

Eigen-Spike Emergence and Quadratic Equivalents for Conjugate Kernels on Nonlinearly Separable Data

Collin Cranston

CCRANSTO@UCSD.EDU

Department of Mathematics, University of California, San Diego

Zhichao Wang

ZHICHAO.WANG@BERKELEY.EDU

ICSI and Department of Statistics, University of California, Berkeley

Todd Kemp

TKEMP@UCSD.EDU

Department of Mathematics, University of California, San Diego

Michael W. Mahoney

MMAHONEY@STAT.BERKELEY.EDU

ICSI, LBNL and Department of Statistics, University of California, Berkeley

Editors: Steve Hanneke and Tor Lattimore

Abstract

Recent work in random matrix theory (RMT) has developed the notion of deterministic equivalents: typically linear surrogate models that approximate the spectral behavior of large nonlinear random matrices, such as nonlinear feature maps in neural networks (NNs). Such equivalents make theoretical predictions tractable by reducing a complex model to a simpler one with properties that fall under the umbrella of classical RMT tools. However, this leaves open the question of whether this idealized linear equivalence remains meaningful for classification of high-dimensional nonlinearly separable data. Motivated by this, we consider the conjugate kernel (CK), which is the nonlinear feature map of a one-layer feedforward NN, under a canonical nonlinearly separable dataset for the XOR problem; and we use the study of informative outlier eigenvalues in the CK and whether their corresponding eigenvectors asymptotically align with XOR labels as a proxy for nonlinear learnability. We develop a robust *quadratic equivalent* of the CK matrix that enables a precise analysis of emergent informative spikes, as one modifies various knobs common in ML practice: sample complexity, signal-to-noise ratio (SNR), nonlinear activation choice, and pretrained features. We identify regimes in which these knobs move the CK beyond the linear equivalent and produce BBP-type transitions to label-aligned outlier eigenspaces. Our analysis helps bring deterministic-equivalence tools from RMT to bear on problems of practical relevance in ML.

Keywords: Conjugate kernel; random features; random matrix theory; BBP phase transition; deterministic equivalents; spectral clustering; Gaussian Equivalence Theory; XOR problem.

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) (Martin and Mahoney, 2021, 2019, 2020; Martin et al., 2021; Yang et al., 2022; Wang et al., 2023; Jacot et al., 2018; Pennington and Worah, 2017; Adlam and Pennington, 2020; Liao and Mahoney, 2021; Sagun et al., 2016; Ghorbani et al., 2019a). For example, eigenvalues of NN weight matrices can capture the strength of learned directions, while the corresponding eigenvectors indicate which data features are amplified or extracted during training (Ba et al., 2022; Damian et al., 2022; Moniri et al., 2023; Dandi et al., 2024). A major direction in deep learning theory is to use spectral tools to understand when NNs can propagate, preserve, transform, or 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 signals 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) phase transitions characterize when an informative outlier separates from the bulk spectrum and when its associated eigenvector becomes correlated with the underlying signal (Baik et al., 2005; Nadler, 2008). Recent work has developed nonlinear counterparts for random-feature and kernel-type matrices, providing sharp predictions for outlier locations and eigenvector alignment (Feldman, 2023; Ba et al., 2022, 2023; Wang et al., 2024). 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) (Jacot et al., 2018; Pennington and Worah, 2017; Adlam and Pennington, 2020) for multi-layer NNs. Their high-dimensional spectra have been studied extensively (Liao and Mahoney, 2025; Wang and Zhu, 2024; Liao et al., 2020). 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 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 paper, the CK is defined by a one-hidden-layer random NN: $\mathbf{Y} = \frac{1}{\sqrt{N}} \sigma(\mathbf{W}\mathbf{X}) \in \mathbb{R}^{N \times n}$, with nonlinear activation σ , random weights $\mathbf{W} \in \mathbb{R}^{N \times d}$, and data $\mathbf{X} \in \mathbb{R}^{d \times n}$. The CK matrix is the Gram matrix of hidden activations, $\mathbf{K} = \mathbf{Y}^\top \mathbf{Y} \in \mathbb{R}^{n \times n}$ for n data points. The spectral distribution of $\mathbf{K}$ is particularly useful for generalization error of random feature regression, as demonstrated in Mei and Montanari (2022); Hu and Lu (2022) for isotropic Gaussian data $\mathbf{X}$. However, real datasets often contain low-rank or nonlinear structure that may be hard to analyze using PCA (Jolliffe, 2002; Schölkopf et al., 1998; Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Coifman and Lafon, 2006; Johnstone and Lu, 2009). To highlight genuinely nonlinear structure, we instead take $\mathbf{X}$ to be the high-dimensional XOR Gaussian mixture dataset (Section 3). This dataset is a balanced four-component Gaussian mixture whose two classes are unions of opposite components (Refinetti et al., 2021). 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 $\mathbf{K}$ 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 paper aims to answer the following spectral question for $\mathbf{K}$:

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 $\mathbf{K}$ 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 $\mathbf{X}$; the pretrained weight matrix $\mathbf{W}$; 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.

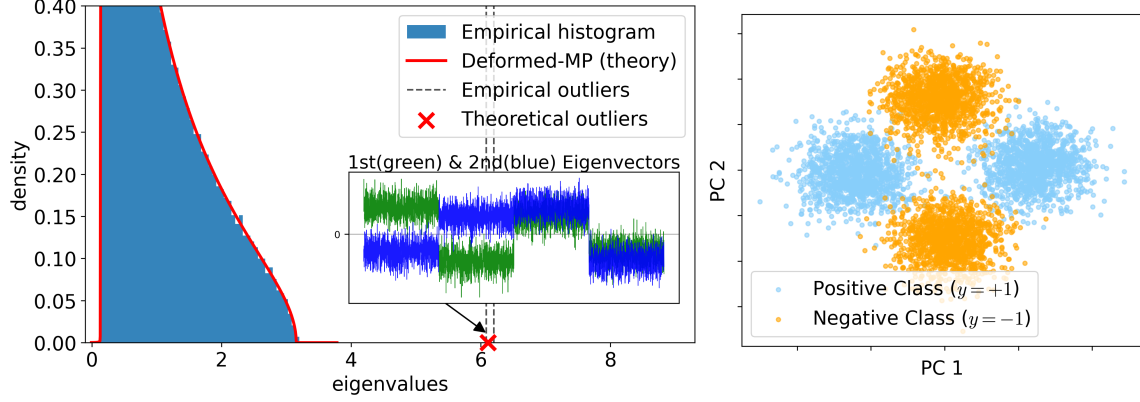


Figure 1: **Finite-SNR proportional regime: linear classification of XOR fails.** Spectrum of the random CK matrix $\mathbf{K}$ built from XOR data with $n = 5000$, $N = d = 15000$, $\text{SNR } r^2 = 36$, and $\sigma \propto \text{ReLU}$. **Left:** empirical eigenvalue histogram of $\mathbf{K}$ (blue) against the deformed MP law predicted by Theorem 3 (red curve); the red $\times$ marks the predicted location of the two (coincident) outlier eigenvalues and the dashed lines mark the empirical values. The subfigure plots the two outlier eigenvectors against sample index, showing four-block pattern of the Gaussian-mixture components but not the binary labels. **Right:** kernel-PCA embedding using the top two principal components (PCs) of $\mathbf{K}$ (one point per sample, colored by its true label); the four clouds form the canonical XOR geometry and are *not* linearly separable. The outlier eigenspace is asymptotically orthogonal to XOR label vector $\mathbf{y}$ in this baseline case.

1.1. Our 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 $\mathbf{Z} \in \mathbb{R}^{d \times n}$. In this setting, a standard tool is a *linear equivalent* (LE) for the nonlinear feature matrix $\sigma(\mathbf{W}\mathbf{Z})$. This approach replaces the nonlinear features with an affine Gaussian surrogate and an independent Gaussian noise term $\mathbf{N}$:

$$\sigma(\mathbf{W}\mathbf{Z}) \approx b_\sigma \mathbf{W}\mathbf{Z} + a_\sigma \mathbf{N}, \quad \mathbf{N}_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad (1.1)$$

for constants a_σ, b_σ determined by σ . This approximation is also called *Gaussian equivalence* in prior work (Mei and Montanari, 2022; Hu and Lu, 2022; Goldt et al., 2021; Bosch et al., 2023). When applying the LE, one can show that the spectra of nonlinear CK matrices are asymptotically equivalent to a linear 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. Figure 1 shows that CK matrix with LE resembles a linear

transform and its top PCs are orthogonal to the XOR labels. See Appendix B for more details of the simulation.

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 (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 $\mathbf{X} = \mathbf{Z} + \mathbf{M}$, where $\mathbf{Z}$ is Gaussian noise and $\mathbf{M}$ 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 (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

$$\sigma(\mathbf{W}\mathbf{X}) \approx \underbrace{\sigma(\mathbf{W}\mathbf{Z})}_{\text{bulk+uninformative spikes}} + \underbrace{\mathbf{T}_1}_{\text{linear spikes}} + \underbrace{\mathbf{T}_2}_{\text{quadratic spikes}} \quad (1.2)$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are finite-rank contributions induced by the signal $\mathbf{M}$ and σ (see (A.1) and Proposition 10). In this QE, there are three terms. The first term, $\sigma(\mathbf{W}\mathbf{Z})$, determines the bulk spectrum; and it may also possibly induce *uninformative outliers*, solely due to the architecture (Benigni and P      , 2022). The second term, $\mathbf{T}_1$, corresponds to a linear transformation of the signal $\mathbf{M}$; 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, $\mathbf{T}_2$, captures quadratic properties: depending on the correct tuning of the ML training “knobs,” this quadratic term $\mathbf{T}_2$ can dominate and create *quadratic informative spikes* (see (A.1)). This effectively provides a “quadratic feature channel” that makes the nonlinear XOR problem analyzable with RMT and linearly separable. In Appendix A, we present the formal statement of this QE and the proof strategy of our main results.

1.2. Our Contributions

We develop a RMT framework for the CK that is based on a QE and that captures when nonlinear learnability emerges for XOR, in regimes where the 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 (see Table 1).

- 1. Finite-SNR proportional limit ($n \asymp d$): negative results.** In the proportional regime with $\text{SNR} = \Theta(1)$, 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 3; Figure 1). We also prove failure for spectral clustering of Euclidean distance kernel.
- 2. Large-SNR proportional limit: emergence of quadratic informative spike.** We prove that with increasing SNR scaling and with proper activation choice, a quadratic informative outlier separates and aligns with the XOR labels, enabling linear spectral clustering/classification (Theorem 7).

- 3. Pretrained/spiked features: test-time BBP transition.** A low-rank perturbation of a random W can yield the emergence of an outlier whose eigenvector aligns with the induced task direction (Theorem 8) even for finite-SNR. Then, empirically, we show that linear spectral clustering occurs for CK matrix with weights from a pretrained NN on CIFAR-2 in Appendix B.
- 4. Quadratic sample-size regime ($n \asymp d^2$): a quadratic kernel.** We prove that in this case the linear-width projected CK model ($n \asymp N$) is determined by a quadratic polynomial kernel and has an informative spike aligned with XOR labels (Theorem 9).

Regime	SNR r^2	Sample	Weights	Linear Label Align?
Case 1 in Section 4.1 (Baseline)	$\Theta(1)$	$n \asymp d$	Random	No
Case 2 in Section 4.2 (Large SNR)	$\Theta(n^{1/2})$	$n \asymp d$	Random	Yes
Case 3 in Section 4.3 (Pretrained weight)	$\Theta(1)$	$n \asymp d$	Spiked	Yes
Case 4 in Section 4.4 (Quad. sample size)	$\Theta(1)$	$n \asymp d^2$	Random	Yes

Table 1: Summary of different knobs. From Cases 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 in all cases. Here ‘Yes’ represents non-trivial label alignment of CK spikes above BBP phase transition.

2. Additional Literature Review

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 (Belrose et al., 2024; Rende et al., 2024). Complementarily, Huh et al. (2023) 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, Wilson (2025) 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 (El Karoui, 2010; Cheng and Singer, 2013; Do and Vu, 2013). In the proportional regime, Cheng and Singer (2013) and Do and Vu (2013) characterize limiting spectra of inner-product kernels. The CK and NTK can be viewed as random kernels, whose spectra were further studied (Fan and Wang, 2020; Adlam and Pennington, 2020; Péché, 2019; Pennington and Worah, 2017; Dabo and Male, 2024; Guionnet and Piccolo, 2026) and connected to generalization in random feature regression (Mei and Montanari, 2022; Adlam and Pennington, 2020) and memorization of spurious feature (Bombari and Mondelli, 2024).

Nonlinear spiked covariance matrices. Spiked models exhibit phase transitions where informative eigenvalues separate from the bulk (Baik et al., 2005; Péché, 2006). Recent work extends this to *nonlinear* random matrices, i.e., matrices whose entries are nonlinear functions of high-dimensional random projections (such as $\sigma(WX)$), leading to nonlinear analogues of BBP transitions and

alignment formulas (Wang and Zhu, 2024; Feldman, 2023; Ba et al., 2023; Wang et al., 2024; Guionnet et al., 2023; Benigni and P  ch  , 2022). Outlier eigenstructure of CK and NTK has been studied Li et al. (2025), using the deterministic equivalence from Gu et al. (2022). Although they can deal with random deep networks and general Gaussian mixture datasets, their CKs and NTKs are expected kernels (namely first take width $N \rightarrow \infty$) 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) (Xiao et al., 2022; Lu and Yau, 2025; Cheng and Montanari, 2024; Misiakiewicz and Saeed, 2024). In the quadratic regime ($k = 2$), nonlinear kernels can reduce in operator norm to quadratic polynomial kernels (Ghorbani et al., 2019b; Pandit et al., 2024), enabling classification of quadratically separable data such as XOR (Refinetti et al., 2021). Wen et al. (2025) 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. [Refinetti et al. \(2021\)](#) 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 ([Liao and Couillet, 2019](#); [Couillet, 2019](#)). [Glasgow \(2023\)](#) showed how training two layer NNs with SGD can learn this XOR problem even with a sample complexity $\tilde{O}(d)$. Recently, [Taheri et al. \(2025\)](#) analyzed continual learning by studying one-hidden-layer quadratic NNs trained with gradient descent on an XOR dataset. [Demir and Dogan \(2025\)](#) analyzed two-layer networks after one gradient step on W with Gaussian-mixture data using higher-order polynomial equivalence. This motivates our spiked feature case in Section 4.3, while our result focuses on label-aligned outliers for linear classification on XOR, unlike trained feature regression tasks in [Demir and Dogan \(2025\)](#).

3. Notations and Preliminaries

Settings of one-layer NN models. We study the outputs of one-layer NNs when fed a dataset $\mathbf{X} \in \mathbb{R}^{d \times n}$. 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 $\mathbf{W} \in \mathbb{R}^{N \times d}$ with $[\mathbf{W}]_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ 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

$$\mathbf{Y} := \frac{1}{\sqrt{N}} \sigma(\mathbf{W}\mathbf{X}), \quad \mathbf{K} := \mathbf{Y}^\top \mathbf{Y}, \quad (3.1)$$

where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation. Our target is the eigenstructure of this CK matrix $\mathbf{K}$ when $\mathbf{X}$ has a nonlinear separable pattern. We now introduce the assumptions of the activation function σ .

Assumption 1 Assume σ is three-times differentiable with $\sup_{x \in \mathbb{R}} |\sigma'(x)|, |\sigma''(x)|, |\sigma'''(x)| \leq \lambda_\sigma$, for some $\lambda_\sigma \in (0, \infty)$, and is centered and normalized with respect to $\xi \sim \mathcal{N}(0, 1)$: $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma(\xi)^2] = 1$. Define the first and second order Hermite coefficients of σ as

$$b_\sigma := \mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi \sigma(\xi)] \in \mathbb{R}, \quad (3.2)$$

$$c_\sigma := \mathbb{E}[\sigma''(\xi)] = \mathbb{E}[(\xi^2 - 1)\sigma(\xi)] \in \mathbb{R}. \quad (3.3)$$

Many smooth activations satisfy Assumption 1 after centering and normalization (e.g., tanh, erf, sigmoid, Softplus, GELU). While ReLU is not C^3 , it can be approximated by smooth functions; we keep Assumption 1 to avoid technical distractions.

XOR dataset. Our dataset $\mathbf{X}$ is a binary class Gaussian Mixture dataset with four balanced clusters. We denote the collection of n data points by $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_{a_i}, \frac{1}{d}\mathbf{I}_d)$ for $i \in [n]$, where $\boldsymbol{\mu}_{a_i} \in \mathbb{R}^d$ is the mean of one of the cluster defined in (3.7) for $a_i \in [4]$. Without loss of generality, we reorder the data points so that we can write $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{M} + \mathbf{Z}, \quad (3.4)$$

where $\mathbf{Z} \in \mathbb{R}^{d \times n}$ has i.i.d. entries with distribution $\mathcal{N}(0, \frac{1}{d})$ and

$$\mathbf{M} = r\sqrt{\frac{n}{2d}}(\mathbf{u}_1 \cdot \mathbf{v}_1^\top + \mathbf{u}_2 \cdot \mathbf{v}_2^\top), \quad (3.5)$$

for some $r \geq 0$, with unit-norm vectors

$$\mathbf{u}_1 := \frac{1}{\sqrt{d}} \begin{bmatrix} \mathbf{1}_{d/2} \\ -\mathbf{1}_{d/2} \end{bmatrix}, \quad \mathbf{u}_2 := \frac{1}{\sqrt{d}} \begin{bmatrix} \mathbf{1}_{d/2} \\ \mathbf{1}_{d/2} \end{bmatrix}, \quad \mathbf{v}_1 = \sqrt{\frac{2}{n}} \begin{bmatrix} \mathbf{1}_{n/4} \\ -\mathbf{1}_{n/4} \\ \mathbf{0}_{n/2} \end{bmatrix}, \quad \mathbf{v}_2 = \sqrt{\frac{2}{n}} \begin{bmatrix} \mathbf{0}_{n/2} \\ \mathbf{1}_{n/4} \\ -\mathbf{1}_{n/4} \end{bmatrix}. \quad (3.6)$$

Here, the signal-to-noise ratio (SNR) is defined as $\text{SNR} := r^2$. In this case, the mean of each cluster is

$$\boldsymbol{\mu}_1 = r \cdot \mathbf{u}_1, \quad \boldsymbol{\mu}_2 = r \cdot \mathbf{u}_2, \quad \boldsymbol{\mu}_3 = -\boldsymbol{\mu}_1, \quad \boldsymbol{\mu}_4 = -\boldsymbol{\mu}_2. \quad (3.7)$$

Following the above definition of $\mathbf{X}$, we consider a binary classification dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i \in \{\pm 1\}$. We assume the dataset is *not* linearly separable in the input space, i.e., there do not exist $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0$, for all $i \in [n]$. For simplicity, in this case, the labels of data $\mathbf{X}$ are defined by

$$\mathbf{y}^\top = \{y_i\}_{i=1}^n = [\mathbf{1}_{n/2}^\top, -\mathbf{1}_{n/2}^\top]. \quad (3.8)$$

Our goal is to study the emergence of outlier eigenvalues in the spectrum of $\mathbf{K} = \mathbf{Y}^\top \mathbf{Y}$ defined by (3.1), and verify whether the corresponding eigenvectors (spikes) are aligned to this label vector $\mathbf{y}$, in order to (partially) recover the classes of $\mathbf{X}$ defined by (3.4).

Marchenko–Pastur law. Given a noise sample covariance $\mathbf{S} = \mathbf{Z}^\top \mathbf{Z} \in \mathbb{R}^{n \times n}$ from (3.4), assume $d, n \rightarrow \infty$ with $\frac{n}{d} \rightarrow \psi \in (0, \infty)$. Then the ESD of $\mathbf{S}$ converges weakly almost surely to the Marchenko–Pastur (MP) law (Marčenko and Pastur, 1967), denoted by 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

$$\rho_\psi^{\text{MP}}(x) := \frac{1}{2\pi\psi x} \sqrt{(b-x)(x-a)} \mathbf{1}_{[a,b]}(x),$$

and, when $\psi > 1$, the density function is $\rho_\psi^{\text{MP}} = \left(1 - \frac{1}{\psi}\right)\delta_0 + \rho_\psi^{\text{MP}}(x) dx$.

Stieltjes transform. The Stieltjes transform provides an analytical tool for studying the limiting ESD in RMT. For $z \in \mathbb{C} \setminus [0, \infty)$, define the Stieltjes transform of ρ_ψ^{MP} by $m_\psi(z) := \int_{\mathbb{R}} \frac{1}{x-z} \rho_\psi^{\text{MP}}(dx)$. This Stieltjes transform fully captures the spectral distribution ρ_ψ^{MP} , as we can recover the density from an inversion formula (Bai and Silverstein, 2010): for $x > 0$

$$\rho_{\phi}^{\text{MP}}(x) = \frac{1}{\pi} \lim_{\eta \downarrow 0} \text{Im } m_{\phi}(x + i\eta).$$

Limiting spectral distribution of CK. From Pennington and Worah (2017); Louart et al. (2018); Benigni and P  ch   (2021); Fan and Wang (2020), we know the limiting ESD of $\mathbf{K}$ is a deformed MP law, under certain conditions of $\mathbf{X}$ and the proportional regime, denoted by

$$\mu := \rho_\phi^{\text{MP}} \boxtimes \nu, \quad \nu := (1 - b_\sigma^2) + b_\sigma^2 \rho_\psi^{\text{MP}}, \quad (3.9)$$

where ρ_ϕ^{MP} is a standard MP law with aspect ratio $\phi \in (0, \infty)$ 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

$$m_\mu(z) = \int \frac{1}{x-z} d\mu(x). \quad (3.10)$$

We refer to (Bai and Silverstein, 2010; Fan and Wang, 2020; Wang et al., 2024) for more details of free multiplicative convolution and these Stieltjes transforms. We denote the following transforms:

$$z(s) = -\frac{1}{s} + \phi \int \frac{x}{1+xs} \nu(dx), \quad \varphi(s) = -\frac{sz'(s)}{z(s)}, \quad (3.11)$$

$$T(s) = \frac{z(s)s^2 - (\phi - 1)s}{\phi}, \quad T^{-1}(t) = -\frac{t(1 - \psi b_\sigma^2 t)}{1 + (1 - \psi b_\sigma^2)t - \psi b_\sigma^2(1 - b_\sigma^2)t^2}, \quad (3.12)$$

which is used to characterize outliers of $\mathbf{K}$. More properties of these transforms are given by Appendix C.3.

4. Main Results

We now present our main results on the outlier eigenvalues and eigenvectors of $\mathbf{K}$ across the regimes summarized in Table 1. In the first three regimes, we consider proportional limit, where $d, n, N \rightarrow \infty$ proportionally: Section 4.1 studies finite SNR = $\Theta(1)$ regime; Section 4.2 analyzes large SNR = $\Theta(n^{1/4})$ case; and Section 4.3 replaces the random weight $\mathbf{W}$ with a pretrained weight. Then, Section 4.4 considers the quadratic sample-size regime $n \asymp d^2$, with finite SNR and random weights. In each case, we predict the emergence of spikes of $\mathbf{K}$ and whether their eigenvectors enable *linear* classification of $\mathbf{X}$.

