
Eigen Analysis of Conjugate Kernel and Neural Tangent Kernel

Xiangchao Li^{*1} Xiao Han^{*2} Qing Yang²

Abstract

In this paper, we investigate deep feedforward neural networks with random weights. The input data matrix $\mathbf{X}$ is drawn from a Gaussian mixture model. We demonstrate that certain eigenvalues of the conjugate kernel and neural tangent kernel may lie outside the support of their limiting spectral measures in the high-dimensional regime. The existence and asymptotic positions of such isolated eigenvalues are rigorously analyzed. Furthermore, we provide a precise characterization of the entrywise limit of the projection matrix onto the eigenspace associated with these isolated eigenvalues. Our findings reveal that the eigenspace captures inherent group features present in $\mathbf{X}$. This study offers a quantitative analysis of how group features from the input data evolve through hidden layers in randomly weighted neural networks.

1. Introduction

In the last decade, deep neural networks (DNNs) have demonstrated dominating performance in various machine-learning tasks such as computer vision (Krizhevsky et al., 2012; He et al., 2016), natural language processing (Amodi et al., 2016; Brown, 2020; Graves et al., 2013; Torfi et al., 2020) and game playing (Silver et al., 2016). One of the empirical findings is their ability to adapt to the features present in the training data, which is considered a fundamental reason for their superior performance (see, for instance, (Bengio et al., 2013; Donahue et al., 2016; Moniri et al., 2023)). Nowadays, many theoretical results have been established to understand the training and generalization of neural networks. In particular, significant research efforts (Abbe et al., 2022; Ba et al., 2022; Bai & Lee, 2019; Damian et al., 2022; Dandi et al., 2024; Mousavi-Hosseini et al., 2022;

Radhakrishnan et al., 2022; Shi et al., 2022) have demonstrated the advantages of feature learning in neural networks. However, understanding the mechanism of DNNs remains challenging and the current theoretical framework is incomplete, largely due to the heavy over-parameterization and high non-convexity of modern neural networks (Allen-Zhu et al., 2019).

Among the theoretical analyses, the Neural Tangent Kernel (NTK, Jacot et al., 2018) has become a powerful tool for understanding neural networks. It is well established that NTK, which is associated with the nonlinear feature map (Bietti & Mairal, 2019), is able to describe the dynamics of heavily over-parameterized neural networks under a specific initialization (Jacot et al., 2018; Du et al., 2019). It also provides insights into the convergence and generalization properties of very wide DNNs (Xie et al., 2017). The NTK theory has also been extended to other network architectures, such as deep residual networks (Belfer et al., 2024) and deep attention networks (Hron et al., 2020).

The eigenvalues and eigenvectors of both the NTK and its closely related counterpart, the Conjugate Kernel (CK), are crucial in understanding the training behavior and generalization performance of the underlying DNN (Fan & Wang, 2020; Yang & Salman, 2019; Yang, 2020; Engel et al., 2023). A recent line of work has analyzed the spectral properties of such kernel matrices. For instance, one may refer to Pennington & Worah (2017); Fan & Wang (2020); Wang et al. (2023); Wang & Zhu (2024); Belfer et al. (2024); Wang et al. (2024); Murray et al. (2022). The spectral properties of these kernel matrices, particularly spiked or extreme eigenvalues, are often closely associated with certain properties of neural networks, including their memorization capacity (Nguyen et al., 2021), training and generalization performance (Liao et al., 2020), and their ability to capture the low-dimensional signal structure inherent in the learning problem (Benigni & P      , 2022; Wang et al., 2024; Ba et al., 2023). We are interested in investigating the limiting properties of the isolated eigenvalues and their corresponding eigenvectors of kernel matrices in the high-dimensional regime. Through this research, we aim to gain deeper insights into how features of the input data propagate through DNNs.

In this paper, we study the CK and NTK matrices of deep fully-connected neural networks. Recent studies have

^{*}Equal contribution ¹School of Management, University of Science and Technology of China ²International Institute of Finance, School of Management, University of Science and Technology of China. Correspondence to: Qing Yang <yangq@ustc.edu.cn>.

demonstrated that, as the network width approaches infinity, the empirical CK and NTK converge to their expected values (Arora et al., 2019; Jacot et al., 2018). Therefore, we focus exclusively on the expected forms of these kernel matrices throughout this paper. We assume that the input data $\mathbf{X} \in \mathbb{R}^{p \times n}$ are independently drawn from a K -class Gaussian mixture model (GMM) and are independent of the random weights. The analysis of class-structured data is common in the literature. For example, Papayan et al. (2020) demonstrated that when the training error reaches zero, the last-layer classifiers collapse to the class-means; Liao & Couillet (2018) conducted a spectral analysis of the Gram matrix associated with random feature mappings, while Couillet et al. (2018) analyzed the asymptotic performance of several classical classification methods under GMM assumptions; Ali et al. (2022) investigated the spectral behavior of the CK in a high-dimensional regime; and Gu et al. (2024) derived the deterministic equivalents for both the CK and NTK. Based on the results established by Gu et al. (2024), we further show that a finite number of eigenvalues of these kernel matrices lie outside the support of the limiting measure. Our analysis reveals a possible phase transition (Baik et al., 2005) that depends on the choice of activation functions and the differences in covariance among different classes. We also precisely determine the positions of these isolated eigenvalues and eigenvector alignments.

In this article, by demonstrating that the eigenvectors associated with the isolated eigenvalues may contain information relevant to unsupervised classification (clustering), we establish a connection between the features of the input data and neural networks. The techniques employed in this paper are grounded in random matrix theory, and from a theoretical perspective, our analysis falls within the framework of finite-rank deformation models in this field.

Notations: Throughout the paper, we use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm for matrices. The spectrum of a matrix $\mathbf{A}$ is denoted by $\text{Spec}(\mathbf{A})$. We denote an all-ones vector of dimension p by $\mathbf{1}_p$ and the identity matrix of size $p \times p$ by $\mathbf{I}_p$. For two sequences of random matrices $\mathbf{A} = \{\mathbf{A}_n\}_{n \geq 1}$ and $\mathbf{B} = \{\mathbf{B}_n\}_{n \geq 1}$, we denote $\mathbf{A} \sim \mathbf{B}$ if

$$\frac{1}{n} \text{tr} \mathbf{D}_n(\mathbf{A}_n - \mathbf{B}_n) \rightarrow 0 \text{ and } \mathbf{u}_1^\top (\mathbf{A}_n - \mathbf{B}_n) \mathbf{u}_2 \rightarrow 0 \text{ a.s.}$$

for all deterministic sequences $\{\mathbf{D}_n\}_{n \geq 1}$ and all deterministic vectors $\mathbf{u}_i \in \mathbb{R}^n, i = 1, 2$ with bounded norms. The symbols $O(\cdot)$ and $o(\cdot)$ stand for the standard big-O and little-o notations. Moreover, if $\|\mathbf{A} - \mathbf{B}\| \rightarrow 0$ almost surely, we write $\mathbf{A} = \mathbf{B} + o_{a.s.}(1)$. The Hadamard product between two matrices of the same size is denoted by $\circ$. The distance between two sets $A, B \subset \mathbb{C}$ is denoted by $\text{dist}(A, B)$. Let $a \otimes \nu \oplus b$ denote the law of $ax + b$, where $x \sim \nu$ is a

random variable (or random vector) following distribution ν . The indicator function is represented by δ . For a function f , we denote its i -th derivative by $f^{(i)}$. Specifically, for $i = 1, 2$, we also use $f' = f^{(1)}$ and $f'' = f^{(2)}$. Additionally, $\|f\| = \sup_x |f(x)|$ denotes the supremum norm of f . The notation 0 may indicate a zero value, a zero vector or a zero matrix in this paper, changing from line to line. We use c and C to denote positive constants, whose values may change from one line to the next.

2. Preliminaries

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be n random vectors independently drawn from one of the K -class Gaussian mixtures $\mathcal{C}_1, \dots, \mathcal{C}_K$, that is

$$\mathbf{x}_i \in \mathcal{C}_a \iff \mathbf{x}_i \sim \mathcal{N}(0, p^{-1} \mathbf{C}_a)$$

for some non-negative definite matrix $\mathbf{C}_a$. For each $a \in \{1, \dots, K\}$, class $\mathcal{C}_a$ has cardinality n_a , satisfying $n_1 + \dots + n_K = n$. Write the input features in a matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$. We define the fully-connected DNN with L hidden layers by

$$\mathbf{X}_\ell = \frac{1}{\sqrt{d_\ell}} \sigma_\ell(\mathbf{W}_\ell \mathbf{X}_{\ell-1}) \in \mathbb{R}^{d_\ell \times n} \quad \text{for } \ell = 1, \dots, L,$$

with weight matrices $\mathbf{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ (with convention $d_0 = p, \mathbf{X}_0 = \mathbf{X}$) and nonlinear activation functions $\sigma_1, \dots, \sigma_L$ applied entrywise. The CK of the ℓ -th layer is given by the Gram matrix

$$\mathbf{K}_{\text{CK}, \ell} = \mathbf{K}_{\text{CK}, \ell}(\mathbf{X}) := \mathbb{E}[\mathbf{X}_\ell^\top \mathbf{X}_\ell] \in \mathbb{R}^{n \times n},$$

where the expectation is taken with respect to the random weights $\mathbf{W}_1, \dots, \mathbf{W}_L$. Following Bietti & Mairal (2019) and Jacot et al. (2018), the CK satisfies that

$$[\mathbf{K}_{\text{CK}, \ell}]_{ij} = \mathbb{E}[\sigma_\ell(u) \sigma_\ell(v)] \quad (1)$$

with

$$[u, v]^\top \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} [\mathbf{K}_{\text{CK}, \ell-1}]_{ii} & [\mathbf{K}_{\text{CK}, \ell-1}]_{ij} \\ [\mathbf{K}_{\text{CK}, \ell-1}]_{ji} & [\mathbf{K}_{\text{CK}, \ell-1}]_{jj} \end{bmatrix}\right),$$

while the NTK denoted by $\mathbf{K}_{\text{NTK}, \ell}$ takes the form

$$\begin{aligned} \mathbf{K}_{\text{NTK}, \ell} &= \mathbf{K}_{\text{CK}, \ell} + \mathbf{K}_{\text{NTK}, \ell-1} \circ \mathbf{K}'_{\text{CK}, \ell}, \\ \mathbf{K}_{\text{NTK}, 0} &= \mathbf{K}_{\text{CK}, 0} = \mathbf{X}^\top \mathbf{X}, \end{aligned} \quad (2)$$

where

$$[\mathbf{K}'_{\text{CK}, \ell}]_{ij} = \mathbb{E}[\sigma'_\ell(u) \sigma'_\ell(v)].$$

We consider the high-dimensional regime where n and p are comparable and assume the following conditions on the input data, weights and activation functions.

Assumption 2.1. As $n \rightarrow \infty$, the following conditions hold:

- The ratios $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$, and $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ for each $a \in \{1, \dots, K\}$.
- Denoting $\mathbf{C}^\circ = \frac{1}{p} \sum_{a=1}^K \frac{n_a}{n} \mathbf{C}_a$, $\mathbf{C}_a^\circ = \mathbf{C}_a - \mathbf{C}^\circ$ and $\tau_0 = \sqrt{\text{tr} \mathbf{C}^\circ / p}$, we have $\lim_{n \rightarrow \infty} \tau_0 \in (0, \infty)$.
- For $a, b \in \{1, \dots, K\}$, it holds that $\|\mathbf{C}_a\| = O(1)$ and $\text{tr} \mathbf{C}_a \mathbf{C}_b = O(p)$.

This assumption ensures the classification task is feasible and non-trivial (see Gu et al. (2024) for example).

Assumption 2.2. The random weights $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times p}, \dots, \mathbf{W}_L \in \mathbb{R}^{d_L \times d_{L-1}}$ are independent. Moreover, the entries of $\mathbf{W}_\ell$ are i.i.d. and satisfy

$$\mathbb{E}[\mathbf{W}_\ell]_{ij} = 0, \quad \mathbb{E}[\mathbf{W}_\ell]_{ij}^2 = 1, \quad \mathbb{E}[\mathbf{W}_\ell]_{ij}^4 < \infty.$$

This assumption on the random weights is relatively mild and holds for many common weights such as i.i.d. standard Gaussian weights.

Assumption 2.3. Let $\xi \sim \mathcal{N}(0, 1)$. The activation functions $\sigma_1, \dots, \sigma_L$ satisfy that

$$\max_{k \in \{0, 1, 2, 3, 4\}} \mathbb{E}[\sigma_\ell^{(k)}(\xi)] < C, \quad 1 \leq \ell \leq L \quad (3)$$

for some universal constant C .

