n}\left(\boldsymbol{q}_\perp^\top \boldsymbol{y}\right)^2 / \|\boldsymbol{q}_\perp\|^2 \to r^4/(4\kappa_\perp)$. Combining with (3.3.209) proves (3.3.210).

Proof of remaining order-one outliers (part (iii)).

The spiked-weight terms $\boldsymbol{P}_s$ and $\boldsymbol{P}_q$ have ranks one, hence they do not alter the existence criteria or locations of the pre-existing $O(1)$ outliers from Theorem 5 beyond $o_{\mathbb{P}}(1)$ perturbations (standard finite-rank stability, cf. Section 3.3.12). All such eigenvectors remain asymptotically orthogonal to $\boldsymbol{y}$ by the same arguments as in Theorem 5 (they align with $\boldsymbol{u}, \boldsymbol{v}_1, \boldsymbol{v}_2$-type directions), while $\boldsymbol{y}$ is orthogonal to those directions. This completes the proof.

3.3.19 Proof of Quadratic Sample-size Regime

In this section we proof Theorem 9 when $n \asymp d^2$ and $N \gg n^2$. Let $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and define the population CK matrix

$$\boldsymbol{\Phi}(\boldsymbol{X}) \;:=\; \mathbb{E}_{\boldsymbol{w}}\left[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X}) \mid \boldsymbol{X}\right] \in \mathbb{R}^{n\times n}, \qquad \boldsymbol{\Phi}(\boldsymbol{X})_{ij} = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)\sigma(\boldsymbol{w}^\top \boldsymbol{x}_j) \mid \boldsymbol{X}].$$

In this section, we consider the empirical CK with projection as

$$\boldsymbol{K}_c := \boldsymbol{P}\boldsymbol{K}\boldsymbol{P},$$

where $\boldsymbol{P} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$.

Proposition 66 (Ultra-wide CK concentration with bounded column norms)**.** *Under the assump-*

tions of Theorem 9, there exist constants $C, c > 0$ such that

$$\mathbb{P}\left(\|\boldsymbol{K} - \boldsymbol{\Phi}(\boldsymbol{X})\| \geq C\sqrt{\frac{n^2}{N}} \,\middle|\, \boldsymbol{X} \right) \leq 2\exp(-cn).$$

Consequently,

$$\|\boldsymbol{K} - \boldsymbol{\Phi}(\boldsymbol{X})\| \xrightarrow{\mathbb{P}} 0, \qquad \|\boldsymbol{K}_c - \boldsymbol{P}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{P}\| \xrightarrow{\mathbb{P}} 0.$$

Proof. Lemma 24 shows that

$$\sum_{\alpha=1}^{n} \left(\|\boldsymbol{x}_\alpha\|_2 - 1\right)^2 \lesssim d,$$

with high probability. Also, we know $\|\boldsymbol{X}\| \leq \|\boldsymbol{M}\| + \|\boldsymbol{Z}\| \prec \sqrt{d}$ where $\|\boldsymbol{M}\| = \theta_{\text{SNR}} \prec \sqrt{d}$ and $\|\boldsymbol{Z}\| \prec \sqrt{n/d} + 1$. Hence, we can directly apply Theorem 2.3 of [WZ21] and conclude the proof. □

3.3.20 Analysis of the Population CK for Quadratic Sample-size Regime

Let $(h_k)_{k\geq 0}$ be the orthonormal Hermite polynomials w.r.t. $\xi \sim \mathcal{N}(0,1)$:

$$\mathbb{E}[h_k(\xi)h_\ell(\xi)] = \delta_{k\ell}$$

with $h_0 \equiv 1$, $h_1(\xi) = \xi$, and $h_2(\xi) = (\xi^2 - 1)/\sqrt{2}$.

For each $i \in [n]$, set $s_i := \|\boldsymbol{x}_i\|$ and define the scaled activation $\sigma_{(i)}(t) := \sigma(s_i t)$. Define its Hermite coefficients

$$\zeta_k^{(i)} := \mathbb{E}[\sigma_{(i)}(\xi)h_k(\xi)], \qquad k \in \mathbb{N},$$

and the diagonal matrices $\boldsymbol{D}_k := \operatorname{diag}(\zeta_k^{(1)}, \ldots, \zeta_k^{(n)})$.

Define the Gram and correlation matrices

$$\boldsymbol{G} := \boldsymbol{X}^\top \boldsymbol{X}, \qquad \boldsymbol{D} := \operatorname{diag}(s_1, \ldots, s_n), \qquad \boldsymbol{R} := \boldsymbol{D}^{-1}\boldsymbol{G}\boldsymbol{D}^{-1}.$$

Note $\boldsymbol{R}_{ii} = 1$ and $\boldsymbol{R}_{ij} = \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j}{s_i s_j}$.

Lemma 67 (Hermite expansion of $\boldsymbol{\Phi}(\boldsymbol{X})$)**.** *Condition on $\boldsymbol{X}$. Then*

$$\boldsymbol{\Phi}(\boldsymbol{X}) \;=\; \sum_{k\geq 0} \boldsymbol{D}_k \boldsymbol{R}^{\odot k} \boldsymbol{D}_k, \qquad (\boldsymbol{R}^{\odot k})_{ij} := (\boldsymbol{R}_{ij})^k.$$

Equivalently, entrywise,

$$\boldsymbol{\Phi}(\boldsymbol{X})_{ij} = \sum_{k\geq 0} \zeta_k^{(i)} \zeta_k^{(j)} (\boldsymbol{R}_{ij})^k.$$

Proof. This lemma is derived in the proof of Lemma 5.2 in [WZ21]. □

Define the Hadamard-square matrix $\boldsymbol{H} := \boldsymbol{G} \odot \boldsymbol{G}$. For $k = 0, 1, 2$, define the following objects:

$$\boldsymbol{\mu} := \boldsymbol{D}_0 \mathbf{1} \in \mathbb{R}^n, \quad \boldsymbol{A} := \operatorname{diag}\Big(\frac{\zeta_1^{(1)}}{s_1}, \ldots, \frac{\zeta_1^{(n)}}{s_n}\Big) \in \mathbb{R}^{n\times n}, \quad \boldsymbol{C} := \operatorname{diag}\Big(\frac{\zeta_2^{(1)}}{s_1^2}, \ldots, \frac{\zeta_2^{(n)}}{s_n^2}\Big) \in \mathbb{R}^{n\times n}.$$

Lemma 68. *Condition on $\boldsymbol{X}$. Then:*

$$\begin{aligned}
\boldsymbol{D}_0 \boldsymbol{R}^{\odot 0} \boldsymbol{D}_0 &= (\boldsymbol{D}_0 \mathbf{1})(\boldsymbol{D}_0 \mathbf{1})^\top = \boldsymbol{\mu}\boldsymbol{\mu}^\top, \\
\boldsymbol{D}_1 \boldsymbol{R} \boldsymbol{D}_1 &= \boldsymbol{A}\boldsymbol{G}\boldsymbol{A}, \\
\boldsymbol{D}_2 \boldsymbol{R}^{\odot 2} \boldsymbol{D}_2 &= \boldsymbol{C}(\boldsymbol{G}\odot\boldsymbol{G})\boldsymbol{C} = \boldsymbol{C}\boldsymbol{H}\boldsymbol{C}.
\end{aligned}$$

Proof. The first identity uses $\boldsymbol{R}^{\odot 0} = \mathbf{1}\mathbf{1}^\top$. For the second,

$$(\boldsymbol{D}_1 \boldsymbol{R} \boldsymbol{D}_1)_{ij} = \zeta_1^{(i)} \zeta_1^{(j)} \frac{\boldsymbol{G}_{ij}}{s_i s_j} = \Big(\frac{\zeta_1^{(i)}}{s_i}\Big)\Big(\frac{\zeta_1^{(j)}}{s_j}\Big)\boldsymbol{G}_{ij} = (\boldsymbol{A}\boldsymbol{G}\boldsymbol{A})_{ij}.$$

Similarly,

$$(\boldsymbol{D}_2 \boldsymbol{R}^{\odot 2} \boldsymbol{D}_2)_{ij} = \zeta_2^{(i)} \zeta_2^{(j)} \frac{\boldsymbol{G}_{ij}^2}{s_i^2 s_j^2} = \Big(\frac{\zeta_2^{(i)}}{s_i^2}\Big)\Big(\frac{\zeta_2^{(j)}}{s_j^2}\Big)\boldsymbol{G}_{ij}^2 = (\boldsymbol{C}\boldsymbol{H}\boldsymbol{C})_{ij}.$$

□

Let $\boldsymbol{t}_i := \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2$ so that $\|\boldsymbol{t}_i\|_2 = 1$ and $\boldsymbol{R}_{ij} = \boldsymbol{t}_i^\top \boldsymbol{t}_j$. For each integer $k \geq 1$, define lifted

features

$$\boldsymbol{v}_i^{(k)} := \boldsymbol{t}_i^{\otimes k} \in \mathbb{R}^{d^k}, \qquad \boldsymbol{V}_k := [\boldsymbol{v}_1^{(k)}, \dots, \boldsymbol{v}_n^{(k)}] \in \mathbb{R}^{d^k \times n}.$$

Then $\boldsymbol{R}^{\odot k} = \boldsymbol{V}_k^\top \boldsymbol{V}_k$ because of kernel trick $(\boldsymbol{t}_i^\top \boldsymbol{t}_j)^k = \langle \boldsymbol{t}_i^{\otimes k}, \boldsymbol{t}_j^{\otimes k} \rangle$.

Lemma 69. *Assume $n \le Cd^2$ and the XOR model with $r = O(1)$. Fix any integer $k \ge 3$. Then, as $d \to \infty$,*

$$\left\| \boldsymbol{P}\boldsymbol{R}^{\odot k}\boldsymbol{P} - \boldsymbol{P} \right\| = o_{\mathbb{P}}(1).$$

Moreover one may take the quantitative bound

$$\left\| \boldsymbol{P}\boldsymbol{R}^{\odot k}\boldsymbol{P} - \boldsymbol{P} \right\| = O_{\mathbb{P}}\Big(\sqrt{\frac{n}{d^k}}\Big) = O_{\mathbb{P}}\big(d^{1-k/2}\big).$$

Proof. The proof is based on the proof of Lemma 5.2 of [WZ21]. We sketch a standard Gram-matrix concentration argument in the regime $d^k \gg n$.

Step 1 (Gram form). As noted, $\boldsymbol{R}^{\odot k} = \boldsymbol{V}_k^\top \boldsymbol{V}_k$ with columns $\boldsymbol{v}_i^{(k)}$ of unit norm. Hence $\boldsymbol{R}^{\odot k} - \boldsymbol{I}_n$ has zero diagonal and off-diagonal entries $(\boldsymbol{t}_i^\top \boldsymbol{t}_j)^k$.