4.1. Proportional Limit with Finite SNR Case

We first consider a baseline regime in which all dimensions grow proportionally and the SNR of $\mathbf{X}$ is finite. We show that the outlier eigenvectors cannot be used to linearly classify the dataset, even partially.

Assumption 2 (Proportional limit) Let $n, d, N \rightarrow \infty$ such that $n/N \rightarrow \phi \in (0, +\infty)$ and $n/d \rightarrow \psi \in (0, +\infty)$ for some fixed constants ϕ and ψ .

We denote some constants used to define the BBP thresholds in the following theorem by

$$\tau := \frac{c_\sigma^2}{2}\psi, \quad \beta_{\text{lin}} := \frac{r^2\psi}{2}b_\sigma^2, \quad \tau_{\text{crit}} := b_\sigma^2\sqrt{\psi}(1 + \sqrt{\psi}). \quad (4.1)$$

Recall the definition of $T(\cdot)$ and $T^{-1}(\cdot)$ in (3.12). When $\tau > 0$ and $\beta_{\text{lin}} > 0$, respectively, we define

$$s_{\text{un}} := T^{-1}(1/\tau), \quad s_{\text{lin}} := T^{-1}(1/\beta_{\text{lin}}).$$

Theorem 3 Under Assumptions 1 and 2, we further assume that $r \geq 0$ is a constant. Consider $\mathbf{K}$ with XOR data $\mathbf{X}$ defined in Section 3. Recall the transforms $z(\cdot)$, $\varphi(\cdot)$ and $T(\cdot)$ defined by (3.11) and (3.12). Denote $\mathbf{u} := \frac{1}{\sqrt{n}}\mathbf{1}_n$, and $\mathbf{m} := \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{Z}) \mid \mathbf{Z}] \in \mathbb{R}^n$ conditioned on $\mathbf{Z}$, where $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$. Assume in addition that $s_{\text{un}} \neq s_{\text{lin}}$ if both are nonzero. Then the CK matrix $\mathbf{K}$ satisfies the following.

- (i) (**Bulk.**) The ESD of $\mathbf{K}$ converges weakly in probability to μ defined in (3.9).
- (ii) (**Uninformative outliers.**) If $c_\sigma \neq 0$ and $z'(s_{\text{un}}) > 0$, then $\mathbf{K}$ has two eigenvalues outside the support of μ satisfying $\hat{\lambda}_{\text{un},1}, \hat{\lambda}_{\text{un},2} = z(s_{\text{un}}) + o_{\mathbb{P}}(1)$. Let $\hat{\mathbf{P}}_{\text{un}}$ be the spectral projector onto this two-dimensional outlier eigenvalues. Then

$$\|\hat{\mathbf{P}}_{\text{un}}\mathbf{u}\|^2 = \varphi(s_{\text{un}}) + o_{\mathbb{P}}(1), \quad (4.2)$$

$$\left\| \hat{\mathbf{P}}_{\text{un}} \frac{\mathbf{m}}{\|\mathbf{m}\|} \right\|^2 = -\frac{\varphi(s_{\text{un}})}{\tau^2 s_{\text{un}}^2 T'(s_{\text{un}})} + o_{\mathbb{P}}(1), \quad (4.3)$$

$$\mathbf{u}^\top \hat{\mathbf{P}}_{\text{un}} \frac{\mathbf{m}}{\|\mathbf{m}\|} = o_{\mathbb{P}}(1). \quad (4.4)$$

Besides, when $b_\sigma = 0$, (4.2) and (4.3) are asymptotically the same. If $c_\sigma = 0$ or $z'(s_{\text{un}}) \leq 0$, then $\mathbf{K}$ has no such uninformative outliers.

- (iii) (**Linear spikes from XOR.**) If $z'(s_{\text{lin}}) > 0$ and $\beta_{\text{lin}} > 0$, then $\mathbf{K}$ has additional two outlier eigenvalues outside the support of μ satisfying $\hat{\lambda}_{\text{lin},1}, \hat{\lambda}_{\text{lin},2} = z(s_{\text{lin}}) + o_{\mathbb{P}}(1)$, and the spectral projector $\hat{\mathbf{V}}_{\text{lin}}$ of the corresponding two-dimensional eigenspace satisfies

$$\|(\mathbf{v}_1 \mathbf{v}_1^\top + \mathbf{v}_2 \mathbf{v}_2^\top) \hat{\mathbf{V}}_{\text{lin}}\|_F^2 = 2 \cdot \left(-\frac{\varphi(s_{\text{lin}})}{\beta_{\text{lin}}^2 s_{\text{lin}}^2 T'(s_{\text{lin}})} \right) + o_{\mathbb{P}}(1). \quad (4.5)$$

If $\beta_{\text{lin}} = 0$ or $z'(s_{\text{lin}}) \leq 0$, then $\mathbf{K}$ has no such linear outliers.

- (iv) (**No label alignment.**) Let $\hat{\mathbf{V}}_{\text{out}}$ denote the orthogonal projector onto the direct sum of all separated outlier clusters described in (ii)–(iii) if they exist. Then, for the XOR labels $\mathbf{y}$,

$$\frac{1}{n} \|\hat{\mathbf{V}}_{\text{out}} \mathbf{y}\|^2 = o_{\mathbb{P}}(1). \quad (4.6)$$

Theorem 3 exhibits a BBP-type phase transition for the CK matrix with XOR data: e.g., $z'(s_{\text{lin}}) > 0$ is the threshold showing when spikes separate from the deformed MP bulk μ ; in the separated case, the associated eigenvectors have nontrivial asymptotic overlap with certain directions. There are up to four candidate spikes outside $\text{supp}(\mu)$: two outliers (Theorem 3(iii)) induced by the linear Hermite component of σ and spike strength s_{lin} depending on r , and two potential “architectural” outliers (Theorem 3(ii)) corresponding to $c_\sigma \neq 0$ (Benigni and P      , 2022). The latter is “uninformative” in the sense that it may persist even in the null model $r = 0$ (pure noise data). See Section B for further discussion. To focus on informative spikes, previous works always impose $c_\sigma = 0$ (Wang et al., 2024). The proof of Theorem 3 is given in Appendix E.

Remark 4 In Theorem 3(ii), there is a necessary condition of the BBP threshold $z'(s_{\text{un}}) > 0$ for the emergence of uninformative spikes: $z'(s_{\text{un}}) > 0 \implies \tau > \tau_{\text{crit}}$. Hence, from (4.1), uninformative spikes can be removed if $\tau \leq \tau_{\text{crit}}$ holds, namely $\sqrt{\psi}(\frac{c_\sigma^2}{2} - b_\sigma^2) \leq b_\sigma^2$. When 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 Appendix C.4), this condition is satisfied for all $\psi > 0$, and thus no uninformative spike emerges.

Remark 5 From (4.6) in Theorem 3(iv), all separated outlier eigenvectors are asymptotically orthogonal to the label vector $\mathbf{y}$. Thus, these outliers cannot be used for linear classification. Although the spikes in Theorem 3(iii) carry information about the within-class structure (alignment with $\mathbf{v}_1, \mathbf{v}_2$ in (4.5)), they cannot recover the XOR labels $\mathbf{y}$ without an additional nonlinear transform on spikes.

Kernel spectral clustering also fails. Previously Couillet and Benaych-Georges (2016) studied this Euclidean distance kernel clustering for general Gaussian mixture data. We can apply this result to the kernel spectral clustering algorithm Ng et al. (2001) and get similar negative results as Theorem 3. Full details are provided in Appendix F.

Theorem 6 (Kernel spectral clustering for XOR) Let Assumption 2 hold, and let $r = \Theta(1)$. Consider the kernel matrix $(\mathbf{K}^{(f)})_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, for $i, j \in [n]$, and some $f(x) \in C^3$ near $x = 2$, with $f(2) > 0$ and $f'(2) \neq 0$, and the normalized Laplacian $\mathbf{L} = n \mathbf{D}^{-1/2} \mathbf{K}^{(f)} \mathbf{D}^{-1/2}$ where $\mathbf{D} = \text{diag}(\mathbf{K}^{(f)} \mathbf{1}_n)$. Then,

- **(Isolated eigenvalues.)** If $r^2 > 2\sqrt{\psi^{-1}}$, then $\mathbf{L}$ has two nontrivial isolated eigenvalues whose limits are given by Theorem 51 in Appendix F. If $r^2 \leq 2\sqrt{\psi^{-1}}$, then $\mathbf{L}$ has no such outliers.
- **(No label alignment.)** Let $\mathbf{\Pi}$ be the orthogonal projector onto the span of the eigenspace associated with the above outliers. Then, $\frac{1}{n} \mathbf{y}^\top \mathbf{\Pi} \mathbf{y} \rightarrow 0$ in probability.

4.2. Large-SNR Proportional Limit Case

Theorem 3 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 needed for XOR. To make XOR linearly learnable, we now increase the SNR so that the quadratic Hermite component of σ becomes non-vanishing. Then two outliers created by the quadratic component become label-informative, enabling linear classification. Numerical simulations of all these cases are in Appendix B.

In this section, we consider large SNR regime by letting r in (3.5) grow with n as follows:

$$r_0 := \lim_{n \rightarrow \infty} \frac{r}{n^{1/4}} \in (0, \infty), \quad \kappa_0 := \frac{r_0^4 \psi^2}{4}, \quad \beta_{\text{quad}} := c_\sigma^2 \kappa_0. \quad (4.7)$$

For $\beta_{\text{quad}} > 0$, define the label-informative factor

$$s_y := T^{-1}(1/\beta_{\text{quad}}). \quad (4.8)$$

Recall τ from (4.1). When $c_\sigma \neq 0$, equivalently $\tau > 0$ and $\beta_{\text{quad}} > 0$, we define non-label factors

$$t_\pm := \frac{1 \pm \sqrt{\beta_{\text{quad}}/(\tau + \beta_{\text{quad}})}}{\tau}, \quad s_\pm := T^{-1}(t_\pm). \quad (4.9)$$

Theorem 7 *Under Assumptions 1 and 2, let $\mathbf{P}_{\text{lin}} = \mathbf{v}_1 \mathbf{v}_1^\top + \mathbf{v}_2 \mathbf{v}_2^\top$, $\mathbf{P}_\perp = \mathbf{I} - \mathbf{P}_{\text{lin}}$, and $\mathbf{K}_\perp := \mathbf{P}_\perp \mathbf{K} \mathbf{P}_\perp$. Then, under the scaling in (4.7), the following statements hold.*

(i) **(Bulk.)** *The ESD of $\mathbf{K}$ converges weakly in probability to μ defined in (3.9).*

(ii) **(Diverging linear outliers.)** *If $b_\sigma \neq 0$, then the top two eigenvalues of $\mathbf{K}$ satisfy*

$$\hat{\lambda}_1(\mathbf{K}), \hat{\lambda}_2(\mathbf{K}) = \frac{b_\sigma^2 r_0^2 \psi}{2} n^{1/2} (1 + o_{\mathbb{P}}(1)), \quad (4.10)$$

and the associated eigenspace projector $\hat{\mathbf{P}}_{\text{lin}}$ satisfies $\|\hat{\mathbf{P}}_{\text{lin}} - \mathbf{P}_{\text{lin}}\|_F \xrightarrow{\mathbb{P}} 0$, and $\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_{\text{lin}} \mathbf{y} \xrightarrow{\mathbb{P}} 0$.

(iii) **(Order-one outliers of $\mathbf{K}_\perp$.)** *Assume $c_\sigma \neq 0$, so $\tau, \beta_{\text{quad}} > 0$. Let $\mathcal{S}_{\text{cand}} := \{s_+, s_-, s_y\}$ with $s_\pm, s_y$ defined by (4.8) and (4.9). An element $s \in \mathcal{S}_{\text{cand}}$ produces a separated order-one outlier of $\mathbf{K}_\perp$ precisely when*

$$z'(s) > 0, \quad z(s) \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\}). \quad (4.11)$$

More precisely, fix a limiting location $\lambda_\star \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$, and define the algebraic multiplicity

$$m_\star := \#\{s \in \mathcal{S}_{\text{cand}} : z(s) = \lambda_\star, z'(s) > 0\}. \quad (4.12)$$

For any sufficiently small $\delta > 0$ and interval $I_\star = (\lambda_\star - \delta, \lambda_\star + \delta)$ containing no other candidate location and disjoint from $\text{supp}(\mu) \cup \{0\}$, $\mathbf{K}_\perp$ has exactly $m_\star$ eigenvalues in $I_\star$, and all of them converge to $\lambda_\star$ in probability. $\mathbf{K}_\perp$ has at most three separated order-one outliers, counting multiplicity.

(iv) **(Label alignment.)** *Suppose that s_y is admissible in the sense of (4.11). Let $\lambda_y := z(s_y)$, and let $\hat{\mathbf{P}}_y^\perp$ be the spectral projector of $\mathbf{K}_\perp$ onto a small interval around λ_y . Then*

$$\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_y^\perp \mathbf{y} \xrightarrow{\mathbb{P}} \gamma_y(s_y) := -\frac{\varphi(s_y)}{\beta_{\text{quad}}^2 s_y^2 T'(s_y)} > 0. \quad (4.13)$$

If $\lambda_\star = z(s_\pm)$ is admissible from (4.11) and distinct from the label candidate λ_y , then the corresponding spectral projector $\hat{\mathbf{P}}_{\text{nl},\star}^\perp$ satisfies $\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_{\text{nl},\star}^\perp \mathbf{y} \xrightarrow{\mathbb{P}} 0$.

In contrast to Theorem 3, when $r = \Theta(n^{1/4})$ the linear outliers characterized by Theorem 3(iii) become diverging and their eigenvectors remain asymptotically orthogonal to the label vector $\mathbf{y}$. Meanwhile, the quadratic component of σ may produce three $O(1)$ outliers (Theorem 7(iii)). Here, (4.13) gains a positive alignment with labels $\mathbf{y}$ because the quadratic features $\mathbf{v}_1^{\odot 2}$ and $\mathbf{v}_2^{\odot 2}$ are preserved. Hence, a linear readout on the λ_y -spike eigenspace in (4.13) can classify XOR. The proof of Theorem 7 is in Appendix G.

When $b_\sigma = 0$, the diverging outliers in (4.10) vanish. If $c_\sigma = 0$, then $\tau = \beta_{\text{quad}} = 0$, only the bulk and the possible two diverging linear outliers in (4.10) remain. The absent quadratic component leads linear classification via these spikes impossible, thus highlighting the role of the activation knob in preserving nonlinear structure. For further discussion about the role of non-linearity choice, see numerical simulations in Appendix B.

4.3. Pretrained Weight Matrix Case

Empirically, weight matrices in well-trained NNs often exhibit spectral spikes (even heavy tails) (Martin and Mahoney, 2021, 2019, 2020; Martin et al., 2021; Yang et al., 2022). Such spectral behavior suggests that training induces nontrivial feature structure in the weights from dataset (Ba et al., 2022, 2023; Demir and Dogan, 2025). However, the above theorems consider only random features, and therefore they do not capture task-dependent structure learned in the pretrained $\mathbf{W}$. As a toy model for feature learning, we consider a simple but nontrivial pretrained model in which the weight is a rank-one perturbation of the random initialization:

$$\mathbf{W}_1 = \mathbf{W} + \theta \mathbf{a} \mathbf{b}^\top, \quad (4.14)$$

where $\mathbf{a}$ and $\mathbf{b}$ are independent of the *test* dataset $\mathbf{X}$. A related rank-one structure arises, for example, after one-step gradient descent in certain student–teacher settings (Ba et al., 2022; Demir and Dogan, 2025), and it is also connected to LoRA (Hu et al., 2021). Below, we analyze a *test-time* BBP transition, i.e., we evaluate the CK built from $\mathbf{W}_1$ on independent XOR test data $\mathbf{X}$, and we characterize the emergence and alignment of the induced spike.

Let μ, z, T, m_μ be the limiting law and transforms defined in Section 3. For any constants $r \geq 0$ and $\theta_0 > 0$, define the following constants:

$$\beta_q := \frac{3c_\sigma^2}{4} \theta_0^4 \phi(3 + 3r^2 + \frac{r^4}{2}). \quad \omega_q := \frac{8 + 8r^2 + r^4}{12 + 12r^2 + 2r^4}, \quad \chi_y := \frac{r^4}{12 + 12r^2 + 2r^4}. \quad (4.15)$$

Recall τ defined in (4.1). We define, for $t \in \mathbb{C}$,

$$F_q(t) := (1 - \tau t)^2 - \beta_q t (1 - \tau \omega_q t) (1 - \frac{2}{3} \tau t). \quad (4.16)$$

Theorem 8 Suppose $r \geq 0$ is a constant and let $\kappa := 1 + \frac{r^2}{2}$. Let pretrained $\mathbf{W}_1$ be defined in (4.14) with $\mathbf{b} = \mathbf{1}_d$, $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N/N)$, and $\theta = \theta_0 N^{1/4}$, where $\mathbf{a}$ is independent of $(\mathbf{W}, \mathbf{X})$, and $\theta_0 > 0$ is a fixed constant. Under Assumptions 1 and 2, consider $\mathbf{Y} := \frac{1}{\sqrt{N}} \sigma(\mathbf{W}_1 \mathbf{X})$, pretrained CK $\mathbf{K} := \mathbf{Y}^\top \mathbf{Y}$, and projected CK $\mathbf{K}_s := \Pi_s \mathbf{K} \Pi_s$ where $\Pi_s := \mathbf{I}_n - \widehat{\mathbf{s}} \widehat{\mathbf{s}}^\top$, $\widehat{\mathbf{s}} := \frac{\mathbf{s}}{\|\mathbf{s}\|}$, $\mathbf{q} := \mathbf{s}^{\odot 2}$, and $\mathbf{s} := \mathbf{X}^\top \mathbf{b}$. Then,

- (i) (**Bulk.**) The ESD of $\mathbf{K}$ converges weakly in probability to μ defined in (3.9).

(ii) (**Diverging spike.**) If $b_\sigma \neq 0$, then the largest eigenvalue of $\mathbf{K}$ satisfies $\hat{\lambda}_1(\mathbf{K}) = b_\sigma^2 \theta_0^2 \phi \kappa \sqrt{N} (1 + o_{\mathbb{P}}(1))$. If $\hat{\mathbf{v}}_1$ is the associated unit eigenvector, then

$$|\langle \hat{\mathbf{v}}_1, \hat{\mathbf{s}} \rangle|^2 \xrightarrow{\mathbb{P}} 1, \quad \frac{1}{n} |\hat{\mathbf{v}}_1^\top \mathbf{y}|^2 \xrightarrow{\mathbb{P}} 0. \quad (4.17)$$

If $b_\sigma = 0$, this diverging spike is absent.

(iii) (**Order-one outliers of $\mathbf{K}_s$.**) All asymptotically separated order-one outliers of $\mathbf{K}_s$ are determined by the roots s of the three equations:

$$1 - \beta_{\text{lin}} T(s) = 0, \quad 1 - \frac{\beta_{\text{lin}}}{\kappa} T(s) = 0, \quad F_q(T(s)) = 0, \quad (4.18)$$

where β_{lin} is defined in (4.1). Fix $\lambda_\star \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$, and let $m_\star$ be the total algebraic multiplicity of all real solutions s to (4.18) such that $z(s) = \lambda_\star$ and $z'(s) > 0$. For any sufficiently small $\delta > 0$ and interval $I_\star = (\lambda_\star - \delta, \lambda_\star + \delta)$, disjoint from $\text{supp}(\mu) \cup \{0\}$ and containing no other candidate location, $\mathbf{K}_s$ has exactly $m_\star$ eigenvalues in $I_\star$, and all of them converge to $\lambda_\star$ in probability. If a real root s satisfies $z'(s) \leq 0$, it does not produce an outlier.

(iv) (**Label alignment.**) Outliers coming only from the first two equations in (4.18) have zero normalized label overlap. Suppose that $\hat{\mathbf{P}}_\star^{(s)}$ is the spectral projector of $\mathbf{K}_s$ onto $I_\star$ where the roots of $F_q(T(s)) = 0$ inside $I_\star$ are simple. Recall m_μ defined in (3.10). Then

$$\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_\star^{(s)} \mathbf{y} \xrightarrow{\mathbb{P}} \sum_{s \in \mathcal{R}_q(\lambda_\star)} -\frac{m_\mu(z(s)) z'(s)}{T'(s)} \frac{\beta_q \chi_y T(s) (1 - \tau T(s)) (1 - \frac{2}{3} \tau T(s))}{F_q'(T(s))} \geq 0 \quad (4.19)$$

where $\mathcal{R}_q(\lambda_\star)$ is the set of admissible roots of $F_q(T(s)) = 0$ with $z(s) = \lambda_\star$, $z'(s) > 0$.

Here $\mathbf{K}_s$ has at most five separated order-one outliers from (4.18), counting multiplicity: two linear informative spikes related to β_{lin} when $b_\sigma \neq 0$ and at most three roots from $F_q(T(s)) = 0$ related to both uninformative and label-aligned spikes when $c_\sigma \neq 0$. This result shows that the large data-dependent spike in the pretrained weight $\mathbf{W}_1$ amplifies the signals in XOR and alters the approximation: the usual LE is replaced by QE to capture the induced quadratic feature. Theorem 8(iv) shows this quadratic informative eigenvector aligns with labels. If $c_\sigma = 0$, so $\beta_q = 0$, then we cannot obtain label alignment. Empirical simulations of test-time BBP are presented in Appendix B. The proof of Theorem 8 is in Appendix H.

4.4. Quadratic Sample-size Case

Sample complexity can determine whether nonlinear structure is visible in high dimensions. In the proportional regime $n = \Theta(d)$, the LE suppresses higher-order structure, making the CK behave like a “linear” model. We now turn to the quadratic sample-size regime $n = \Theta(d^2)$, where the CK behaves like a quadratic polynomial kernel (Xiao et al., 2022; Wen et al., 2025) and becomes label-informative for XOR even at finite SNR.

Theorem 9 Under Assumption 1 with $c_\sigma \neq 0$, consider the XOR model (3.4)–(3.8) with finite $r \in (0, \infty)$. Assume that $\frac{n}{p} \rightarrow \gamma \in (0, \infty)$ and $\frac{n}{N} \rightarrow \phi \in (0, \infty)$, where $p := \frac{d(d+1)}{2}$. Denote $\ell := \frac{r^4}{4}$ and $\lambda_{\text{out}}(\gamma, \ell) := 1 + \gamma + \gamma\ell + \frac{1}{\ell}$. Then there exists a projection $\mathbf{P}_\sigma$ defined in (I.11) such that the following hold for $\mathbf{K}_\sigma := \mathbf{P}_\sigma \mathbf{K} \mathbf{P}_\sigma$.