The boundness of $\mathbb{E}[\sigma_\ell^{(k)}(\xi)]$ in (3) is needed to derive the deterministic equivalents for the CK and NTK. Moreover, as noted by Gu et al. (2024), for non-differentiable functions, Assumption 2.3 holds if $|\sigma_\ell|$ is bounded above by some polynomial function through the application of Gaussian integration by parts: $\mathbb{E}[\sigma'_\ell(\xi)] = \mathbb{E}[\xi \sigma_\ell(\xi)]$. Consequently, Assumption 2.3 accommodates commonly used activation functions such as ReLU, Sigmoid, and Tanh, provided that these activation functions are centered and normalized to ensure compliance with (3).

Before presenting the preliminary theoretical results, it is necessary to introduce several key expressions:

$$\begin{aligned} \boldsymbol{\psi} &= \{\psi_i\}_{i=1}^n := \{\|\mathbf{x}_i\|^2 - \mathbb{E}\|\mathbf{x}_i\|^2\}_{i=1}^n \in \mathbb{R}^n, \\ \mathbf{t} &:= \left\{ \frac{1}{\sqrt{p}} \text{tr} \mathbf{C}_a^\circ \right\}_{a=1}^K \in \mathbb{R}^K, \\ \mathbf{T} &:= \left\{ \frac{1}{p} \text{tr} \mathbf{C}_a \mathbf{C}_b \right\}_{a,b=1}^K \in \mathbb{R}^{K \times K}. \end{aligned} \quad (4)$$

The following two lemmas, proved by Gu et al. (2024), provide asymptotic equivalents for the CK and NTK.

Lemma 2.4. (Asymptotic spectral equivalent for the CK (Gu et al., 2024)). Suppose Assumptions 2.1-2.3 hold. Let $\tau_0, \tau_1, \dots, \tau_L$ be a sequence of non-negative numbers satisfying the recursion:

$$\tau_\ell = \sqrt{\mathbb{E}[\sigma_\ell^2(\tau_{\ell-1}\xi)]}, \quad \ell \in \{1, \dots, L\}.$$

We further assume $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$. Then as $n \rightarrow \infty$, we have

$$\begin{aligned} \|\mathbf{K}_{\text{CK},\ell}(\mathbf{X}) - \widetilde{\mathbf{K}}_{\text{CK},\ell}(\mathbf{X})\| &\rightarrow 0 \text{ a.s.}, \\ \widetilde{\mathbf{K}}_{\text{CK},\ell} &:= \alpha_{\ell,1} \mathbf{X}^\top \mathbf{X} + \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top + \alpha_{\ell,0} \mathbf{I}_n, \end{aligned} \quad (5)$$

where $\alpha_{\ell,0} = \tau_\ell^2 - \tau_0^2 \alpha_{\ell,1} \geq 0$,

$$\begin{aligned} \mathbf{V} &:= [\mathbf{J} / \sqrt{p}, \boldsymbol{\psi}] \in \mathbb{R}^{n \times (K+1)}, \\ \mathbf{A}_\ell &:= \begin{bmatrix} \alpha_{\ell,2} \mathbf{t} \mathbf{t}^\top + \alpha_{\ell,3} \mathbf{T} & \alpha_{\ell,2} \mathbf{t} \\ \alpha_{\ell,2} \mathbf{t}^\top & \alpha_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1) \times (K+1)}, \end{aligned} \quad (6)$$

for class label vectors $\mathbf{J} = [\mathbf{j}_1, \dots, \mathbf{j}_K] \in \mathbb{R}^{n \times K}$ with $\mathbf{j}_a = \{\delta_{x_i \in \mathcal{C}_a}\}_{i=1}^n$. The non-negative scalars $\alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3}$ satisfy the following recursions:

$$\begin{aligned} \alpha_{\ell,1} &= \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,1}, \\ \alpha_{\ell,2} &= \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,2} + \frac{1}{4} \mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,4}, \\ \alpha_{\ell,3} &= \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,3} + \frac{1}{2} \mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,1} \end{aligned} \quad (7)$$

with

$$\alpha_{\ell,4} = \alpha_{\ell-1,4} \mathbb{E}[(\sigma'_\ell(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi) \sigma''_\ell(\tau_{\ell-1}\xi)],$$

$$\alpha_{0,1} = \alpha_{0,4} = 1, \quad \alpha_{0,2} = \alpha_{0,3} = 0.$$

Lemma 2.5. (Asymptotic spectral equivalent for the NTK (Gu et al., 2024)). Suppose Assumptions 2.1-2.3 hold under the same notation and settings as in Lemma 2.4. Let $\dot{\tau}_0 = 0, \dot{\tau}_1, \dots, \dot{\tau}_L \geq 0$ be a sequence non-negative numbers such that

$$\dot{\tau}_\ell = \sqrt{\mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2}, \quad \ell \in \{1, \dots, L\}$$

and let $\kappa_\ell^2 = \tau_\ell^2 + \dot{\tau}_\ell^2$ with $\kappa_0 = \tau_0$. Then as $n \rightarrow \infty$, we have

$$\begin{aligned} \|\mathbf{K}_{\text{NTK},\ell}(\mathbf{X}) - \widetilde{\mathbf{K}}_{\text{NTK},\ell}(\mathbf{X})\| &\rightarrow 0 \text{ a.s.}, \\ \widetilde{\mathbf{K}}_{\text{NTK},\ell} &:= \beta_{\ell,1} \mathbf{X}^\top \mathbf{X} + \mathbf{V} \mathbf{B}_\ell \mathbf{V}^\top + \beta_{\ell,0} \mathbf{I}_n, \end{aligned} \quad (8)$$

where $\beta_{\ell,0} = \kappa_\ell^2 - \tau_0^2 \beta_{\ell,1} \geq 0$ and

$$\mathbf{B}_\ell := \begin{bmatrix} \beta_{\ell,2} \mathbf{t} \mathbf{t}^\top + \beta_{\ell,3} \mathbf{T} & \beta_{\ell,2} \mathbf{t} \\ \beta_{\ell,2} \mathbf{t}^\top & \beta_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1) \times (K+1)}.$$

The non-negative scalars $\beta_{\ell,1}, \beta_{\ell,2}, \beta_{\ell,3}$ satisfy

$$\begin{aligned} \beta_{\ell,1} &= \alpha_{\ell,1} + \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \beta_{\ell-1,1}, \\ \beta_{\ell,2} &= \alpha_{\ell,2} + \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \beta_{\ell-1,2}, \\ \beta_{\ell,3} &= \alpha_{\ell,3} + \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2 \beta_{\ell-1,3} \\ &\quad + \mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2 \alpha_{\ell-1,1} \beta_{\ell-1,1} \end{aligned} \quad (9)$$

with $\beta_{0,1} = 1, \beta_{0,2} = \beta_{0,3} = 0$.

For an $p \times p$ matrix M with real eigenvalues $\lambda_1^M \geq \dots \geq \lambda_p^M$, its empirical spectral distribution is defined as

$$F^M(x) = \frac{1}{p} \#\{i = 1, \dots, p, \lambda_i^M \leq x\},$$

and the probability measure induced by F^M is denoted by μ_M . We define the Stieltjes transform of a probability measure ν as

$$m_\nu(z) = \int \frac{1}{\lambda - z} d\nu(\lambda), \quad z \in \mathbb{C}^+ = \{z : z \in \mathbb{C}, \Im z > 0\}.$$

It admits a natural extension to the lower-half of the complex space by the fact that

$$m_G(z) = \overline{m_G(\bar{z})}, \quad z \in \mathbb{C}^- = \{z : z \in \mathbb{C}, \Im z < 0\}.$$

Prior to presenting our results, we need the following lemma, which guarantees the existence of the limiting spectral distribution of $\mathbf{X}^\top \mathbf{X}$. This lemma is established by [Benaych-Georges & Couillet \(2016\)](#).

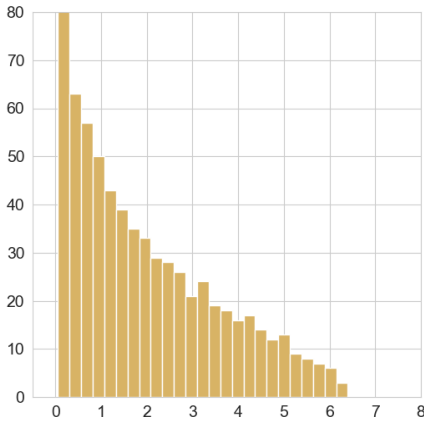


Figure 1. Spectrum of $\widetilde{\mathbf{K}}_{\text{CK},3}$ under the parameter settings: $n = 1200$, $p = 600$, $c_1 = c_3 = 0.3$, $c_2 = 0.4$ and $\mathbf{C}_a = (1 + 2(a-1)/\sqrt{p})\mathbf{I}_p$ for $a = 1, 2, 3$. The activation functions for the three layers are $\sigma_1 = \sigma_2 = \sigma_3 = \text{Poly}$, where $\text{Poly}(t) = 0.2t^2 + t$. The weights $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times p}$, $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_1}$ and $\mathbf{W}_3 \in \mathbb{R}^{d_3 \times d_2}$ consist of i.i.d. standard normal entries, where $d_1 = d_2 = 2000$, $d_3 = 1000$.

Lemma 2.6. Suppose Assumption 2.1 holds. For $z \in \mathbb{C}$, we define the resolvents

$$\mathbf{G}(z) = (\mathbf{X}^\top \mathbf{X} - z\mathbf{I}_n)^{-1}, \quad \underline{\mathbf{G}}(z) = (\mathbf{X}\mathbf{X}^\top - z\mathbf{I}_p)^{-1}.$$

Then as $n \rightarrow \infty$, we have

$$\mathbf{G}(z) \sim \mathbf{Q}_1(z) := c_0 \text{diag} \{m_a(z)1_{n_a}\}_{a=1}^K \quad (10)$$

and

$$\underline{\mathbf{G}}(z) \sim \mathbf{Q}_2(z) := -\frac{1}{z} \left(\mathbf{I}_p + \sum_{a=1}^K c_a m_a(z) \mathbf{C}_a \right)^{-1}, \quad (11)$$

where $\{m_a(z)\}_{a=1}^K \in \mathbb{C}^K$ is a unique vector such that

$$\Im z \Im m_a(z) \geq 0, \quad \Im z \Im (z m_a(z)) \geq 0, \quad c_0 |m_a(z)| \leq (\Im z)^{-1}$$

and

$$\begin{aligned} c_0 m_a(z) &= -\frac{1}{z} \frac{1}{1 + \widetilde{m}_a(z)}, \\ \widetilde{m}_a(z) &= -\frac{1}{z} \frac{1}{p} \text{tr} \mathbf{C}_a \left(\mathbf{I}_p + \sum_{b=1}^K c_b m_b(z) \mathbf{C}_b \right)^{-1} \\ &= \frac{1}{p} \text{tr} \mathbf{C}_a \mathbf{Q}_2(z). \end{aligned} \quad (12)$$

Besides, $c_0 m_1(z), \dots, c_0 m_K(z)$ are Stieltjes transforms of some $\mathbb{R}^+$ -compactly supported probability measures $\nu_1, \dots, \nu_K$. The probability measure μ defined by the Stieltjes transform

$$m_\mu(z) = c_0 \sum_{a=1}^K c_a m_a(z) \quad (13)$$

is a deterministic probability measure with compact support $\mathcal{S} = \bigcup_{a=1}^K \text{supp}(\nu_a)$ such that

$$\mu_{\mathbf{X}^\top \mathbf{X}} \rightarrow \mu \text{ in distribution}, \quad (14)$$

$$\text{dist}(\text{Spec}(\mathbf{X}^\top \mathbf{X}), \mathcal{S} \cup \{0\}) \rightarrow 0 \quad (15)$$

almost surely. For $x \in \mathbb{R}$ such that $\text{dist}(x, \mathcal{S} \cup \{0\}) > c$ for some positive constant c , (12) also holds.

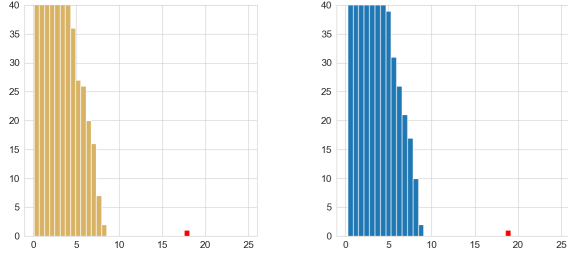
According to Lemma 2.4 and Lemma 2.5, the ℓ -th layer of kernel matrices can be approximated by a linear combination of $\mathbf{X}^\top \mathbf{X}$, $\mathbf{I}_n$ and a low rank matrix. Based on above results and Lemma A.4, it follows immediately that $\|F^{\mathbf{K}_{\text{CK},\ell}} - F^{\widetilde{\mathbf{K}}_{\text{CK},\ell}}\| \rightarrow 0$ ($\|F^{\mathbf{K}_{\text{NTK},\ell}} - F^{\widetilde{\mathbf{K}}_{\text{NTK},\ell}}\| \rightarrow 0$). Therefore, the LSD of $\mathbf{K}_{\text{CK},\ell}(\mathbf{K}_{\text{NTK},\ell})$ can be written as $\mu_{\text{CK},\ell} = \alpha_{\ell,1} \otimes \mu \oplus \alpha_{\ell,0}$ ($\mu_{\text{NTK},\ell} = \beta_{\ell,1} \otimes \mu \oplus \beta_{\ell,0}$).