Step 2 (nearly isotropic, high dimension). Under the XOR model with $r = O(1)$, the normalized directions $\boldsymbol{t}_i$ are i.i.d. and (uniformly in d) subgaussian with nearly isotropic covariance. Consequently, the tensor features $\boldsymbol{v}_i^{(k)} = \boldsymbol{t}_i^{\otimes k}$ are also i.i.d., subgaussian in dimension d^k, with

$$\mathbb{E}[\boldsymbol{v}_i^{(k)} (\boldsymbol{v}_i^{(k)})^\top] \approx \frac{1}{d^k} \boldsymbol{I}_{d^k}.$$

(One can formalize this by restricting to the symmetric tensor subspace; the effective dimension is still $\asymp d^k$.)

Step 3 (operator-norm concentration). For i.i.d. subgaussian isotropic vectors in dimension $p = d^k$, a standard sample-covariance concentration inequality implies

$$\left\| \frac{1}{n} \boldsymbol{V}_k \boldsymbol{V}_k^\top - \frac{1}{d^k} \boldsymbol{I}_{d^k} \right\| = O_{\mathbb{P}}\left(\sqrt{\frac{d^k}{n}} \right) \cdot \frac{1}{d^k} = O_{\mathbb{P}}\left(\sqrt{\frac{1}{nd^k}} \right).$$

Multiplying by n and using that nonzero eigenvalues of $\boldsymbol{V}_k^\top \boldsymbol{V}_k$ coincide with those of $\boldsymbol{V}_k \boldsymbol{V}_k^\top$, we obtain

$$\|\boldsymbol{V}_k^\top \boldsymbol{V}_k - \boldsymbol{I}_n\| = O_{\mathbb{P}}\Big(\sqrt{\frac{n}{d^k}}\Big).$$

Finally, since $\boldsymbol{P}(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{V}_k^\top \boldsymbol{V}_k - \boldsymbol{I})\boldsymbol{P}$ and $\|\boldsymbol{P}\| = 1$,

$$\|\boldsymbol{P}\boldsymbol{R}^{\odot k}\boldsymbol{P} - \boldsymbol{P}\| \le \|\boldsymbol{R}^{\odot k} - \boldsymbol{I}\| = O_{\mathbb{P}}\Big(\sqrt{\frac{n}{d^k}}\Big) = O_{\mathbb{P}}(d^{1-k/2}),$$

which tends to 0 for $k \ge 3$ since $n \le Cd^2$. □

Define the *variance-1 Hermite weights*

$$\alpha_0 := 1 - b_\sigma^2 - \frac{c_\sigma^2}{2}, \qquad \alpha_2 := \frac{c_\sigma^2}{2}.$$

(Recall $\mathbb{E}[\sigma(\xi)^2] = 1$ and $b_\sigma = \zeta_1(\sigma)$, $\zeta_2(\sigma) = c_\sigma/\sqrt{2}$.)

Lemma 70 (Tail ($k \ge 3$) collapses to $\alpha_0 \boldsymbol{P}$ after centering)**.** *Assume the XOR model with $n \le Cd^2$ and $r = O(1)$, and σ is Lipschitz with $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$. Then*

$$\left\| \boldsymbol{P}\left(\sum_{k\ge 3} \boldsymbol{D}_k \boldsymbol{R}^{\odot k} \boldsymbol{D}_k \right) \boldsymbol{P} - \alpha_0 \boldsymbol{P} \right\| = o_{\mathbb{P}}(1).$$

Proof. Let $\boldsymbol{T}_{\ge 3} := \sum_{k\ge 3} \boldsymbol{D}_k \boldsymbol{R}^{\odot k} \boldsymbol{D}_k$. Write

$$\boldsymbol{T}_{\ge 3} = \sum_{k\ge 3} \boldsymbol{D}_k^2 \; + \; \sum_{k\ge 3} \boldsymbol{D}_k (\boldsymbol{R}^{\odot k} - \boldsymbol{I}) \boldsymbol{D}_k.$$

We treat the two pieces.

Step 1 (diagonal piece). By Parseval in $L^2(\mathcal{N}(0,1))$,

$$\sum_{k\ge 0} (\zeta_k^{(i)})^2 = \mathbb{E}[\sigma(s_i \xi)^2].$$

Since σ is Lipschitz and $s_i = \|\boldsymbol{x}_i\|_2 = 1 + o_{\mathbb{P}}(1)$ uniformly (by Lemma 24), we have

$$\max_i |\mathbb{E}[\sigma(s_i\xi)^2] - 1| = o_{\mathbb{P}}(1)$$

Also $\zeta_0^{(i)} = \mathbb{E}[\sigma(s_i\xi)]$ and $\zeta_1^{(i)} = \mathbb{E}[\sigma(s_i\xi)\xi]$ and $\zeta_2^{(i)} = \mathbb{E}[\sigma(s_i\xi)h_2(\xi)]$ are uniformly $O_{\mathbb{P}}(1)$. Therefore

$$\max_{i\in[n]} \left| \sum_{k\geq 3} (\zeta_k^{(i)})^2 - \alpha_0 \right| = o_{\mathbb{P}}(1),$$

which implies

$$\left\| \boldsymbol{P}\left(\sum_{k\geq 3} \boldsymbol{D}_k^2\right)\boldsymbol{P} - \alpha_0 \boldsymbol{P} \right\| \leq \left\| \sum_{k\geq 3} \boldsymbol{D}_k^2 - \alpha_0 \boldsymbol{I} \right\| = o_{\mathbb{P}}(1),$$

since the LHS is a centered diagonal perturbation and the operator norm of a diagonal matrix is its maximum entry.

Step 2 (off-diagonal piece). Fix an integer $K \geq 3$ (later $K = K(d) \to \infty$ slowly). Split

$$\sum_{k\geq 3} \boldsymbol{D}_k(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{D}_k = \sum_{k=3}^{K} \boldsymbol{D}_k(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{D}_k + \sum_{k>K} \boldsymbol{D}_k(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{D}_k.$$

For the finite sum: $\|\boldsymbol{D}_k\| \leq \max_i |\zeta_k^{(i)}| \leq \max_i \|\sigma(s_i\xi)\|_{L^2} = O_{\mathbb{P}}(1)$, and Lemma 69 implies $\|\boldsymbol{P}(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{P}\| = O_{\mathbb{P}}(d^{1-k/2})$. Hence, for fixed K,

$$\left\| \boldsymbol{P} \sum_{k=3}^{K} \boldsymbol{D}_k(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{D}_k \boldsymbol{P} \right\| \leq \sum_{k=3}^{K} O_{\mathbb{P}}(d^{1-k/2}) = O_{\mathbb{P}}\Big(\frac{K}{\sqrt{d}}\Big).$$

Choose $K(d) = \lfloor \log d \rfloor$ to get $K/\sqrt{d} \to 0$.

For the tail $k > K(d)$, use the crude bound $|\boldsymbol{R}_{ij}| \leq \max_{i\neq j} |\boldsymbol{R}_{ij}| \leq \tau_n$ w.h.p. and $\tau_n \to 0$ (Lemma 24), so entrywise $|(\boldsymbol{R}_{ij})^k| \leq \tau_n^k$ for $i \neq j$. Since $\boldsymbol{R}^{\odot k} - \boldsymbol{I}$ has zero diagonal, Gershgorin yields

$$\|\boldsymbol{R}^{\odot k} - \boldsymbol{I}\| \leq n\tau_n^k.$$

Also $\|\boldsymbol{D}_k\| \le O_{\mathbb{P}}(1)$. Hence

$$\left\| \boldsymbol{P} \sum_{k>K(d)} \boldsymbol{D}_k(\boldsymbol{R}^{\odot k} - \boldsymbol{I})\boldsymbol{D}_k \boldsymbol{P} \right\| \le O_{\mathbb{P}}(1) \sum_{k>K(d)} n\tau_n^k = O_{\mathbb{P}}(1) \cdot \frac{n\tau_n^{K(d)+1}}{1-\tau_n} = o_{\mathbb{P}}(1),$$

because $K(d) \asymp \log d$ and $\tau_n \to 0$.

Combining Steps 1–2 yields the claim. □

Theorem 11 (QE for the centered population CK, general $b_\sigma \neq 0$)**.** *Assume the XOR model with $n \le Cd^2$ and $r = O(1)$, and σ is Lipschitz with $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$. Let $\boldsymbol{\mu}$, $\boldsymbol{A}$, $\boldsymbol{C}$ be as in Lemma 68, and let α_0, α_2 be as above. Then*

$$\left\| \boldsymbol{P}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{P} - \left(\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A}\boldsymbol{G}\boldsymbol{A}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}\boldsymbol{P} + \alpha_0 \boldsymbol{P} \right) \right\| = o_{\mathbb{P}}(1).$$

Proof. By Lemma 67,

$$\boldsymbol{P}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{P} = \boldsymbol{P}\left(\sum_{k=0}^{2} \boldsymbol{D}_k \boldsymbol{R}^{\odot k} \boldsymbol{D}_k \right)\boldsymbol{P} + \boldsymbol{P}\left(\sum_{k\ge 3} \boldsymbol{D}_k \boldsymbol{R}^{\odot k} \boldsymbol{D}_k \right)\boldsymbol{P}.$$

The $k = 0, 1, 2$ part equals $\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A}\boldsymbol{G}\boldsymbol{A}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}\boldsymbol{P}$ by Lemma 68. The $k \ge 3$ part equals $\alpha_0 \boldsymbol{P} + o_{\mathbb{P}}(1)$ by Lemma 70. Combining yields the claim. □

The following proposition is the formal statement of Proposition 36.