- (i) (**Bulk.**) The ESD of $\mathbf{K}_\sigma$ converges weakly to $\mu_q := \rho_\phi^{\text{MP}} \boxtimes \nu_q$ where $\nu_q := (1 - b_\sigma^2 - \frac{c_\sigma^2}{2}) + \frac{c_\sigma^2}{2} \rho_\gamma^{\text{MP}}$.
- (ii) (**Spike of the population CK.**) If $\ell > \gamma^{-1/2}$, then the population CK matrix $\mathbb{E}[\mathbf{K}_\sigma | \mathbf{X}]$ has a label-aligned outlier $\widehat{\Lambda}_y = \Lambda_y + o_{\mathbb{P}}(1)$ where $\Lambda_y := (1 - b_\sigma^2 - \frac{c_\sigma^2}{2}) + \frac{c_\sigma^2}{2} \lambda_{\text{out}}(\gamma, \ell) \notin \text{supp}(\nu_q)$. In this case, its associated eigenvector is aligned with label $\mathbf{y}$. Otherwise, no such outlier exists.
- (iii) (**Spike of the linear-width CK.**) If $\ell > \gamma^{-1/2}$ and $z'(-1/\Lambda_y) > 0$, then for every sufficiently small deterministic interval I_y around

$$\lambda_y = z\left(-\frac{1}{\Lambda_y}\right) := \Lambda_y \left(1 + \phi \int \frac{t}{\Lambda_y - t} \nu_q(dt)\right), \quad (4.20)$$

which is disjoint from $\text{supp}(\mu_q)$, $\mathbf{K}_\sigma$ has exactly one eigenvalue in I_y which converges in probability to λ_y . If $\widehat{\mathbf{v}}_y$ is the associated unit-norm eigenvector, then

$$\left| \left\langle \widehat{\mathbf{v}}_y, \frac{\mathbf{y}}{\sqrt{n}} \right\rangle \right|^2 \xrightarrow{\mathbb{P}} \varphi\left(-\frac{1}{\Lambda_y}\right) \frac{\gamma \ell^2 - 1}{\gamma \ell(\ell + 1)}, \quad \varphi\left(-\frac{1}{\Lambda_y}\right) := \frac{1 - \phi \int \frac{t^2}{(\Lambda_y - t)^2} \nu_q(dt)}{1 + \phi \int \frac{t}{\Lambda_y - t} \nu_q(dt)}. \quad (4.21)$$

If either $\ell \leq \gamma^{-1/2}$ or $z'(-1/\Lambda_y) \leq 0$, then no separated label-aligned eigenvalue of $\mathbf{K}_\sigma$ is produced.

Theorem 9 only characterizes the label-aligned outlier generated by the XOR. The full CK may contain additional outliers which are removed by the projection $\mathbf{P}_\sigma$: a mean outlier, d linear-term outliers when $b_\sigma \neq 0$, and possibly $o(n)$ quadratic nuisance outliers caused by the rank- $O(d)$ non-isotropic covariance component in $\mathbb{E}[\mathbf{K} | \mathbf{X}]$. These outliers do not affect the limiting ESD and have asymptotically zero overlap with XOR label. The $\mathbf{P}_\sigma$ defined in (I.11) only depends on σ and $\mathbf{X}$. The proof of Theorem 9 is in Appendix I.

5. Conclusion

We introduce a quadratic equivalent framework and illustrate 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, increasing SNR, moving to the quadratic sample-size regime, or adding pretrained structure to the weights can make the quadratic feature non-negligible and produce 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 QE to more general scalings, deeper networks, and other forms of structured data is an interesting direction.

Acknowledgments

CC and TK were supported in part by NSF grant DMS-2400246. ZW and MWM would like to acknowledge the NSF and the DARPA DIAL and DARPA AIQ programs for partial support of this work.

References

- Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.
- 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.
- 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.
- 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.
- 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.
- Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.
- 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.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.
- Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. Neural networks learn statistics of increasing complexity. *arXiv preprint arXiv:2402.04362*, 2024.
- 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.
- Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.
- Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*, 2022.
- 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. doi: 10.1214/EJP.v19-3054.
- 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. URL <https://proceedings.mlr.press/v235/bombari24a.html>.
- 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.
- Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879–2912, 2024.
- Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.
- Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.
- 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.
- Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393 – 1454, 2016. doi: 10.1214/16-EJS1144. URL <https://doi.org/10.1214/16-EJS1144>.
- Issa Dabo and Camille Male. A traffic approach for profiled pennington-worah matrices. *arXiv preprint arXiv:2409.13433*, 2024.
- Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.
- Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue M Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. *arXiv preprint arXiv:2410.18938*, 2024.
- Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
- 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. URL <https://openreview.net/forum?id=tNn6Hskmti>.
- DLMF. *NIST Digital Library of Mathematical Functions*. <https://dlmf.nist.gov/>, Release 1.2.4 of 2025-03-15. URL <https://dlmf.nist.gov/>. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.
- 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.

- Nouredine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1): 1–50, 2010.
- 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.
- Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.
- Zhou Fan, Renyuan Ma, Elliot Paquette, and Zhichao Wang. Anisotropic local law for non-separable sample covariance matrices. *arXiv preprint arXiv:2602.17960*, 2026.
- Michael J Feldman. Spectral properties of elementwise-transformed spiked matrices. *arXiv preprint arXiv:2311.02040*, 2023.
- 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.
- 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, 2019a. URL <https://proceedings.mlr.press/v97/ghorbani19b.html>.
- Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019b.
- 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.
- 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.
- Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α -sub-exponential random variables. *Electronic Journal of Probability*, 26(none):1 – 22, 2021. doi: 10.1214/21-EJP606. URL <https://doi.org/10.1214/21-EJP606>.
- 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. URL <https://openreview.net/forum?id=NaW6T93F34m>.
- Alice Guionnet and Vanessa Piccolo. Global law of conjugate kernel random matrices with heavy-tailed weights. *Electronic Journal of Probability*, 31:1–58, 2026.
- 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.

- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- 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.
- Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.
- 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. ISSN 2835-8856. URL <https://openreview.net/forum?id=bCiNWDmly2>.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
- 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. doi: 10.1198/jasa.2009.0121.
- I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 2002. doi: 10.1007/b98835.
- Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017.
- Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.
- 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.
- 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. doi: 10.1109/CAMSAP45676.2019.9022455.
- Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.
- Zhenyu Liao and Michael W Mahoney. Random matrix theory for deep learning: Beyond eigenvalues of linear models. *arXiv preprint arXiv:2506.13139*, 2025.
- 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.
- Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

- Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *The Annals of Applied Probability*, 35(4):2411–2470, 2025.
- Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.
- 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.
- 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.
- 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.
- 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.
- Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.
- 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.
- Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.
- Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 12 2008. doi: 10.1214/08-AOS618.
- 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. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.
- Ivan Nourdin, Giovanni Peccati, and Gesine Reinert. Invariance principles for homogeneous sums: Universality of Gaussian Wiener chaos. *The Annals of Probability*, 38(5):1947 – 1985, 2010. doi: 10.1214/10-AOP531. URL <https://doi.org/10.1214/10-AOP531>.

- 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. URL <https://arxiv.org/abs/2408.01062>.
- Sandrine Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.
- Sandrine Péché. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.
- Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.
- 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.
- 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.
- 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.
- Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323.
- Levent Sagun, Léon Bottou, and Yann LeCun. Singularity of the hessian in deep learning. *CoRR*, abs/1611.07476, 2016. URL <http://arxiv.org/abs/1611.07476>.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467.
- 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.
- Hossein Taheri, Avishek Ghosh, and Arya Mazumdar. On the theory of continual learning with gradient descent for neural networks. *arXiv preprint arXiv:2510.05573*, 2025.
- Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.
- Francesco Giacomo Tricomi, Arthur Erdélyi, et al. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math*, 1(1):133–142, 1951.
- Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.
- Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability*, 34(2):1896–1947, 2024.
- 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. URL <https://openreview.net/forum?id=gpqBGyKeKH>.
- Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4891–4957. PMLR, 2024.
- Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.
- 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.
- Andrew Gordon Wilson. Position: Deep learning is not so mysterious or different. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=42Au7FoD8F>.
- 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.
- 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, 2022.

Contents

1	Introduction	1
1.1	Our Approach	3
1.2	Our Contributions	4
2	Additional Literature Review	5
3	Notations and Preliminaries	6
4	Main Results	8
4.1	Proportional Limit with Finite SNR Case	8
4.2	Large-SNR Proportional Limit Case	10
4.3	Pretrained Weight Matrix Case	12
4.4	Quadratic Sample-size Case	13
5	Conclusion	14
A	Proof Strategy for Main Results	23
B	Numerical Simulations: Varying Knobs	24
B.1	Knob I: SNR of the XOR Dataset	24
B.2	Knob II: Activation Functions	24
B.3	Knob III: Weight Structure	27
B.4	Knob IV: Sample Size	28
C	Preliminaries	30
C.1	Properties of XOR Data	30
C.2	Properties of CK Matrix	34
C.3	Properties of z-transform and T-transform	38
C.4	Centered and Normalized GELU Function	41
D	Quadratic Equivalence for Proportional Limit	42
E	Proof for Finite-SNR and Proportional Limit Regime	44
E.1	Analysis of the Population Covariance Matrix	44
E.2	Analysis of Spike in Population Covariance	45
E.3	BBP Transition for the Null Model	47
E.4	BBP Transition for Rank-three Additive Deformation	48
E.4.1	Orthogonal Decomposition of Weight	49
E.4.2	Mean Spike Strength and Orthogonality	50
E.4.3	Resolvent Quadratic Forms for the Spike Analysis	51
E.4.4	Analyses of Additive Spikes and Alignments	56
E.5	Proof of Theorem 3	68
F	Kernel Spectral Clustering for XOR	68

G	Proof for Large-SNR Regime	70
G.1	Projected QE Model	70
G.2	The Covariance Spike and the Projected Resolvents	72
G.3	Master Determinant Equation for Coupled Outliers	73
G.4	Proof of Theorem 7	75
H	Proof for Pretrained Weight Regime	76
H.1	Moment Estimates and Deterministic Directions	76
H.2	QE for Spiked Weight	78
H.3	The Compressed Covariance Spike	79
H.4	Proof of Giant Spike and Bulk Limit	81
H.5	Proof of Order-one Outliers and Label Alignment	81
I	Proof for Quadratic Sample-size Regime	83
I.1	The Projection Preserves the XOR Label	85
I.2	QE of the Population Covariance Matrix	87
I.2.1	The Quadratic Term	87
I.2.2	The Cubic and Quartic Terms	92
I.3	BBP Transition for the Projected Population CK	95
I.4	BBP Transition for the Projected CK	98

Appendix A. Proof Strategy for Main Results

Dataset $\mathbf{X}$ in (3.4) is decomposed as a rank-2 XOR signal $\mathbf{M}$ and isotropic noise $\mathbf{Z}$. In the proportional limit, for Theorems 3, 7, and 8, we expand $\sigma(\mathbf{W}(\mathbf{Z} + \mathbf{M}))$ around $\mathbf{W}\mathbf{Z}$ and retain Hermite components of σ up to second order. This yields a bulk term $\sigma(\mathbf{W}\mathbf{Z})$ and a low-rank spike decomposition for $\mathbf{Y}$. Let $\mathbf{Y}_0 := \frac{1}{\sqrt{N}}\sigma(\mathbf{W}\mathbf{Z})$ and $\theta_{\text{SNR}} := r\sqrt{\frac{n}{2d}}$. Consider Gaussian random vectors $\mathbf{g}_1 = \mathbf{W}\mathbf{u}_1$, $\mathbf{g}_2 = \mathbf{W}\mathbf{u}_2$. Theorems 3, 7, and 8 repeatedly use the following QE formulation, a formal approximation of (1.2):

$$\mathbf{Y}_{\text{QE}} := \mathbf{Y}_0 + \underbrace{\frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}} (\mathbf{g}_1 \mathbf{v}_1^\top + \mathbf{g}_2 \mathbf{v}_2^\top)}_{\mathbf{T}_1 \text{ (linear spikes)}} + \underbrace{\frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N}} (\mathbf{g}_1^{\odot 2} \mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2} \mathbf{v}_2^{\odot 2\top})}_{\mathbf{T}_2 \text{ (quadratic spikes)}}. \quad (\text{A.1})$$

Proposition 10 (QE for CK in Proportional Limit) *Assume $r = O(n^{1/4})$ and Assumptions 1 and 2 hold. Let $\mathbf{Y}_{\text{QE}}$ be (A.1). Then, in probability, $\|\mathbf{Y} - \mathbf{Y}_{\text{QE}}\| \rightarrow 0$. Consequently, if $\|\mathbf{Y}\|, \|\mathbf{Y}_{\text{QE}}\| \lesssim 1$, then $\mathbf{K} = \mathbf{Y}^\top \mathbf{Y}$ and $\mathbf{K}_{\text{QE}} := \mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}}$ have asymptotically the same outlier eigenvalues and eigenvector alignments.*

With this proposition, we can analyze informative spikes from $\mathbf{T}_1$ and $\mathbf{T}_2$, although we have to carefully consider the uninformative spikes from the null model $\mathbf{Y}_0$, which gives more technical difficulty. The magnitude of $\mathbf{T}_1$ scales like $\theta_{\text{SNR}} b_\sigma$, while $\mathbf{T}_2$ scales like $\theta_{\text{SNR}}^2 c_\sigma / \sqrt{n}$. For finite SNR (or when $c_\sigma = 0$), the quadratic spikes $\mathbf{T}_2$ are negligible relative to the bulk. In this case, only the linear spikes in $\mathbf{T}_1$ contribute; but these are useless for linear clustering. In the large-SNR regime (with $c_\sigma \neq 0$), $\mathbf{T}_2$ becomes comparable to the bulk. Then since $\mathbf{y} \in \text{span}\{\mathbf{v}_1^{\odot 2}, \mathbf{v}_2^{\odot 2}\}$, the corresponding outlier eigenvectors become label-informative. Proposition 10 fails when we consider

the quadratic sample-size regime $n = \Theta(d^2)$ in Section 4.4. In Appendix I.2, we develop a different QE for the population level of CK matrix in this case, which captures the concentration of the random kernel around a deterministic quadratic polynomial kernel.

Appendix B. Numerical Simulations: Varying Knobs

We now present numerical simulations that validate and complement the theoretical predictions in Section 4.

Baseline case: proportional limit with finite SNR. Figure 1 shows the random CK spectra for the XOR model in the finite-SNR regime and proportional limit (Assumption 2), together with the theoretical spike predictions from Theorem 3. For the activation function used in this simulation, uninformative spikes described in Theorem 3(ii) do not appear. Consistent with Theorem 3(iii)–(iv), the two leading eigenvalues are outliers, and their eigenvectors exhibit nontrivial alignment with v_1 and v_2 , defined in (3.6), while remaining asymptotically orthogonal to the XOR label vector y . Consequently, although one could potentially apply downstream nonlinear transformations to these leading eigenvectors to classify the data, direct linear classification of the XOR problem from the CK eigenvectors fails in this regime.

This setting serves as our baseline. The theoretical results in Section 4 identify four primary control parameters, or “knobs,” that govern the emergence of CK outliers and the alignment of their eigenvectors with the XOR labels y : (1) the SNR r^2 ; (2) the choice of nonlinearity σ , through the constants (b_σ, c_σ) defined in (3.2) and (3.3); (3) the weight structure, comparing random and pretrained weights; and (4) the asymptotic scaling regime relating (n, d, N) . Below, we vary these knobs one at a time to isolate their effects on the CK spectrum relative to this baseline.

B.1. Knob I: SNR of the XOR Dataset

Theorem 7 shows that increasing the SNR to $r^2 = \Theta(n^{1/2})$ can lead to the emergence of a label-informative order-one CK outlier. In this regime, the quadratic component of σ is no longer negligible if $c_\sigma \neq 0$, making linear classification of XOR possible through (4.13) in Theorem 7(iv). When moving from Figure 1 to Figure 2, we only increase the SNR from 36 to 1225. This change produces two additional outliers, one of whose eigenvectors asymptotically aligns with the XOR labels y .

Figures 1 and 2 use the centered and normalized ReLU activation, for which both b_σ and c_σ are nonzero. In Figure 3, we show the same SNR-driven phase transition using a centered and normalized activation $\sigma(x) \propto \sqrt{x^2 + 1} - 1$, for which $b_\sigma = 0$ and $c_\sigma \neq 0$. Since $b_\sigma = 0$, the diverging linear outliers predicted by Theorem 3(iii) and Theorem 7(ii) are absent. As the SNR increases from $r^2 = 36$ to 625, the second eigenvector develops a two-class structure, and XOR becomes linearly classifiable using CK eigenvectors.

B.2. Knob II: Activation Functions

Figures 1, 2, and 3 already show that changing the activation function can lead to qualitatively different CK spectral behavior. We now present two additional examples: the emergence of uninformative spikes and the absence of informative spikes.

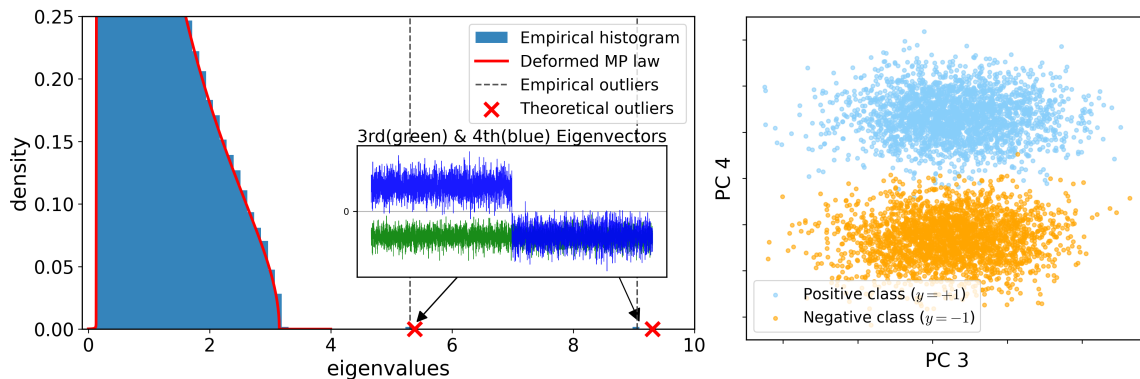


Figure 2: **Large-SNR proportional regime: informative CK spikes recover the XOR labels.** We consider the same setting as in Figure 1 only increasing SNR to $r = 35^2$. **Left:** spectrum of $\mathbf{K}$ (blue) together with the predicted curve from Theorem 7 (red curve). To focus on the order-one eigenvalues, we do not display the first two extremely large eigenvalues $\hat{\lambda}_1 = 153.9$ and $\hat{\lambda}_2 = 153.1$ predicted by (4.10). The red $\times$ are the theoretical locations of the order-one outliers. The subfigure presents the 3rd & 4th eigenvectors against sample index. The 4th outlier eigenvector of $\mathbf{K}$ is informative for linear classification of XOR. **Right:** kernel-PCA visualization using the 3rd & 4th PCs of $\mathbf{K}$. In contrast to Figure 1, these two spectral coordinates clearly separate the positive and negative classes by a linear boundary.

Uninformative spikes can emerge even in the null model. Theorem 3(ii) describes uninformative spikes of the CK matrix even when $r = 0$, namely for an isotropic Gaussian dataset. These uninformative spikes are caused by the nonzero quadratic coefficient $c_\sigma \neq 0$. Such spikes were studied by Benigni and P  ch   (2022) for more general datasets, while our result gives a precise characterization of their locations for Gaussian data. To illustrate their emergence in the CK matrix, we consider the single-parameter family of activation functions from Benigni and P  ch   (2022),

$$f_{\alpha}(x) = \frac{\cos(\alpha x) - e^{-\alpha^2/2}}{\sqrt{e^{-\alpha^2}(\cosh(\alpha^2) - 1)}}. \quad (\text{B.1})$$

These functions satisfy Assumption 1. Moreover, $b_\sigma = 0$ for all choices of α , while c_σ varies with α . Figure 4 plots the CK spectra under f_α in the null model SNR = 0 for $\alpha = 2$ and $\alpha = 1.5$. As α decreases, two outlier spikes emerge in the CK spectrum. These two uninformative spikes arise from the population covariance of the CK and from the CK mean; Theorem 3(ii) predicts that their locations are asymptotically close.

XOR remains linearly nonseparable at large SNR when $c_\sigma = 0$. If the activation function σ satisfies $c_\sigma = 0$, then the quadratic component in the CK is absent. As discussed in Theorem 7, this implies that the CK cannot be used to linearly classify XOR for any signal-to-noise ratio. Figure 5 follows the same setting as the bottom panel of Figure 3, changing only the activation to $\sigma \propto \tanh$ with $c_\sigma = 0$. Here $\text{SNR} = 625$. By Theorem 7, only two diverging linear outliers separate from the bulk. Moreover, their associated spectral coordinates are not linearly separable with respect to the XOR labels.

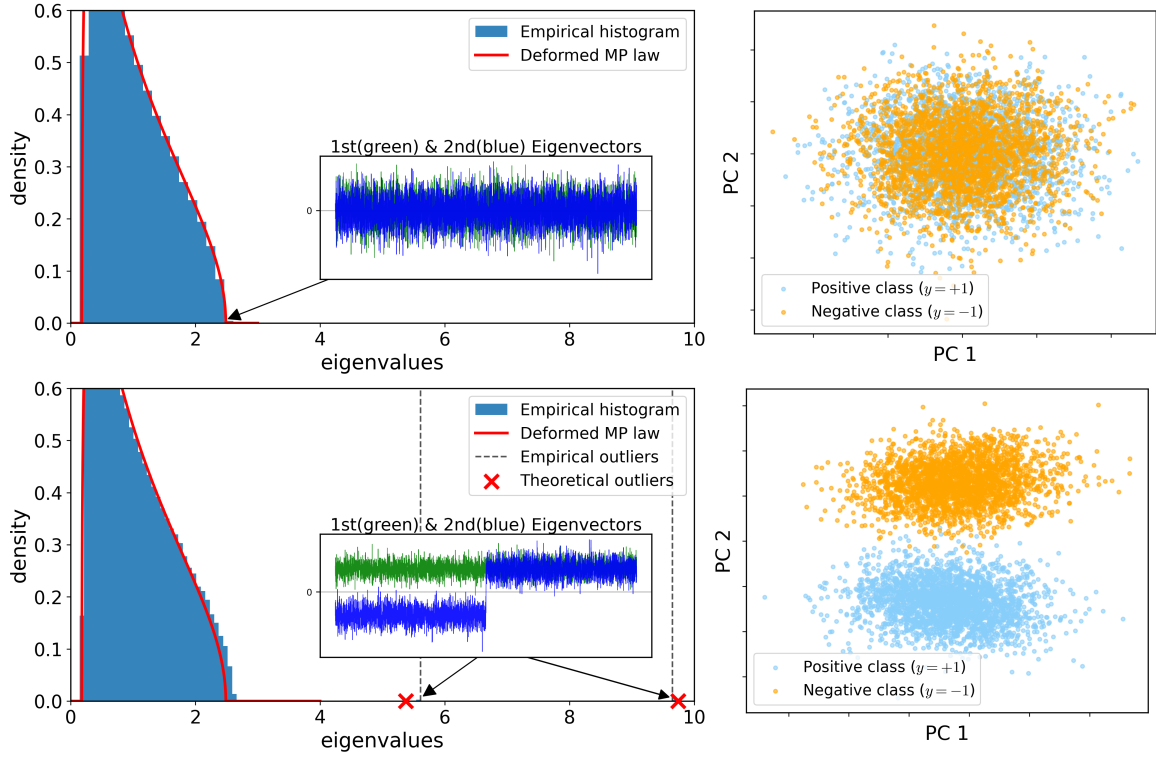


Figure 3: **CK spectral phase transition driven by SNR. Top:** CK spectrum in the finite-SNR regime with $r = 6$; no eigenvalue separates from the bulk. **Bottom:** CK spectrum in the large-SNR regime with $r = 25$. Two quadratic eigenvalues separate from the bulk, and the second leading eigenvector aligns with the XOR labels. Here the activation $\sigma(x) \propto \sqrt{x^2 + 1} - 1$, $n = 5000$, and $N = d = 15000$.