To visualize the ESD of the CK, we present the empirical spectral distribution of $\widetilde{\mathbf{K}}_{\text{CK},3}$ in Figure 1, obtained using the same polynomial activation function across all layers. As shown, this ESD does not exhibit any isolated eigenvalues. In the following section, we will investigate scenarios in which eigenvalues detach from the bulk.

3. Main results: eigen analysis of the CK and NTK

In this section, we follow the notation established in Lemmas 2.4-2.5, and without loss of generality, assume that

$$\mathbf{x}_i \in \mathcal{C}_a \quad \text{for} \quad 1 + \sum_{j=1}^{a-1} n_j \leq i \leq \sum_{j=1}^a n_j.$$



(a) Eigenvalues histogram of $\widetilde{\mathbf{K}}_{\text{CK},3}$ (b) Eigenvalues histogram of $\mathbf{K}_{\text{CK},3}$

Figure 2. Spectra of $\widetilde{\mathbf{K}}_{\text{CK},3}$ and $\mathbf{K}_{\text{CK},3}$ (with the expectation estimated from 500 realizations of weights).

This assumption is also adopted in the simulations. We denote the supports of the limiting spectral measures of $\mathbf{K}_{\text{CK},\ell}$ and $\mathbf{K}_{\text{NTK},\ell}$ by $\mathcal{S}_{\text{CK},\ell}$ and $\mathcal{S}_{\text{NTK},\ell}$, respectively. According to Lemma 2.4, Lemma 2.5 and Lemma 2.6, we have

$$\mathcal{S}_{\text{CK},\ell} = \alpha_{\ell,1}\mathcal{S} + \alpha_{\ell,0}, \quad \mathcal{S}_{\text{NTK},\ell} = \beta_{\ell,1}\mathcal{S} + \beta_{\ell,0}.$$

3.1. Isolated eigenvalues

This subsection provides a precise characterization of the isolated eigenvalues of the kernel matrices. Based on the findings from the previous section, it suffices to focus on the isolated eigenvalues of $\widetilde{\mathbf{K}}_{\text{CK},\ell}$ and $\widetilde{\mathbf{K}}_{\text{NTK},\ell}$. To illustrate this, we visualize the spectra of $\widetilde{\mathbf{K}}_{\text{CK},\ell}$ and $\mathbf{K}_{\text{CK},\ell}$ in Figure 2, for the case where $\mathbf{C}_a = (1 + 8(a-1)/\sqrt{p})\mathbf{I}_p$, with all other settings kept consistent with those in Figure 1.

We now present the results for the eigenvalues of the CK and NTK that detach from the bulk of their spectrum, focusing on the non-trivial case where $\alpha_{\ell,1}$ ($\beta_{\ell,1}$) is bounded away from zero. The following theorem establishes the asymptotic behavior of the isolated eigenvalues of $\mathbf{K}_{\text{CK},\ell}$.

Theorem 3.1. (Isolated eigenvalues of the CK). Define the function $h_\ell^1(z) = 1 + \frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \frac{1}{p} \sum_{a=1}^K c_a m_a(z) \text{tr} \mathbf{C}_a^2$ and the set $\mathcal{H}_{1,\ell}^p = \{z \in \mathbb{C} | h_\ell^1(z) = 0\}$. Denote

$$\mathcal{S}_{1,\ell} = \mathcal{S} \cup \mathcal{H}_{1,\ell}^p \cup \{0\},$$

and

$$\underline{\mathbf{H}}_1(z) = \left[h_\ell^1(z) \mathbf{I}_K + \frac{\alpha_{\ell,3}}{\alpha_{\ell,1}} h_\ell^1(z) \mathbf{T} \mathbf{T}^\top(z) + \frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \mathbf{t} \mathbf{t}^\top \mathbf{T}^\top(z) \right],$$

where $\mathbf{T}$ is defined in (4) and

$$\mathbf{T}(z) = \text{diag}\{c_a m_a(z)\}_{a=1}^K.$$

Under Assumptions 2.1-2.3, if $\underline{\mathbf{H}}_1(\rho)$ has a zero eigenvalue with multiplicity k_ρ and $\text{dist}(\rho, \mathcal{S}_{1,\ell}) > C$ for some positive

constant C , then $\mathbf{K}_{\text{CK},\ell}$ has k_ρ eigenvalues

$$\lambda_j^{\text{CK}} \geq \dots \geq \lambda_{j+k_\rho-1}^{\text{CK}}$$

outside $\mathcal{S}_{\text{CK},\ell}$ such that

$$\max_{0 \leq i \leq k_\rho-1} |\lambda_{j+i}^{\text{CK}} - (\alpha_{\ell,1}\rho + \alpha_{\ell,0})| \rightarrow 0 \quad \text{a.s.} \quad (16)$$

Moreover, if there exists a $\rho_+ \in \mathcal{S}^c$ satisfying $\det \mathbf{H}_1(\rho_+) = 0$ and $\rho_+ \rightarrow \rho \in \mathcal{H}_{1,\ell}^p$, where $\mathbf{H}_1(z)$ is defined in (28), then there are k_ρ^+ eigenvalues of $\mathbf{K}_{\text{CK},\ell}$,

$$\lambda_j^{\text{CK}} \geq \dots \geq \lambda_{j+k_\rho^+-1}^{\text{CK}}$$

outside $\mathcal{S}_{\text{CK},\ell}$ such that

$$\max_{0 \leq i \leq k_\rho^+-1} |\lambda_{j+i}^{\text{CK}} - (\alpha_{\ell,1}\rho + \alpha_{\ell,0})| \rightarrow 0 \quad \text{a.s.}, \quad (17)$$

where k_ρ^+ is the multiplicity of zero as an eigenvalue of $\mathbf{H}_1(\rho_+)$.

Lemma 2.5 shows that the asymptotic properties of the eigenvalues of the CK matrices also hold for the NTK, up to a change of the associated coefficients $\alpha_{\ell,i}$ to $\beta_{\ell,i}$ for $i = 1, 2, 3, 4$. Therefore, we can immediately obtain the following theorem, which describes the behaviors of the isolated eigenvalues of the NTK.

Theorem 3.2. (Isolated eigenvalues of the NTK). Define $h_\ell^2(z) := 1 + \frac{\beta_{\ell,3}}{\beta_{\ell,1}} \frac{1}{p} \sum_{a=1}^K c_a m_a(z) \text{tr} \mathbf{C}_a^2$ and the set $\mathcal{H}_{2,\ell}^p := \{z \in \mathbb{C} | h_\ell^2(z) = 0\}$. Denote $\mathcal{S}_{2,\ell} = \mathcal{S} \cup \mathcal{H}_{2,\ell}^p \cup \{0\}$ and

$$\underline{\mathbf{H}}_2(z) := \left[h_\ell^2(z) \mathbf{I}_K + \frac{\beta_{\ell,3}}{\beta_{\ell,1}} h_\ell^2(z) \mathbf{T} \mathbf{T}^\top(z) + \frac{\beta_{\ell,2}}{\beta_{\ell,1}} \mathbf{t} \mathbf{t}^\top \mathbf{T}^\top(z) \right].$$

Suppose Assumptions 2.1-2.3 hold, for ρ being a solution to $\det \underline{\mathbf{H}}_2(\rho) = 0$ with multiplicity k_ρ and $\text{dist}(\rho, \mathcal{S}_{2,\ell}) > C$ for some positive constant C , we conclude that there exists k_ρ eigenvalues of $\mathbf{K}_{\text{NTK},\ell}$,

$$\lambda_j^{\text{NTK}} \geq \dots \geq \lambda_{j+k_\rho-1}^{\text{NTK}}$$

outside $\mathcal{S}_{\text{NTK},\ell}$ such that

$$\max_{0 \leq i \leq k_\rho-1} |\lambda_{j+i}^{\text{NTK}} - (\beta_{\ell,1}\rho + \beta_{\ell,0})| \rightarrow 0 \quad \text{a.s.},$$

where the scalars $\beta_{\ell,0}$ and $\beta_{\ell,1}$ are defined in (9) and k_ρ is the multiplicity of zero as an eigenvalue of $\underline{\mathbf{H}}_2(\rho)$. We denote $\underline{\mathbf{H}}_2(z)$ by replacing the α 's in $\mathbf{H}_1(z)$ with β 's. For $\rho_+ \in \mathcal{S}^c$ being a solution to $\det \underline{\mathbf{H}}_2(z) = 0$ and $\rho_+ \rightarrow \rho \in \mathcal{H}_{2,\ell}^p$, there are k_ρ^+ eigenvalues of $\mathbf{K}_{\text{NTK},\ell}$,

$$\lambda_j^{\text{NTK}} \geq \dots \geq \lambda_{j+k_\rho^+-1}^{\text{NTK}}$$

such that

$$\max_{0 \leq i \leq k_\rho^+-1} |\lambda_{j+i}^{\text{NTK}} - (\beta_{\ell,1}\rho + \beta_{\ell,0})| \rightarrow 0 \quad \text{a.s.},$$

where k_ρ^+ is the multiplicity of zero as an eigenvalue of $\underline{\mathbf{H}}_2(\rho_+)$.

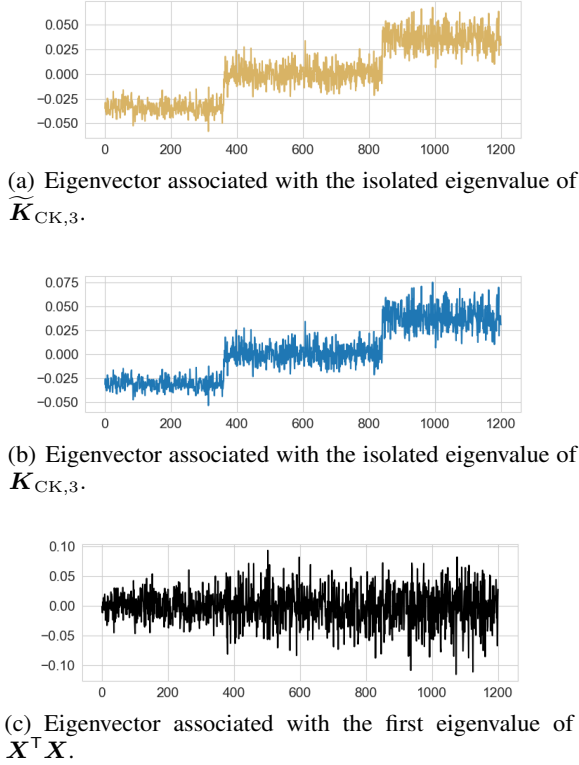


Figure 3. Visualization of the eigenvectors corresponding to the first eigenvalues of matrices $\widetilde{K}_{CK,\ell}$, $K_{CK,\ell}$ and $X^T X$ under the setting of Figure 2. In (a), $\omega_1 = [-0.643, 0.005, 0.693]$ while in (c), $\omega_1 = [-0.008, 0.036, 0.008]$.

Theorem 3.1 and Theorem 3.2 establish the conditions for the occurrence of isolated eigenvalues and provide their asymptotic positions.

3.2. Behaviors of the eigenvectors

The preceding subsection studied the eigenvalues of the kernel matrices. We now turn our attention to the asymptotic properties of the eigenvectors associated with the isolated eigenvalues. Specifically, we provide a detailed analysis of the CK matrices, as the NTK matrices exhibit a similar pattern.

We begin with a toy empirical example. Under the same setting as Figure 2, Figures 3(a)-(b) display the eigenvectors associated with the isolated eigenvalues of $\widetilde{K}_{CK,3}$ and $K_{CK,3}$, respectively. Figure 3(c) shows the eigenvector corresponding to the leading eigenvalue of $X^T X$.

From these figures, we can reasonably infer that the eigenvectors corresponding to the isolated eigenvalues (Figures 3(a)-(b)) are more informative, as they likely capture relevant information for clustering. In contrast, the eigenvector depicted in Figure 3 (c) can be reasonably regarded as pure

noise. Inspired by these observations, an interesting question arises:

How can we understand the structure of the eigenvector (or eigenspace) associated with an isolated eigenvalue in a high-dimensional regime?

We use $\text{Span}(\mathbf{J})$ to denote the space spanned by $\mathbf{j}_a$, $a = 1, \dots, K$. Let $\mathbf{P}_J$ be the projection onto $\text{Span}(\mathbf{J})$. The eigenvector associated with the i th isolated eigenvalue λ_i^{CK} is denoted by $\hat{\mathbf{u}}_i$. This eigenvector can be decomposed into a signal part and a random part:

$$\hat{\mathbf{u}}_i = \mathbf{P}_J \hat{\mathbf{u}}_i + (\mathbf{I}_n - \mathbf{P}_J) \hat{\mathbf{u}}_i = \sum_{a=1}^K \left(\omega_{ia} \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sigma_a \mathbf{v}_a \right), \quad (18)$$

where $\mathbf{v}_a$ is a random unit vector orthogonal to the $\mathbf{j}_a$, and its entries are identically distributed. The scalar $\omega_{ia} = \hat{\mathbf{u}}_i^T \frac{\mathbf{j}_a}{\sqrt{n_a}}$ measures the cosine between $\hat{\mathbf{u}}_i$ and $\frac{\mathbf{j}_a}{\sqrt{n_a}}$, and σ_a quantifies the extent of fluctuations in $\omega_{ia} \frac{\mathbf{j}_a}{\sqrt{n_a}}$. We denote $\omega_i = \{\omega_{ia}\}_{a=1}^K$.