Proposition 71 (QE for centered empirical CK, general $b_\sigma \neq 0$)**.** *Assume the hypotheses of Theorem 11 and additionally $n^2/N \to 0$. Then, with $\boldsymbol{K}_c = \boldsymbol{P}\boldsymbol{K}\boldsymbol{P}$,*

$$\left\| \boldsymbol{K}_c - \left(\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A}\boldsymbol{G}\boldsymbol{A}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{C}\boldsymbol{H}\boldsymbol{C}\boldsymbol{P} + \alpha_0 \boldsymbol{P} \right) \right\| = o_{\mathbb{P}}(1).$$

Proof. Triangle inequality:

$$\begin{aligned}&\left\|\boldsymbol{K}_c - \left(\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{P} + \boldsymbol{PAGAP} + \boldsymbol{PCHCP} + \alpha_0\boldsymbol{P}\right)\right\| \\ \le &\|\boldsymbol{K}_c - \boldsymbol{P}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{P}\| + \left\|\boldsymbol{P}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{P} - \left(\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{P} + \boldsymbol{PAGAP} + \boldsymbol{PCHCP} + \alpha_0\boldsymbol{P}\right)\right\|.\end{aligned}$$

The first term is $o_{\mathbb{P}}(1)$ by Proposition 66 (since $n/N \to 0$). The second term is $o_{\mathbb{P}}(1)$ by Theorem 11. Hence we complete the proof. □

Corollary 72 (QE in the quadratic regime when $b_\sigma = 0$)**.** *Assume the hypotheses of Proposition 71. Assume in addition that*

$$\mathbb{E}[\xi\sigma(\xi)] = b_\sigma = 0 \quad \textit{and} \quad \mathbb{E}[\xi\sigma(s\xi)] = 0 \textit{ for all } s \in [1-\delta, 1+\delta]$$

for some fixed $\delta > 0$ (e.g. this holds if σ is even*). Then $\boldsymbol{A} = \boldsymbol{0}$ and*

$$\left\|\boldsymbol{K}_c - \left(\boldsymbol{P}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{P} + \boldsymbol{PCHCP} + \alpha_0\boldsymbol{P}\right)\right\| = o_{\mathbb{P}}(1).$$

If furthermore $\mathbb{E}[\sigma(s\xi)] = 0$ for all $s \in [1-\delta, 1+\delta]$ (e.g. σ is odd*), then $\boldsymbol{\mu} = \boldsymbol{0}$ and*

$$\|\boldsymbol{K}_c - (\boldsymbol{PCHCP} + \alpha_0\boldsymbol{P})\| = o_{\mathbb{P}}(1).$$

Finally, by Lemma 24, $\boldsymbol{C} = \zeta_2(\sigma)\boldsymbol{I}_n + o_{\mathbb{P}}(1)$ and $\boldsymbol{PCHCP} = \alpha_2\boldsymbol{H}_c + o_{\mathbb{P}}(1)$ with $\boldsymbol{H}_c := \boldsymbol{PHP}$, yielding the simplified QE

$$\|\boldsymbol{K}_c - (\alpha_0\boldsymbol{P} + \alpha_2\boldsymbol{H}_c)\| = o_{\mathbb{P}}(1).$$

Proof. If $\mathbb{E}[\xi\sigma(s\xi)] = 0$ for all relevant s, then $\zeta_1^{(i)} = 0$ for all i, hence $\boldsymbol{D}_1 = \boldsymbol{0}$ and $\boldsymbol{A} = \boldsymbol{0}$, so Proposition 71 reduces to the first displayed statement.

If also $\mathbb{E}[\sigma(s\xi)] = 0$ for all relevant s, then $\zeta_0^{(i)} = 0$ for all i, hence $\boldsymbol{\mu} = \boldsymbol{0}$, giving the second statement.

Finally, if $\max_i |s_i - 1| = o_{\mathbb{P}}(1)$, then by Lipschitz continuity of $s \mapsto \zeta_2(\sigma(s\cdot))$ in L^2, we have $\max_i |\zeta_2^{(i)} - \zeta_2(\sigma)| = o_{\mathbb{P}}(1)$ and $\max_i |s_i^{-2} - 1| = o_{\mathbb{P}}(1)$, hence $\|\boldsymbol{C} - \zeta_2(\sigma)\boldsymbol{I}\| = o_{\mathbb{P}}(1)$. Since $\|\boldsymbol{H}_c\| = O_{\mathbb{P}}(1)$ in the quadratic regime (it is a centered Gram matrix in dimension $\asymp d^2$),

$$\|\boldsymbol{PCHCP} - \zeta_2(\sigma)^2 \boldsymbol{PHP}\| \le \|\boldsymbol{C} - \zeta_2(\sigma)\boldsymbol{I}\| \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1),$$

and $\zeta_2(\sigma)^2 = \alpha_2 = c_\sigma^2/2$. This yields the last claim. □

Remark. Lemma 5.2 of [WZ21] shows $\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\| \lesssim B\,\varepsilon_n^2\sqrt{n}$ under (ε_n, B)-orthonormality and bounded $\|\boldsymbol{X}\| \le B$, where $\boldsymbol{\Phi}_0$ is a degree-3 Hermite-truncation polynomial in $\boldsymbol{X}^\top\boldsymbol{X}$ (with $f_k(x) = x^k$ entrywise). In our quadratic regime, by Lemma 24, $\varepsilon_n \asymp \sqrt{\log(n)/d}$ and $B = \sqrt{d}$, which is increading as $d \to \infty$. Hence, we instead: (i) work with the correlation matrix $\boldsymbol{\bar{R}}$ (bounded diagonals), (ii) keep the degree-2 term explicitly, and (iii) show degrees $k \ge 3$ collapse to $\boldsymbol{P}$ on the centered subspace. This is the regime-adapted version of the same Hermite-expansion idea.

3.3.21 Deterministic Equivalent of Spiked Kernel Model

Let $\mathrm{svec} : \mathbb{S}^d \to \mathbb{R}^p$ be the Frobenius-isometric symmetric vectorization, with $p = d(d+1)/2$ and

$$\langle \mathrm{svec}(\boldsymbol{A}), \mathrm{svec}(\boldsymbol{B}) \rangle = \langle \boldsymbol{A}, \boldsymbol{B} \rangle_F.$$

Define the lifted data matrix

$$\boldsymbol{Q} := [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n] \in \mathbb{R}^{p \times n}, \qquad \boldsymbol{q}_i := \mathrm{svec}(\boldsymbol{x}_i \boldsymbol{x}_i^\top).$$

Then

$$\boldsymbol{H} = \boldsymbol{Q}^\top \boldsymbol{Q}, \qquad \boldsymbol{H}_c = (\boldsymbol{QP})^\top (\boldsymbol{QP}).$$

This above formula is also called kernel trick, see Section 3.7 of [Ver18]. Kernel trick helps us make the nonlinear matrix (quadratic kernel matrix) into a linear sample covariance matrix/Gram

matrix. Then we can apply standard random matrix results to analyze spikes and ESD.

Lemma 73 (Unique label-dependent direction in the quadratic lift for XOR)**.** *Define*

$$\boldsymbol{s} := \frac{1}{\sqrt{2}}\mathrm{svec}(\boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top) \in \mathbb{R}^p, \qquad \|\boldsymbol{s}\|_2 = 1.$$

Then there exists a deterministic $\bar{\boldsymbol{q}} \in \mathbb{R}^p$ and a scalar $\delta_d > 0$ such that for all i,

$$\mathbb{E}[\boldsymbol{q}_i \mid y_i] = \bar{\boldsymbol{q}} + y_i \frac{\delta_d}{\sqrt{p}}\boldsymbol{s}, \qquad \delta_d = \frac{r^2}{2}\sqrt{1+\frac{1}{d}}.$$

In particular, $\delta_d \to \delta := r^2/2$ as $d \to \infty$, and one may write

$$\mathbb{E}[\boldsymbol{q}_i \mid y_i] = \bar{\boldsymbol{q}} + y_i \frac{\delta}{\sqrt{p}}\boldsymbol{s} + \boldsymbol{r}_{i,d}, \qquad \|\boldsymbol{r}_{i,d}\|_2 \le C\frac{1}{d\sqrt{p}}.$$

Moreover, centering removes $\bar{\boldsymbol{q}}$:

$$\boldsymbol{Q}\boldsymbol{P} = (\boldsymbol{Q} - \bar{\boldsymbol{q}}\,\mathbf{1}^\top)\boldsymbol{P}.$$

Proof. Fix i and condition on y_i (equivalently, on $a(i) \in \{1,2\}$). Expand

$$\boldsymbol{x}_i\boldsymbol{x}_i^\top = \boldsymbol{z}_i\boldsymbol{z}_i^\top + \boldsymbol{z}_i\boldsymbol{m}_i^\top + \boldsymbol{m}_i\boldsymbol{z}_i^\top + \boldsymbol{m}_i\boldsymbol{m}_i^\top.$$

Taking conditional expectation given y_i:

- $\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^\top] = \frac{1}{d}\boldsymbol{I}_d$.
- Since $\boldsymbol{z}_i$ is mean-zero and independent of $\boldsymbol{m}_i$, $\mathbb{E}[\boldsymbol{z}_i\boldsymbol{m}_i^\top + \boldsymbol{m}_i\boldsymbol{z}_i^\top \mid y_i] = \mathbf{0}$.
- Since $\boldsymbol{m}_i = \eta_i \frac{r}{\sqrt{d}}\boldsymbol{u}_{a(i)}$ and $\eta_i^2 = 1$,

$$\mathbb{E}[\boldsymbol{m}_i\boldsymbol{m}_i^\top \mid y_i] = \frac{r^2}{d}\,\boldsymbol{u}_{a(i)}\boldsymbol{u}_{a(i)}^\top = \frac{r^2}{2d}\big(\boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top\big) + y_i\frac{r^2}{2d}\big(\boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top\big).$$

Hence

$$\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top \mid y_i] = \underbrace{\Big(\frac{1}{d}\boldsymbol{I}_d + \frac{r^2}{2d}\big(\boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top\big)\Big)}_{=:\,\bar{\boldsymbol{\Sigma}}} + y_i \underbrace{\Big(\frac{r^2}{2d}\big(\boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top\big)\Big)}_{=:\,\Delta\boldsymbol{\Sigma}}.$$

Apply the linear map svec:

$$\mathbb{E}[\boldsymbol{q}_i \mid y_i] = \operatorname{svec}(\bar{\boldsymbol{\Sigma}}) + y_i \operatorname{svec}(\Delta\boldsymbol{\Sigma}).$$

Define $\bar{\boldsymbol{q}} := \operatorname{svec}(\bar{\boldsymbol{\Sigma}})$. For the label-dependent part,

$$\operatorname{svec}(\Delta\boldsymbol{\Sigma}) = \frac{r^2}{2d}\operatorname{svec}(\boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{u}_2\boldsymbol{u}_2^\top) = \frac{r^2}{2d}\sqrt{2}\,\boldsymbol{s} = \frac{r^2}{\sqrt{2}\,d}\,\boldsymbol{s}.$$

Since $p = d(d+1)/2$, we have

$$\frac{1}{\sqrt{p}} = \frac{\sqrt{2}}{d\sqrt{1+1/d}}, \qquad \text{so} \qquad \frac{r^2}{\sqrt{2}\,d} = \frac{r^2}{2}\sqrt{1+\frac{1}{d}}\cdot\frac{1}{\sqrt{p}} = \frac{\delta_d}{\sqrt{p}}.$$

This proves the exact identity with δ_d. Since $\delta_d - \delta = O(1/d)$ and $\|\boldsymbol{s}\|_2 = 1$, the residual $\boldsymbol{r}_{i,d}$ bound follows.