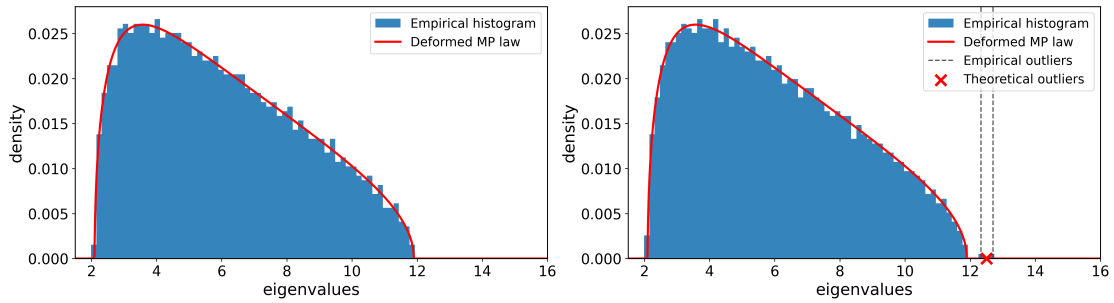


Figure 4: **Emergence of uninformative spikes when SNR = 0.** CK spectra under the activation function f_α defined in (B.1), with $n = 6000$, $N = d = 1000$, and $r = 0$. **Left:** $\alpha = 2$. **Right:** $\alpha = 1.5$. Two uninformative spikes emerge and are close to the theoretical predictions from Theorem 3(ii), shown by red $\times$ markers.

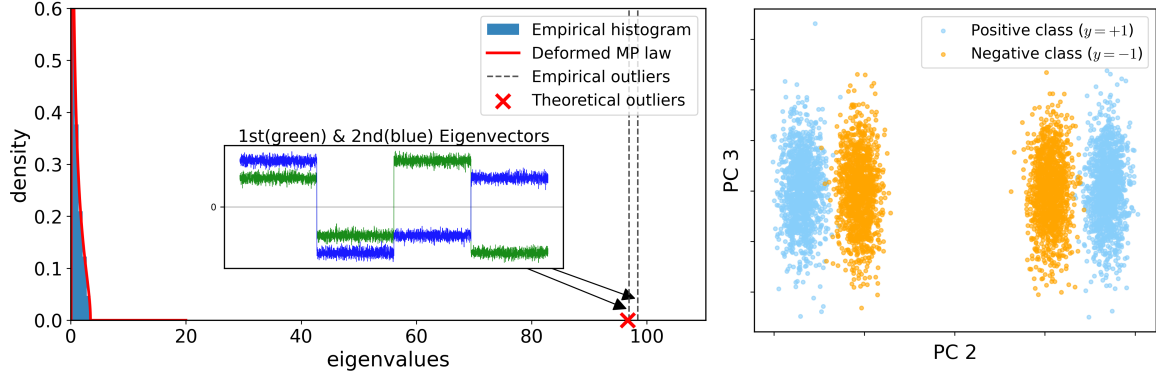


Figure 5: **Large-SNR regime with $c_\sigma = 0$: linear classification of XOR still fails.** Here $n = 5000$, $N = d = 15000$, $r = 25$, and $\sigma(x) \propto \tanh(x)$, so $c_\sigma = 0$. **Left:** spectrum of K (blue), together with the theoretical locations of the two diverging outliers predicted by Theorem 7(ii), shown by red $\times$ markers. The outlier eigenvectors, displayed in the inset, exhibit a four-block pattern associated with the four Gaussian-mixture components but remain asymptotically orthogonal to the XOR label vector y . **Right:** kernel-PCA visualization of K . Thus the leading nontrivial spectral coordinates are not linearly separable with respect to the XOR labels.

B.3. Knob III: Weight Structure

Following Theorem 8, we generate a rank-one perturbation of the randomly initialized weights, $W_1 = W + \theta ab^\top$, as a model of pretrained weights. In Figure 6, we plot the spectrum of the projected CK matrix K_s and its PCs for the centered and normalized ReLU activation, an XOR dataset with SNR $r^2 = 36$, and spike strength $\theta = 5$. The rank-one structure in the pretrained weights produces informative spikes. As predicted by Theorem 8, two order-one outlier eigenvalues emerge, and the second leading eigenvector exhibits a two-class structure that enables linear classification of XOR.

Pretrained weights from feature learning on a real-world dataset. The pretrained-weight model in (4.14), used in Theorem 8, is an idealized model for our analysis. We now show that real pretrained weights obtained during training can exhibit similar label-aligned CK spikes. A recent line of work (Martin and Mahoney, 2021, 2019; Wang et al., 2023) has tracked the evolution of the empirical spectral distribution of weight matrices throughout the training process of deep NNs. We focus on an early-training phase in which the weight spectrum is well described by an “MP bulk+spikes” structure. The emergence of spikes in the weight spectrum is strongly related to feature learning in NNs (Frei et al., 2022) and suggests that these spikes carry task-relevant information.

We consider the CK matrix formed using pretrained weights extracted from the first-layer weight matrix of a 20-layer fully connected NN trained on CIFAR-2. Each hidden layer has width $N = 512$, and the input dimension is $d = 3072$. We focus on the first 50 steps of SGD training, using learning rate 0.05, momentum 0.5, and batch size 128. At step 50, the test accuracy is 80.4%, and the training accuracy is 91.41%. Figure 7 shows the spectra of the first-layer trained weights at steps 1, 2, and 50 (bottom row), together with the spectra of CK matrices evaluated on XOR data using these pretrained

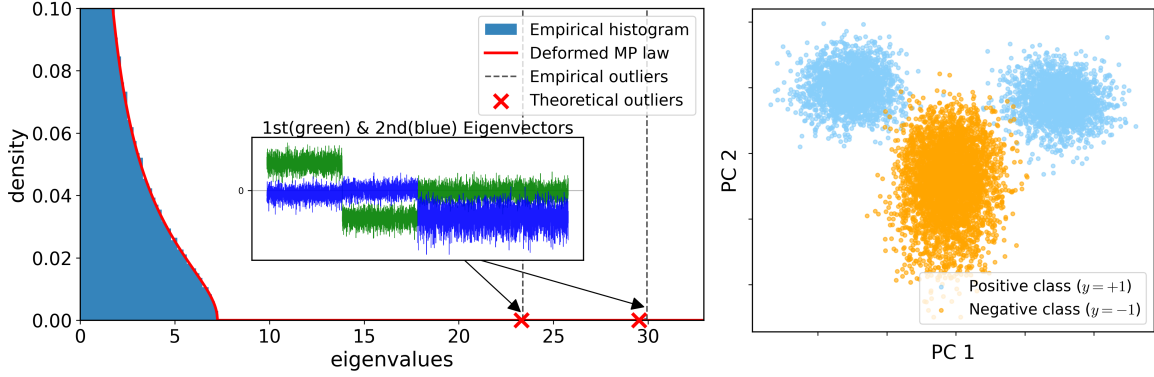


Figure 6: **Spiked-weight regime: a weight-induced spike aligns with XOR labels.** Here $N = n = 8000$, $d = 4000$, the SNR is $r^2 = 36$, $\sigma(x) \propto \text{ReLU}(x)$, and the weight spike strength in (4.14) is $\theta = 5.0$. **Left:** spectrum of K_s (blue), as defined in Theorem 8, together with the isolated eigenvalue locations predicted by Theorem 8, shown by red $\times$ markers. The top two empirical eigenvectors are shown in the insets; the second one enables linear classification of XOR. **Right:** kernel-PCA visualization of the samples using the second and first PCs of K_s .

weights (top row). The emergence of spikes in the weight spectra during training suggests that the model is learning task-dependent features.

For the CK matrices in this experiment, $n = 8000$, $N = 512$, $d = 3072$, and the SNR is $r^2 = 36$. In contrast to the prediction of Theorem 3 for randomly initialized weights, the pretrained CK at step 50 exhibits four spikes. We omit the top outlier in Figure 7 because it is a large-order outlier. The fourth leading eigenvector aligns with the XOR labels. After sufficient training, this eigenvector transitions from pure noise to near-perfect alignment with the XOR labels, thereby enabling linear classification.

These experiments suggest that the LE approximation of the CK can break down even during the early stages of training. They also indicate that, under suitable training regimes and parameter choices, there can be a critical phase transition in which the emergence of weight spikes makes quadratic channels in the CK nonnegligible. One possible explanation is that training the NN induces spiked eigenvalues with eigenvectors that align with XOR labels, even though the model is not trained directly on the XOR task. Theorem 8 captures this mechanism through an idealized model in which a large spike is artificially introduced into the weights, producing a spiked CK eigenvalue with an XOR-label-aligned eigenvector.

B.4. Knob IV: Sample Size

Following Section 4.4, we study CK spectra under the quadratic sample-size scaling $n = \Theta(d^2)$. Figure 8 shows the spectrum and PCs of the projected CK matrix for the centered and normalized ReLU activation and an XOR dataset with finite SNR $r^2 = 6.25$. Here $n = 15000$, $N = 20000$, and $d = 150$. In this scaling, many outliers emerge in the unprojected CK spectrum. We therefore focus on the projected CK matrix K_σ defined in Theorem 9, which isolates the label-aligned spike. We

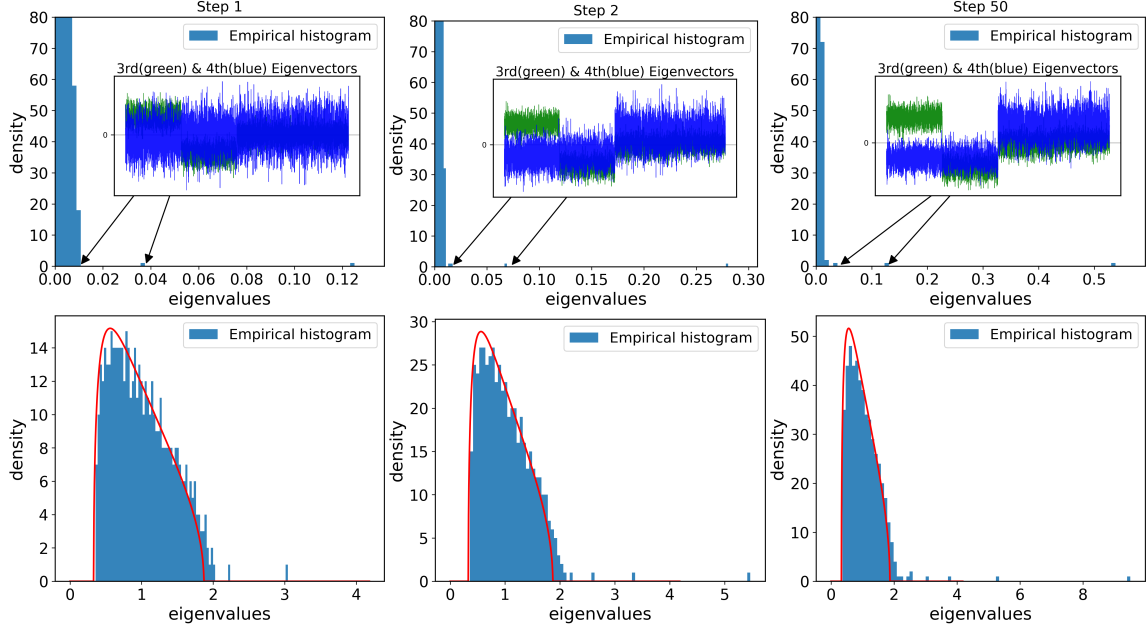


Figure 7: **Pretrained weights help linear classification of XOR.** **Bottom:** spectra of the first-layer weight matrix of a NN trained on CIFAR-2 at steps 1, 2, and 50. The red curves are MP laws. **Top:** CK spectra with pretrained weights at training steps 1, 2, and 50. For these CK matrices, $n = 8000$, $N = 512$, $d = 3072$, the SNR is $r^2 = 36$, and $\sigma(x) \propto \text{ReLU}(x)$.

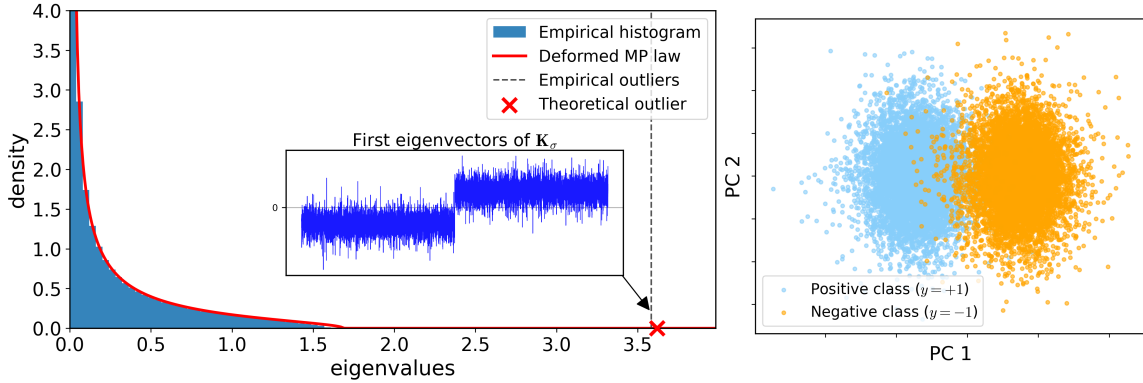


Figure 8: **Quadratic sample-size regime: an informative CK spike aligns with XOR labels.** Here $n = 15000$, $N = 20000$, $d = 150$, the SNR is $r^2 = 6.25$, and $\sigma(x) \propto \text{ReLU}(x)$. **Left:** spectrum of the projected CK matrix K_σ , defined in Theorem 9, together with the asymptotic spike prediction from Theorem 9, shown by a red $\times$ marker. There is a single isolated spike, whose eigenvector is displayed in the inset. Its blockwise sign pattern aligns with the binary XOR partition, explaining the linear class separation seen on the right. **Right:** kernel-PCA visualization of the samples using the top two PCs of K_σ .

observe a single eigenvalue outside the bulk of the deformed MP law μ_q , and its eigenvector aligns with the XOR labels $\mathbf{y}$, even at an SNR smaller than that used in Figure 1.

Appendix C. Preliminaries

Throughout this paper, we utilize the stochastic domination notation introduced by Erdős et al. (2013).

Definition 11 (Stochastic domination) Consider two families of random variables

$$X = \{X_n(u) : n \in \mathbb{N}, u \in U_n\}, \quad Y = \{Y_n(u) : n \in \mathbb{N}, u \in U_n\},$$

with parameter $u \in U_n$ where U_n is a possibly n -dependent deterministic set. Let $\mathcal{E} = \{E_n\}_{n \in \mathbb{N}}$ 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 $\epsilon > 0$ and large $D > 0$, there exists some $N_0(\epsilon, D)$, such that for all $n \geq N_0(\epsilon, D)$,

$$\sup_{u \in U_n} \mathbb{P}(|X_n(u)| \geq n^\epsilon Y_n(u)) \leq n^{-D}.$$

2. For a family of random matrices $\mathbf{A}$ and a family of non-negative random variables ζ , $\mathbf{A} = O_{\prec}(\zeta)$ represents $|\langle \mathbf{v}, \mathbf{A} \mathbf{u} \rangle| \prec \zeta \|\mathbf{v}\| \|\mathbf{u}\|$ uniformly for all deterministic vectors $\mathbf{v}$ and $\mathbf{u}$.

If $\{\mathbf{A}_n\}$ are random matrices, we interpret $\|\mathbf{A}_n\|_{\text{op}} = o_{\mathbb{P}}(1)$ if and only if $\|\mathbf{A}_n\|_{\text{op}} \xrightarrow{\mathbb{P}} 0$. For simplicity, in the proofs, we use $\|\mathbf{A}_n\|$ to represent operator norm $\|\mathbf{A}_n\|_{\text{op}}$.

C.1. Properties of XOR Data

We introduce the following notion of (τ_n, B) -orthonormality from (Fan and Wang, 2020).

Definition 12 A matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ is (τ_n, B) -orthonormal if the columns $\{\mathbf{x}_\alpha\}_{\alpha=1}^n$ of $\mathbf{X}$ satisfy

$$\left| \|\mathbf{x}_\alpha\| - 1 \right| \leq \tau_n, \quad \left| \|\mathbf{x}_\beta\| - 1 \right| \leq \tau_n, \quad \left| \mathbf{x}_\alpha^\top \mathbf{x}_\beta \right| \leq \tau_n \quad (\text{C.1})$$

for all pairs $\alpha \neq \beta \in [n]$, $\sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\| - 1)^2 \leq B^2$ and $\|\mathbf{X}\| \leq B$.

Lemma 13 Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be the XOR data matrix defined in (3.4). 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 $1 - n^{-c}$, $\mathbf{X}$ satisfies (C.1) with

$$\tau_n = C \left(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d} \right). \quad (\text{C.2})$$

Moreover, on the same event,

$$\sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\| - 1)^2 \leq \sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\|^2 - 1)^2 \leq C \left(\frac{n}{d} + \frac{nr^2}{d^2} + \frac{nr^4}{d^2} \right), \quad (\text{C.3})$$

and also $\|\mathbf{X}\| \leq C\left(1 + (1+r)\sqrt{\frac{n}{d}}\right)$. 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 $\mathbf{X}$ is (τ_n, B) -orthonormal with probability at least $1 - n^{-c}$.

Proof. Recall that $\theta_{\text{SNR}} := r\sqrt{\frac{n}{2d}}$, and $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ are orthonormal vectors defined by (3.6). This proof is basically following the proof of Proposition 3.3 of Fan and Wang (2020). Write the columns as $\mathbf{x}_\alpha = \mathbf{z}_\alpha + \mathbf{m}_\alpha$, where $\mathbf{z}_\alpha$ is the α -th column of $\mathbf{Z}$ and $\mathbf{m}_\alpha$ is the α -th column of $\mathbf{M}$. Since $\mathbf{M} = \theta_{\text{SNR}} \mathbf{u}_1 \mathbf{v}_1^\top + \theta_{\text{SNR}} \mathbf{u}_2 \mathbf{v}_2^\top$ and $\text{supp}(\mathbf{v}_1) \cap \text{supp}(\mathbf{v}_2) = \emptyset$, for each α exactly one of $\mathbf{v}_1[\alpha], \mathbf{v}_2[\alpha]$ is nonzero, and $\mathbf{m}_\alpha = \theta_{\text{SNR}} \mathbf{u}_1 \mathbf{v}_1[\alpha]$ or $\theta_{\text{SNR}} \mathbf{u}_2 \mathbf{v}_2[\alpha]$. Using $|\mathbf{v}_1[\alpha]| = |\mathbf{v}_2[\alpha]| = \sqrt{2/n}$ on the support, we obtain for all α

$$\|\mathbf{m}_\alpha\| = \theta_{\text{SNR}} \sqrt{\frac{2}{n}} = r \sqrt{\frac{n}{2d}} \sqrt{\frac{2}{n}} = \frac{r}{\sqrt{d}}, \quad \|\mathbf{m}_\alpha\|^2 = \frac{r^2}{d}, \quad \|\mathbf{m}_\alpha\|^4 = \frac{r^4}{d^2}. \quad (\text{C.4})$$

Now we verify the four requirements in the (τ_n, B) -orthonormal definition:

Uniform control of column norms. For each $\alpha \in [n]$,

$$\|\mathbf{x}_\alpha\|^2 = \|\mathbf{z}_\alpha\|^2 + 2\mathbf{z}_\alpha^\top \mathbf{m}_\alpha + \|\mathbf{m}_\alpha\|^2. \quad (\text{C.5})$$

Since $\mathbf{z}_\alpha \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/d)$, we have $d\|\mathbf{z}_\alpha\|^2 \sim \chi_d^2$. By Lemma 1 of Laurent and Massart (2000),

$$\mathbb{P}\left(\left|\|\mathbf{z}_\alpha\|^2 - 1\right| \geq 2\sqrt{\frac{t}{d}} + 2\frac{t}{d}\right) \leq 2e^{-t} \quad (\text{C.6})$$

for all $t > 0$. Taking $t = C_0 \log n$ and union bounding over $\alpha \in [n]$ yields: with probability at least $1 - 2n \cdot n^{-C_0}$, for C_0 large enough,

$$\max_{\alpha \in [n]} \left| \|\mathbf{z}_\alpha\|^2 - 1 \right| \leq C \sqrt{\frac{\log n}{d}}. \quad (\text{C.7})$$

In particular, on the same event, $\max_{\alpha} \|\mathbf{z}_\alpha\|^2 \leq 2$ for all large n . Conditioning on $\mathbf{m}_\alpha, \mathbf{z}_\alpha^\top \mathbf{m}_\alpha \mid \mathbf{m}_\alpha \sim \mathcal{N}(0, r^2/d^2)$, using (C.4). Hence for $t = C_0 \log n$,

$$\mathbb{P}\left(|\mathbf{z}_\alpha^\top \mathbf{m}_\alpha| \geq \frac{r}{d} \sqrt{2t}\right) \leq 2e^{-t}.$$

A union bound over $\alpha \in [n]$ yields: with probability at least $1 - n^{-c}$,

$$\max_{\alpha \in [n]} |\mathbf{z}_\alpha^\top \mathbf{m}_\alpha| \leq C \frac{r\sqrt{\log n}}{d}. \quad (\text{C.8})$$

Plugging (C.7), (C.8), and $\|\mathbf{m}_\alpha\|^2 = r^2/d$ into (C.5) gives, with probability at least $1 - n^{-c}$,

$$\max_{\alpha \in [n]} \left| \|\mathbf{x}_\alpha\|^2 - 1 \right| \leq C \left(\sqrt{\frac{\log n}{d}} + \frac{r\sqrt{\log n}}{d} + \frac{r^2}{d} \right). \quad (\text{C.9})$$

We now simplify the RHS under regimes (A1) and (A2). In both regimes, for large n , $\frac{r\sqrt{\log n}}{d} \leq \sqrt{\frac{\log n}{d}} + \frac{r^2}{d}$. Therefore (C.9) implies

$$\max_{\alpha \in [n]} |\|\mathbf{x}_\alpha\|^2 - 1| \leq C \left(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d} \right). \quad (\text{C.10})$$

Finally, notice that

$$|\|\mathbf{x}_\alpha\| - 1| = \frac{|\|\mathbf{x}_\alpha\|^2 - 1|}{\|\mathbf{x}_\alpha\| + 1} \leq |\|\mathbf{x}_\alpha\|^2 - 1|, \quad (\text{C.11})$$

so (C.10) yields

$$\max_{\alpha \in [n]} |\|\mathbf{x}_\alpha\| - 1| \leq C \left(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d} \right). \quad (\text{C.12})$$

This verifies the column-norm part of (τ_n, B) -orthonormality with τ_n as in (C.2).