If the isolated eigenvalue λ_i^{CK} is simple, it follows from $\omega_{ia} = \hat{\mathbf{u}}_i^T \frac{\mathbf{j}_a}{\sqrt{n_a}}$ that

$$\frac{1}{n_a} [\mathbf{J}^T \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \mathbf{J}]_{aa} = \frac{\mathbf{j}_a^T \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \mathbf{j}_a}{n_a} = \omega_{ia}^2. \quad (19)$$

Thus, if ρ identified in Theorem 3.1 satisfies that $\underline{H}_1(\rho)$ has a simple (multiplicity one) zero eigenvalue, we can explicitly evaluate ω_a . Moreover, we say an eigenvector $\hat{\mathbf{u}}_i$ is non-informative for clustering (i.e., it does not contain any information about $\mathbf{J}$) if $\max_a \hat{\mathbf{u}}_i^T \frac{\mathbf{j}_a}{\sqrt{n_a}} = o_{a.s.}(1)$, meaning that $\hat{\mathbf{u}}_i$ becomes asymptotically orthogonal to $\text{Span}(\mathbf{J})$. It is important to note that the eigenvector is not unique, and essentially, our analysis concerns the asymptotic behavior of the matrix $\frac{1}{n_a} \mathbf{J}^T \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T \mathbf{J}$, as described in (19).

Let $\lambda_j, \dots, \lambda_{j+k_\rho-1}$ be a group of isolated eigenvalues of K_{CK} that converge to the same limit $\alpha_{\ell,1}\rho + \alpha_{\ell,0}$, where ρ is identified in Theorem 3.1. As per Lemma 2.4, the eigenspace

$$\text{Span}\{\hat{\mathbf{u}}_i \in \mathbb{S}^{n-1} \mid K_{CK,\ell} \hat{\mathbf{u}}_i = \lambda_{j+i-1} \hat{\mathbf{u}}_i\} \quad (20)$$

is asymptotically equivalent to the eigenspace associated with eigenvalues of $X^T X + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^T$ that have a deterministic limit ρ . This relationship is also illustrated in Figure 3. In Definition 3.3 below, we present a precise mathematical criterion for determining when an eigenspace is considered informative.

Definition 3.3. (Informative eigenspace). We say the eigenspace defined in (20) is informative if there is a non zero matrix $\mathbf{A}(\mathbf{J})$ depending on $\mathbf{J}$ such that

$$\left\| \frac{1}{p} \mathbf{J}^T \hat{\Pi}_\rho \mathbf{J} - \mathbf{A}(\mathbf{J}) \right\| = o_{a.s.}(1).$$

Otherwise if $\mathbf{A}(\mathbf{J}) = 0$, then the eigenspace is non-informative.

Denote by $\hat{\Pi}_\rho = \sum_{i=1}^{k_\rho} \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$ the projection onto the eigenspace defined in (20). Then we can write $\frac{1}{p} \mathbf{J}_a^\top \hat{\Pi}_\rho \mathbf{J}_b$ as

$$- \frac{1}{2\pi i} \oint_{\partial\gamma_\rho} \frac{1}{p} \mathbf{J}_a^\top (\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n)^{-1} \mathbf{J}_b dz + o_{a.s.}(1), \quad (21)$$

where γ_ρ is an open disc such that ρ belongs to the interior of γ_ρ and $\partial\gamma_\rho$ is a positively oriented closed circle. Therefore, it suffices to investigate the asymptotic properties of the right hand side (RHS) of (21). We establish the following Theorem 3.4, which provides the asymptotic behavior of eigenspace $\hat{\Pi}_\rho$ characterized by $\mathbf{u}^\top \hat{\Pi}_\rho \mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are non-random unit vectors.

Theorem 3.4. *Suppose the assumptions in Lemma 2.4 hold. Let $\lambda_j^{CK}, \dots, \lambda_{j+k_\rho-1}^{CK}$ be a group of isolated eigenvalues of $\mathbf{K}_{CK,\ell}$ converging to the same limit $\alpha_{\ell,1}\rho + \alpha_{\ell,0}$, where ρ is defined in Theorem 3.1. For $\underline{\mathbf{H}}_1(\rho)$, we denote the left and right eigenvectors corresponding to 0 as $(\mathbf{U}_{l,\rho})_i$ and $(\mathbf{U}_{r,\rho})_i$ respectively, where i ranges from 1 to k_ρ . If $\text{dist}(\rho, \mathcal{S}_{1,\ell})$ is bounded away from 0, then for any non-random unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have*

$$\mathbf{u}^\top \hat{\Pi}_\rho \mathbf{v} = -h_\ell^1(\rho) \mathbf{u}^\top \mathbf{Q}_1(\rho) \mathbf{J} \mathbf{F}(\rho) \Gamma^{-1}(\rho) \mathbf{J}^\top \mathbf{Q}_1(\rho) \mathbf{v} + o_{a.s.}(1),$$

where

$$\mathbf{F}(\rho) = \sum_{i=1}^{k_\rho} \frac{(\mathbf{U}_{r,\rho})_i (\mathbf{U}_{l,\rho})_i^\top}{(\mathbf{U}_{l,\rho})_i^\top [\partial_z \underline{\mathbf{H}}_1(z)]_{z=\rho} (\mathbf{U}_{r,\rho})_i}.$$

By noticing that $\frac{1}{p} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{J} = \Gamma(z)$, one may immediately derive the following corollary.

Corollary 3.5. *Under the same conditions as those in Theorem 3.4, we have*

$$\frac{1}{p} \mathbf{J}^\top \hat{\Pi}_\rho \mathbf{J} = -h_\ell^1(\rho) \Gamma(\rho) \mathbf{F}(\rho) + o_{a.s.}(1). \quad (22)$$

This result identifies the conditions under which the eigenspace corresponding to the isolated eigenvalues of the CK matrix are informative. Discussions on non-informative eigenspace can be found in Remark 3.6 below.

Remark 3.6. For ρ_+ satisfying $\det \underline{\mathbf{H}}_1(\rho_+) = 0$ and $\rho_+ \rightarrow \rho \in \mathcal{H}_{1,\ell}^p$, it follows that for any non-random unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{u}^\top \hat{\Pi}_{\rho_+} \mathbf{v} = o_{a.s.}(1). \quad (23)$$

When λ_{j+i} in Equation (20) has multiplicity one, $\|\omega_i\|$ can be regarded as a measure of the alignment between

$\hat{\mathbf{u}}_i$ and $\text{Span}(\mathbf{J})$. Furthermore, if $h_\ell^1(\rho_+) \rightarrow 0$, then $\|\omega_i\|$ approaches 0 by letting $\mathbf{u}, \mathbf{v} \in \{\mathbf{j}_1/\sqrt{p}, \dots, \mathbf{j}_K/\sqrt{p}\}$, indicating that it can always be considered as a non-informative eigenvector.

The theoretical results presented above remain valid for the NTK, as summarized in Theorem 3.7 and Corollary 3.8.

Theorem 3.7. *Suppose the assumptions in Lemma 2.5 hold. Let $\lambda_j^{NTK}, \dots, \lambda_{j+k_\rho-1}^{NTK}$ be a group of isolated eigenvalues of $\mathbf{K}_{NTK,\ell}$ that converge to the same limit $\beta_{\ell,1}\rho + \beta_{\ell,0}$, and let $\tilde{\Pi}_\rho$ denote the projection onto the eigenspace spanned by the eigenvectors associated with these eigenvalues. Here, ρ is defined in Theorem 3.2. Suppose Assumptions 2.1-2.3 hold. For $\underline{\mathbf{H}}_2(\rho)$, we denote the left and right eigenvectors corresponding to 0 as $(\mathbf{V}_{l,z})_i$ and $(\mathbf{V}_{r,z})_i$, respectively, where $i = 1, \dots, k_\rho$. If $\text{dist}(\rho, \mathcal{S}_{2,\ell})$ is bounded away from 0, then we have*

$$\mathbf{u}^\top \tilde{\Pi}_\rho \mathbf{v} = -h_\ell^2(\rho) \mathbf{u}^\top \mathbf{Q}_1(\rho) \mathbf{J} \tilde{\mathbf{F}}(\rho) \Gamma^{-1}(\rho) \mathbf{J}^\top \mathbf{Q}_1(\rho) \mathbf{v} + o_{a.s.}(1),$$

where

$$\tilde{\mathbf{F}}(z) = \sum_{i=1}^{k_\rho} \frac{(\mathbf{V}_{r,\rho})_i (\mathbf{V}_{l,\rho})_i^\top}{(\mathbf{V}_{l,\rho})_i^\top [\partial_z \underline{\mathbf{H}}_2(z)]_{z=\rho} (\mathbf{V}_{r,\rho})_i}.$$

Corollary 3.8. *Under the same conditions as those in Theorem 3.7, we have*

$$\frac{1}{p} \mathbf{J}^\top \tilde{\Pi}_\rho \mathbf{J} = -\frac{1}{p} h_\ell^2(\rho) \Gamma(\rho) \tilde{\mathbf{F}} + o_{a.s.}(1).$$

Remark 3.9. For the NTK, analogous to Remark 3.6, if ρ_+ satisfies $\det \underline{\mathbf{H}}_2(\rho_+) = 0$ and $\rho_+ \rightarrow \rho \in \mathcal{H}_{2,\ell}^p$, then its associated eigenspace is non-informative.

Remark 3.10. We note that Wang et al. (2024) investigated how spiked eigenstructures in the input data propagate through the hidden layers of a neural network by analyzing the spectrum of the CK, under the assumptions that the input data contains spiked eigenvalues and that the same activation function is applied at every layer. In contrast, our work examines both the CK and NTK. Our analysis demonstrates that isolated eigenvalues in the kernel matrices may arise from the underlying group structure of the input features, with the corresponding eigenvectors carrying useful information, even in the absence of a spiked structure in the original data as required in Wang et al. (2024). Additionally, our framework allows for different activation functions across layers and investigates their effects on the spectral properties of the kernel matrices.

3.3. Additional simulations and real data analysis

In this subsection, we present additional simulations on both synthetic GMM and real data to further support our theoretical findings.

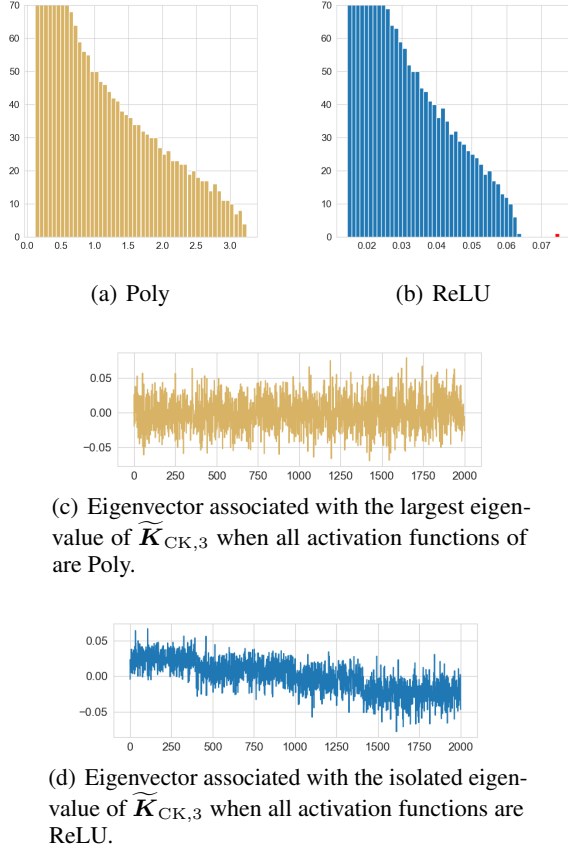


Figure 4. Eigenvalue histograms ((a)-(b)) and eigenvectors ((c)-(d)) corresponding to the isolated (largest) eigenvalues of $\tilde{\mathbf{K}}_{\text{CK},3}$ obtained with Poly and ReLU activation functions. Parameters settings: $n = 2000$, $p = 3600$, $c_1 = c_3 = 0.2$, $c_2 = c_4 = 0.3$ and $\mathbf{C}_a = (1 + 8(a - 1)/\sqrt{p})\mathbf{I}_p$ for $a = 1, 2, 3, 4$. The width d_1, d_2, d_3 are identical to those in Figure 1.