Finally, $\boldsymbol{P}\mathbf{1} = \mathbf{0}$ implies $\bar{\boldsymbol{q}}\mathbf{1}^\top\boldsymbol{P} = \mathbf{0}$, hence

$$(\boldsymbol{Q} - \bar{\boldsymbol{q}}\mathbf{1}^\top)\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{P}.$$

□

Lemma 74 (Lifted noise is asymptotically isotropic)**.** *Define the centered lifted noise*

$$\boldsymbol{\varepsilon}_i := \boldsymbol{q}_i - \mathbb{E}[\boldsymbol{q}_i \mid y_i] = \operatorname{svec}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top \mid y_i]\Big).$$

Then $(\boldsymbol{\varepsilon}_i)_{i=1}^n$ *are independent, mean-zero, and for each fixed* i*,*

$$\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top \mid y_i] = \frac{2}{d^2} \boldsymbol{I}_p + \frac{4r^2}{d^2} \boldsymbol{U}_{a(i)},$$

where $\boldsymbol{U}_{a(i)} \in \mathbb{R}^{p \times p}$ *is a deterministic PSD matrix (depending on* $\boldsymbol{u}_{a(i)}$*) satisfying*

$$\operatorname{rank}(\boldsymbol{U}_{a(i)}) \le d, \qquad \|\boldsymbol{U}_{a(i)}\|_{\mathrm{op}} \le 1.$$

Equivalently,

$$\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top \mid y_i] = \frac{1}{p} \boldsymbol{I}_p + \frac{2r^2}{p} \boldsymbol{U}_{a(i)} + O\Big(\frac{1}{pd}\Big) \quad \text{in } \|\cdot\|_{\mathrm{op}}.$$

In particular, the covariance is isotropic up to a rank-$O(d)$ perturbation *of size* $\Theta(1/p)$ *in operator norm. If* $\boldsymbol{\Pi}_{a(i)}$ *denotes the orthogonal projector onto* $\operatorname{Im}(\boldsymbol{U}_{a(i)})$*, then on the orthogonal complement* $\operatorname{Im}(\boldsymbol{U}_{a(i)})^\perp$*,*

$$(\boldsymbol{I}_p - \boldsymbol{\Pi}_{a(i)})\, \mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top \mid y_i]\, (\boldsymbol{I}_p - \boldsymbol{\Pi}_{a(i)}) = \frac{2}{d^2} (\boldsymbol{I}_p - \boldsymbol{\Pi}_{a(i)}).$$

Proof. Step 1 (independence and centering). The columns $\boldsymbol{x}_i$ are independent across i, hence so are $\boldsymbol{q}_i$ and $\boldsymbol{\varepsilon}_i$. By construction, $\mathbb{E}[\boldsymbol{\varepsilon}_i \mid y_i] = \boldsymbol{0}$.

Step 2 (explicit centered matrix representation). Fix i and condition on y_i (so $\boldsymbol{u} := \boldsymbol{u}_{a(i)}$ is fixed) and on η_i. Write $\boldsymbol{m} := \eta_i \frac{r}{\sqrt{d}} \boldsymbol{u}$ and $\boldsymbol{z} := \boldsymbol{z}_i$ so that $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{m}$. Using $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mid y_i] = \frac{1}{d} \boldsymbol{I}_d + \frac{r^2}{d} \boldsymbol{u}\boldsymbol{u}^\top$, we obtain

$$\boldsymbol{x}\boldsymbol{x}^\top - \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mid y_i] = \Big(\boldsymbol{z}\boldsymbol{z}^\top - \frac{1}{d} \boldsymbol{I}_d\Big) + \Big(\boldsymbol{z}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{z}^\top\Big).$$

Therefore,

$$\boldsymbol{\varepsilon}_i = \operatorname{svec}\Big(\boldsymbol{z}\boldsymbol{z}^\top - \frac{1}{d} \boldsymbol{I}_d\Big) + \operatorname{svec}\Big(\boldsymbol{z}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{z}^\top\Big).$$

Step 3 (covariance in test-matrix form). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d$ and set $\boldsymbol{a} := \operatorname{svec}(\boldsymbol{A})$, $\boldsymbol{b} := \operatorname{svec}(\boldsymbol{B})$.

By the svec isometry,

$$\langle \boldsymbol{\varepsilon}_i, \boldsymbol{a} \rangle = \left\langle \boldsymbol{z}\boldsymbol{z}^\top - \frac{1}{d}\boldsymbol{I}_d, \boldsymbol{A} \right\rangle_F + \langle \boldsymbol{z}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{z}^\top, \boldsymbol{A} \rangle_F.$$

Since $\boldsymbol{A}$ is symmetric,

$$\langle \boldsymbol{z}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{z}^\top, \boldsymbol{A} \rangle_F = \mathrm{tr}((\boldsymbol{z}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{z}^\top)\boldsymbol{A}) = 2\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{m}.$$

Hence

$$\langle \boldsymbol{\varepsilon}_i, \boldsymbol{a} \rangle = \underbrace{\left(\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} - \frac{1}{d}\mathrm{tr}(\boldsymbol{A}) \right)}_{=: T_2(\boldsymbol{A})} + \underbrace{2\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{m}}_{=: T_1(\boldsymbol{A})}.$$

Here $T_2(\boldsymbol{A})$ is a centered quadratic form in $\boldsymbol{z}$ (2nd chaos) and $T_1(\boldsymbol{A})$ is linear in $\boldsymbol{z}$ (1st chaos). By symmetry of Gaussian moments, $\mathbb{E}[T_2(\boldsymbol{A})T_1(\boldsymbol{B})] = 0$ for all $\boldsymbol{A}, \boldsymbol{B}$. Therefore

$$\mathbb{E}[\langle \boldsymbol{\varepsilon}_i, \boldsymbol{a} \rangle \langle \boldsymbol{\varepsilon}_i, \boldsymbol{b} \rangle \mid y_i] = \mathbb{E}[T_2(\boldsymbol{A})T_2(\boldsymbol{B})] + \mathbb{E}[T_1(\boldsymbol{A})T_1(\boldsymbol{B})].$$

Step 4 (compute each covariance term). Since $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$, a standard Wick/Isserlis computation gives

$$\mathbb{E}[T_2(\boldsymbol{A})T_2(\boldsymbol{B})] = \mathrm{Cov}(\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z}, \boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z}) = \frac{2}{d^2}\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}).$$

Also,

$$\mathbb{E}[T_1(\boldsymbol{A})T_1(\boldsymbol{B})] = 4\,\mathbb{E}[(\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{m})(\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{m})] = 4\,(\boldsymbol{A}\boldsymbol{m})^\top \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]\,(\boldsymbol{B}\boldsymbol{m}) = \frac{4}{d}\boldsymbol{m}^\top \boldsymbol{A}\boldsymbol{B}\boldsymbol{m}.$$

Since $\boldsymbol{m} = \eta_i \frac{r}{\sqrt{d}}\boldsymbol{u}$ and $\eta_i^2 = 1$,

$$\frac{4}{d}\boldsymbol{m}^\top \boldsymbol{A}\boldsymbol{B}\boldsymbol{m} = \frac{4r^2}{d^2}\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{B}\boldsymbol{u}.$$

Altogether,

$$\mathbb{E}[\langle \boldsymbol{\varepsilon}_i, \boldsymbol{a}\rangle \langle \boldsymbol{\varepsilon}_i, \boldsymbol{b}\rangle \mid y_i] = \frac{2}{d^2}\mathrm{tr}(\boldsymbol{AB}) + \frac{4r^2}{d^2}\,\boldsymbol{u}^\top \boldsymbol{ABu}.$$

Step 5 (identify $\boldsymbol{U}_{a(i)}$ and its rank bound). Define the linear operator $\mathscr{U}_{\boldsymbol{u}} : \mathbb{S}^d \to \mathbb{S}^d$ by

$$\langle \mathscr{U}_{\boldsymbol{u}}(\boldsymbol{A}), \boldsymbol{B}\rangle_F := \boldsymbol{u}^\top \boldsymbol{ABu} \qquad (\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d),$$

and let $\boldsymbol{U}_{a(i)}$ be its matrix representation in svec coordinates: for $\boldsymbol{a} = \mathrm{svec}(\boldsymbol{A})$, $\boldsymbol{b} = \mathrm{svec}(\boldsymbol{B})$,

$$\boldsymbol{a}^\top \boldsymbol{U}_{a(i)} \boldsymbol{b} = \boldsymbol{u}^\top \boldsymbol{ABu}.$$

Then the covariance identity becomes exactly

$$\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top \mid y_i] = \frac{2}{d^2}\boldsymbol{I}_p + \frac{4r^2}{d^2}\boldsymbol{U}_{a(i)}.$$

The map $\boldsymbol{A} \mapsto \boldsymbol{Au}$ has range in $\mathbb{R}^d$, so $\mathscr{U}_{\boldsymbol{u}}$ has rank at most d, hence $\mathrm{rank}(\boldsymbol{U}_{a(i)}) \le d$.

Moreover, for any $\boldsymbol{A} \in \mathbb{S}^d$ with $\|\boldsymbol{A}\|_F = 1$,

$$\boldsymbol{u}^\top \boldsymbol{A}^2 \boldsymbol{u} = \|\boldsymbol{Au}\|_2^2 \le \|\boldsymbol{A}\|_F^2 \|\boldsymbol{u}\|_2^2 = 1,$$

so $\|\mathscr{U}_{\boldsymbol{u}}\|_{\mathrm{op}} \le 1$ as an operator on $(\mathbb{S}^d, \|\cdot\|_F)$, which implies $\|\boldsymbol{U}_{a(i)}\|_{\mathrm{op}} \le 1$ in svec coordinates.

Finally, since $p = d(d+1)/2$, we have $\frac{2}{d^2} = \frac{1}{p} \cdot \frac{d+1}{d} = \frac{1}{p} + O(\frac{1}{pd})$, yielding the stated equivalent form. □

Proposition 75 (Decomposition of the centered lifted matrix)**.** *Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be balanced ($\boldsymbol{y}^\top \boldsymbol{1} = 0$), and define*

$$\boldsymbol{E} := [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]\boldsymbol{P} \in \mathbb{R}^{p \times n}.$$

Then

$$\boldsymbol{QP} = \frac{\delta}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + \boldsymbol{E} + \boldsymbol{R}_Q, \qquad \|\boldsymbol{R}_Q\|_{\mathrm{op}} = o_{\mathbb{P}}(1),$$

where $\delta = r^2/2$ *and* $\boldsymbol{s}$ *is as in Lemma 73. In fact, one may take the explicit remainder*

$$\boldsymbol{R}_Q = \Big(\frac{\delta_d - \delta}{\sqrt{p}}\Big)\boldsymbol{s}\boldsymbol{y}^\top, \qquad \|\boldsymbol{R}_Q\|_{\mathrm{op}} \le \Big|\frac{\delta_d - \delta}{\sqrt{p}}\Big| \, \|\boldsymbol{s}\|_2 \, \|\boldsymbol{y}\|_2 \lesssim \frac{1}{d} \cdot \sqrt{\frac{n}{p}} = o(1),$$

since $\delta_d - \delta = O(1/d)$ *and* $n/p \to \gamma \in (0, \infty)$.