Uniform control of pairwise inner products. Fix $\alpha \neq \beta \in [n]$. Expand

$$\mathbf{x}_\alpha^\top \mathbf{x}_\beta = \mathbf{z}_\alpha^\top \mathbf{z}_\beta + \mathbf{z}_\alpha^\top \mathbf{m}_\beta + \mathbf{m}_\alpha^\top \mathbf{z}_\beta + \mathbf{m}_\alpha^\top \mathbf{m}_\beta. \quad (\text{C.13})$$

The scalar $\mathbf{z}_\alpha^\top \mathbf{z}_\beta$ is not exactly Gaussian, but it is conditionally Gaussian: conditioning on $\mathbf{z}_\alpha$, $\mathbf{z}_\alpha^\top \mathbf{z}_\beta \mid \mathbf{z}_\alpha \sim \mathcal{N}(0, \|\mathbf{z}_\alpha\|^2/d)$. On the event $\max_\alpha \|\mathbf{z}_\alpha\|^2 \leq 2$ we have

$$\mathbb{P}\left(|\mathbf{z}_\alpha^\top \mathbf{z}_\beta| \geq t \mid \mathbf{z}_\alpha\right) \leq 2 \exp\left(-\frac{dt^2}{4\|\mathbf{z}_\alpha\|^2}\right) \leq 2 \exp\left(-c dt^2\right).$$

Taking $t = C\sqrt{\frac{\log n}{d}}$ and union bounding over all $\alpha \neq \beta \in [n]$ yields with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\mathbf{z}_\alpha^\top \mathbf{z}_\beta| \leq C\sqrt{\frac{\log n}{d}}. \quad (\text{C.14})$$

For $\alpha \neq \beta$, conditional on $\mathbf{m}_\beta$, $\mathbf{z}_\alpha^\top \mathbf{m}_\beta \sim \mathcal{N}(0, r^2/d^2)$, and similarly $\mathbf{m}_\alpha^\top \mathbf{z}_\beta \sim \mathcal{N}(0, r^2/d^2)$. A union bound over all $\alpha \neq \beta$ gives, with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\mathbf{z}_\alpha^\top \mathbf{m}_\beta| \leq C \frac{r\sqrt{\log n}}{d}, \quad \max_{\alpha \neq \beta} |\mathbf{m}_\alpha^\top \mathbf{z}_\beta| \leq C \frac{r\sqrt{\log n}}{d}. \quad (\text{C.15})$$

Each $\mathbf{m}_\alpha$ equals $\pm \frac{r}{\sqrt{d}} \mathbf{u}_1$ or $\pm \frac{r}{\sqrt{d}} \mathbf{u}_2$. By Cauchy–Schwarz inequality, $|\mathbf{m}_\alpha^\top \mathbf{m}_\beta| \leq \|\mathbf{m}_\alpha\| \|\mathbf{m}_\beta\| = r^2/d$. Hence

$$\max_{\alpha \neq \beta} |\mathbf{m}_\alpha^\top \mathbf{m}_\beta| \leq \frac{r^2}{d}. \quad (\text{C.16})$$

Combining (C.13)–(C.16), we obtain with probability at least $1 - n^{-c}$,

$$\max_{\alpha \neq \beta} |\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq C \left(\sqrt{\frac{\log n}{d}} + \frac{r^2}{d} \right). \quad (\text{C.17})$$

This verifies the pairwise-inner-product part of (τ_n, B) -orthonormality with τ_n as in (C.2).

Sum of squared squared-norm deviations. From (C.5), $\|\mathbf{x}_\alpha\|^2 - 1 = (\|\mathbf{z}_\alpha\|^2 - 1) + 2\mathbf{z}_\alpha^\top \mathbf{m}_\alpha + \|\mathbf{m}_\alpha\|^2$. Using $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, we get

$$(\|\mathbf{x}_\alpha\|^2 - 1)^2 \leq 3(\|\mathbf{z}_\alpha\|^2 - 1)^2 + 12(\mathbf{z}_\alpha^\top \mathbf{m}_\alpha)^2 + 3\|\mathbf{m}_\alpha\|^4.$$

Summing over α and using (C.4) gives

$$\sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\|^2 - 1)^2 \leq 3 \sum_{\alpha=1}^n (\|\mathbf{z}_\alpha\|^2 - 1)^2 + 12 \sum_{\alpha=1}^n (\mathbf{z}_\alpha^\top \mathbf{m}_\alpha)^2 + 3n \frac{r^4}{d^2}. \quad (\text{C.18})$$

Set $g_\alpha := \sqrt{d} \mathbf{z}_\alpha \sim \mathcal{N}(0, I_d)$ so that $\|\mathbf{z}_\alpha\|^2 = \|g_\alpha\|^2/d$ and $\|g_\alpha\|^2 \sim \chi_d^2$. Define $U_\alpha := \|g_\alpha\|^2 - d$, and $X_\alpha := (\|\mathbf{z}_\alpha\|^2 - 1)^2 = \frac{U_\alpha^2}{d^2}$. Then $\mathbb{E}[X_\alpha] = \text{Var}(\|\mathbf{z}_\alpha\|^2) = 2/d$, hence $\mathbb{E} \sum_{\alpha=1}^n X_\alpha = 2n/d$. Write $U_\alpha = \sum_{i=1}^d (\xi_{i\alpha}^2 - 1)$ with $\xi_{i\alpha} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. For any integer $p \geq 2$, Rosenthal's inequality (Theorem 3 of Rosenthal (1970)) yields

$$\mathbb{E}|U_\alpha|^{2p} \leq C_p \left(d \mathbb{E}|\xi_{11}^2 - 1|^{2p} + (d \text{Var}(\xi_{11}^2 - 1))^p \right) \leq C'_p d^p, \quad (\text{C.19})$$

where C_p, C'_p depend only on p (not on d, n). Consequently,

$$\mathbb{E}|X_\alpha|^p = \frac{\mathbb{E}|U_\alpha|^{2p}}{d^{2p}} \leq \frac{C'_p}{d^p}, \quad \mathbb{E}|X_\alpha - \mathbb{E}X_\alpha|^p \leq \frac{C''_p}{d^p}. \quad (\text{C.20})$$

Let $\tilde{X}_\alpha := X_\alpha - \mathbb{E}X_\alpha$ so that $\mathbb{E}\tilde{X}_\alpha = 0$. Applying Rosenthal's inequality again (now over $\alpha = 1, \dots, n$) gives, for any integer $p \geq 2$,

$$\mathbb{E} \left| \sum_{\alpha=1}^n \tilde{X}_\alpha \right|^p \leq C_p \left(n \mathbb{E}|\tilde{X}_1|^p + (n \mathbb{E}\tilde{X}_1^2)^{p/2} \right) \leq C'_p \left(\frac{n}{d^2} \right)^{p/2},$$

using (C.20) and $\mathbb{E}\tilde{X}_1^2 \lesssim d^{-2}$. Therefore, by Markov's inequality, for any fixed $t > 0$,

$$\mathbb{P} \left(\sum_{\alpha=1}^n X_\alpha \geq \frac{2n}{d} + t \frac{n}{d} \right) = \mathbb{P} \left(\sum_{\alpha=1}^n \tilde{X}_\alpha \geq t \frac{n}{d} \right) \leq \frac{\mathbb{E} \left| \sum_{\alpha=1}^n \tilde{X}_\alpha \right|^p}{(tn/d)^p} \leq C''_p t^{-p} n^{-p/2}.$$

Taking $t = 1$ and p large enough (so that $p/2 \geq c$) yields, for some constants $c, C > 0$,

$$\sum_{\alpha=1}^n (\|\mathbf{z}_\alpha\|^2 - 1)^2 \leq C \frac{n}{d} \quad \text{with probability at least } 1 - n^{-c}. \quad (\text{C.21})$$

This bound is uniform in (n, d) and in particular covers both regimes $n \asymp d$ and $n = O(d^2)$.

Conditioning on $\mathbf{m}_\alpha$, the random variable $\mathbf{z}_\alpha^\top \mathbf{m}_\alpha$ is Gaussian $\mathcal{N}(0, r^2/d^2)$, hence $(\mathbf{z}_\alpha^\top \mathbf{m}_\alpha)^2$ is sub-exponential with mean r^2/d^2 . By the Bernstein inequality,

$$\sum_{\alpha=1}^n (\mathbf{z}_\alpha^\top \mathbf{m}_\alpha)^2 \leq C \frac{nr^2}{d^2} \quad \text{with probability at least } 1 - n^{-c}. \quad (\text{C.22})$$

Plugging (C.21), (C.22) into (C.18) gives, with probability at least $1 - n^{-c}$,

$$\sum_{\alpha=1}^n (\|\mathbf{x}_\alpha\|^2 - 1)^2 \leq C \left(\frac{n}{d} + \frac{nr^2}{d^2} + \frac{nr^4}{d^2} \right).$$

This proves (C.3) because of (C.11).

Operator norm bound. We bound $\|\mathbf{X}\| \leq \|\mathbf{Z}\| + \|\mathbf{M}\|$. A standard Gaussian operator norm bound states that for all $t \geq 0$,

$$\mathbb{P}\left(\|\mathbf{Z}\| \geq 1 + \sqrt{\frac{n}{d}} + t\right) \leq 2 \exp(-cdt^2).$$

Taking $t = \sqrt{\frac{\log n}{d}}$ yields $\|\mathbf{Z}\| \leq C(1 + \sqrt{n/d})$ with probability at least $1 - n^{-c}$. Since $\mathbf{u}_1, \mathbf{u}_2$ are orthonormal and $\mathbf{v}_1, \mathbf{v}_2$ are orthonormal, $\|\mathbf{M}\| = \theta_{\text{SNR}} = r\sqrt{\frac{n}{2d}} \leq Cr\sqrt{\frac{n}{d}}$. Therefore, with probability at least $1 - n^{-c}$,

$$\|\mathbf{X}\| \leq C\left(1 + \sqrt{\frac{n}{d}}\right) + Cr\sqrt{\frac{n}{d}} \leq C\left(1 + (1+r)\sqrt{\frac{n}{d}}\right).$$

This completes the proof. ■

C.2. Properties of CK Matrix

The following intermediary results are necessary to construct the QE in Appendix D.

Proposition 14 *Let χ_d denote the chi distribution with d degrees of freedom. Then, as $d \rightarrow \infty$,*

$$\frac{\mathbb{E}[\chi_d]}{\sqrt{d}} = 1 - \frac{1}{4d} + \frac{1}{32d^2} + \frac{5}{128d^3} + O(d^{-4}). \quad (\text{C.23})$$

Proof. It is classical that $\mathbb{E}[\chi_d] = \sqrt{2} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}$. (See, e.g., (Abramowitz and Stegun, 1964, Ch. 6), (DLMF, 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 $z \rightarrow \infty$ in $|\arg z| < \pi$ (DLMF, (5.11.2)):

$$\log \Gamma(z) = (z - \frac{1}{2}) \log z - z + \frac{1}{2} \log(2\pi) + \frac{1}{12z} - \frac{1}{360z^3} + O(z^{-5}). \quad (\text{C.24})$$

Write $\log Q_m = (\log \Gamma(m + \frac{1}{2}) - \log \Gamma(m)) - \frac{1}{2} \log m$. Using (C.24) with $z = m + \frac{1}{2}$ and $z = m$, we obtain

$$\begin{aligned} \log Q_m &= m \log(m + \frac{1}{2}) - (m + \frac{1}{2}) - (m - \frac{1}{2}) \log m + m \\ &\quad + \frac{1}{12(m + \frac{1}{2})} - \frac{1}{12m} - \frac{1}{360(m + \frac{1}{2})^3} + \frac{1}{360m^3} - \frac{1}{2} \log m + O(m^{-5}). \end{aligned}$$

The logarithmic terms simplify to $m \log(1 + \frac{1}{2m}) - \frac{1}{2}$. 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

$$m \log\left(1 + \frac{1}{2m}\right) = \frac{1}{2} - \frac{1}{8m} + \frac{1}{24m^2} - \frac{1}{64m^3} + O(m^{-4}), \quad (\text{C.25})$$

$$\frac{1}{12(m + \frac{1}{2})} - \frac{1}{12m} = -\frac{1}{24m^2} + \frac{1}{48m^3} + O(m^{-4}). \quad (\text{C.26})$$

Further, the difference $-\frac{1}{360(m+\frac{1}{2})^3} + \frac{1}{360m^3}$ is $O(m^{-4})$. Collecting terms, we have $\log Q_m = -\frac{1}{8m} + \frac{1}{192m^3} + O(m^{-4})$. 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 (C.23) by substituting $m = d/2$. ■

Notice that the same expansion follows from the general ratio formula [Tricomi et al. \(1951\)](#); [DLMF](#), for $z \rightarrow \infty$

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} \sim z^{a-b} \left(1 + \frac{(a-b)(a+b-1)}{2z} + \dots \right).$$

When the dataset is pure standard noise, let us define the null model as

$$\mathbf{Y}_0 := \frac{1}{\sqrt{N}} \sigma(\mathbf{W}\mathbf{Z}). \quad (\text{C.27})$$

Lemma 15 *Under the Assumption 1, we know that $\|\mathbb{E}[\mathbf{Y}_0]\| \lesssim \frac{\sqrt{n}}{d}$ as $n, d \rightarrow \infty$.*

Proof. Notice that $\mathbb{E}[\mathbf{Y}_0] = \frac{C_d}{\sqrt{N}} \mathbf{1}_N \mathbf{1}_n^\top$ where $C_d := \mathbb{E}[\sigma(\frac{1}{\sqrt{d}} \mathbf{w}^\top \mathbf{z})]$ where $\mathbf{z}, \mathbf{w} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Hence, $\|\mathbb{E}[\mathbf{Y}_0]\| = \sqrt{n}|C_d|$. Denote $\xi \sim \mathcal{N}(0, 1)$ which is independent with $\mathbf{w}$. 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

$$\begin{aligned} C_d &= \mathbb{E}[\sigma(\frac{1}{\sqrt{d}} \|\mathbf{w}\|\xi)] - \mathbb{E}[\sigma(\xi)] \\ &= \mathbb{E}[\sigma'(\xi)\xi(\frac{1}{\sqrt{d}} \|\mathbf{w}\| - 1)] + \mathbb{E}[\sigma''(\eta)\xi^2(\frac{1}{\sqrt{d}} \|\mathbf{w}\| - 1)^2] \\ &= c_\sigma(\mathbb{E}[\frac{1}{\sqrt{d}} \|\mathbf{w}\|] - 1) + \mathbb{E}[\sigma''(\eta)\xi^2(\frac{1}{\sqrt{d}} \|\mathbf{w}\| - 1)^2], \end{aligned}$$

for some η between ξ and $\frac{1}{\sqrt{d}} \|\mathbf{w}\|\xi$. Notice that $\frac{1}{\sqrt{d}} \|\mathbf{w}\| \stackrel{\mathcal{D}}{\sim} \frac{1}{\sqrt{d}} \chi_d$ where χ_d is chi distribution with degree d . Thus,

$$\mathbb{E}[\frac{1}{\sqrt{d}} \|\mathbf{w}\|] = \frac{\sqrt{2}\Gamma(\frac{1}{2}(d+1))}{\sqrt{d}\Gamma(\frac{1}{2}d)} = \sqrt{1 - \frac{1}{2d}} + o(d^{-1}) = 1 - \frac{1}{4d} + o(d^{-1}). \quad (\text{C.28})$$

Then, we can get

$$|C_d| \lesssim \frac{1}{d} + \mathbb{E}[(\frac{1}{\sqrt{d}} \|\mathbf{w}\| - 1)^2] \lesssim \frac{1}{d},$$

where we apply the Assumption 1 and the fact $\mathbb{E}[(\frac{1}{\sqrt{d}} \|\mathbf{w}\| - 1)^2] = 2 - 2\mathbb{E}[\frac{1}{\sqrt{d}} \|\mathbf{w}\|] = O(\frac{1}{d})$. ■

Lemma 16 *Let $\mathbf{z}, \mathbf{w} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Under the Assumption 1, we know that*

$$|\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}} \mathbf{w}^\top \mathbf{z})] - b_\sigma| \lesssim 1/d, \quad |\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}} \mathbf{w}^\top \mathbf{z})] - c_\sigma| \lesssim 1/\sqrt{d}. \quad (\text{C.29})$$

Proof. Denote $\xi \sim \mathcal{N}(0, 1)$ which is independent with $\mathbf{w}$. 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 15, we can apply a Taylor approximation of σ' at ξ :

$$|\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\mathbf{w}^\top \mathbf{z})] - b_\sigma| = |\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\|\mathbf{w}\|\xi)] - \mathbb{E}[\sigma'(\xi)]| \quad (\text{C.30})$$

$$\leq |\mathbb{E}[\sigma''(\xi)\xi]| \cdot |\frac{1}{\sqrt{d}}\mathbb{E}[\|\mathbf{w}\|] - 1| + |\mathbb{E}[\sigma'''(\eta)\xi^2]| \mathbb{E}[(\frac{1}{\sqrt{d}}\|\mathbf{w}\| - 1)^2] \quad (\text{C.31})$$

$$\lesssim |\frac{1}{\sqrt{d}}\mathbb{E}[\chi_d] - 1| + \mathbb{E}[(\frac{1}{\sqrt{d}}\chi_d - 1)^2] \lesssim \frac{1}{d}, \quad (\text{C.32})$$

for some η between ξ and $\frac{1}{\sqrt{d}}\|\mathbf{w}\|\xi$. Similar proof can be applied for σ'' to complete the proof. ■

Lemma 17 Consider $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_N)$. Then we have $\|\mathbf{g}\|_\infty \prec \sqrt{2 \log N}$.

Proof. Apply Proposition 2.7.6 of Vershynin (2018) and use the fact that $\mathbb{E}[\|\mathbf{g}\|_\infty] = \mathbb{E}[\max_{i \in [N]} |g_i|]$. ■

Lemma 18 Consider a matrix $\mathbf{M} \in \mathbb{R}^{N \times n}$ and two vectors $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^n$. We have the following identity

$$\mathbf{M} \odot (\mathbf{a}\mathbf{b}^\top) = \text{diag}(\mathbf{a})\mathbf{M} \text{diag}(\mathbf{b}).$$

This is a classical linear algebra result, so we will omit the proof here.

Lemma 19 Assume σ is globally L -Lipschitz for a constant $L > 0$. Then there exist absolute constants $C, c > 0$ such that for every $t > 0$, with probability at least $1 - 6e^{-ct^2}$,

$$\left\| \frac{1}{\sqrt{N}} \left(\sigma(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma(\mathbf{W}\mathbf{Z})] \right) \right\| \leq CL \left(1 + \sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}} \right) \left(1 + \sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}} \right).$$

Proof. Let $\mathbf{z}_j \in \mathbb{R}^d$ denote the j th column of $\mathbf{Z}$. By exchangeability, $\mathbb{E}\mathbf{Y}_0 = \frac{m}{\sqrt{N}} \mathbf{1}_N \mathbf{1}_n^\top$ where $\mathbf{Y}_0$ is defined in (C.27) and $m := \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d/d})} [\sigma(\mathbf{w}^\top \mathbf{z})]$. Conditioned on $\mathbf{Z}$, define $\boldsymbol{\mu}(\mathbf{Z}) \in \mathbb{R}^n$ whose entries are

$$\mu_j(\mathbf{Z}) := \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} [\sigma(\mathbf{w}^\top \mathbf{z}_j)], \quad j = 1, \dots, n.$$

Then each row of $\mathbf{Y}_0$ has conditional mean $\boldsymbol{\mu}(\mathbf{Z})^\top$, hence $\mathbb{E}_{\mathbf{W}}[\mathbf{Y}_0 | \mathbf{Z}] = \frac{1}{\sqrt{N}} \mathbf{1}_N \boldsymbol{\mu}(\mathbf{Z})^\top$. Write

$$\|\mathbf{Y}_0 - \mathbb{E}\mathbf{Y}_0\| = \underbrace{\|(\mathbf{Y}_0 - \mathbb{E}_{\mathbf{W}}[\mathbf{Y}_0 | \mathbf{Z}])\|}_{T_1} + \underbrace{\|(\mathbb{E}_{\mathbf{W}}[\mathbf{Y}_0 | \mathbf{Z}] - \mathbb{E}\mathbf{Y}_0)\|}_{T_2}. \quad (\text{C.33})$$

We then control T_1 and T_2 separately.