Figures 4(a)-(b) visualize the eigenvalues of $\tilde{\mathbf{K}}_{\text{CK},3}$ under different activation function settings. In Figure 4(a), all activation functions are set to Poly, consistent with those used in Figure 2, whereas in Figure 4(b), all activation functions are set to ReLU. The ReLU activation functions have been normalized and centered to satisfy the conditions specified in Assumption 2.3. Unlike the results shown in Figure 2(a), no isolated eigenvalues are observed when all activation functions are Poly. In contrast, when all activation functions are ReLU, an isolated eigenvalue emerges. Figures 4(c)-(d) show the eigenvectors corresponding to the largest eigenvalue in Figure 4(a) and the isolated eigenvalue in Figure 4(b), respectively. The former appears to be non-informative, whereas the latter is notably informative.

These observations underscore the critical role of activation function selection in determining the spectral properties of the kernel matrix.

Next, we turn to real data analysis. The input data consists

of 1600 randomly selected images from each of the digit classes 1 and 7 in the MNIST dataset, with the class-specific mean subtracted from each group. Figure 5(a) displays the spectrum of the CK, obtained using a three-layer neural network. The eigenvectors associated with the isolated eigenvalues are shown in Figure 5(b). The spectrum reveals four isolated eigenvalues, with the eigenvector corresponding to the largest eigenvalue being informative, while the other three appear non-informative. These observations are in line with our theoretical findings.

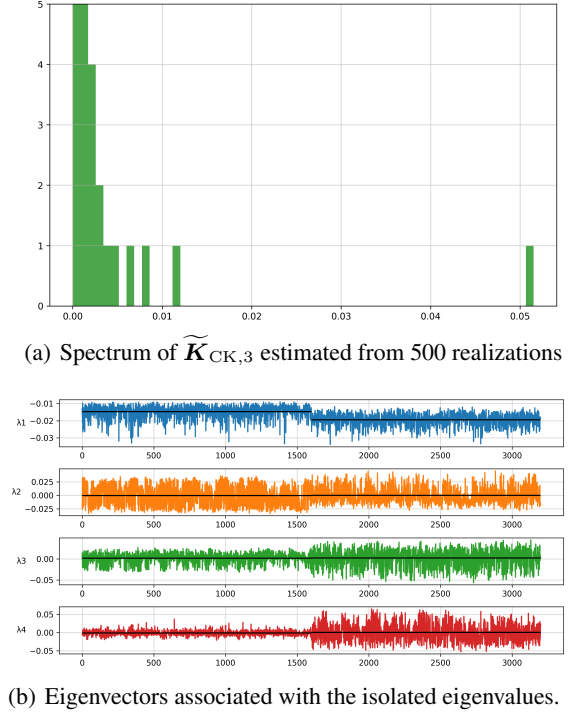


Figure 5. Eigenvalue histograms (a) and eigenvectors (b) corresponding to the isolated eigenvalue of $\tilde{\mathbf{K}}_{\text{CK},3}$. Activations=[Sin, ReLU/10, Sin]. The weights $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times p}$, $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, $\mathbf{W}_3 \in \mathbb{R}^{d_3 \times d_2}$ consist of i.i.d. standard normal entries, where $d_1 = 2000$, $d_2 = d_3 = 1000$.

4. Theoretical insights for machine learning applications

It is known that the NTK theory leads to concrete convergence and generalization results (Bai & Lee, 2019) of neural networks. Let $\mathbf{y}$ be the true label vector of the training data and $\hat{\mathbf{y}}(t)$ be the prediction at time t . In an ultra-wide neural network with training loss $\frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}(t)\|^2$, the time evolution of the residual $\mathbf{y} - \hat{\mathbf{y}}(t)$ during early training is approximately described by the following ordinary differential equation (Jacot et al., 2018; Du et al., 2018; 2019):

$$\frac{d}{dt}\hat{\mathbf{y}}(t) = \mathbf{K}_{\text{NTK}}(\mathbf{y} - \hat{\mathbf{y}}(t)). \quad (24)$$

This paper investigates the spectral properties of the conjugate kernel and neural tangent kernel, revealing the evolution of inherent group features through hidden layers. From Theorem 3.1 and Theorem 3.2, it can be seen that the occurrence of the isolated eigenvalues depends on two key factors: (1) differences in the covariance matrices between different classes and (2) the choice of activation functions. The former shapes the vector $\mathbf{t}$ and the matrix $\mathbf{T}$, while the latter determines the coefficients α 's (β 's). When isolated eigenvalues are present, these factors also determine the asymptotic behaviors of the corresponding eigenvectors. This conclusion is further supported by comparing the simulation results presented in Figures 1-4.

Acknowledgements

Impact Statement

References

- 9

- low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.
- Murray, M., Jin, H., Bowman, B., and Montufar, G. Characterizing the spectrum of the ntk via a power series expansion. *arXiv preprint arXiv:2211.07844*, 2022.
- Nguyen, Q., Mondelli, M., and Montufar, G. F. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pp. 8119–8129. PMLR, 2021.
- Papayan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.
- Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.
- Radhakrishnan, A., Beaglehole, D., Pandit, P., and Belkin, M. Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.
- Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.
- Wang, Z. and Zhu, Y. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability*, 34(2):1896–1947, 2024.
- Wang, Z., Engel, A., Sarwate, A. D., Dumitriu, I., and Chiang, T. Spectral evolution and invariance in linear-width neural networks. *Advances in neural information processing systems*, 36:20695–20728, 2023.
- Wang, Z., Wu, D., and Fan, Z. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4891–4957. PMLR, 2024.
- Xie, B., Liang, Y., and Song, L. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224. PMLR, 2017.
- Yang, G. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.
- Yang, G. and Salman, H. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

A. Appendix

In the later proof, the following lemmas are needed.

Lemma A.1. (*Couillet & Benaych-Georges, 2016*) Suppose Assumption 2.1 holds and let $z \in \mathbb{C}$ be a point that is at least a distance $c > 0$ away from $\mathcal{S}$. We have

$$\mathbb{E}\psi^\top \mathbf{G}(z)\psi = \frac{1}{p} \sum_{a=1}^K c_a m_a(z) \text{tr} \mathbf{C}_a^2 + O\left(\frac{1}{p}\right), \quad \mathbb{E}|\psi^\top \mathbf{G}(z)\psi - \mathbb{E}\psi^\top \mathbf{G}(z)\psi|^m = O(p^{-\frac{m}{2}}).$$

Lemma A.2. Under Assumption 2.1, for any deterministic $\mathbf{u} = \{u_i\}_{i=1}^n$ with finite Euclidean norm, we have

$$\mathbf{u}^\top \mathbf{G}(z)\psi \rightarrow 0 \quad \text{a.s.} \quad \text{for } z \text{ such that } \text{dist}(\mathcal{S} \cup \{0\}, z) > c,$$

where c is a positive constant.

We postpone the proof of Lemma A.2 to the end of this section.

A.1. Proof of Theorem 3.1 and Theorem 3.2

Noting that the asymptotic spectral equivalents for the CK and NTK are identical up to a change of the associated coefficients α 's to β 's, we only provide the proof for Theorem 3.1 here and the proof for Theorem 3.2 is exactly the same.

We devote the majority of this section to finding the isolated eigenvalues of $\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top$, which, up to multiplication by $\alpha_{\ell,1}$ and an addition of $\alpha_{\ell,0}$, constitute the eigenvalues of $\widetilde{\mathbf{K}}_{\text{CK},\ell}$. In order to find the isolated eigenvalues of $\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top$, one only need to solve the determinant equation

$$\det(\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n) = 0 \quad (25)$$

for z away from $\mathcal{S}_{1,\ell}$. An application of Lemma A.8 yields that (25) shares the same solutions as

$$\det(\alpha_{\ell,1}^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{V} + \mathbf{I}_{K+1}) = 0. \quad (26)$$

According to Lemma 2.6 and the fact that

$$\frac{1}{p} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{J} = \frac{c_0}{p} \text{diag}\{n_a m_a(z)\}_{a=1}^K = \text{diag}\{c_a m_a(z)\}_{a=1}^K = \mathbf{\Gamma}(z),$$

we have

$$\frac{1}{p} \mathbf{J}^\top \mathbf{G}(z) \mathbf{J} = \frac{1}{p} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{J} + o_{a.s.}(1) = \mathbf{\Gamma}(z) + o_{a.s.}(1).$$

Lemma A.1 gives

$$\psi^\top \mathbf{G}(z) \psi = \frac{1}{p} \sum_{a=1}^K c_a m_a(z) \text{tr} \mathbf{C}_a^2 + o_{a.s.}(1).$$

From Lemma A.2, it can be seen that the cross term $\frac{1}{\sqrt{p}} \mathbf{J}^\top \mathbf{G}(z) \psi$ vanishes as $n \rightarrow \infty$. Therefore, we can write

$$\mathbf{V}^\top \mathbf{G}(z) \mathbf{V} = \begin{bmatrix} \frac{1}{p} \mathbf{J}^\top \mathbf{G}(z) \mathbf{J} & \frac{1}{\sqrt{p}} \mathbf{J}^\top \mathbf{G}(z) \psi \\ \frac{1}{\sqrt{p}} \psi^\top \mathbf{G}(z) \mathbf{J} & \psi^\top \mathbf{G}(z) \psi \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}(z) & 0 \\ 0 & \frac{1}{p} \sum_{a=1}^K c_a m_a(z) \text{tr} \mathbf{C}_a^2 \end{bmatrix} + o_{a.s.}(1)$$

as well as

$$\mathbf{I}_{K+1} + \alpha_{\ell,1}^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{V} = \mathbf{H}_1(z) + o_{a.s.}(1), \quad (27)$$

where

$$\mathbf{H}_1(z) = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} := \begin{bmatrix} \mathbf{I}_K + \alpha_{\ell,1}^{-1} [\alpha_{\ell,2} \mathbf{t} \mathbf{t}^\top + \alpha_{\ell,3} \mathbf{T}] \mathbf{\Gamma}(z) & [h_\ell^1(z) - 1] \mathbf{t} \\ \alpha_{\ell,1}^{-1} \alpha_{\ell,2} \mathbf{t}^\top \mathbf{\Gamma}(z) & h_\ell^1(z) \end{bmatrix}. \quad (28)$$

Therefore, it suffices to find the solution of

$$\det \mathbf{H}_1(z) = 0 \quad (29)$$

on the real line by (27) and Lemma A.4.

We first consider the case where $h_\ell^1(z)$ is bounded away from 0. Using the Schur complement formula, we have

$$\begin{aligned} \det \mathbf{H}_1 &= \mathbf{H}_{22} \det (\mathbf{H}_{11} - \mathbf{H}_{22}^{-1} \mathbf{H}_{12} \mathbf{H}_{21}) \\ &= \mathbf{H}_{22}^{1-K} \det (\mathbf{H}_{11} \mathbf{H}_{22} - \mathbf{H}_{12} \mathbf{H}_{21}) = \mathbf{H}_{22}^{1-K} \det \underline{\mathbf{H}}_1. \end{aligned} \quad (30)$$

Therefore, if ρ satisfies that $\text{dist}(\rho, \mathcal{S}_{1,\ell}) > C$ for some constant C and makes $\underline{\mathbf{H}}_1(\rho)$ a singular matrix, with the multiplicity of the eigenvalue 0 denoted by k_ρ , then according to (5), there exists at least one eigenvalue of $\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top$ that converges to ρ almost surely.

If $h_\ell^1(z) \rightarrow 0$, studying (30) is not suitable because $\mathbf{H}_{22}^{1-K} = h_\ell^1(z)^{1-K}$ tends to infinity. We instead consider the matrix $\mathbf{H}_1(z)$ directly. If (29) has a solution ρ_+ such that $\rho_+ \rightarrow \rho \in \mathcal{H}_{1,\ell}^p$, then ρ is a deterministic limit of certain eigenvalues of $\widetilde{\mathbf{K}}_{\text{CK},\ell}$.