Proof. For each i, by definition of $\boldsymbol{\varepsilon}_i$,

$$\boldsymbol{q}_i = \mathbb{E}[\boldsymbol{q}_i \mid y_i] + \boldsymbol{\varepsilon}_i.$$

By Lemma 73,

$$\mathbb{E}[\boldsymbol{q}_i \mid y_i] = \bar{\boldsymbol{q}} + y_i \frac{\delta_d}{\sqrt{p}}\boldsymbol{s}.$$

Stacking columns yields

$$\boldsymbol{Q} = \bar{\boldsymbol{q}}\,\mathbf{1}^\top + \frac{\delta_d}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n].$$

Right-multiply by $\boldsymbol{P}$. Using $\mathbf{1}^\top \boldsymbol{P} = \mathbf{0}^\top$ and $\boldsymbol{y}^\top \boldsymbol{P} = \boldsymbol{y}^\top$ (balanced labels),

$$\boldsymbol{QP} = \frac{\delta_d}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]\boldsymbol{P} = \frac{\delta}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + \boldsymbol{E} + \Big(\frac{\delta_d - \delta}{\sqrt{p}}\Big)\boldsymbol{s}\boldsymbol{y}^\top.$$

Define $\boldsymbol{R}_Q := \big(\frac{\delta_d - \delta}{\sqrt{p}}\big)\boldsymbol{s}\boldsymbol{y}^\top$. The displayed operator-norm bound follows since $\|\boldsymbol{s}\|_2 = 1$, $\|\boldsymbol{y}\|_2 = \sqrt{n}$, $\sqrt{p} \asymp d$, and $\delta_d - \delta = O(1/d)$. □

From the deterministic equivalence in Proposition 75, we obtain the spiked-rectangular form

$$\boldsymbol{QP} = \frac{\delta}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + \boldsymbol{E} + o_{\mathbb{P}}(1), \qquad \boldsymbol{E} := [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]\boldsymbol{P}.$$

Thus $\boldsymbol{H}_c = (\boldsymbol{QP})^\top(\boldsymbol{QP})$ behaves like a rank-one deformation of an isotropic Wishart Gram

matrix.

3.3.22 BBP Transition for Quadratic Regime

Let

$$p = \frac{d(d+1)}{2}, \qquad \gamma := \frac{n}{p} \to 2\kappa, \qquad \ell := \delta^2 = \frac{r^4}{4}, \qquad \theta := \delta\sqrt{\gamma} = \sqrt{\gamma\ell}.$$

Let $\lambda_+(\gamma) := (1+\sqrt{\gamma})^2$ denote the MP upper edge.

Proof of Theorem 9. By Lemmas 73–74, $\boldsymbol{QP} = \frac{\delta}{\sqrt{p}}\boldsymbol{s}\boldsymbol{y}^\top + \boldsymbol{E} + o_{\mathbb{P}}(1)$, with $\boldsymbol{E}$ having asymptotically isotropic columns. Therefore $\boldsymbol{H}_c = (\boldsymbol{QP})^\top(\boldsymbol{QP})$ is an $o_{\mathbb{P}}(1)$ perturbation of an isotropic spiked-Wishart Gram matrix. Hence, we can consider an equivalent model, spiked rectangular random matrix model, $\widetilde{\boldsymbol{Q}} = (1/\sqrt{p})\boldsymbol{G} + \theta\,\boldsymbol{s}(\boldsymbol{y}/\sqrt{n})^\top$. Hence, by [PWZ24], the limiting ESD of $\boldsymbol{K}$ and $\boldsymbol{H}$ will become MP law $\rho_\gamma^{\mathrm{MP}}$ whose upper edge is $\lambda_+(\gamma) := (1+\sqrt{\gamma})^2$. Also, from [BGN12], if $\theta > \gamma^{1/4}$, with outlier location $\lambda_{\mathrm{out}}(\gamma,\ell)$ and overlap $\mathrm{Align}(\gamma,\ell)$. Then applying standard Weyl's inequality yields the eigenvalue convergence; and Davis–Kahan Theorem yields the eigenvector overlap stability in the supercritical regime (where the outlier is separated from the bulk by a constant gap).

By Corollary 72,

$$\boldsymbol{K}_c = \alpha_0\boldsymbol{P} + \alpha_2\boldsymbol{H}_c + o_{\mathbb{P}}(1),$$

so on $\mathrm{Im}(\boldsymbol{P})$ this is an affine spectral map and preserves eigenvectors up to $o_{\mathbb{P}}(1)$. This completes the proof of $\boldsymbol{K}_c$. □

3.4 Numerical Experiments

The theorems presented in this chapter highlight several control parameters ("knobs") that govern whether outlier eigenvalues appear and whether the associated eigenvectors align with the XOR labels $\boldsymbol{y}$: *(1)* SNR r, *(2)* nonlinearity σ through parameters (b_σ, c_σ) in (3.2.3)

and (3.2.4), *(3)* weight structure $\boldsymbol{W}$ and $\boldsymbol{W}_1$ (random/pretrained), and *(4)* scaling regime relating (n,d,N). In this section, we present the numerical experiments and figures verifying each of the BBP phase transition results presented in Section 3.2.2. We also include an experiment with pre-trained CIFAR-10 features fed into the CK and used to linearly classify the XOR dataset.

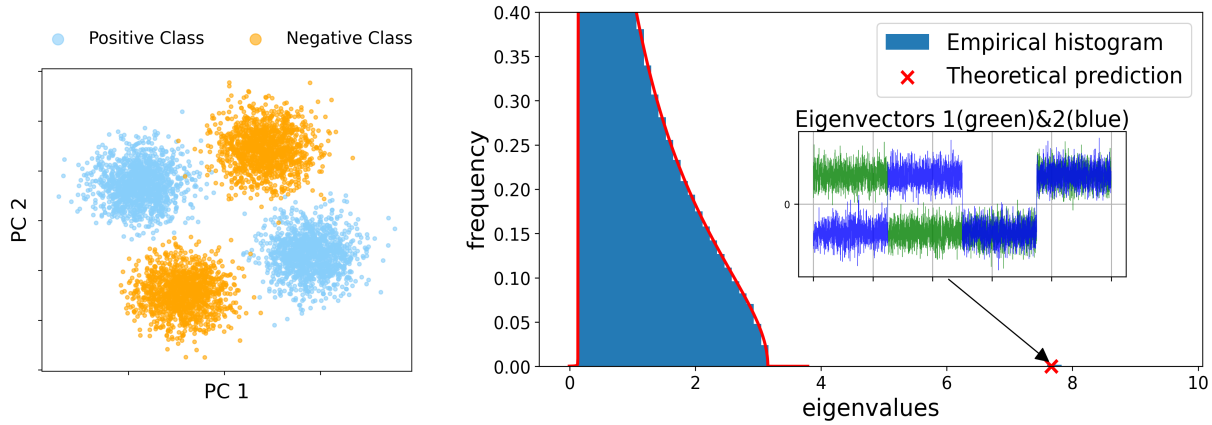


Figure 3.1. Finite-SNR proportional regime: failure of linear classification for XOR. We consider the CK matrix $\boldsymbol{K}$ in the regime $n \asymp d \asymp N$ with a finite SNR. **Left:** kernel-PCA visualization of the samples using the top two principle components (PCs) of $\boldsymbol{K}$ (each point is one sample, colored by its true binary label). The four visible clouds form the canonical XOR geometry: positive and negative classes occupy alternating clusters, so the leading two-dimensional spectral representation is still not linearly separable. **Right:** empirical eigenvalue histogram of $\boldsymbol{K}$ (blue bars) together with the theoretical prediction by Theorem 5 (red curve). The red $\times$ marks the predicted location of the two outlier eigenvalues of $\boldsymbol{K}$ with the same value by Theorem 5; the arrow links the sub-figure of the first two outlier eigenvectors plotted against sample index, whose four-block pattern tracks mixture components rather than the binary labels. Thus the outlier eigenspace is asymptotically orthogonal to the XOR label vector $\boldsymbol{y}$. Simulation parameters: $n = 5000$, $N = d = 15000$, SNR $r = 6.0$, and activation $\sigma(x) \propto \mathrm{ReLU}$.

3.4.1 Baseline Case in Section 3.2.3

Finite-SNR regime with random weights and proportional limits.

Figure 3.1 depicts the random CK spectra with XOR in the finite-SNR regime and proportional limit (Assumption 4), with theoretical spike predictions from Theorem 5. Notice that in this simulation, because of the choice of σ, there are no uninformative spikes from Theorem 5 (ii) and(iii). Consistent with Theorem 5 (iv) and(v), two leading eigenvectors are outliers and the sub-figure of Figure 3.1 exhibits non-trivial alignment with $\boldsymbol{v}_1, \boldsymbol{v}_2$ defined in (3.2.7), but remains orthogonal to the XOR-labels $\boldsymbol{y}$. Consequently, although we can apply

additional nonlinear transforms for leading eigenvectors to classify the data, linear classification of XOR via the eigenvectors of the CK fails in this regime. This case is our baseline and then we vary one knob at a time, reporting how the CK spectra change.

3.4.2 Knob I: Increase SNR

From Theorem 7, we generate the CK of XOR with $r = \Theta(n^{1/4})$ under $\sigma(x) \propto \sqrt{x^2+1} - 1$, which satisfies $b_\sigma = 0, c_\sigma \neq 0$. Figure 3.2 depicts the CK spectra and the principal component plots. CK has two $O(1)$ quadratic outliers. Since $b_\sigma = 0$, the diverging linear outliers are not present. Moreover, the second eigenvector exhibits two-class structure, and XOR becomes linearly classifiable via the CK eigenvectors.

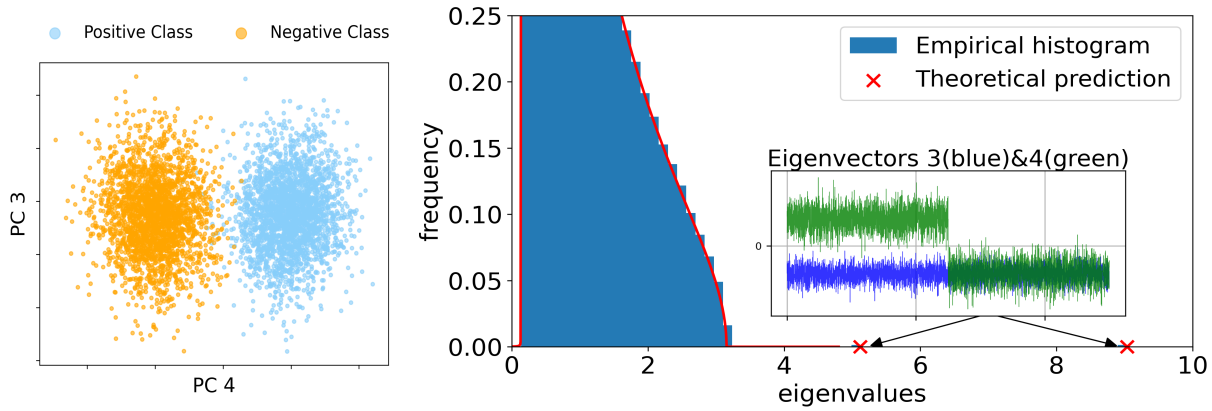


Figure 3.2. Large-SNR proportional regime: informative CK spikes recover the XOR labels. We consider the same proportional limit as in Figure 3.1, but now at large SNR $r = 35$, where the spectral predictions of Theorem 7 imply the emergence of quadratic informative outliers. All other parameters are the same as in Figure 3.1. **Left:** kernel-PCA visualization of the samples using the *3rd* and *4th* PCs of CK matrix $\boldsymbol{K}$ (each point is one sample, colored by its true binary label). In contrast to the finite-SNR case, these two spectral coordinates clearly separate the positive and negative classes by a linear boundary, showing that the informative outlier eigenspace now aligns with the XOR labels. **Right:** empirical eigenvalue histogram of $\boldsymbol{K}$ (blue bars) together with the predicted curve from Theorem 7 (red curve). The red $\times$ markers indicate the theoretical locations of the order-one informative outliers, and the arrows link the sub-figure of the 3rd and 4th empirical eigenvectors (blue and green) plotted against sample index. The block structure of the 4th eigenvector of $\boldsymbol{K}$ matches the binary label split, explaining why the associated outlier eigenspace is informative for linear classification of XOR. To focus on the order-one eigenvalues, we do not display the first two extremely large eigenvalues here for $\boldsymbol{K}$: $\widehat{\lambda}_1 = 155.6265$ and $\widehat{\lambda}_2 = 152.6605$.