Bound T_2 . We have $T_2 = \|\boldsymbol{\mu}(\mathbf{Z}) - m\mathbf{1}_n\|$. Define $\mu(\mathbf{z}) := \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{z})]$. For any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$,

$$|\mu(\mathbf{z}) - \mu(\mathbf{z}')| \leq \mathbb{E}_{\mathbf{w}} |\sigma(\mathbf{w}^\top \mathbf{z}) - \sigma(\mathbf{w}^\top \mathbf{z}')| \leq L \mathbb{E}_{\mathbf{w}} |\mathbf{w}^\top (\mathbf{z} - \mathbf{z}')| = L \sqrt{\frac{2}{\pi}} \|\mathbf{z} - \mathbf{z}'\|_2.$$

Thus $\mu(\cdot)$ is $(L\sqrt{2/\pi})$ -Lipschitz. Write $\mathbf{z}_j = \mathbf{g}_j/\sqrt{d}$ with $\mathbf{g}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then $f(\mathbf{g}) := \mu(\mathbf{g}/\sqrt{d})$ is Lipschitz with constant $O(L/\sqrt{d})$. By Gaussian concentration for Lipschitz functions (e.g., Chapter 5 of (Vershynin, 2018)), each coordinate $\mu(\mathbf{z}_j) - \mathbb{E}\mu(\mathbf{z}_j) = \mu(\mathbf{z}_j) - m$ is subgaussian with subgaussian norm

$$\|\mu(\mathbf{z}_j) - m\|_{\psi_2} \leq C \frac{L}{\sqrt{d}}.$$

Since $(\mathbf{z}_j)$ are independent across j , the coordinates of $\boldsymbol{\mu}(\mathbf{Z}) - m\mathbf{1}_n$ are independent subgaussian. A standard concentration bound for the Euclidean norm of an i.i.d. subgaussian vector (e.g., Section 3.1 of (Vershynin, 2018)) yields: for every $t \geq 0$, with probability at least $1 - 2e^{-ct^2}$,

$$T_2 = \|\boldsymbol{\mu}(\mathbf{Z}) - m\mathbf{1}_n\|_2 \leq C \frac{L}{\sqrt{d}}(\sqrt{n} + t) = CL \left(\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}} \right). \quad (\text{C.34})$$

Bound T_1 . Let $\mathbf{A} := \sigma(\mathbf{W}\mathbf{Z}) - \mathbf{1}_N \boldsymbol{\mu}(\mathbf{Z})^\top$. Then $T_1 = \|\mathbf{A}/\sqrt{N}\|$. Conditional on $\mathbf{Z}$, the rows $\mathbf{A}_i \in \mathbb{R}^n$ are i.i.d. (over rows $\mathbf{w}_i$ of $\mathbf{W}$) and mean-zero. Fix any $\mathbf{x} \in \mathbb{S}^{n-1}$ and define the scalar function of $\mathbf{w} \in \mathbb{R}^d$ $F_{\mathbf{x}}(\mathbf{w}) := \mathbf{x}^\top \sigma(\mathbf{w}^\top \mathbf{Z})$. Then $\mathbf{A}_i^\top \mathbf{x} = F_{\mathbf{x}}(\mathbf{w}_i) - \mathbb{E}[F_{\mathbf{x}}(\mathbf{w}_i) \mid \mathbf{Z}]$. Moreover, $F_{\mathbf{x}}$ is Lipschitz with constant at most $L\|\mathbf{Z}\|$:

$$\begin{aligned} |F_{\mathbf{x}}(\mathbf{w}) - F_{\mathbf{x}}(\mathbf{w}')| &\leq \sum_{j=1}^n |x_j| |\sigma(\mathbf{w}^\top \mathbf{z}_j) - \sigma(\mathbf{w}'^\top \mathbf{z}_j)| \leq L \sum_{j=1}^n |x_j| |(\mathbf{w} - \mathbf{w}')^\top \mathbf{z}_j| \\ &\leq L\|\mathbf{x}\|_2 \cdot \|\mathbf{Z}^\top(\mathbf{w} - \mathbf{w}')\|_2 \leq L\|\mathbf{Z}\| \|\mathbf{w} - \mathbf{w}'\|_2, \end{aligned}$$

for any $\mathbf{w}, \mathbf{w}'$, since $\|\mathbf{x}\|_2 = 1$. By Gaussian concentration for Lipschitz functions, conditional on $\mathbf{Z}$ we have

$$\mathbb{P}\left(|\mathbf{A}_i^\top \mathbf{x}| \geq u \mid \mathbf{Z}\right) \leq 2 \exp\left(-c \frac{u^2}{L^2 \|\mathbf{Z}\|^2}\right) \quad (\text{C.35})$$

for all $u \geq 0$. Now fix $\mathbf{x} \in \mathbb{S}^{n-1}$ and $\mathbf{y} \in \mathbb{S}^{N-1}$. Then $\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^N y_i (\mathbf{A}_i^\top \mathbf{x})$ is a sum of independent centered subgaussian random variables. Using (C.35) and $\|\mathbf{y}\|_2 = 1$, we obtain the conditional tail bound

$$\mathbb{P}\left(|\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle| \geq u \mid \mathbf{Z}\right) \leq 2 \exp\left(-c \frac{u^2}{L^2 \|\mathbf{Z}\|^2}\right). \quad (\text{C.36})$$

Next, take $1/4$ -nets $\mathcal{N}_n \subset \mathbb{S}^{n-1}$ and $\mathcal{N}_N \subset \mathbb{S}^{N-1}$ with $|\mathcal{N}_n| \leq 9^n$ and $|\mathcal{N}_N| \leq 9^N$. We have

$$\|\mathbf{A}\| = \sup_{\mathbf{x} \in \mathbb{S}^{n-1}, \mathbf{y} \in \mathbb{S}^{N-1}} \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle \leq 2 \max_{\mathbf{x} \in \mathcal{N}_n, \mathbf{y} \in \mathcal{N}_N} \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle. \quad (\text{C.37})$$

Combining (C.36)–(C.37) with a union bound yields, for any $u \geq 0$,

$$\mathbb{P}\left(\|\mathbf{A}\| \geq 2u \mid \mathbf{Z}\right) \leq 2|\mathcal{N}_n||\mathcal{N}_N| \exp\left(-c \frac{u^2}{L^2 \|\mathbf{Z}\|^2}\right) \leq 2 \exp\left((n+N) \log 9 - c \frac{u^2}{L^2 \|\mathbf{Z}\|^2}\right).$$

Choosing $u = CL\|\mathbf{Z}\|(\sqrt{n} + \sqrt{N} + t)$ and taking C large enough (absorbing $(n+N) \log 9$) gives: for every $t \geq 0$, with conditional probability at least $1 - 2e^{-ct^2}$,

$$\|\mathbf{A}\| \leq CL\|\mathbf{Z}\|(\sqrt{n} + \sqrt{N} + t), \quad \text{hence} \quad T_1 = \left\| \frac{\mathbf{A}}{\sqrt{N}} \right\| \leq CL\|\mathbf{Z}\| \left(1 + \sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}} \right).$$

Since the bound holds conditionally on $\mathbf{Z}$ with failure probability $\leq 2e^{-ct^2}$, it also holds unconditionally.

Bound $\|\mathbf{Z}\|$. By the norm bound for Gaussian matrices, with probability at least $1 - 2e^{-ct^2}$,

$$\|\mathbf{Z}\| \leq 1 + \sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}.$$

On the intersection of the high-probability events (C.34), the T_1 bound above, and the $\|\mathbf{Z}\|$ bound, we get

$$T_1 \leq CL \left(1 + \sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right) \left(1 + \sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}\right), \quad (\text{C.38})$$

$$T_2 \leq CL \left(\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right) \leq CL \left(1 + \sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right) \left(1 + \sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}\right), \quad (\text{C.39})$$

and then (C.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. $\blacksquare$

C.3. Properties of z-transform and T-transform

Let ν be a compactly supported probability measure on $[0, \infty)$ and $\phi \in (0, \infty)$. For $s \in \mathbb{R} \setminus \{0\}$ such that $1 + \lambda s \neq 0$ for all $\lambda \in \text{supp}(\nu)$,

$$z(s) = -\frac{1}{s} + \phi \int \frac{\lambda}{1 + \lambda s} \nu(d\lambda), \quad z'(s) = \frac{1}{s^2} - \phi \int \frac{\lambda^2}{(1 + \lambda s)^2} \nu(d\lambda), \quad (\text{C.40})$$

are well-defined and finite. Then the alignment map $\varphi(s)$ and $T(s)$ in (3.11) and (3.12) respectively, are also well-defined and finite for such s . $z(s)$ is called the z-transform of the deformed MP law $\rho_\phi^{\text{MP}} \boxtimes \nu$.

Lemma 20 *Let ν be a compactly supported probability measure on $[0, \infty)$ and let $\phi \in (0, \infty)$. Let $\mu = \rho_\phi^{\text{MP}} \boxtimes \nu$ 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 $\mathbb{C} \setminus \text{supp}(\mu)$ and $\mathbb{C} \setminus \text{supp}(\tilde{\mu})$ respectively. Then for all $z \in \mathbb{C} \setminus (\text{supp}(\mu) \cup \{0\})$, $m_{\tilde{\mu}}(z) = \tilde{m}_\mu(z)$ where*

$$\tilde{m}_\mu(z) := \frac{\phi - 1}{z} + \phi m_\mu(z). \quad (\text{C.41})$$

Define the set of singularities $\mathcal{T} := \{0\} \cup \{-1/\lambda : \lambda \in \text{supp}(\nu)\}$. For $s \in \mathbb{C} \setminus \mathcal{T}$, $z(s)$ and $z'(s)$ in (C.40) are well defined. Then, we have

(i) *For any real $\lambda \in \mathbb{R} \setminus \text{supp}(\tilde{\mu})$, the quantity $s := m_{\tilde{\mu}}(\lambda)$ belongs to $\mathbb{R} \setminus \mathcal{T}$ and satisfies*

$$\lambda = z(s), \quad \text{and} \quad z'(s) > 0. \quad (\text{C.42})$$

Equivalently, since $\text{supp}(\mu)$ and $\text{supp}(\tilde{\mu})$ may differ only at $\{0\}$, the same conclusion holds for any real $\lambda \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$ with $s = \tilde{m}_\mu(\lambda)$.

(ii) Conversely, if $s \in \mathbb{R} \setminus \mathcal{T}$ satisfies $z'(s) > 0$, then $\lambda := z(s)$ belongs to $\mathbb{R} \setminus \text{supp}(\tilde{\mu})$ and $m_{\tilde{\mu}}(\lambda) = s$. Moreover, let I be a maximal open interval contained in $\mathbb{R} \setminus \mathcal{T}$ on which $z'(s) > 0$. Then z is strictly increasing on I and maps I bijectively onto a connected component of $\mathbb{R} \setminus \text{supp}(\tilde{\mu})$ (an exterior component or an internal gap). Finally, every nonzero boundary point $\lambda_* \neq 0$ of $\text{supp}(\tilde{\mu})$ can be written as $\lambda_* = z(s_*)$ for some $s_* \in \mathbb{R} \setminus \mathcal{T}$ satisfying $z'(s_*) = 0$.

Proof. This is exactly the classical result of Silverstein–Choi (Theorems. 4.1–4.2 of [Silverstein and Choi \(1995\)](#)). ■

Lemma 21 With the same notation as Lemma 20, recall the T -transform defined by (3.12). For any real $\lambda \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$, let $s = \tilde{m}_\mu(\lambda) = m_{\tilde{\mu}}(\lambda)$. Then $\lambda = z(s)$ and

$$\lambda s m_\mu(\lambda) = T(s). \quad (\text{C.43})$$

Proof. Since $\lambda \neq 0$ and $s = \tilde{m}_\mu(\lambda) = \frac{\phi-1}{\lambda} + \phi m_\mu(\lambda)$, we can get $m_\mu(\lambda) = \frac{1}{\phi} \left(s - \frac{\phi-1}{\lambda} \right)$. Therefore,

$$\lambda s m_\mu(\lambda) = \lambda s \cdot \frac{1}{\phi} \left(s - \frac{\phi-1}{\lambda} \right) = \frac{\lambda s^2 - (\phi-1)s}{\phi}.$$

By Lemma 20(i), $\lambda = z(s)$, so the right-hand side equals $T(s)$ by definition. This proves (C.43). ■

Lemma 22 (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 $\lambda \in \text{supp}(\nu)$, the function T is C^1 and strictly decreasing:

$$T'(s) < 0, \quad s \in I. \quad (\text{C.44})$$

Hence, T is invertible on this interval.

Proof. For $s \in I$, we define

$$I_1(s) := \int \frac{\lambda}{1 + \lambda s} \nu(d\lambda), \quad I_2(s) := \int \frac{\lambda^2}{(1 + \lambda s)^2} \nu(d\lambda).$$

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 (3.12) and (C.40)) we compute

$$T'(s) = -1 + 2s I_1(s) - s^2 I_2(s). \quad (\text{C.45})$$

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 $T'(s) = -1 + 2s \mathbb{E}[F] - s^2 \mathbb{E}[F^2] = -(1 - s \mathbb{E}[F])^2 - s^2 \text{Var}(F)$. If ν is not a single atom, then $\text{Var}(F) > 0$ 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 (C.44). ■

Remark 23 (Relation to the classical D -transform in Benaych-Georges and Nadakuditi (2012)) If one defines $D(\lambda) := \lambda m_\mu(\lambda) \tilde{m}_\mu(\lambda)$, then by Lemma 21, $D(\lambda) = T(s)$ 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 Benaych-Georges and Nadakuditi (2012) to characterize the outlier singular values in additive spiked random rectangular matrices.

Lemma 24 (Closed-form inverse of T for a shifted MP bulk) Fix $\psi \in (0, \infty)$ and let $b_\sigma^2 \in [0, 1]$. Let $m_{\text{MP},\psi}(\cdot)$ denote the Stieltjes transform of ρ_ψ^{MP} . Define the shifted MP law $\nu := (1 - b_\sigma^2) + b_\sigma^2 \rho_\psi^{\text{MP}}$, and for any $s \in \mathbb{C} \setminus \{0\}$ such that $1 + \lambda s \neq 0$ for all $\lambda \in \text{supp}(\nu)$, T transform defined by (3.12) can be written as

$$T(s) = -s + s^2 \int \frac{\lambda}{1 + \lambda s} \nu(d\lambda). \quad (\text{C.46})$$

(i) If $b_\sigma = 0$, then $T(s) = -s/(1 + s)$. Otherwise, let $\xi(s) := -\frac{1+(1-b_\sigma^2)s}{b_\sigma^2 s}$ and we have

$$T(s) = -\frac{1}{b_\sigma^2} m_{\text{MP},\psi}(\xi(s)). \quad (\text{C.47})$$

(ii) Recall the definition of $T^{-1}(t)$ in (3.12) for $t \in \mathbb{C}$ such that the denominator in (3.12) is nonzero. Then $T(T^{-1}(t)) = t$. In particular, on any real monotone branch of T (e.g., $s < 0$), (3.12) is the functional inverse of T .

Proof. Write $\alpha := 1 - b_\sigma^2$ and $\beta := b_\sigma^2$ for short, so that $\nu = \alpha + \beta \rho_\psi^{\text{MP}}$. Let $X \sim \rho_\psi^{\text{MP}}$ and set $\lambda = \alpha + \beta X$. Then $\int \frac{\lambda}{1 + \lambda s} \nu(d\lambda) = \mathbb{E} \left[\frac{\alpha + \beta X}{1 + s(\alpha + \beta X)} \right]$. Now note the following exact algebraic identity:

$$\frac{\alpha + \beta X}{1 + s(\alpha + \beta X)} = \frac{1}{s} \left(1 - \frac{1}{1 + s(\alpha + \beta X)} \right). \quad (\text{C.48})$$

Taking expectations in (C.48) gives

$$\int \frac{\lambda}{1 + \lambda s} \nu(d\lambda) = \frac{1}{s} \left(1 - \mathbb{E} \left[\frac{1}{1 + s(\alpha + \beta X)} \right] \right). \quad (\text{C.49})$$

Insert (C.49) into the definition (C.46) to get

$$T(s) = -s \mathbb{E} \left[\frac{1}{1 + s\alpha + s\beta X} \right]. \quad (\text{C.50})$$

Now rewrite the expectation into Stieltjes transform. Therefore

$$\mathbb{E} \left[\frac{1}{1 + s\alpha + s\beta X} \right] = \frac{1}{s\beta} \mathbb{E} \left[\frac{1}{X - \xi(s)} \right], \quad \xi(s) = -\frac{1 + s\alpha}{s\beta}.$$

By definition of the MP Stieltjes transform, $\mathbb{E} \left[\frac{1}{1 + s\alpha + s\beta X} \right] = \frac{1}{s\beta} m_{\text{MP},\psi}(\xi(s))$. Plug this into (C.50) to obtain (C.47).

Lemma 22 shows T is invertible on certain intervals. Let $t = T(s)$. By (C.47), this is equivalent to $m_{\text{MP},\psi}(\xi(s)) = -\beta t$. Set $m := -\beta t$. For MP, $m = m_{\text{MP},\psi}(\xi)$ satisfies the exact quadratic identity

$$\psi \xi m^2 + (\xi + \psi - 1)m + 1 = 0. \quad (\text{C.51})$$

Solve (C.51) for ξ in terms of m to obtain $\xi = -\frac{1+(\psi-1)m}{m(1+\psi m)}$. Thus, with $m = -\beta t$, we have

$$\xi = -\frac{1 + (\psi - 1)(-\beta t)}{(-\beta t)(1 + \psi(-\beta t))} = \frac{1 - (\psi - 1)\beta t}{\beta t(1 - \psi\beta t)}. \quad (\text{C.52})$$

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 (C.52) 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 (3.12). Let $s = T^{-1}(t)$ given by (3.12). The above computation is reversible: from t we obtain ξ via (C.52), then s via $s = -1/(\alpha + \beta\xi)$, and finally $m_{\text{MP},\psi}(\xi) = -\beta t$ by construction. Plugging into (C.47) gives $T(s) = -(1/\beta)m_{\text{MP},\psi}(\xi) = t$. This shows $T(T^{-1}(t)) = t$. $\blacksquare$

C.4. Centered and Normalized GELU Function

Let $\xi \sim \mathcal{N}(0, 1)$ and denote the standard normal PDF and CDF by

$$\varphi(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad \Phi(x) := \int_{-\infty}^x \varphi(t) dt.$$

The GELU activation function (Hendrycks and Gimpel, 2016) is defined by $g(x) = x\Phi(x)$. Since $\mathbb{E}[\sigma(\xi)] = 0$ and $\mathbb{E}[\sigma(\xi)^2] = 1$, we construct a centered and normalized version of GELU $\sigma(x) := \frac{g(x)-\mu_g}{v_g}$ by $\mu_g := \mathbb{E}[g(\xi)]$ and $v_g^2 := \text{Var}(g(\xi))$.

Assumption 1 is satisfied for $\sigma(\xi)$. A direct calculation gives

$$g'(x) = \Phi(x) + x\varphi(x), \quad g''(x) = (2-x^2)\varphi(x), \quad g^{(3)}(x) = (x^3-4x)\varphi(x).$$

Since $0 \leq \Phi(x) \leq 1$ 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 1.

Compute the centering constant $\mu_g = \mathbb{E}[g(\xi)]$. Using $\varphi'(x) = -x\varphi(x)$, we write $x\varphi(x) dx = -d\varphi(x)$ and integrate by parts to get $\mu_g = \int_{\mathbb{R}} \varphi(x)^2 dx$. Since $\varphi(x)^2 = \frac{1}{2\pi}e^{-x^2}$, we obtain

$$\mu_g = \int_{\mathbb{R}} \frac{1}{2\pi}e^{-x^2} dx = \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}}. \quad (\text{C.53})$$

Compute the variance $v_g^2 = \text{Var}(g(\xi))$. Let $m_{2,g} := \mathbb{E}[g(\xi)^2] = \int_{\mathbb{R}} x^2\Phi(x)^2\varphi(x) dx$. Define the antiderivative $V(x) := \int_{-\infty}^x t^2\varphi(t) dt$. A standard identity gives $V(x) = -x\varphi(x) + \Phi(x)$. Integrating by parts yields

$$\begin{aligned} m_{2,g} &= \left[\Phi(x)^2 V(x) \right]_{-\infty}^{\infty} - \int_{\mathbb{R}} V(x) d(\Phi(x)^2) \\ &= 1 - 2 \int_{\mathbb{R}} V(x) \Phi(x) \varphi(x) dx = 1 - 2 \int_{\mathbb{R}} (-x\varphi(x) + \Phi(x)) \Phi(x) \varphi(x) dx \\ &= 1 + 2 \int_{\mathbb{R}} x\Phi(x)\varphi(x)^2 dx - 2 \int_{\mathbb{R}} \Phi(x)^2 \varphi(x) dx. \end{aligned}$$

The last term equals $\mathbb{E}[\Phi(\xi)^2]$. Since $U := \Phi(\xi) \sim \text{Unif}(0, 1)$, we have

$$\int_{\mathbb{R}} \Phi(x)^2 \varphi(x) dx = \mathbb{E}[U^2] = \frac{1}{3}. \quad (\text{C.54})$$

With integration by parts, we can also get

$$\int_{\mathbb{R}} x\Phi(x)\varphi(x)^2 dx = \frac{1}{2\pi} \int_{\mathbb{R}} x\Phi(x)e^{-x^2} dx = \frac{1}{4\pi\sqrt{3}}. \quad (\text{C.55})$$

Plugging (C.54)–(C.55) into the expression for $m_{2,g}$ gives $m_{2,g} = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}}$. Finally, by (C.53), $v_g^2 = \text{Var}(g(\xi)) = m_{2,g} - \mu_g^2 = \frac{1}{3} + \frac{1}{2\pi\sqrt{3}} - \frac{1}{4\pi}$.

Compute b_σ and c_σ . Since

$$\sigma'(x) = \frac{g'(x)}{v_g} = \frac{\Phi(x) + x\varphi(x)}{v_g}, \quad \sigma''(x) = \frac{g''(x)}{v_g} = \frac{(2 - x^2)\varphi(x)}{v_g}.$$

Hence, using the standard Gaussian integrals,

$$b_\sigma = \mathbb{E}[\sigma'(\xi)] = \frac{1}{v_g} \mathbb{E}[\Phi(\xi) + \xi\varphi(\xi)] = \frac{1}{v_g} (\mathbb{E}[\Phi(\xi)] + \mathbb{E}[\xi\varphi(\xi)]) = \frac{1}{v_g} \cdot \frac{1}{2} = \frac{1}{2v_g}, \quad (\text{C.56})$$

$$c_\sigma = \mathbb{E}[\sigma''(\xi)] = \frac{1}{v_g} \mathbb{E}[(2 - \xi^2)\varphi(\xi)] = \frac{1}{v_g} \int_{\mathbb{R}} (2 - x^2)\varphi(x)^2 dx = \frac{3}{4\sqrt{\pi}v_g}. \quad (\text{C.57})$$

Numerically, $b_\sigma \approx 0.85$ and $c_\sigma \approx 0.72$. Thus GELU does not have uninformative spikes in Theorem 3(ii) since $\tau \leq \tau_{\text{crit}}$ and $\sqrt{\psi}(\frac{c_\sigma^2}{2} - b_\sigma^2) \leq b_\sigma^2$.

Appendix D. Quadratic Equivalence for Proportional Limit

Let us recall the null model $\mathbf{Y}_0$ in (C.27) and QE model $\mathbf{Y}_{\text{QE}}$ of the random feature matrix on XOR dataset defined in (A.1). Below we control the difference between $\mathbf{Y}_{\text{QE}}$ and $\mathbf{Y}$.