Next, we investigate the number of eigenvalues of $\mathbf{K}_{\text{CK},\ell}$ that converge to the limits identified above. By Lemma 2.6, we see that the eigenvalues of $\mathbf{X}^\top \mathbf{X}$ asymptotically do not escape $\mathcal{S} \cup \{0\}$. Therefore, one may find a compact interval $\mathcal{I} = [a_1, b_1] \subset (a, b)$ such that, for large enough n , $\mathbf{X}^\top \mathbf{X}$ has no eigenvalue in $\mathcal{I}$ with probability one and

$$\{\rho \mid \det \mathbf{H}_1(\rho) = 0\} \cap \{a_1, b_1\} = \emptyset. \quad (31)$$

We use $D_{\mathcal{I}}^\circ$ to denote the open disc centered at $(a_1 + b_1)/2$ and of diameter $(b_1 - a_1)$. Let

$$L_1 := \#\{\text{zeros of } \det(\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n) \text{ in } D_{\mathcal{I}}^\circ\}, \quad L_2 := \#\{\rho \mid \rho \in D_{\mathcal{I}}^\circ \text{ and } \det \mathbf{H}_1(\rho) = 0\}.$$

Note that the functions $\det(\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n)$, $\det \mathbf{G}(z)$ and $\det \mathbf{H}_1(z)$ are analytic on $D_{\mathcal{I}}^\circ$ and non-vanishing on $\mathbb{C} \setminus \mathbb{R}$. By Lemma A.5 and the argument principle, we obtain

$$\begin{aligned} L_1 &= \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \frac{\partial_z \det(\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n)}{\det(\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n)} \\ &= \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \frac{\partial_z \det(\mathbf{I}_n + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z))}{\det(\mathbf{I}_n + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z))} + \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \frac{\partial_z \det \mathbf{G}(z)}{\det \mathbf{G}(z)} \\ &= \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \frac{\partial_z \det(\mathbf{I}_{K+1} + \alpha_{\ell,1}^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{V})}{\det(\mathbf{I}_{K+1} + \alpha_{\ell,1}^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{V})} \rightarrow \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \frac{\partial_z \det \mathbf{H}_1(z)}{\det \mathbf{H}_1(z)} = L_2 \end{aligned}$$

almost surely, where $\partial D_{\mathcal{I}}^\circ$ is seen as a positively oriented contour. Since both L_1 and L_2 are integers, the multiplicity of an isolated eigenvalue of $\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top$, which has a deterministic limit ρ , is the same as that of 0 as an eigenvalue of $\mathbf{H}_1(\rho)$. It is straightforward to see that if ρ_+ defined in Theorem 3.1 satisfies $h_\ell^1(\rho_+) \rightarrow 0$, then the multiplicity of ρ_+ denoted by k_ρ^+ is equal to the multiplicity of 0 as an eigenvalue of $\mathbf{H}_1(\rho_+)$. In the case that $h_\ell^1(\rho)$ is away from 0, we have that $\underline{\mathbf{H}}_1(\rho)$ has a 0 eigenvalue of multiplicity k_ρ by (30). This concludes the proof of Theorem 3.1.

A.2. Proof of Theorem 3.4 and Theorem 3.7

For any closed interval $\mathcal{I} \subset \mathbb{R}$ defined above (31), we have

$$-\frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^\circ} \mathbf{u}^\top \mathbf{G}(z) \mathbf{v} dz = 0$$

almost surely. Lemma A.2 implies that, for any deterministic $\mathbf{u}$ and $\mathbf{v}$ with finite Euclidean norm,

$$\mathbf{u}^\top \mathbf{G}(z) \mathbf{V} = \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{V} + o_{a.s.}(1), \quad \mathbf{v}^\top \mathbf{G}(z) \mathbf{V} = \mathbf{v}^\top \mathbf{Q}_1(z) \mathbf{V} + o_{a.s.}(1).$$

Then by using Woodbury's identity, we obtain

$$\begin{aligned}
 \mathbf{u}^\top \widehat{\Pi}_{\mathcal{I}} \mathbf{v} &= -\frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^o} \mathbf{u}^\top (\mathbf{X}^\top \mathbf{X} + \alpha_{\ell,1}^{-1} \mathbf{V} \mathbf{A}_\ell \mathbf{V}^\top - z \mathbf{I}_n)^{-1} \mathbf{v} dz \\
 &= -\frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^o} \mathbf{u}^\top \mathbf{G}(z) \mathbf{v} dz \\
 &\quad + \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^o} \alpha_{\ell,1}^{-1} \mathbf{u}^\top \mathbf{G}(z) \mathbf{V} (\mathbf{I}_{k+1} + \alpha_{\ell,1}^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{V})^{-1} \mathbf{A}_\ell \mathbf{V}^\top \mathbf{G}(z) \mathbf{v} dz \\
 &= \frac{1}{2\pi i} \oint_{\partial D_{\mathcal{I}}^o} \alpha_{\ell,1}^{-1} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{V} \mathbf{H}_1^{-1}(z) \mathbf{A}_\ell \mathbf{V}^\top \mathbf{Q}_1(z) \mathbf{v} dz + o_{a.s.}(1),
 \end{aligned} \tag{32}$$

where $\widehat{\Pi}_{\mathcal{I}}$ is the projection onto the eigenspace spanned by the eigenvectors corresponding to the eigenvalues of $\mathbf{K}_{\text{CK},\ell}$ that lie inside $\mathcal{I}$. Thus what remains to find the limit of (32) is to obtain the deterministic equivalent of $\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell$. Since the zeros of $h_\ell(z)$ are away from $\bar{D}_{\mathcal{I}}^o$, we obtain

$$\mathbf{H}_1^{-1}(z) = \begin{bmatrix} \mathbf{H}_{22} \mathbf{H}_1^{-1}(z) & -\mathbf{H}_1^{-1}(z) \mathbf{H}_{12} \\ -\mathbf{H}_{12} \mathbf{H}_1^{-1} & \mathbf{H}_{22}^{-1} + \mathbf{H}_{22}^{-1} \mathbf{H}_{21} \mathbf{H}_1^{-1}(z) \mathbf{H}_{12} \end{bmatrix}.$$

All notations presented here are defined in (28). Therefore, we can write

$$\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell = \begin{bmatrix} [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{11} & [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{12} \\ [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{21} & [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{22} \end{bmatrix},$$

where

$$\begin{aligned}
 [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{11} &= \alpha_{\ell,3} h_\ell^1(z) \underline{\mathbf{H}}_1^{-1}(z) \mathbf{T} + \alpha_{\ell,2} \underline{\mathbf{H}}_1^{-1}(z) \mathbf{t} \mathbf{t}^\top, \\
 [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{12} &= \alpha_{\ell,2} \underline{\mathbf{H}}_1^{-1}(z) \mathbf{t}, \\
 [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{21} &= -\frac{\alpha_{\ell,2} \alpha_{\ell,3}}{\alpha_{\ell,1}} \mathbf{t}^\top \Gamma(z) \underline{\mathbf{H}}_1^{-1}(z) \mathbf{T} + \frac{\alpha_{\ell,2} \mathbf{t}^\top}{h_\ell^1(z)} + \frac{1}{h_\ell^1(z)} \frac{\alpha_{\ell,2}^2}{\alpha_{\ell,1}} \mathbf{t}^\top \Gamma(z) \underline{\mathbf{H}}_1^{-1}(z) \mathbf{t} \mathbf{t}^\top, \\
 [\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{22} &= \frac{\alpha_{\ell,2}}{h_\ell^1(z)} - \frac{1}{h_\ell^1(z)} \frac{\alpha_{\ell,2}^2}{\alpha_{\ell,1}} \mathbf{t}^\top \Gamma(z) \underline{\mathbf{H}}_1^{-1}(z) \mathbf{t}.
 \end{aligned}$$

Observing that $\mathbf{u}^\top \mathbf{G}(z) \mathbf{V} = \frac{1}{\sqrt{p}} [\mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} \quad 0] + o_{a.s.}(1)$ ($\mathbf{v}^\top \mathbf{G}(z) \mathbf{V} = \frac{1}{\sqrt{p}} [\mathbf{v}^\top \mathbf{Q}_1(z) \mathbf{J} \quad 0] + o_{a.s.}(1)$), it becomes evident that only $[\mathbf{H}_1^{-1}(z) \mathbf{A}_\ell]_{11}$ is needed. Then we deduce that

$$\begin{aligned}
 &\alpha_{\ell,1}^{-1} \mathbf{u}^\top \mathbf{Q}(z) \mathbf{V} \mathbf{H}_1^{-1}(z) \mathbf{A}_\ell \mathbf{V}^\top \mathbf{Q}(z) \mathbf{v} \\
 &= \frac{1}{p} \alpha_{\ell,1}^{-1} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} [\alpha_{\ell,3} h_\ell^1(z) \underline{\mathbf{H}}_1^{-1}(z) \mathbf{T} + \alpha_{\ell,2} \underline{\mathbf{H}}_1^{-1}(z) \mathbf{t} \mathbf{t}^\top] \mathbf{J} \mathbf{Q}_1(z) \mathbf{v} + o_{a.s.}(1).
 \end{aligned}$$

We denote the spectral decomposition of $\underline{\mathbf{H}}_1(z)$ as

$$\underline{\mathbf{H}}_1(z) = \mathbf{U}_{r,z} \Lambda_z \mathbf{U}_{l,z}^\top.$$

For $\rho \in \mathcal{I}$ such that $\underline{\mathbf{H}}_1(\rho)$ is singular and k_ρ denotes the multiplicity of 0 as an eigenvalue of $\underline{\mathbf{H}}_1(\rho)$, we denote the left and right eigenvectors corresponding to 0 as $(\mathbf{U}_{l,\rho})_i$ and $(\mathbf{U}_{r,\rho})_i$ respectively, where i ranges from 1 to k_ρ . For a matrix $\mathbf{M}(z) = \{M_{ij}(z)\}_{i,j=1}^K$, denote $\text{Res}(\mathbf{M}) = \{\text{Res}(M_{ij}(z))\}_{i,j=1}^K$. Then we have

$$\begin{aligned}
 \lim_{z \rightarrow \rho} (z - \rho) \underline{\mathbf{H}}_1^{-1}(z) &= \lim_{z \rightarrow \rho} (z - \rho) \mathbf{U}_{r,z} \Lambda_z^{-1} \mathbf{U}_{l,z}^\top \\
 &= \lim_{z \rightarrow \rho} (z - \rho) \sum_{i=1}^{k_\rho} \frac{(\mathbf{U}_{r,z})_i (\mathbf{U}_{l,z})_i^\top}{(\mathbf{U}_{l,z})_i^\top \underline{\mathbf{H}}_1(z) (\mathbf{U}_{r,z})_i} \\
 &= \lim_{z \rightarrow \rho} \sum_{i=1}^{k_\rho} \frac{(\mathbf{U}_{r,z})_i (\mathbf{U}_{l,z})_i^\top}{\partial_z (\mathbf{U}_{l,z})_i^\top \underline{\mathbf{H}}_1(z) (\mathbf{U}_{r,z})_i},
 \end{aligned}$$

where the last equation is obtained by applying the L'Hôpital's rule. Since

$$(\mathbf{U}_{l,\rho})_i^\top \underline{\mathbf{H}}_1(\rho) = 0 \quad \text{and} \quad \underline{\mathbf{H}}_1(\rho)(\mathbf{U}_{r,\rho})_i = 0,$$

we obtain

$$\begin{aligned} & [\partial_z (\mathbf{U}_{l,z})_i^\top \underline{\mathbf{H}}_1(z) (\mathbf{U}_{r,z})_i]_{z=\rho} \\ &= [\partial_z (\mathbf{U}_{l,z})_i^\top]_{z=\rho} \underline{\mathbf{H}}_1(z) (\mathbf{U}_{r,z})_i + (\mathbf{U}_{l,z})_i^\top [\partial_z \underline{\mathbf{H}}_1(z)]_{z=\rho} (\mathbf{U}_{r,z})_i + (\mathbf{U}_{l,z})_i^\top \underline{\mathbf{H}}_1(z) [\partial_z ((\mathbf{U}_{r,z})_i)]_{z=\rho} \\ &= (\mathbf{U}_{l,z})_i^\top [\partial_z \underline{\mathbf{H}}_1(z)]_{z=\rho} (\mathbf{U}_{r,z})_i. \end{aligned}$$

It is easy to see that

$$\begin{aligned} & \frac{1}{p} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} \left\{ \underline{\mathbf{H}}_1^{-1}(z) \left[\frac{\alpha_{\ell,3}}{\alpha_{\ell,1}} h_\ell^1(z) \mathbf{T} + \frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \mathbf{t} \mathbf{t}^\top \right] \right\} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{v} \\ &= \frac{1}{p} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} \left\{ \underline{\mathbf{H}}_1^{-1}(z) \left[\frac{\alpha_{\ell,3}}{\alpha_{\ell,1}} h_\ell^1(z) \mathbf{T} + \frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \mathbf{t} \mathbf{t}^\top \right] \Gamma(z) \Gamma^{-1}(z) \right\} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{v} \\ &= \frac{1}{p} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} \left\{ \underline{\mathbf{H}}_1^{-1}(z) [\underline{\mathbf{H}}_1(z) - h_\ell^1(z) \Gamma^{-1}(z)] \Gamma^{-1}(z) \right\} \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{v} \\ &= \frac{1}{p} \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} [\Gamma^{-1}(z) - h_\ell^1(z) \underline{\mathbf{H}}_1^{-1}(z) \Gamma^{-1}(z)] \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{v}. \end{aligned}$$

Now one can conclude by (32) that

$$\begin{aligned} \mathbf{u}^\top \hat{\Pi}_{\mathcal{I}} \mathbf{v} &= -\frac{1}{p} \sum_{z \in \mathcal{I}} \text{Res}(h_\ell^1(z) \mathbf{u}^\top \mathbf{Q}_1(z) \mathbf{J} \underline{\mathbf{H}}_1^{-1}(z) \mathbf{J}^\top \mathbf{Q}_1(z) \mathbf{v}) + o_{a.s.}(1) \\ &= -\frac{1}{p} h_\ell^1(\rho) \mathbf{u}^\top \mathbf{Q}_1(\rho) \mathbf{J} \sum_{i=1}^{k_\rho} \frac{(\mathbf{U}_{r,\rho})_i (\mathbf{U}_{l,\rho})_i^\top}{(\mathbf{U}_{l,\rho})_i^\top [\partial_z \underline{\mathbf{H}}_1(z)]_{z=\rho} (\mathbf{U}_{r,\rho})_i} \mathbf{J}^\top \mathbf{Q}_1(\rho) \mathbf{v} + o_{a.s.}(1). \end{aligned}$$

Thus the proof of Theorem 3.4 is completed. Theorem 3.7 can be proved by the same lines as that of Theorem 3.4.