Spectral/algorithmic phase transition as r increases.

The results of Section 3.2.4 imply that there is a phase transition when the SNR surpasses the critical value of $r = \Theta(n^{\frac{1}{4}})$. In this case, the quadratic component of σ is no longer negligible, and two further eigenvalues emerge from the bulk with eigenvectors that asymptotically align with XOR-labels. To support this, we simulate the CK under increasing SNR in Figure 3.3. Plots are drawn with $\sigma(x) \propto \sqrt{x^2+1} - 1$, which has $b_\sigma = 0$ and $c_\sigma \neq 0$. Since $b_\sigma = 0$, the two linear eigenvalues are not present. Instead only the two $O(1)$ quadratic eigenvalues isolate from the bulk. Figure 3.3 depicts the CK principal component plots of XOR exhibiting a phase transition in SNR, during which XOR becomes linearly separable. This transition roughly coincides with the two quadratic eigenvalues with XOR-label aligned eigenvectors emerging from the bulk of the CK, depicted on the right of Figure 3.3.

3.4.3 Knob II: Activation nonlinearity σ

Varying c_σ with $b_\sigma = 0$: uninformative spikes can emerge.

In order to emphasize the role of non-linearity σ on the possible emergence of uninformative spikes in the CK, we consider the single-parameter family of activation functions

$$f_\alpha(x) = \frac{\cos(\alpha x) - e^{-\alpha^2/2}}{\sqrt{e^{-\alpha^2}\left((\cosh(\alpha^2) - 1\right)}} \tag{3.4.1}$$

from [BP22]. These functions satisfy Assumption 3, with $b_\sigma = 0$ for all choices of α while c_σ varies with α. Figure 3.4 plots the spectra of the CK under f_α and the null model $r = 0$ for decreasing values of α. As α decreases two outlier spikes emerge in the CK. These two uninformative spikes correspond to the covariance spike and mean spike from Theorem 5. The bulk distribution remains MP, which is plotted in red.

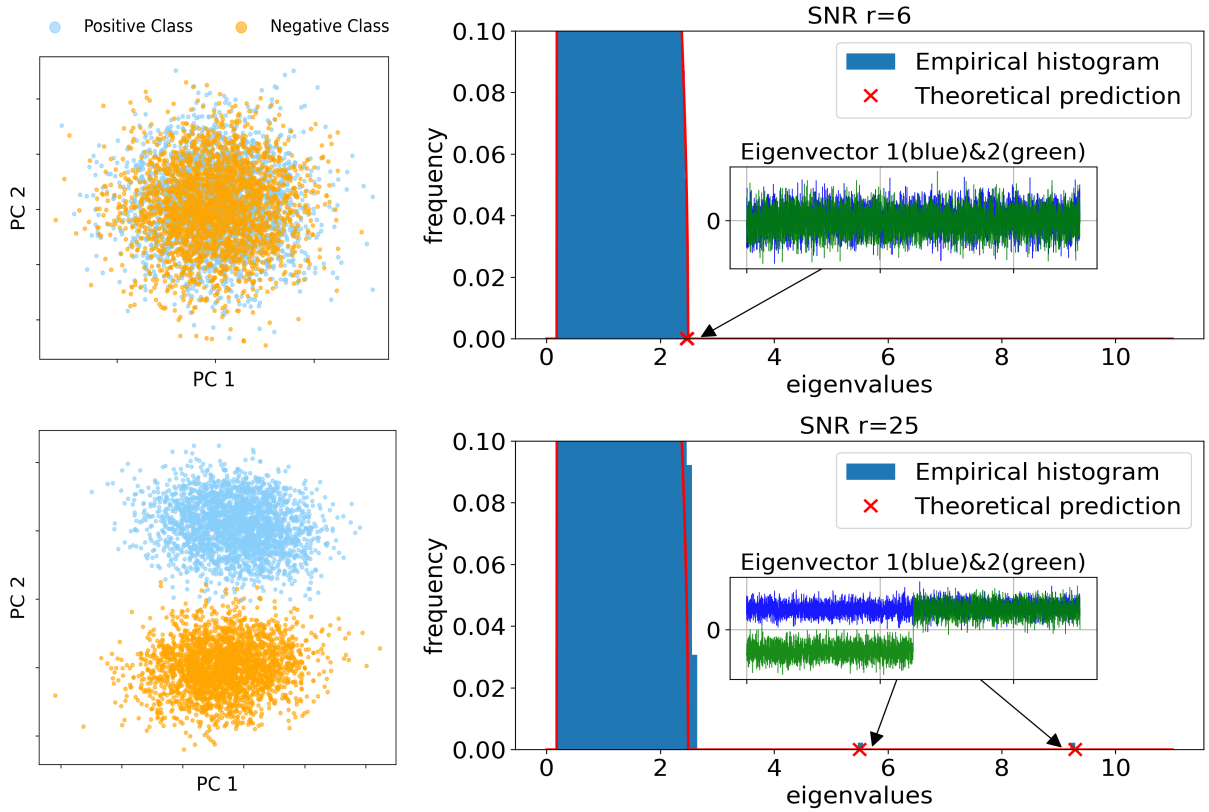


Figure 3.3. CK Spectra Phase Transition: **(Top)** CK spectra under finite SNR regime $(r = 6)$, no eigenvalues separate from the bulk. **(Bottom)** CK spectra under large SNR regime $(r = 25)$. Two quadratic eigenvalues separate from bulk, with the second leading eigenvector aligning with XOR-labels. Plots generated with $n = 5000, N = d = 15000$ and $\sigma(x) \propto \sqrt{x^2+1} - 1$, which satisfies $b_\sigma = 0, c_\sigma \neq 0$.

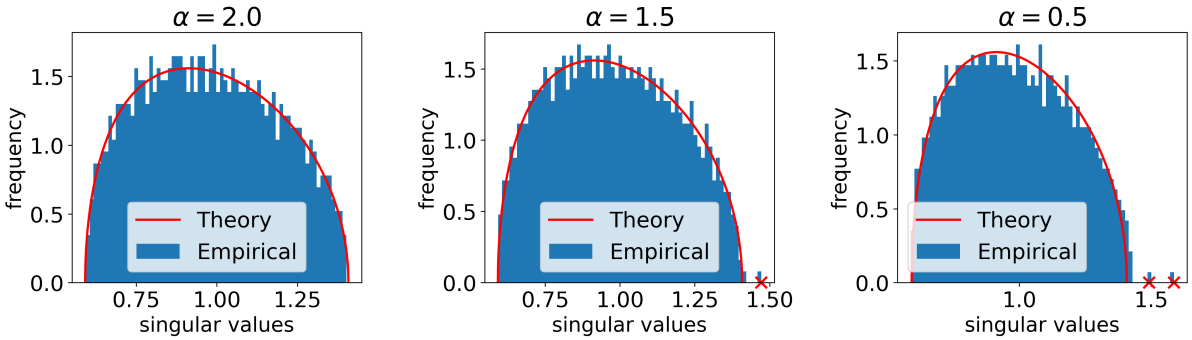


Figure 3.4. CK spectra under activation function $f_\alpha(x)$ defined in (3.4.1) and null model SNR $r = 0$. As α decreases, c_σ increases and two (uninformative) spikes emerge. The bulk remains close to MP (plotted in red). Plots are generated with $n = 6000, N = d = 1000$.

Switching off the quadratic channel ($c_\sigma = 0$): XOR remains non-separable even at large SNR.

On the other hand, if the activation function σ satisfies $c_\sigma = 0$ the quadratic feature channel in the CK is absent. As discussed in Theorem 7, this implies the CK cannot be used to linearly classify XOR for any SNR. Figure 3.5 depicts the principal components of the CK under $\sigma \propto \tanh$ for increasing SNR. Even in the large-SNR regime, the two classes in the CK space do not linearly separate. Indeed by Theorem 7, only two eigenvalues from the linear channel of the CK will isolate from the bulk in the large SNR regime. Moreover, their eigenvectors remain asymptotically orthogonal to XOR-labels $\boldsymbol{y}$, as depicted in the right side of Figure 3.5.

3.4.4 Knob III: Weight structure

Spiked weight case: controlled feature injection into $\boldsymbol{W}_1$.

We generate a rank-one spike to the random initialized weights $\boldsymbol{W}_1 = \boldsymbol{W} + \theta \boldsymbol{a}\boldsymbol{b}^\top$, and plot the CK and its principal components in Figure 3.6, under $\sigma(x) \propto$ ReLU and XOR with SNR $r = 6.0$. As Theorem 8 predicts, the leading eigenvalue is $O(\sqrt{N})$. The second and third spiked eigenvalues remain $O(1)$, with the second leading eigenvector exhibiting two-class structure, which allows for linear classification of XOR in the CK space.