Proof of Proposition 10. In this proof, we aim to prove a more precise statement:

$$\|\mathbf{Y} - \mathbf{Y}_{\text{QE}}\| \prec n^{-1/4}. \quad (\text{D.1})$$

Recall that $\mathbf{Y} = \frac{1}{\sqrt{N}}\sigma(\mathbf{W}\mathbf{Z} + \mathbf{W}\mathbf{M})$ where $\mathbf{Z}$ and $\mathbf{M}$ are defined by (3.5). Then, we can take Taylor approximations for each entry of $\mathbf{Y}$ around $\mathbf{Y}_0$ to obtain

$$\mathbf{Y} = \mathbf{Y}_0 + \frac{1}{\sqrt{N}}\sigma'(\mathbf{W}\mathbf{Z}) \odot (\mathbf{W}\mathbf{M}) + \frac{1}{2\sqrt{N}}\sigma''(\mathbf{W}\mathbf{Z}) \odot (\mathbf{W}\mathbf{M})^{\odot 2} + \frac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\mathbf{W}\mathbf{M})^{\odot 3} \quad (\text{D.2})$$

for some random matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N \times n}$ such that $\Theta_{i,j}$ is between $[\mathbf{W}\mathbf{Z}]_{i,j}$ and $[\mathbf{W}\mathbf{X}]_{i,j}$ for all $i \in [N]$ and $j \in [n]$. Notice that

$$\mathbf{W}\mathbf{M} = \theta_{\text{SNR}}(\mathbf{g}_1\mathbf{v}_1^\top + \mathbf{g}_2\mathbf{v}_2^\top), \quad (\text{D.3})$$

and $\mathbf{g}_1, \mathbf{g}_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_N)$. By the definitions of $\mathbf{v}_1$ and $\mathbf{v}_2 \in \mathbb{R}^n$, we have

$$(\mathbf{W}\mathbf{M})^{\odot 2} = \theta_{\text{SNR}}^2(\mathbf{g}_1^{\odot 2}\mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2}\mathbf{v}_2^{\odot 2\top} + (\mathbf{g}_1 \odot \mathbf{g}_2)(\mathbf{v}_1 \odot \mathbf{v}_2)^\top) = \theta_{\text{SNR}}^2(\mathbf{g}_1^{\odot 2}\mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2}\mathbf{v}_2^{\odot 2\top}). \quad (\text{D.4})$$

Thus, from (D.2), we have that

$$\mathbf{Y} = \mathbf{Y}_0 + \frac{\mathbb{E}[\sigma'(\frac{1}{\sqrt{d}}\mathbf{w}^\top \mathbf{z})]\theta_{\text{SNR}}}{\sqrt{N}}(\mathbf{g}_1\mathbf{v}_1^\top + \mathbf{g}_2\mathbf{v}_2^\top) + \frac{\mathbb{E}[\sigma''(\frac{1}{\sqrt{d}}\mathbf{w}^\top \mathbf{z})]\theta_{\text{SNR}}^2}{2\sqrt{N}}(\mathbf{g}_1^{\odot 2}\mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2}\mathbf{v}_2^{\odot 2\top}) \quad (\text{D.5})$$

$$+ \frac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\mathbf{W}\mathbf{M})^{\odot 3} + \frac{1}{\sqrt{N}}(\sigma'(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma'(\mathbf{W}\mathbf{Z})]) \odot (\mathbf{W}\mathbf{M}) \quad (\text{D.6})$$

$$+ \frac{1}{2\sqrt{N}}(\sigma''(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma''(\mathbf{W}\mathbf{Z})]) \odot (\mathbf{W}\mathbf{M})^{\odot 2}, \quad (\text{D.7})$$

where $\mathbf{z}, \mathbf{w} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. Next, we address the terms on the right-hand side of (D.7) individually.

(a) Since $\|\sigma'''\|_\infty = O(1)$, we have that

$$\left\| \frac{1}{3!\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \odot (\mathbf{W}\mathbf{M})^{\odot 3} \right\| \lesssim \theta^3 \left\| \frac{1}{\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \right\| \sum_{i,j,k=1}^2 \|\mathbf{g}_i\|_\infty \|\mathbf{g}_j\|_\infty \|\mathbf{g}_k\|_\infty \|\mathbf{v}_i\|_\infty \|\mathbf{v}_j\|_\infty \|\mathbf{v}_k\|_\infty \quad (\text{D.8})$$

$$\lesssim n^{-3/4} \left\| \frac{1}{\sqrt{N}}\sigma'''(\boldsymbol{\Theta}) \right\|_F \|\mathbf{g}_1\|_\infty^3 \prec \frac{1}{n^{1/4}}, \quad (\text{D.9})$$

where we apply Lemmas 17 and 18.

(b) Next, we have that

$$\left\| \frac{1}{\sqrt{N}}(\sigma'(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma'(\mathbf{W}\mathbf{Z})]) \odot (\mathbf{W}\mathbf{M}) \right\| \quad (\text{D.10})$$

$$\leq \left\| \frac{1}{\sqrt{N}}(\sigma'(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma'(\mathbf{W}\mathbf{Z})]) \right\| \cdot \theta_{\text{SNR}} \cdot (\|\mathbf{g}_1\|_\infty \|\mathbf{v}_1\|_\infty + \|\mathbf{g}_2\|_\infty \|\mathbf{v}_2\|_\infty) \quad (\text{D.11})$$

$$\lesssim \frac{\sqrt{\log n}}{n^{1/4}} \left\| \frac{1}{\sqrt{N}}(\sigma'(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma'(\mathbf{W}\mathbf{Z})]) \right\| \prec \frac{1}{n^{1/4}} \quad (\text{D.12})$$

where we utilize Lemma 19 in the last line since σ' is Lipschitz.

(c) Analogously, we have that

$$\frac{1}{2\sqrt{N}}\|(\sigma''(\mathbf{W}\mathbf{Z}) - \mathbb{E}[\sigma''(\mathbf{W}\mathbf{Z})]) \odot (\mathbf{W}\mathbf{M})^{\odot 2}\| \prec \frac{1}{\sqrt{n}}.$$

(d) Lastly, we apply Lemma 16 to obtain that

$$\mathbf{Y} = \mathbf{Y}_0 + \frac{\theta_{\text{SNR}}b_\sigma}{\sqrt{N}}(\mathbf{g}_1\mathbf{v}_1^\top + \mathbf{g}_2\mathbf{v}_2^\top) + \frac{\theta_{\text{SNR}}^2c_\sigma}{2\sqrt{N}}(\mathbf{g}_1^{\odot 2}\mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2}\mathbf{v}_2^{\odot 2\top}) + O_{\prec}(n^{-1/4}). \quad (\text{D.13})$$

Combining all these together, we proved (D.1). Then, if $\|\mathbf{Y}\|, \|\mathbf{Y}_{\text{QE}}\| \lesssim 1$, we can conclude that

$$\|\mathbf{K} - \mathbf{K}_{\text{QE}}\| \prec n^{-1/4}$$

and they have asymptotically the same outlier eigenvalues and eigenvector alignments. $\blacksquare$

Appendix E. Proof for Finite-SNR and Proportional Limit Regime

In this section, we present the proof for Theorem 3 when we consider proportional limit and finite SNR regime. In this case, the rank-two XOR signal satisfies $M = \theta_{\text{SNR}} (\mathbf{u}_1 \mathbf{v}_1^\top + \mathbf{u}_2 \mathbf{v}_2^\top)$, where $\theta_{\text{SNR}} := r \sqrt{\frac{n}{2d}} \rightarrow r \sqrt{\frac{\psi}{2}}$. We know that both $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ are deterministic orthonormal vectors satisfying

$$\mathbf{v}_1^\top \mathbf{v}_2 = 0, \quad \|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1, \quad \mathbf{u}^\top \mathbf{v}_1 = \mathbf{u}^\top \mathbf{v}_2 = 0, \quad \mathbf{u} := \frac{\mathbf{1}_n}{\sqrt{n}}. \quad (\text{E.1})$$

Recall that $\mathbf{Z} \in \mathbb{R}^{d \times n}$ has i.i.d. $\mathcal{N}(0, 1/d)$ entries, and $\mathbf{W} \in \mathbb{R}^{N \times d}$ has i.i.d. $\mathcal{N}(0, 1)$ entries, independent of $\mathbf{Z}$. Recall the null feature matrix $\mathbf{Y}_0$ in (C.27) and its columnwise conditional mean $\mathbf{m} := \mathbb{E}[\sigma(\mathbf{w}^\top \mathbf{Z}) \mid \mathbf{Z}] \in \mathbb{R}^n$, whose entries are

$$m_j = F(\|\mathbf{z}_j\|^2), \quad F(s) := \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma(\sqrt{s} \xi)], \quad j \in [n]. \quad (\text{E.2})$$

Here $\mathbf{z}_j$ denotes the j -th column of $\mathbf{Z}$, and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ independent of $\mathbf{Z}$. Notice that conditioning on $\mathbf{Z}$, $\mathbf{Y}_0$ has independent and identically distributed rows. However, when $c_\sigma \neq 0$, the mean of rows in $\mathbf{Y}_0$, $\mathbf{m}$, is not vanishing which will potentially give us some uninformative spike, see (Benigni and P      , 2022; Wang et al., 2024). Hence, we first need to subtract the mean vector $\mathbf{m}$ to make all rows centered and to apply theorems in Wang et al. (2024). Then we consider the effect of the mean vector $\mathbf{m}$. Define the centered null feature matrix by

$$\mathbf{G} := \mathbf{Y}_0 - \frac{1}{\sqrt{N}} \mathbf{1}_N \mathbf{m}^\top \quad (\text{E.3})$$

and the null model of CK matrix by $\mathbf{K}_0 := \mathbf{G}^\top \mathbf{G}$.

E.1. Analysis of the Population Covariance Matrix

Conditional on $\mathbf{Z}$, the rows of $\mathbf{G}$ are i.i.d. centered vectors in $\mathbb{R}^n$. Let $\bar{\mathbf{g}}^\top := \sigma(\mathbf{w}^\top \mathbf{Z}) - \mathbf{m}^\top$ for $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ independent of $\mathbf{Z}$, and define the conditional population covariance

$$\Sigma := \mathbb{E}[\bar{\mathbf{g}} \bar{\mathbf{g}}^\top \mid \mathbf{Z}] \in \mathbb{R}^{n \times n}. \quad (\text{E.4})$$

Lemma 25 (Approximation of Σ) *Let $\mathbf{S} := \mathbf{Z}^\top \mathbf{Z}$. Under Assumptions 1 and 2, we have*

$$\Sigma = (1 - b_\sigma^2) \mathbf{I}_n + b_\sigma^2 \mathbf{S} + \frac{c_\sigma^2}{2} (\mathbf{S}^{\odot 2} - \mathbf{I}_n) + \mathbf{R}_\Sigma, \quad (\text{E.5})$$

such that $\|\mathbf{R}_\Sigma\| = \tilde{O}_{\mathbb{P}}(n^{-1/2})$.

Proof. With Lemma 5.2 of Wang and Zhu (2024) and Lemma 13, we have

$$\Sigma = (1 - b_\sigma^2) \mathbf{I}_n + b_\sigma^2 \mathbf{S} + \frac{c_\sigma^2}{2} (\mathbf{S}^{\odot 2} - \mathbf{I}_n) + d_\sigma^2 (\mathbf{S}^{\odot 3} - \mathbf{I}_n) + O_{\mathbb{P}}(n^{-1/2}),$$

where d_σ is the third-order Hermite coefficient of σ . Note that under event of Lemma 13,

$$\begin{aligned} \|(\mathbf{Z}^\top \mathbf{Z})^{\odot 3} - \mathbf{I}_n\| &\leq \|\text{diag}((\mathbf{Z}^\top \mathbf{Z})^{\odot 3} - \mathbf{I}_n)\| + \|\text{offdiag}(\mathbf{Z}^\top \mathbf{Z})^{\odot 3}\|_F \\ &\leq \max_{\alpha \in [n]} |\|\mathbf{z}_\alpha\|^6 - 1| + n \cdot \max_{\alpha \neq \beta \in [n]} |\mathbf{z}_\alpha^\top \mathbf{z}_\beta|^3 \leq C(\tau_n + n\tau_n^3) \leq 2C\sqrt{(\log^3 n)/n} \end{aligned} \quad (\text{E.6})$$

for some constant $C \geq 0$. This completes the proof. $\blacksquare$

Lemma 26 Let $S = Z^\top Z$ and $\mathbf{u} = \mathbf{1}_n/\sqrt{n}$. Then $S^{\odot 2} - \mathbf{I}_n = \frac{n}{d} \mathbf{u} \mathbf{u}^\top + \mathbf{E}_2$ with $\|\mathbf{E}_2\| = o_{\mathbb{P}}(1)$.

Proof. The proof is directly derived from the proof of Theorem 2.1 of [El Karoui \(2010\)](#), where $S^{\odot 2}$ is the quadratic term in the Taylor expansion of the inner-product kernel of Z . $\blacksquare$

Corollary 27 Define $\Sigma_0 := (1 - b_\sigma^2) \mathbf{I}_n + b_\sigma^2 S$, and $\mathbf{u} := \frac{\mathbf{1}_n}{\sqrt{n}}$. Let $\tau := \frac{c_\sigma^2}{2} \frac{n}{d} \rightarrow \frac{c_\sigma^2}{2} \psi$. Then

$$\Sigma = \Sigma_0 + \tau \mathbf{u} \mathbf{u}^\top + \mathbf{R}, \quad \|\mathbf{R}\| = o_{\mathbb{P}}(1). \quad (\text{E.7})$$

Moreover the limiting ESD of Σ_0 (and of Σ) equals

$$\nu = (1 - b_\sigma^2) + b_\sigma^2 \rho_\psi^{\text{MP}}. \quad (\text{E.8})$$

Thus, the right edge of ν is $\lambda_+^\nu = (1 - b_\sigma^2) + b_\sigma^2(1 + \sqrt{\psi})^2$.

Proof. Combine Lemmas 25 and 26 and absorb all $o_{\mathbb{P}}(1)$ terms into $\mathbf{R}$. The rank-one term $\tau \mathbf{u} \mathbf{u}^\top$ does not change the limiting ESD. The limiting ESD of $\Sigma_0 = (1 - b_\sigma^2) \mathbf{I} + b_\sigma^2 Z^\top Z$ is the shifted MP law ν in (E.8). $\blacksquare$

E.2. Analysis of Spike in Population Covariance

Lemma 28 (Population outlier from $\Sigma_0 + \tau \mathbf{u} \mathbf{u}^\top$) Let $\mathbf{A} := \Sigma_0$ and consider $\mathbf{A} + \tau \mathbf{u} \mathbf{u}^\top$. For $\tau \geq 0$, any deterministic limit point $\Lambda_\tau > \lambda_+^\nu$ of an isolated eigenvalue of $\mathbf{A} + \tau \mathbf{u} \mathbf{u}^\top$ must satisfy

$$1 + \tau m_\nu(\Lambda_\tau) = 0, \quad (\text{E.9})$$

where m_ν is the Stieltjes transform of ν . Moreover, such a solution exists iff

$$\tau > \tau_{\text{crit}} := b_\sigma^2 \sqrt{\psi} (1 + \sqrt{\psi}). \quad (\text{E.10})$$

Proof. For any $\lambda \notin \text{spec}(\mathbf{A})$, the rank-one perturbation identity gives

$$\lambda \in \text{spec}(\mathbf{A} + \tau \mathbf{u} \mathbf{u}^\top) \setminus \text{spec}(\mathbf{A}) \iff 1 + \tau \mathbf{u}^\top (\mathbf{A} - \lambda \mathbf{I})^{-1} \mathbf{u} = 0. \quad (\text{E.11})$$

Write $\mathbf{A} = (1 - b_\sigma^2) \mathbf{I} + b_\sigma^2 S$ with $S = Z^\top Z$. 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 $(\mathbf{A} - \lambda \mathbf{I})^{-1} = \frac{1}{b_\sigma^2} (S - \zeta \mathbf{I})^{-1}$. The isotropic MP local law ([Bloemendal et al., 2014](#)) yields, uniformly for such ζ , $\mathbf{u}^\top (S - \zeta \mathbf{I})^{-1} \mathbf{u} = m_{\text{MP}, \psi}(\zeta) + o_{\mathbb{P}}(1)$, hence

$$\mathbf{u}^\top (\mathbf{A} - \lambda \mathbf{I})^{-1} \mathbf{u} = \frac{1}{b_\sigma^2} m_{\text{MP}, \psi}(\zeta) + o_{\mathbb{P}}(1) = m_\nu(\lambda) + o_{\mathbb{P}}(1). \quad (\text{E.12})$$

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 $\lim_{\lambda \rightarrow \infty} F(\lambda) = 1$. 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 (E.10). $\blacksquare$

Lemma 29 (Closed form for Λ_τ when ν is shifted MP) *Let $\nu = (1 - b_\sigma^2) + b_\sigma^2 \rho_\psi^{\text{MP}}$ and assume $\tau > \tau_{\text{crit}}$. Then the unique solution $\Lambda_\tau > \lambda_+^\nu$ of $1 + \tau m_\nu(\Lambda_\tau) = 0$ equals*

$$\Lambda_\tau = (1 - b_\sigma^2) + \tau + \frac{b_\sigma^2 \tau}{\tau - b_\sigma^2 \psi}. \quad (\text{E.13})$$

Proof. Write $\Lambda = (1 - b_\sigma^2) + b_\sigma^2 \zeta$. Then $m_\nu(\Lambda) = \frac{1}{b_\sigma^2} m_{\text{MP},\psi}(\zeta)$, and

$$m_{\text{MP},\psi}(\zeta) = -\frac{b_\sigma^2}{\tau}. \quad (\text{E.14})$$

The Stieltjes transform $m = m_{\text{MP},\psi}(\zeta)$ satisfies the quadratic equation

$$\psi \zeta m^2 + (\zeta + \psi - 1)m + 1 = 0. \quad (\text{E.15})$$

Solving (E.15) for ζ in terms of m yields

$$\zeta = -\frac{1 + (\psi - 1)m}{m(1 + \psi m)} = -\frac{1}{m} + \frac{1}{1 + \psi m}. \quad (\text{E.16})$$

Substitute $m = -b_\sigma^2/\tau$ from (E.14) into (E.16). Therefore, we obtain (E.13). Uniqueness follows from strict monotonicity of $m_\nu(\lambda)$ on (λ_+^ν, ∞) . $\blacksquare$

Corollary 30 (Subcritical case: no population outlier) *If $\tau \leq \tau_{\text{crit}}$, then Σ has no isolated eigenvalue outside $\text{supp}(\nu)$ and $\lambda_{\text{max}}(\Sigma) = \lambda_+^\nu + o_{\mathbb{P}}(1)$.*

Proof. If $\tau \leq \tau_{\text{crit}}$, then Lemma 28 shows the limiting secular equation has no root in (λ_+^ν, ∞) . Using (E.11)–(E.12) and Bai and Silverstein (1998), no separated eigenvalue exists and $\lambda_{\text{max}}(\Sigma) = \lambda_+^\nu + o_{\mathbb{P}}(1)$. $\blacksquare$

Lemma 31 *Recall the shifted MP law $\nu = (1 - b_\sigma^2) + b_\sigma^2 \rho_\psi^{\text{MP}}$ and define interval $I_{\text{un}} := \left(-\frac{1}{\lambda_+^\nu}, 0\right)$ where $\lambda_+^\nu = (1 - b_\sigma^2) + b_\sigma^2(1 + \sqrt{\psi})^2$. Recall τ and τ_{crit} defined in (4.1), and Λ_τ in (E.13). Then the following hold.*

(i) *For every $\Lambda > \lambda_+^\nu$,*

$$T\left(-\frac{1}{\Lambda}\right) = -m_\nu(\Lambda). \quad (\text{E.17})$$

(ii) *The map T is strictly decreasing on I_{un} and maps I_{un} bijectively onto $(0, 1/\tau_{\text{crit}})$.*

(iii) *For $\tau > \tau_{\text{crit}}$ there exists a unique $s_{\text{un}} \in I_{\text{un}}$ such that $T(s_{\text{un}}) = \frac{1}{\tau}$. Moreover, $s_{\text{un}} = -\frac{1}{\Lambda_\tau}$. Equivalently, on the uninformative branch one has the exact identity*

$$s_{\text{mean}} := T^{-1}(1/\tau) = s_{\text{un}} = s_{\text{cov}} := -1/\Lambda_\tau. \quad (\text{E.18})$$

(iv) *If $s_{\text{un}} \in I_{\text{un}}$ and $z'(s_{\text{un}}) > 0$, then necessarily $\tau > \tau_{\text{crit}}$.*

Proof. For $s \in I_{\text{un}}$ write $\Lambda := -1/s$. Then $\Lambda > \lambda_+^\nu$ and

$$T(s) = -s + s^2 \int \frac{\lambda}{1 + \lambda s} \nu(d\lambda) = -s \int \frac{1}{1 + \lambda s} \nu(d\lambda)$$

is well-defined. Substituting $s = -1/\Lambda$ gives

$$T\left(-\frac{1}{\Lambda}\right) = \frac{1}{\Lambda} \int \frac{1}{1 - \lambda/\Lambda} \nu(d\lambda) = \int \frac{1}{\Lambda - \lambda} \nu(d\lambda) = -m_\nu(\Lambda),$$

which proves (E.17). Since $m_\nu(\Lambda)$ is strictly increasing and negative on (λ_+^ν, ∞) , the map $-m_\nu(\Lambda)$ is strictly decreasing in Λ ; equivalently, T is strictly decreasing in $s \in I_{\text{un}}$. Moreover,

$$\lim_{\Lambda \rightarrow \infty} -m_\nu(\Lambda) = 0, \quad \lim_{\Lambda \downarrow \lambda_+^\nu} -m_\nu(\Lambda) = -m_\nu(\lambda_+^\nu) = \frac{1}{\tau_{\text{crit}}},$$

where the last identity is exactly the edge computation in Lemma 28. This proves item (ii).

Item (iii) now follows immediately: $T(s) = 1/\tau$ has a unique solution $s \in I_{\text{un}}$ iff $1/\tau \in (0, 1/\tau_{\text{crit}})$, i.e. iff $\tau > \tau_{\text{crit}}$. For such τ , let s_{un} be the unique solution. Setting $\Lambda_\tau := -1/s_{\text{un}}$ and using (E.17), $\frac{1}{\tau} = T(s_{\text{un}}) = -m_\nu(\Lambda_\tau)$, so $1 + \tau m_\nu(\Lambda_\tau) = 0$. By Lemma 29, the unique solution of this secular equation above λ_+^ν is precisely Λ_τ from (E.13), hence $s_{\text{un}} = -1/\Lambda_\tau$. Finally, item (iv) is immediate from item (iii): existence of $s_{\text{un}} \in I_{\text{un}}$ already forces $\tau > \tau_{\text{crit}}$. $\blacksquare$

E.3. BBP Transition for the Null Model

Let z, φ be defined from (ν, ϕ) by (3.11) where ν is defined by (E.8). In this section, we analyze the uninformative spike of $\mathbf{K}_0$ derived from the potential spike in the population covariance matrix Σ .

Lemma 32 *Let $\mathbf{K}_0 = \mathbf{G}^\top \mathbf{G}$. The following hold.*

- (i) **(Bulk.)** *The ESD of $\mathbf{K}_0$ 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 Σ in (E.13). Then:*

$$\text{If additionally } z'\left(-\frac{1}{\Lambda_\tau}\right) > 0, \quad \mathbf{K}_0 \text{ has a separated outlier at } \hat{\lambda}_\tau = z\left(-\frac{1}{\Lambda_\tau}\right) + o_{\mathbb{P}}(1),$$

where $\hat{\lambda}_\tau$ lies in a connected component of $(0, \infty) \setminus \text{supp}(\mu)$. If $z'(-1/\Lambda_\tau) \leq 0$ or $\tau \leq \tau_{\text{crit}}$, the corresponding eigenvalue sticks to $\text{supp}(\mu)$ and $\mathbf{K}_0$ has no outlier.