A.3. Proof of Equation (23)

Now, we consider the case that $\rho_+ \rightarrow \rho \in \mathcal{H}_{1,\ell}^p$, with ρ being at a distance from $\mathcal{S}_\ell$ and $\det \mathbf{H}_1(\rho_+) = 0$. Suppose $\|\mathbf{t}\|$ is bounded away from 0, we note that $\lim_{z \rightarrow \rho} \underline{\mathbf{H}}_1(z) = \frac{\alpha_{\ell,2}}{\alpha_{\ell,1}} \mathbf{t} \mathbf{t}^\top \Gamma(z)$ and the RHS is a rank-1 matrix. Hence there are $K-1$ eigenvalues of $\underline{\mathbf{H}}_1(\rho_+)$ converging to 0 almost surely. It is natural to see that the derivative $\partial_z \underline{\mathbf{H}}_1(z)|_{z=\rho_+}$ is well defined and not close to 0. Based on the above facts and applying the L'Hôpital's rule again, we have

$$\begin{aligned} \lim_{z \rightarrow \rho_+} (z - \rho_+) h_\ell^1(z) \underline{\mathbf{H}}_1^{-1}(z) &= \lim_{z \rightarrow \rho_+} (z - \rho_+) h_\ell^1(z) \mathbf{U}_{r,z} \Lambda_z^{-1} \mathbf{U}_{l,z}^\top \\ &= \lim_{z \rightarrow \rho_+} h_\ell^1(z) \lim_{z \rightarrow \rho_+} \sum_{i=1}^{K-1} (z - \rho_+) \frac{(\mathbf{U}_{r,\rho})_i (\mathbf{U}_{l,\rho})_i^\top}{(\mathbf{U}_{l,\rho_+})_i^\top \underline{\mathbf{H}}_1(z) (\mathbf{U}_{r,\rho_+})_i} \\ &= \lim_{z \rightarrow \rho_+} h_\ell^1(z) \sum_{i=1}^{K-1} \frac{(\mathbf{U}_{r,\rho_+})_i (\mathbf{U}_{l,\rho_+})_i^\top}{(\mathbf{U}_{l,\rho_+})_i^\top [\partial_z \underline{\mathbf{H}}_1(z)]_{z=\rho_+} (\mathbf{U}_{r,\rho_+})_i} \\ &= 0. \end{aligned}$$

If $\|\mathbf{t}\| \rightarrow 0$, then ρ_+ becomes a removable singularity for $h_\ell^1(z) \underline{\mathbf{H}}_1^{-1}(z)$, thereby ensuring that the residue remains zero. The conclusion can be obtained by repeating the proof of Theorem 3.4.

A.4. Proof of Lemma A.2

We divide the proof into three steps:

Step 1: Fix $\eta > 0$. For any $z \in \mathbb{C}$ with $\Im z \geq \eta$, the expectation $\mathbb{E} \mathbf{u}^\top \mathbf{G}(z) \mathbf{v} = o(1)$.

Step 2: Convergence of $\mathbf{u}^\top \mathbf{G}(z) \psi$ to $\mathbb{E} \mathbf{u}^\top \mathbf{G}(z) \psi$ almost surely.

Step 3: For any z such that $\text{dist}(\mathcal{S} \cup \{0\}, z) > c$ for some constant c , we also have $\mathbf{u}^\top \mathbf{G}(z) \psi \rightarrow 0$ a.s.

We assume $\Im z \geq \eta > 0$ if $\Re z = E > 0$. For the case $\Im z \leq -\eta < 0$, one only needs to replace η by $|\eta|$ in the following proof. Define

$$\hat{\mathbf{X}}_j = \mathbf{X} - \mathbf{x}_j \mathbf{e}_j^\top, \quad \hat{\mathbf{G}}_j(z) = (\hat{\mathbf{X}}_j^\top \hat{\mathbf{X}}_j - z \mathbf{I}_n)^{-1},$$

where $\mathbf{e}_j \in \mathbb{R}^n$ denotes the vector with the j th element being 1 and otherwise being 0. Note that $\psi = \sum_{j=1}^n \psi_j \mathbf{e}_j$. We can write

$$\mathbf{u}^\top \mathbf{G}(z) \psi = \sum_{j=1}^n \mathbf{u}^\top \mathbf{G}(z) \psi_j \mathbf{e}_j.$$

Decompose the term $\mathbb{E} \mathbf{u}^\top \mathbf{G}(z) \psi_j \mathbf{e}_j$ as

$$\mathbf{u}^\top \mathbb{E} \mathbf{G}(z) \psi_j \mathbf{e}_j = \mathbf{u}^\top \mathbb{E} [\mathbf{G}(z) - \hat{\mathbf{G}}_j(z)] \psi_j \mathbf{e}_j + \mathbf{u}^\top \mathbb{E} \hat{\mathbf{G}}_j(z) \psi_j \mathbf{e}_j.$$

Since $\hat{\mathbf{G}}_j(z)$ is independent of ψ_j , we have $\mathbb{E} \hat{\mathbf{G}}_j(z) \psi_j \mathbf{e}_j = \mathbb{E} \hat{\mathbf{G}}_j(z) \mathbb{E} \psi_j \mathbf{e}_j = 0$. Let

$$\mathbf{X}_j = [\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n], \quad \mathbf{G}_j(z) = (\mathbf{X}_j^\top \mathbf{X}_j - z \mathbf{I}_{n-1})^{-1},$$

$$\beta_j = (\mathbf{x}_j^\top \mathbf{x}_j - z - \mathbf{x}_j^\top \hat{\mathbf{X}}_j \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j)^{-1}, \quad b_j = \left(\frac{\text{tr} \mathbf{C}_{g(j)}}{p} - z - \frac{\mathbb{E} \text{tr} \mathbf{C}_{g(j)} \mathbf{X}_j \mathbf{G}_j(z) \mathbf{X}_j^\top}{p} \right)^{-1},$$

where $g(j) \in \{1, \dots, K\}$ denotes the class to which $\mathbf{x}_j$ belongs. All of these quantities are bounded in absolute value by η^{-1} . Then from Lemma A.6, we get for $m \geq 2$,

$$\begin{aligned} \mathbb{E} |\beta_j - b_j|^m &\leq \frac{1}{\eta^{2m}} \mathbb{E} \left| \mathbf{x}_j^\top \mathbf{x}_j - \mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j - \left(\frac{\text{tr} \mathbf{C}_{g(j)}}{p} - z - \frac{\mathbb{E} \text{tr} \mathbf{C}_{g(j)} \mathbf{X}_j \mathbf{G}_j(z) \mathbf{X}_j^\top}{p} \right) \right|^m \\ &= O((p\eta^6)^{-\frac{m}{2}}), \end{aligned} \quad (33)$$

and

$$\mathbb{E} \psi_j^m = \frac{1}{p^m} \mathbb{E} (\mathbf{x}_j^\top \mathbf{x}_j - \text{tr} \mathbf{C}_{g(j)})^m = O(p^{-\frac{m}{2}}), \quad m \geq 2. \quad (34)$$

By the Schur complement formula, we obtain

$$\psi_j \mathbf{u}^\top [\mathbf{G}(z) - \hat{\mathbf{G}}_j(z)] \mathbf{e}_j = -\psi_j [(\beta_j - z^{-1}) u_j - \beta_j \mathbf{u}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j],$$

where $\mathbf{u}_j = [u_1, \dots, u_{j-1}, u_{j+1}, \dots, u_n]$. It follows from (33) and (34) that

$$\begin{aligned} |\mathbb{E} \psi_j \beta_j u_j| &= |\mathbb{E} \psi_j (\beta_j - b_j) u_j| \leq |u_j| \mathbb{E} |\psi_j (\beta_j - b_j)| \\ &\leq |u_j| \sqrt{\mathbb{E} |\psi_j|^2 \mathbb{E} |\beta_j - b_j|^2} = |u_j| O\left(\frac{1}{\eta^3 p}\right). \end{aligned}$$

Thus we have

$$\sum_{j=1}^n \mathbb{E} \psi_j \beta_j u_j \leq \sum_{j=1}^n |\mathbb{E} \psi_j \beta_j u_j| \leq \frac{c}{\eta^3 p} \sum_{j=1}^n |u_j| = O\left(\frac{1}{\eta^3 p^{1/2}}\right) \rightarrow 0. \quad (35)$$

We write $\mathbf{x}_j = \frac{1}{\sqrt{p}} \mathbf{C}_{g(j)}^{1/2} \mathbf{z}_j$, where $\mathbf{z}_j = \{z_i^j\}_{i=1}^p$ is a random vector with i.i.d. standard Gaussian entries. Denote $\mathbf{y}_j^\top = \mathbf{u}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{C}_{g(j)}^{1/2} = \{y_i^j\}_{i=1}^p$. It is easy to see that $\|\mathbf{y}_j\| \leq \|\mathbf{u}_j\| \|\mathbf{G}_j(z) \mathbf{X}_j\| \|\mathbf{C}_{g(j)}^{1/2}\| = O(\eta^{-1})$. For $m \geq 1$, by Lemma A.7, it is true that

$$\begin{aligned} \mathbb{E} |\mathbf{u}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j|^{2m} &= \mathbb{E} \left| \frac{1}{\sqrt{p}} \mathbf{y}_j^\top \mathbf{z}_j \right|^{2m} = \frac{1}{p^m} \mathbb{E} \mathbb{E} [|\mathbf{y}_j^\top \mathbf{z}_j|^{2m} | \mathbf{y}_j] = \frac{1}{p^m} \mathbb{E} \mathbb{E} \left[\left| \sum_{i=1}^p y_i^j z_i^j \right|^{2m} | \mathbf{y}_j \right] \\ &\leq \mathbb{E} \left\{ c \left(\sum_{i=1}^p \frac{|y_i^j|^2}{p} \right)^m + c \mathbb{E} \left[\sum_{i=1}^p \left| y_i^j \frac{z_i^j}{\sqrt{p}} \right|^{2m} | \mathbf{y}_j \right] \right\} \\ &\leq \frac{C}{\eta^{2m} p^m} + C \mathbb{E} \frac{\sum_{i=1}^p |y_i^j|^{2m}}{p^m} = O\left(\frac{1}{\eta^{2m} p^m}\right), \end{aligned} \quad (36)$$

where we use the inequality

$$\sum_{i=1}^p |y_i^j|^{2m} \leq \left(\sum_{i=1}^p |y_i^j|^2 \right)^m = \|\mathbf{y}_j\|^{2m} = O(\eta^{-2m}).$$

Write $\mathbb{E}\psi_j \mathbf{u}_j^\top \beta_j \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j = \mathbb{E}\psi_j \mathbf{u}_j^\top (\beta_j - b_j) \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j$. It can be seen from (33), (34) and (36) that

$$\begin{aligned} |\mathbb{E}\psi_j \mathbf{u}_j^\top \beta_j \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j| &\leq \mathbb{E}|\psi_j \mathbf{u}_j^\top (\beta_j - b_j) \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j| \leq \mathbb{E}|\psi_j| |\beta_j - b_j| |\mathbf{u}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{x}_j| \\ &\leq (\mathbb{E}|\psi_j|^4 \mathbb{E}|\beta_j - b_j|^4)^{\frac{1}{4}} \sqrt{\mathbb{E}|\mathbf{y}_j^\top \mathbf{x}_j|^2} \\ &\leq \frac{c}{\eta^4 p} \sqrt{\mathbb{E}|\mathbf{y}^\top \mathbf{z}_j|^2} = O\left(\frac{1}{\eta^5 p^{3/2}}\right). \end{aligned}$$

This, in conjunction with (35), establishes **Step 1**.