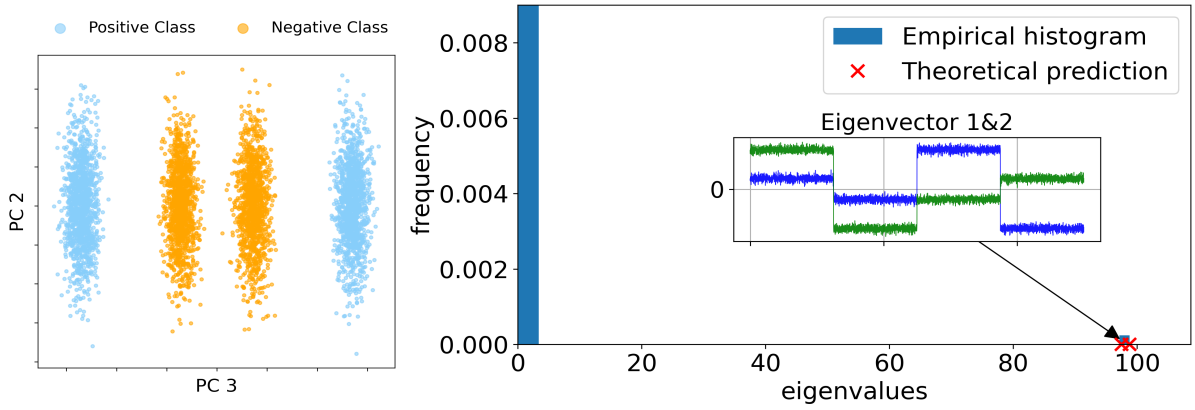


Figure 3.5. Large-SNR regime with an odd activation: separated CK spikes can appear without revealing the XOR labels. We consider the CK matrix $\boldsymbol{K}$ with activation $\sigma(x) \propto \tanh(x)$, for which the quadratic Hermite coefficient vanishes ($c_\sigma = 0$). All settings are the same as in Figure 3.2 except σ. In this case Theorem 7 predicts only *linear* outliers, with no quadratic label-aligned spike. **Left:** kernel-PCA visualization of the samples using the *2nd* and *3rd* PCs of $\boldsymbol{K}$ (one point per sample, colored by its binary XOR label). The samples split into four vertical clusters: the two outer clusters belong to the positive class and the two inner clusters belong to the negative class. Thus the leading nontrivial spectral coordinates remain *not* linearly separable with respect to the XOR labels, despite the presence of isolated spikes. **Right:** empirical eigenvalue histogram of $\boldsymbol{K}$ (blue bars) together with the theoretical locations of the two isolated linear outliers predicted by the large-SNR asymptotic theory (red $\times$ markers).Two spikes appears with very large order and the arrows connect these predicted outlier locations to the corresponding empirical outlier eigenvectors plotted against sample index. Their approximately piecewise-constant four-block pattern reflects the four Gaussian-mixture components (equivalently, the two underlying coordinate-sign directions) rather than the binary XOR partition. Consequently, the outlier eigenspace is spectrally visible but asymptotically orthogonal to the XOR label vector $\boldsymbol{y}$.

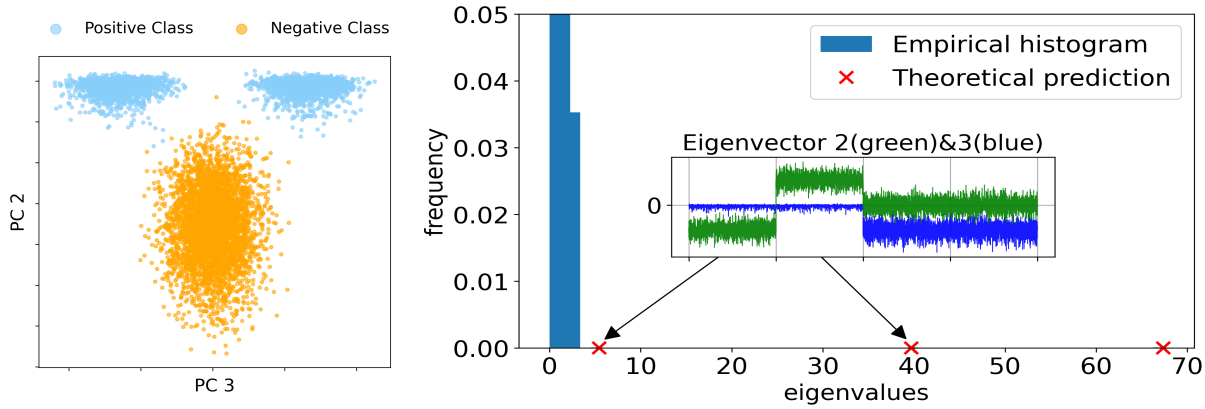


Figure 3.6. Spiked weight regime: a dominant weight-induced spike coexists aligned with XOR labels. We consider the CK matrix $\boldsymbol{K}$ built from the spiked weight $\boldsymbol{W}_1 = \boldsymbol{W} + \theta \boldsymbol{a}\boldsymbol{b}^\top$. where the rank-one perturbation creates an additional large spectral direction for XOR linear classification. **Left:** kernel-PCA visualization of the samples using the *2nd* and *3rd* PCs of K (each point is one sample, colored by its true binary label). These two coordinates linearly separate the positive and negative classes, indicating that the informative label-aligned structure is carried by the second and third spectral directions. **Right:** empirical eigenvalue histogram of $\boldsymbol{K}$ (blue bars) together with the isolated eigenvalue locations predicted by Theorem 8 (red $\times$ markers). The arrows link the two informative outliers to the empirical eigenvectors shown in the sub-figure: empirical eigenvectors 2 (green) and 3 (blue) plotted against sample index. The far-right red marker corresponds to the largest spike induced by the rank-one perturbation in the weights; this dominant direction is separated from the two lower informative spikes and is therefore not used in the left-panel embedding. Simulation parameters: $N = n = 8000$, $d = 4000$, SNR $r = 6.0$, activation $\sigma(x) \propto$ ReLU, and weight spike $\theta = 18.0$, $b = \mathbf{1}_d$, $\boldsymbol{a} \sim \mathcal{N}(0, \frac{1}{N}\boldsymbol{I}_N)$.

CIFAR-10 Pretrained Weights Experiment

From the spiked weight studied in Section 3.2.5, we consider the CK under pretrained weights extracted from a 20-layer NN trained on the CIFAR-2 (40 epochs with final test accuracy 88%). Figure 3.7 depicts 1st layer trained weight and its CK spectra tested on XOR. The emergence of spikes in the spectra of weights during the training process suggests the model is learning task-dependent features, a phenomena known as feature learning [FCB22] in NNs. In contrast to Theorem 5, the pretrained CK exhibits four spikes, with the fourth leading eigenvector aligning with XOR-labels. After sufficient training, the fourth leading eigenvector transitions from pure noise to near-perfect alignment with XOR-labels, enabling linear classification. Our experiment suggests that under certain training dynamics and parameter tuning, there exists a critical phase transition in learning parallel to a spectral transition.

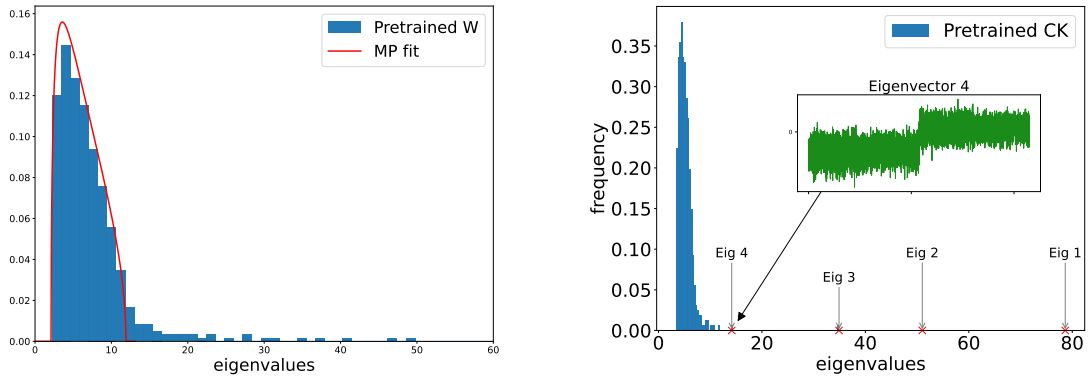


Figure 3.7. CIFAR-2 pretrained model: **(Left)** ESD of layer FC1 weights after 40 epochs. Overlaid in red is the MP distribution. **(Right)** ESD of CK with pretrained weights. Plots in this section are generated with $n = 10000, d = 3072, N = 512$ and XOR with SNR $r = 5.25$ under $\sigma(x) \propto$ ReLU.

A recent line of works [MM21, MM19a] have tracked the evolution of the Empirical Spectral Distribution (ESD) of weight matrices throughout the training process of Deep Neural Networks. We focus on two main phases of ESD during training: the so-called "Bulk+Spikes" model, and Bulk-Decay model. The emergence of spikes in the weight spectra is strongly related to the feature-learning ability of neural networks [FCB22], and suggests the spikes carry task-relevant information.

In order to study the spectra of the CK under pretrained weights, we trained a $L = 20$-

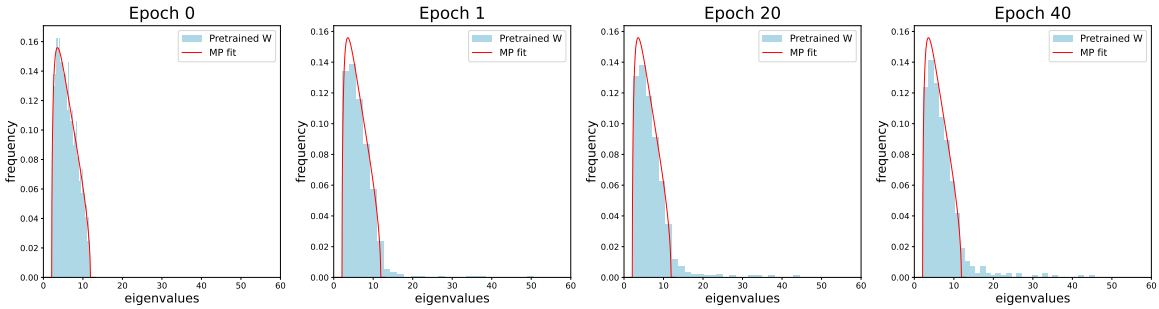


Figure 3.8. Spectra over training: Zoomed in ESD of the CIFAR-2 trained weights $\widetilde{\boldsymbol{W}}_1$ tracked through training from Epoch 0 to Epoch 40. MP distribution overlaid in red. By epoch 40, significant spikes begin to bleed out from the bulk distribution. Plots in this section are generated with $n = 10000, d = 3072, N = 512$.

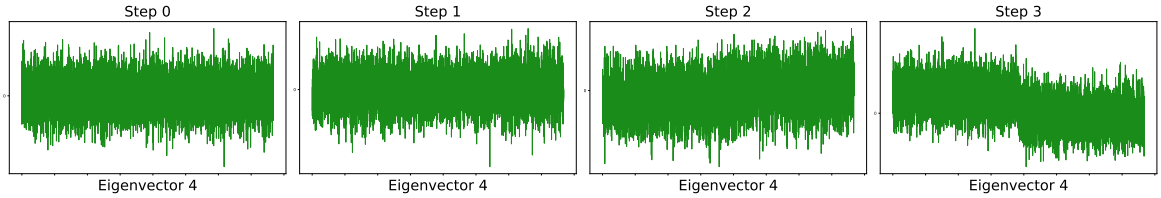


Figure 3.9. Eigenvector alignment over training: Fourth leading eigenvector of CK with pre-trained weights $\widetilde{\boldsymbol{W}}_1$ and XOR with SNR $r = 6$ plotted over training. Weights are frozen after each step of gradient descent, to emphasize phase transition that occurs around step 3, leading to non-trivial alignment with XOR labels $\boldsymbol{y}$.

layer fully connected Neural Network on the CIFAR image dataset, with input dimension $d_0 = 3072$ and hidden dimensions $N_\ell = 512$, for $\ell = 1, \cdots, L$. To extract binary-classification relevant features from the trained weights, we restrict the CIFAR-10 dataset to two classes. The model was trained using stochastic gradient descent (SGD) with momentum, with a learning rate $\text{lr} = 0.05$ and batch size $b = 128$. Our model achieved approximately 88% accuracy on an independent test set after 40 epochs of training. Figure 3.8 depicts the spectra of FC1 layer weights, tracked over each pass of the training dataset. Early phases of training exhibit the "Bulk+Spikes" model: the bulk distribution remains close to random MP initialization, with several outlier spikes beginning to emerge past the right edge of the MP distribution. In later training phases, the bulk begins to bleed out from the MP distribution, with significant outlier spikes emerging.