According to the Borel-Cantelli lemma, it suffices to prove that for any $\varepsilon > 0$, the probability $\mathbb{P}(|\mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} - \mathbb{E} \mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi}| > \varepsilon) = O(p^{-1-c})$ to establish **Step 2**, where c is a positive constant. Let $\mathbb{E}_0$ denote expectation and $\mathbb{E}_j$ denote conditional expectation with respect to the σ -field generated by $\mathbf{x}_1, \dots, \mathbf{x}_j$. Using the Schur complement formula again, we write

$$\begin{aligned} \mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} - \mathbb{E} \mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} &= \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} \\ &= \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) [\mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} - \mathbf{u}^\top \hat{\mathbf{G}}_j(z) (\boldsymbol{\psi} - \mathbf{e}_j \psi_j)] \\ &= \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) [-u_j \beta_j \psi_j + u_j \beta_j \mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j + \psi_j \beta_j \mathbf{u}_j^\top \mathbf{G}_j \mathbf{X}_j^\top \mathbf{x}_j + \beta_j \mathbf{u}_j^\top \mathbf{G}_j \mathbf{X}_j^\top \mathbf{x}_j \mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j] \\ &= \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) [\gamma_{j1} + \gamma_{j2} + \gamma_{j3} + \gamma_{j4}], \end{aligned}$$

where $\boldsymbol{\psi}_j = [\psi_1, \dots, \psi_{j-1}, \psi_{j+1}, \dots, \psi_n]$. According to Lemma A.7, we have

$$\begin{aligned} \mathbb{E} \left| \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j1} \right|^4 &\leq c \mathbb{E} \left(\sum_{j=1}^n \mathbb{E} |\gamma_{j1}|^2 \right)^2 + c \sum_{j=1}^n \mathbb{E} |\gamma_{j1}|^4 \\ &\leq \frac{c}{\eta^4} \left(\sum_{j=1}^n u_j^2 \mathbb{E} \psi_j^2 \right)^2 + \frac{c}{\eta^4} \sum_{j=1}^n u_j^4 \mathbb{E} \psi_j^4 \\ &= O\left(\frac{1}{\eta^4 p^2}\right). \end{aligned}$$

Similarly, we obtain

$$\mathbb{E} \left| \sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j2} \right|^4 \leq \frac{c}{\eta^4} \left(\sum_{j=1}^n u_j^2 \mathbb{E} |\mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j|^2 \right)^2 + \frac{c}{\eta^4} \sum_{j=1}^n u_j^4 \mathbb{E} |\mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j|^4.$$

Denote $\mathbf{r}_j^\top = \boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top$. It is apparent that

$$\|\mathbf{r}_j\|^2 = \mathbf{r}_j^\top \bar{\mathbf{r}}_j = \boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{X}_j \mathbf{G}_j(\bar{z}) \boldsymbol{\psi}_j.$$

According to Lemma A.1 and the identity

$$\mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{X}_j \mathbf{G}_j(\bar{z}) = \mathbf{G}_j(\bar{z}) + z \mathbf{G}_j(z) \mathbf{G}_j(\bar{z}) = \mathbf{G}_j(\bar{z}) + z \frac{\mathbf{G}_j(z) - \mathbf{G}_j(\bar{z})}{z - \bar{z}},$$

we get for $m \geq 1$,

$$\begin{aligned}
 \mathbb{E}\|\mathbf{r}_j\|^{2m} &\leq c\mathbb{E}\left(\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{X}_j \mathbf{G}_j(\bar{z}) \boldsymbol{\psi}_j - \mathbb{E}\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{X}_j \mathbf{G}_j(\bar{z}) \boldsymbol{\psi}_j\right)^m + c(\mathbb{E}\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \mathbf{X}_j^\top \mathbf{X}_j \mathbf{G}_j(\bar{z}) \boldsymbol{\psi}_j)^m \\
 &\leq C\mathbb{E}\left(|\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \boldsymbol{\psi}_j - \mathbb{E}\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \boldsymbol{\psi}_j|^m + \frac{1}{\eta^m} |\boldsymbol{\psi}_j^\top [\mathbf{G}_j(z) - \mathbf{G}_j(\bar{z})] \boldsymbol{\psi}_j - \mathbb{E}\boldsymbol{\psi}_j^\top [\mathbf{G}_j(z) - \mathbf{G}_j(\bar{z})] \boldsymbol{\psi}_j|^m\right) \\
 &\quad + C\left|\mathbb{E}\boldsymbol{\psi}_j^\top \frac{\mathbf{G}_j(z) - \mathbf{G}_j(\bar{z})}{z - \bar{z}} \boldsymbol{\psi}_j\right|^m + C|\mathbb{E}\boldsymbol{\psi}_j^\top \mathbf{G}_j(z) \boldsymbol{\psi}_j|^m \\
 &= O(1),
 \end{aligned}$$

where we use Lemma A.1 and the observations that

$$\begin{aligned}
 \mathbb{E}\boldsymbol{\psi}_j^\top \frac{\mathbf{G}_j(z) - \mathbf{G}_j(\bar{z})}{z - \bar{z}} \boldsymbol{\psi}_j &= \frac{\sum_{a=1}^K c_a [m_a(z) - m_a(\bar{z})]}{p[z - \bar{z}]} \text{tr} \mathbf{C}_a^2 + o(1), \\
 c_0 \frac{m_a(z) - m_a(\bar{z})}{z - \bar{z}} &= \int \frac{1}{|\lambda - z|^2} d\nu_a(\lambda) \geq \frac{1}{\text{dist}^2(\mathcal{S}, z)}.
 \end{aligned}$$

Following a similar argument as in (36), we have for $m \geq 1$,

$$\mathbb{E}|\mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j|^{2m} = \mathbb{E}|\mathbf{r}_j^\top \mathbf{x}_j|^{2m} = \mathbb{E}[\|\mathbf{r}_j^\top \mathbf{x}_j\|^{2m} | \mathbf{X}_j] \leq c\mathbb{E} \frac{\|\mathbf{r}_j\|^m + \|\mathbf{r}_j\|^{2m}}{p^m} = O\left(\frac{1}{p^m}\right).$$

Therefore, we have $\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j2} \rightarrow 0$ a.s. by showing that

$$\mathbb{E}\left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j2}\right|^4 \leq \frac{c}{\eta^4} \left[\left(\sum_{j=1}^n \frac{u_j^2}{p}\right)^2 + c \sum_{j=1}^n \frac{u_j^4}{p^2}\right] = O\left(\frac{1}{\eta^4 p^2}\right).$$

Similarly, for $m \geq 1$, it can be obtained that

$$\mathbb{E}|\gamma_{j3}|^{2m} = \mathbb{E}|\boldsymbol{\psi}_j \beta_j \mathbf{u}_j^\top \mathbf{G}_j \mathbf{X}_j^\top \mathbf{x}_j|^{2m} \leq \frac{1}{\eta^{2m}} \sqrt{\mathbb{E} \boldsymbol{\psi}_j^{4m} \mathbb{E} |\mathbf{y}_j^\top \mathbf{x}_j|^{4m}} = O\left(\frac{1}{\eta^{6m} p^{2m}}\right) \quad (37)$$

and

$$\mathbb{E}|\gamma_{j4}|^{2m} = \mathbb{E}|\beta_j \mathbf{u}_j^\top \mathbf{G}_j \mathbf{X}_j^\top \mathbf{x}_j \mathbf{x}_j^\top \mathbf{X}_j \mathbf{G}_j \boldsymbol{\psi}_j|^{2m} \leq \frac{1}{\eta^{2m}} \sqrt{\mathbb{E} |\mathbf{y}_j^\top \mathbf{x}_j|^{4m} \mathbb{E} |\mathbf{r}_j^\top \mathbf{x}_j|^{4m}} = O\left(\frac{1}{\eta^{10m} p^{2m}}\right). \quad (38)$$

One can conclude by Lemma A.7 that

$$\left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j3}\right|^4 \leq c \left(\sum_{j=1}^n \mathbb{E} |\gamma_{j3}|^2\right)^2 + c \sum_{j=1}^n \mathbb{E} |\gamma_{j3}|^4 = O\left(\frac{1}{p^2 \eta^{12}}\right)$$

and

$$\left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{j4}\right|^4 \leq c \left(\sum_{j=1}^n \mathbb{E} |\gamma_{j4}|^2\right)^2 + c \sum_{j=1}^n \mathbb{E} |\gamma_{j4}|^4 = O\left(\frac{1}{p^2 \eta^{20}}\right).$$

Combining the results above, one can conclude that

$$\begin{aligned}
 \mathbb{P}\left(\left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) [\gamma_{j1} + \gamma_{j2} + \gamma_{j3} + \gamma_{j4}]\right| > \varepsilon\right) &\leq \varepsilon^{-4} \mathbb{E} \left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) [\gamma_{j1} + \gamma_{j2} + \gamma_{j3} + \gamma_{j4}]\right|^4 \\
 &\leq C \varepsilon^{-4} \sum_{k=1}^4 \mathbb{E} \left|\sum_{j=1}^n (\mathbb{E}_j - \mathbb{E}_{j-1}) \gamma_{jk}\right|^4 \\
 &= O\left(\frac{1}{\eta^{20} p^2}\right).
 \end{aligned}$$

Hence, the proof of **Step 2** is completed. Moreover, for any z with $\Re z < 0$, we can obtain the result by replacing η by $|\Re z|$ in the proof. Although we assume z is fixed, the same conclusion also holds for $\eta = \eta_n$ converging to 0 sufficiently slowly such as $\eta_n = (\log n)^{-1}$, when $\Re z$ is bounded away from $\mathcal{S} \cup \{0\}$. We notice that $\mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi}$ is an analytic function on any open set $O \subset \mathbb{C}$ if $\text{dist}(\text{Spec}(\mathbf{X}^\top \mathbf{X}), O)$ is away from 0. Recall that $\text{dist}(\text{Spec}(\mathbf{X}^\top \mathbf{X}), \mathcal{S} \cup \{0\}) \rightarrow 0$ almost surely, for $z \in O$ with $0 \leq |\Im z| \leq \eta_n$ and $\text{dist}(\mathcal{S} \cup \{0\}, \Re z) > c$ for some positive constant c . It is clear that

$$|\mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi}| \leq |\mathbf{u}^\top \mathbf{G}(z) \boldsymbol{\psi} - \mathbf{u}^\top \mathbf{G}(z + i\eta_n) \boldsymbol{\psi}| + |\mathbf{u}^\top \mathbf{G}(z + i\eta_n) \boldsymbol{\psi}| \rightarrow 0 \quad a.s.$$

Therefore, we conclude **Step 3** and complete the proof of Lemma A.2.

A.5. Auxiliary lemmas

This section contains several established results from previous work which are needed in our proof.

Lemma A.3. (Theorem A.43 in [Bai & Silverstein \(2010\)](#)). Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ Hermitian matrices, and let $F^{\mathbf{A}}$ and $F^{\mathbf{B}}$ denote their respective empirical spectral distributions. Then

$$\|F^{\mathbf{A}} - F^{\mathbf{B}}\| \leq \frac{1}{n} \text{rank}(\mathbf{A} - \mathbf{B}).$$

Lemma A.4. (Theorem A.46 in [Bai & Silverstein \(2010\)](#)). Let $\mathbf{A}$ and $\mathbf{B}$ be two Hermitian matrices. Then

$$\max_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|.$$

Lemma A.5. Let $f_1, f_2, \dots$ be analytic on D , a connected open set of $\mathbb{C}$, satisfying $|f_n(z)| \leq M$ for every n and z in D , and $f_n(z)$ converges, as $n \rightarrow \infty$ for each z in a subset of D having a limit point in D . Then there exists a function f , analytic in D for which $f_n(z) \rightarrow f(z)$ and $f'_n(z) \rightarrow f'(z)$ for all $z \in D$. Moreover, on any set bounded by a contour interior to D the convergence is uniform and $f_n(z)$ is uniformly bounded by $2M/\epsilon$, where ϵ is the distance between the contour and the boundary of D .

The proof of Lemma A.5 can be found in [Bai & Silverstein \(2004\)](#).

Lemma A.6. (Lemma 2.2 in [Bai & Silverstein \(2004\)](#)). For $\mathbf{z} = [z_1, \dots, z_n]^\top$ with i.i.d. standardized random entries, $\mathbf{C}$ is an $n \times n$ matrix, we have, for any $m \geq 2$,

$$\mathbb{E} |\mathbf{z}^\top \mathbf{C} \mathbf{z} - \text{tr} \mathbf{C}|^m \leq C[(\mathbb{E} |z_1|^4 \text{tr} \mathbf{C} \mathbf{C}^*)^{\frac{m}{2}} + \mathbb{E} |z_1|^{2m} \text{tr}(\mathbf{C} \mathbf{C}^*)^{\frac{m}{2}}].$$

Lemma A.7. (Burkholder's inequality). Let $\{x_k\}$ be a complex martingale difference sequence with respect to the increasing σ -field $\{\mathcal{F}_k\}$. Then for any $m \geq 2$,

$$\mathbb{E} \left| \sum x_k \right|^m \leq c_m \mathbb{E} \left(\sum \mathbb{E}(|x_k|^2 | \mathcal{F}_k) \right)^{\frac{m}{2}} + c_m \sum \mathbb{E} |x_k|^m.$$

Lemma A.8. (Sylvester's determinant identity). If $\mathbf{A}$ and $\mathbf{B}$ are matrices of sizes $m \times n$ and $n \times m$, then

$$\det(\mathbf{I}_m + \mathbf{A} \mathbf{B}) = \det(\mathbf{I}_n + \mathbf{B} \mathbf{A}).$$