Our results empirically suggest that the LE approximation of the CK breaks down even during the early stages of training. Indeed, XOR becomes almost linearly separable in the CK component space after just one epoch of training. After further training, a spiked eigenvalue emerges with XOR-label aligned eigenvector. Figure 3.9 depicts the fourth leading eigenvector of the CK under weights frozen after each epoch of training, with a phase transition occurring between Epoch 20 and 30. The results of these experiments suggests that under certain training regimes and parameter tuning, there exists a critical phase transition in the learning process for which the emergence of spikes in weights becomes sufficient for quadratic channels in the CK to become non-negligible. One possible explanation is that training the neural network induces spiked eigenvalues with eigenvectors that align with XOR labels, even though the model is not trained on the XOR problem directly. Theorem 8 models this by artificially introducing a large spike in the weights, which induces a spiked CK eigenvalue with XOR-label aligned eigenvector.

In conclusion, we have introduced a quadratic equivalent framework and illustrated its use by characterizing when nonlinear structure becomes spectrally visible in the CK matrix. For the nonlinear XOR problem, we show that in the proportional limit with finite SNR, the CK can exhibit outliers, but their eigenvectors are asymptotically unaligned with labels. In contrast,

by increasing SNR, sample complexity, or adding pretrained structure in weight, the quadratic nonlinearity contributes emergence of quadratic informative outliers that enable linear spectral classification. Our results highlight the complex and subtle relationship between common ML training "knobs" and the emergence of spikes in the spectrum of ML models. Extending these quadratic equivalents to more general scalings, deeper networks, and other forms of structured data is an interesting direction.

Acknowledgments

Chapter 3 is extracted from the paper "Collin Cranston, Zhichao Wang, Todd Kemp, Michael W. Mahoney. Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data, Proceedings of the 39th Annual Conference on Learning Theory, *Proceedings of Machine Learning Research (PMLR)*. arXiv:2605.29669" To appear. The dissertation author is the co-author of this paper.

Bibliography

[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.

[ALP19] Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

[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.

[AS64] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

[AZVP26] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4):043404, April 2026.

[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.

[BAP24] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws, 2024.

[BCP21] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks, 2021.

[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.

[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), 2019.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BM24] Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and NTK features. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4267–4299. PMLR, 21–27 Jul 2024.

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[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. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4132–4179. PMLR, 2023.

[BPQ+24] Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. Neural networks learn statistics of increasing complexity. *arXiv preprint arXiv:2402.04362*, 2024.

[Bre73] Richard P. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, NJ, 1973. Reprinted by Dover Publications, 2002.

[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.

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

[CBG16] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393 – 1454, 2016.

[CCBG22] Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks, 2022.

[CDV07a] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[CDV07b] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[CL06] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[CL22] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. https://zhenyu-liao.github.io/book/.

[CLKZ22] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114004, November 2022.

[CM24] Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879–2912, 2024.

[COB19] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in dif-

ferentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

[Cou19] Romain Couillet. High dimensional robust classification: A random matrix analysis. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 420–424. IEEE, 2019.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[DD25] Samet Demir and Zafer Dogan. Asymptotic analysis of two-layer neural networks after one gradient step under gaussian mixtures data with structure. In *International Conference on Learning Representations*, 2025.

[DHK22] Bruce K. Driver, Brian C. Hall, and Todd Kemp. The brown measure of the free multiplicative brownian motion. *Probab. Theory Relat. Fields 184, 209–273 (2022)*, 2022.

[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.

[DL20] Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

[DLM] *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.4 of 2025-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, ...

[DLM24] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[DM24] Issa Dabo and Camille Male. A traffic approach for profiled pennington-worah matrices. *arXiv preprint arXiv:2409.13433*, 2024.

[DV13] Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.

[DZPS19] 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. *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.

[Eva10] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, 2010.

[FCB22] Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *arXiv preprint arXiv:2202.07626*, 2022.

[Fel23] Michael J Feldman. Spectral properties of elementwise-transformed spiked matrices. *arXiv preprint arXiv:2311.02040*, 2023.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, 2019.

[FMPW26] Zhou Fan, Renyuan Ma, Elliot Paquette, and Zhichao Wang. Anisotropic local law for non-separable sample covariance matrices. *arXiv preprint arXiv:2602.17960*, 2026.

[FW20] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

[GDZ+22] Lingyu Gu, Yongqi Du, Yuan Zhang, Di Xie, Shiliang Pu, Robert C Qiu, and Zhenyu Liao. "lossless" compression of deep neural networks: A high-dimensional neural tangent kernel approach. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[GHW79] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[Gir01] Vyacheslav L. Girko. *Theory of stochastic canonical equations*. 'Springer Dordrecht', 2001.

[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.

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and

Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 2019.

[Gla23] Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

[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.

[GLR+22] 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. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9111–9121, 2019.

[GP26] Alice Guionnet and Vanessa Piccolo. Global law of conjugate kernel random matrices with heavy-tailed weights. *Electronic Journal of Probability*, 31:1–58, 2026.

[GSS21] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26:1 – 22, 2021.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[HG16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[HL22a] Hong Hu and Yue M Lu. Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv preprint arXiv:2205.06798*, 2022.

[HL22b] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[HMZ+23] Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023.

[HSW+21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 10, 2021.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome 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.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8580–8589, 2018.

[JKN26] David A. Jekel, Todd A. Kemp, and Evangelos A. Nikitopoulos. A martingale approach to noncommutative stochastic calculus. *Journal of Functional Analysis*, 290(2):111200, January 2026.

[JL09] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 2002.

[JSS+20] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578, 2020.

[Kem11] Todd Kemp. Introduction to random matrix theory. Lecture notes, Math 247A,

University of California, San Diego, 2011. https://mathweb.ucsd.edu/~tkemp/RM.LectureNotes.pdf.

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[KY17] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017.

[LBN+18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes, 2018.

[LBOM12] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[LC18] 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. *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 579–583, 12 2019.

[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. *Advances in Neural Information Processing Systems*, 33:13939–13950, 2020.

[LGC+21] 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*, 34:18137–18151, 2021.

[LHY25] Xiangchao Li, Xiao Han, and Qing Yang. Eigen analysis of conjugate kernel and neural tangent kernel. In *Forty-second International Conference on Machine Learning*, 2025.

[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.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

[LM21] Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

[LM25] Zhenyu Liao and Michael W Mahoney. Random matrix theory for deep learning: Beyond eigenvalues of linear models. *arXiv preprint arXiv:2506.13139*, 2025.

[LXS+20] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002, December 2020.

[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.

[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.

[Mis22] Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

[MJBM22] Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the NTK via a power series expansion. *arXiv preprint arXiv:2211.07844*, 2022.

[MM19a] C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4284–4293, 2019.

[MM19b] 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*, 2019.

[MM20] C. H. Martin and M. W. Mahoney. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 20th SIAM International Conference on Data Mining*, 2020.

[MM21] C. H. Martin and M. 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.

[MM23] Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural

networks, 2023.

[MMM22] 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*, 59:3–84, 2022.

[MP67] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[MPM21] C. H. Martin, T. S. Peng, and M. W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(4122):1–13, 2021.

[MRS22] Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws, 2022.

[MS17] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[MS22] Andrea Montanari and Basil Saeed. Universality of empirical risk minimization, 2022.

[MS24a] Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator, 2024.

[MS24b] Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

[Nad08] Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 12 2008.

[Nea95] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1995.

[NJW01] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[NM20] Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *34th Conference on Neural Information Processing Systems*, volume 33, 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.

[Øks98] Bernt K. Øksendal. *Stochastic differential equations: an introduction with applications*. Springer, Berlin, 1998.

[Péc06] Sandrine Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.

[Péc19] S Péché. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24, 2019.

[PPXP25] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws, 2025.

[PRT22] Pratik Patil, Alessandro Rinaldo, and Ryan Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 6087–6120. PMLR, 2022.

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

[PWRT21] Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3178–3186. PMLR, 13–15 Apr 2021.

[PWZ24] Parthe Pandit, Zhichao Wang, and Yizhe Zhu. Universality of kernel random matrices and kernel regression in the quadratic regime. *arXiv preprint arXiv:2408.01062*, 2024.

[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.

[RGLG24] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. A distributional simplicity bias in the learning dynamics of transformers. *Advances in Neural Information Processing Systems*, 37:96207–96228, 2024.

[Ros70] Haskell P Rosenthal. On the subspaces of l^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

[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.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[Rud97] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997.

[SB95] Jack W. Silverstein and Zhidong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192, 1995.

[SBL16] Levent Sagun, Léon Bottou, and Yann LeCun. Singularity of the hessian in deep learning. *CoRR*, abs/1611.07476, 2016.

[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. In *International Conference on Machine Learning*, pages 30285–30320. PMLR, 2023.

[SK20] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold, 2020.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[TE+51] Francesco Giacomo Tricomi, Arthur Erdélyi, et al. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math*, 1(1):133–142, 1951.

[TGM25] Hossein Taheri, Avishek Ghosh, and Arya Mazumdar. On the theory of con-

tinual learning with gradient descent for neural networks. *arXiv preprint arXiv:2510.05573*, 2025.

[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.

[Wed72] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[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.

[WHL+25] Garrett G Wen, Hong Hu, Yue M Lu, Zhou Fan, and Theodor Misiakiewicz. When does gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling. *arXiv preprint arXiv:2512.03325*, 2025.

[WHS22] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.

[Wil96] Christopher Williams. Computing with infinite networks. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[Wil25] Andrew Gordon Wilson. Position: Deep learning is not so mysterious or different. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[WL25] Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: spectrum and generalization, 2025.

[WWF24] Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks, 2024.

[WZ21] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv:2109.09304*, 2021.

[WZ23] Zhichao Wang and Yizhe Zhu. Overparameterized random feature regression with nearly orthogonal data. In *International Conference on Artificial Intelligence and Statistics*, pages 8463–8493. PMLR, 2023.

[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.

[XMRH21] Ji Xu, Arian Maleki, Kamiar Rahnama Rad, and Daniel Hsu. Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030, 2021.

[YTH+22] Y. Yang, R. Theisen, L. Hodgkinson, J. E. Gonzalez, K. Ramchandran, C. H. Martin, and M. W. Mahoney. Test accuracy vs. generalization gap: Model selection in nlp without accessing training or testing data (evaluating natural language processing models with generalization metrics that do not need access to any training or testing data). Technical Report Preprint: arXiv:2202.02842, arXiv, 2022.