- (iii) **(Eigenvector alignment.)** *In the above separated case, if $\hat{\mathbf{v}}_\tau$ is a unit eigenvector of $\mathbf{K}_0$ associated to $\hat{\lambda}_\tau$, then $|\hat{\mathbf{v}}_\tau^\top \mathbf{u}|^2 \xrightarrow{\mathbb{P}} \varphi\left(-\frac{1}{\Lambda_\tau}\right)$. In general, any fixed unit vectors $\mathbf{v} \in \mathbb{R}^n$ (potentially $\mathbf{Z}$ dependent),*

$$|\hat{\mathbf{v}}_\tau^\top \mathbf{v}|^2 - \varphi\left(-\frac{1}{\Lambda_\tau}\right) |\mathbf{u}^\top \mathbf{v}|^2 \xrightarrow{\mathbb{P}} 0. \quad (\text{E.19})$$

Proof. This lemma is an application of Theorem 12 of Wang et al. (2024). It suffices to verify Assumptions 5 and 6 of Wang et al. (2024). Assumption 6 is verified by Lemmas 28 and 29, and Corollary 30: Σ has bulk law ν and at most one population spike Λ_τ (in the $\mathbf{u}$ direction), present iff $\tau > \tau_{\text{crit}}$. Notice that from Corollary 3.5 of Wang and Zhu (2024), we have

$$\bar{\mathbf{g}}^\top \mathbf{A} \bar{\mathbf{g}} - \text{Tr} \mathbf{A} \Sigma \prec \|\mathbf{A}\|_F,$$

for any deterministic matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$. Then, Assumption 5 can be verified based on Lemmas 32, 33, and 34 in Wang et al. (2024). We ignore the details here for simplicity. $\blacksquare$

E.4. BBP Transition for Rank-three Additive Deformation

First, we prove that, for finite SNR $r = \Theta(1)$, the explicit quadratic term in Proposition 10 is negligible.

Lemma 33 (Quadratic spike is negligible) *Assume $r = \Theta(1)$ and the assumptions of Proposition 10. Recall*

$$\mathbf{T}_2 = \frac{\theta_{\text{SNR}}^2 c_\sigma}{\sqrt{N}} (\mathbf{g}_1^{\odot 2} \mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2} \mathbf{v}_2^{\odot 2\top}).$$

Then $\|\mathbf{T}_2\| = O_\prec(n^{-1/2})$. Furthermore, $\|\mathbf{Y}_{\text{QE}}\| \prec 1$ and

$$\|(\mathbf{Y}_{\text{QE}} - \mathbf{T}_2)^\top (\mathbf{Y}_{\text{QE}} - \mathbf{T}_2) - \mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}}\| \prec n^{-1/2}.$$

Proof. We bound

$$\|\mathbf{T}_2\| \leq \frac{|\theta_{\text{SNR}}^2 c_\sigma|}{\sqrt{N}} \sum_{\ell=1}^2 \|\mathbf{g}_\ell^{\odot 2}\| \|\mathbf{v}_\ell^{\odot 2}\|.$$

Since $\mathbf{g}_\ell \sim \mathcal{N}(0, \mathbf{I}_N)$, $\|\mathbf{g}_\ell^{\odot 2}\|^2 = \sum_{i=1}^N g_{\ell,i}^4 = O_\prec(N)$, hence $\|\mathbf{g}_\ell^{\odot 2}\| = O_\prec(\sqrt{N})$. Also $\|\mathbf{v}_\ell^{\odot 2}\|^2 = \sum_{j=1}^n v_{\ell,j}^4 = 2/n$ for XOR sign vectors, so $\|\mathbf{v}_\ell^{\odot 2}\| = \sqrt{2/n}$. Therefore $\|\mathbf{T}_2\| = O_\prec(n^{-1/2})$.

By Lemmas 15 and 19, we have $\sigma(\mathbf{W}\mathbf{Z})/\sqrt{N} \prec 1$. And since $\mathbf{g}_1, \mathbf{g}_2 \prec \sqrt{N}$ and $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$, we have that $\mathbf{Y}_{\text{QE}} \prec 1$ by the definition of $\mathbf{Y}_{\text{QE}}$ in (A.1). The Gram difference bound follows by expansion:

$$(\mathbf{Y}_{\text{QE}} - \mathbf{T}_2)^\top (\mathbf{Y}_{\text{QE}} - \mathbf{T}_2) - \mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}} = -\mathbf{Y}_{\text{QE}}^\top \mathbf{T}_2 - \mathbf{T}_2^\top \mathbf{Y}_{\text{QE}} + \mathbf{T}_2^\top \mathbf{T}_2,$$

hence, applying $\|\mathbf{Y}_{\text{QE}}\| \prec 1$ and $\|\mathbf{T}_2\| = O_\prec(n^{-1/2})$ we can get the final result. $\blacksquare$

In the next proposition, we show that, in addition to the potential spike in $\mathbf{G}$ characterized in Lemma 32, the matrix $\mathbf{Y}$ may exhibit further potential spikes arising from a rank-three additive perturbation of $\mathbf{G}$.

Proposition 34 (Rank-three additive deformation of $\mathbf{G}$) *Assume $r = \Theta(1)$. Define*

$$\mathbf{Y}^\sharp := \underbrace{\mathbf{G} + \frac{1}{\sqrt{N}} \mathbf{1}_N \mathbf{m}^\top}_{\text{mean spike}} + \underbrace{\frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}} (\mathbf{g}_1 \mathbf{v}_1^\top + \mathbf{g}_2 \mathbf{v}_2^\top)}_{\text{two XOR linear spikes}}. \quad (\text{E.20})$$

Denote a new kernel matrix by $\mathbf{K}^\sharp := \mathbf{Y}^{\sharp\top} \mathbf{Y}^\sharp$. Then $\|\mathbf{Y} - \mathbf{Y}^\sharp\| \prec n^{-1/4}$, and hence $\|\mathbf{K} - \mathbf{K}^\sharp\| \prec n^{-1/4}$. In particular, all the isolated eigenvalues and their eigenvectors of $\mathbf{K}$ and $\mathbf{K}^\sharp$ are asymptotically the same.

Proof. By the proof of Proposition 10 and Lemma 33, in finite SNR, we can further drop T_2 to get $\|Y - Y^\sharp\| \prec n^{-1/4}$. From the proof of Lemma 33, we also know that $\|Y\|, \|Y_{QE}\|, \|Y^\sharp\| \prec 1$. Hence we have

$$\|K - K^\sharp\| = \|Y^\top Y - Y^{\sharp\top} Y^\sharp\| \leq \|Y - Y^\sharp\|(\|Y\| + \|Y^\sharp\|) \prec n^{-1/4},$$

which implies asymptotic matching of isolated eigenvalues/eigenspaces by Weyl's inequality and Davis–Kahan Theorem (Davis and Kahan, 1970; Wedin, 1972). $\blacksquare$

E.4.1. ORTHOGONAL DECOMPOSITION OF WEIGHT

Next, to analyze two XOR linear spikes in (E.20), we need to make bulk matrix G independent with the vectors g_1, g_2 by the following decomposition.

Lemma 35 (Orthogonal decomposition and independence) *Let $P_\perp := I_d - u_1 u_1^\top - u_2 u_2^\top$ and define*

$$W_\perp := W P_\perp, \quad g_1 = W u_1, \quad g_2 = W u_2.$$

$$(i) \quad W = W_\perp + g_1 u_1^\top + g_2 u_2^\top.$$

$$(ii) \quad g_1, g_2 \text{ are independent vectors with distribution } \mathcal{N}(0, I_N), \text{ and are independent of } W_\perp.$$

Proof. For each row $w_i^\top$ of W , $w_i = (w_i^\top u_1)u_1 + (w_i^\top u_2)u_2 + P_\perp w_i \in \mathbb{R}^d$. Stacking over i yields $W = W_\perp + g_1 u_1^\top + g_2 u_2^\top$. Since $w_i \sim \mathcal{N}(0, I_d)$, the scalar components $w_i^\top u_1, w_i^\top u_2$, and the vector $P_\perp w_i$ are independent Gaussians because they are orthogonal linear functionals of a Gaussian vector. Stacking over i gives independence of g_1, g_2 and $W_\perp$. $\blacksquare$

Lemma 36 (The null bulk is asymptotically independent of g_1, g_2) *For $\tilde{w} \sim \mathcal{N}(0, P_\perp)$, define*

$$\tilde{Y}_0 := \frac{1}{\sqrt{N}} \sigma(W_\perp Z), \quad \tilde{m} := \mathbb{E}[\sigma(\tilde{w}^\top Z) \mid Z], \quad \tilde{G} := \tilde{Y}_0 - \frac{1}{\sqrt{N}} \mathbf{1}_N \tilde{m}^\top.$$

Then $\tilde{G}$ is independent of (g_1, g_2) and $\|m - \tilde{m}\|, \|G - \tilde{G}\| \prec 1/\sqrt{N}$.

Proof. By Lemma 35, we have the decomposition $WZ = W_\perp Z + g_1(u_1^\top Z) + g_2(u_2^\top Z)$. Let $t_1^\top := u_1^\top Z$ and $t_2^\top := u_2^\top Z$. Each $t_k \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0, 1/d)$ entries and is independent of W . Set $A := W_\perp Z$ and $\Delta := g_1 t_1^\top + g_2 t_2^\top$, so $\sigma(WZ) = \sigma(A + \Delta)$ and $\sigma(W_\perp Z) = \sigma(A)$.

Following the same proof of Proposition 10, we can apply entrywisely 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

$$Y_0 - \tilde{Y}_0 = \frac{1}{\sqrt{N}} (\sigma(A + \Delta) - \sigma(A)) = \tilde{T}_1 + \tilde{T}_2,$$

where $\tilde{T}_1 := \frac{1}{\sqrt{N}} (\sigma'(A) \odot \Delta)$ and $\tilde{T}_2 := \frac{1}{2\sqrt{N}} (\sigma''(\Xi) \odot \Delta^{\odot 2})$. First, we have

$$\|\tilde{T}_1\| \leq \frac{1}{\sqrt{N}} \sum_{i=1}^2 \|\text{diag}(g_i)\| \|\sigma'(A)\| \|\text{diag}(t_i)\| \prec 1/\sqrt{n},$$

where we apply Lemmas 17 and 18, and $\|\sigma'(A)\| = O_{\prec}(\sqrt{N})$ because σ' is bounded and A has i.i.d. bounded sub-Gaussian norm of rows. Similarly we have $\|\tilde{T}_2\| \prec 1/\sqrt{n}$ since $\|\sigma'; (\Xi)\| = O_{\prec}(1)$. A parallel argument applies to the conditional mean vectors, giving $\|\mathbf{m} - \widetilde{\mathbf{m}}\| \prec 1/\sqrt{n}$ since $\|\mathbf{m} - \widetilde{\mathbf{m}}\| \leq \mathbb{E}[\|\mathbf{Y}_0 - \widetilde{\mathbf{Y}}_0\| | \mathbf{Z}]$. Therefore we complete the proof. $\blacksquare$

E.4.2. MEAN SPIKE STRENGTH AND ORTHOGONALITY

Lemma 37 (Mean spike strength and orthogonality) *Under the Assumptions 1 and 2, we have*

$$\|\mathbf{m}\|^2 \xrightarrow{\mathbb{P}} \beta_{\text{mean}} := \frac{c_{\sigma}^2}{2} \psi, \quad (\text{E.21})$$

and

$$\mathbf{u}^{\top} \mathbf{m} = o_{\mathbb{P}}(1), \quad \mathbf{v}_1^{\top} \mathbf{m} = o_{\mathbb{P}}(1), \quad \mathbf{v}_2^{\top} \mathbf{m} = o_{\mathbb{P}}(1). \quad (\text{E.22})$$

Proof. Fix $j \in [n]$ and condition on $\mathbf{z}_j$. Since $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ is independent of $\mathbf{z}_j$, $\mathbf{w}^{\top} \mathbf{z}_j \sim \mathcal{N}(0, \|\mathbf{z}_j\|^2)$. Define for $s > 0$ $F(s) := \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma(\sqrt{s}\xi)]$, so that $m_j = F(\|\mathbf{z}_j\|^2)$. By $\mathbb{E}[\sigma(\xi)] = 0$ we have $F(1) = 0$.

Under Assumption 1, F is C^2 on a neighborhood of 1 and we can differentiate under the expectation

$$F'(s) = \frac{1}{2\sqrt{s}} \mathbb{E}[\sigma'(\sqrt{s}\xi) \xi].$$

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

$$F(s) = \frac{c_{\sigma}}{2}(s-1) + R(s), \quad |R(s)| \leq C(s-1)^2, \quad (\text{E.23})$$

for some constant C depending only on $\sup_{t \in [1-\delta, 1+\delta]} |F''(t)|$.

Then let $\Delta_j := \|\mathbf{z}_j\|^2 - 1$. By (E.2) and (E.23), $m_j = \frac{c_{\sigma}}{2} \Delta_j + R_j$, where $|R_j| \leq C \Delta_j^2$ on the event $\|\mathbf{z}_j\|^2 \in [1-\delta, 1+\delta]$, which holds w.h.p. since $\|\mathbf{z}_j\|^2 = (1/d)\chi_d^2$. Then

$$m_j^2 = \frac{c_{\sigma}^2}{4} \Delta_j^2 + \underbrace{\frac{c_{\sigma}}{1} \Delta_j R_j + R_j^2}_{=: e_j}, \quad |e_j| \leq C(|\Delta_j|^3 + \Delta_j^4).$$

Summing over j gives

$$\|\mathbf{m}\|^2 = \frac{c_{\sigma}^2}{4} \sum_{j=1}^n \Delta_j^2 + \sum_{j=1}^n e_j. \quad (\text{E.24})$$

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 $n/d \rightarrow \psi$, we have

$$\mathbb{E} \left[\sum_{j=1}^n |e_j| \right] \leq Cn \left(\mathbb{E}|\Delta_j|^3 + \mathbb{E}[\Delta_j^4] \right) = O \left(\frac{n}{d^{3/2}} + \frac{n}{d^2} \right) = o(1),$$

hence $\sum_{j=1}^n e_j = o_{\mathbb{P}}(1)$ by Markov inequality. Thus (E.24) reduces to

$$\|\mathbf{m}\|^2 = \frac{c_{\sigma}^2}{4} \sum_{j=1}^n \Delta_j^2 + o_{\mathbb{P}}(1). \quad (\text{E.25})$$

Lastly, we give the concentration of $\sum \Delta_j^2$. Because $\mathbf{z}_j \sim \mathcal{N}(0, \mathbf{I}_d/d)$, we have $\Delta_j = (\chi_{d,j}^2 - d)/d$ with i.i.d. $\chi_{d,j}^2$. Then

$$\mathbb{E}[\Delta_j^2] = \text{Var}(\|\mathbf{z}_j\|^2) = \frac{2}{d}, \quad \text{Var}(\Delta_j^2) = O(d^{-2}).$$

Therefore $\text{Var}\left(\sum_{j=1}^n \Delta_j^2\right) = O\left(\frac{1}{d}\right) \rightarrow 0$, so the Chebyshev inequality implies $\sum_{j=1}^n \Delta_j^2 = \frac{2n}{d} + o_{\mathbb{P}}(1)$. Plugging into (E.25) yields (E.21).

Let $\mathbf{v} \in \mathbb{R}^n$ be any deterministic unit vector (in particular $\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2$). Since m_j are independent and $\text{Var}(m_j) = O(1/d)$ (from $m_j = (c_\sigma/2)\Delta_j + O(\Delta_j^2)$ and $\text{Var}(\Delta_j) = 2/d$), we have $\text{Var}(\mathbf{v}^\top \mathbf{m}) = O(1/d) \rightarrow 0$. Also $\mathbb{E}[m_j] = O(1/d)$ because $\mathbb{E}[\Delta_j] = 0$ and the remainder is $O(\Delta_j^2)$. Hence

$$|\mathbb{E}[\mathbf{v}^\top \mathbf{m}]| = \left| \sum_{j=1}^n v_j \mathbb{E}[m_j] \right| \leq \|\mathbf{v}\|_1 O(1/d) \leq \sqrt{n} O(1/d) = O(d^{-1/2}) \rightarrow 0.$$

Thus $\mathbf{v}^\top \mathbf{m} \rightarrow 0$ in L^2 and hence in probability. Taking $\mathbf{v} = \mathbf{u}, \mathbf{v}_1, \mathbf{v}_2$ yields (E.22). $\blacksquare$

Lemma 38 *Let $\mathbf{g}_k = \mathbf{W}\mathbf{u}_k$ for fixed orthonormal $\mathbf{u}_1, \mathbf{u}_2$. Then*

$$\frac{\|\mathbf{g}_k\|^2}{N} - 1 \prec \frac{1}{\sqrt{N}}, \quad \frac{\mathbf{g}_1^\top \mathbf{g}_2}{N} \prec \frac{1}{\sqrt{N}}.$$

Consequently, the two XOR additive spikes in (E.20) have asymptotic strength $\theta_{\text{SNR}}^2 b_\sigma^2 \rightarrow r^2 b_\sigma^2 \psi/2 = \beta_{\text{lin}}$.

Proof. Each $\mathbf{g}_k \sim \mathcal{N}(0, \mathbf{I}_N)$ and $(\mathbf{g}_1, \mathbf{g}_2)$ is jointly Gaussian with $\text{Cov}(\mathbf{g}_1, \mathbf{g}_2) = \mathbf{0}$. Therefore, we can directly apply Proposition 2.6.6 and Theorem 3.1.1 of Vershynin (2018) to get conclusion. $\blacksquare$

E.4.3. RESOLVENT QUADRATIC FORMS FOR THE SPIKE ANALYSIS

Fix any compact set $\mathcal{D} \subset \mathbb{R}$ such that $\mathcal{D} \cap \text{supp}(\mu) = \emptyset$. For $\lambda \in \mathcal{D}$, define $s(\lambda) \in \mathbb{R}$ by the unique relation

$$\lambda = z(s(\lambda)), \quad z'(s(\lambda)) > 0, \quad (\text{E.26})$$

which is well-defined by Lemma 20. Define the left and right resolvents

$$\mathbf{Q}_L(\lambda) := (\mathbf{G}\mathbf{G}^\top - \lambda\mathbf{I}_N)^{-1}, \quad \mathbf{Q}_R(\lambda) := (\mathbf{K}_0 - \lambda\mathbf{I}_n)^{-1}. \quad (\text{E.27})$$

Lemma 39 (Left resolvent limits for quadratic forms) *Uniformly for $\lambda \in \mathcal{D} \subset \mathbb{R} \setminus \{0\}$,*

$$\frac{1}{N} \mathbf{1}_N^\top \mathbf{Q}_L(\lambda) \mathbf{1}_N = s(\lambda) + o_{\mathbb{P}}(1), \quad (\text{E.28})$$

$$\frac{1}{N} \mathbf{g}_k^\top \mathbf{Q}_L(\lambda) \mathbf{g}_\ell = \delta_{k\ell} s(\lambda) + o_{\mathbb{P}}(1), \quad (\text{E.29})$$

$$\frac{1}{N} \mathbf{1}_N^\top \mathbf{Q}_L(\lambda) \mathbf{g}_k = o_{\mathbb{P}}(1). \quad (\text{E.30})$$

Proof. First, for (E.28), we can directly utilize Theorem 2.10 of Fan et al. (2026) to derive the anisotropic local law of $\mathbf{G}\mathbf{G}^\top$ and obtain

$$\frac{1}{N} \mathbf{1}_N^\top \mathbf{Q}_L(\lambda) \mathbf{1}_N - s(\lambda) \prec \frac{1}{\sqrt{N}} \quad (\text{E.31})$$

uniformly for $\lambda \in \mathcal{D} \subset \mathbb{R} \setminus \{0\}$ where $\mathcal{D} \subset \mathbb{R}$ is any fixed compact set such that $\mathcal{D} \cap (\text{supp}(\mu) \cup \{0\}) = \emptyset$.

Second, we can replace $\mathbf{G}$ by an independent copy of the bulk. By Lemma 36, $\|\mathbf{G} - \tilde{\mathbf{G}}\| \prec 1/\sqrt{n}$, hence $\|\mathbf{G}\mathbf{G}^\top - \tilde{\mathbf{G}}\tilde{\mathbf{G}}^\top\| \prec 1/\sqrt{n}$. Resolvent stability gives, uniformly on $\mathcal{D}$,

$$\|\mathbf{Q}_L(\lambda) - \tilde{\mathbf{Q}}_L(\lambda)\| \prec 1/\sqrt{n}, \quad \tilde{\mathbf{Q}}_L(\lambda) := (\tilde{\mathbf{G}}\tilde{\mathbf{G}}^\top - \lambda \mathbf{I})^{-1}.$$

Thus it suffices to prove (E.29)–(E.30) with $\mathbf{Q}_L$ replaced by $\tilde{\mathbf{Q}}_L$. By Lemma 35, $\mathbf{g}_1, \mathbf{g}_2$ are independent standard Gaussian random vectors and independent of $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{Q}}_L(\lambda)$. Conditional on $\tilde{\mathbf{Q}}_L(\lambda)$, for $k \neq \ell$, $\mathbf{g}_k^\top \tilde{\mathbf{Q}}_L(\lambda) \mathbf{g}_\ell$ is centered with variance $O(N)$, so $N^{-1} \mathbf{g}_k^\top \tilde{\mathbf{Q}}_L(\lambda) \mathbf{g}_\ell = o_{\mathbb{P}}(1)$. For $k = \ell$, $\mathbb{E} \left[\frac{1}{N} \mathbf{g}_k^\top \tilde{\mathbf{Q}}_L(\lambda) \mathbf{g}_k \mid \tilde{\mathbf{Q}}_L(\lambda) \right] = \frac{1}{N} \text{Tr} \tilde{\mathbf{Q}}_L(\lambda)$ and the Hanson–Wright inequality yields concentration around this trace at $O_{\mathbb{P}}(N^{-1/2})$, uniformly on $\mathcal{D}$ since $\|\tilde{\mathbf{Q}}_L(\lambda)\| \leq 1/\text{dist}(\lambda, \text{spec}(\tilde{\mathbf{G}}\tilde{\mathbf{G}}^\top)) \leq C(\mathcal{D})$. Similarly, $N^{-1} \mathbf{1}^\top \tilde{\mathbf{Q}}_L(\lambda) \mathbf{g}_k = o_{\mathbb{P}}(1)$. $\blacksquare$

Lemma 40 (Wishart resolvent quadratic form with a column-dependent vector) *Let $\mathbf{Z} \in \mathbb{R}^{d \times n}$ have i.i.d. entries $Z_{\ell j} \sim \mathcal{N}(0, 1/d)$, and set $\mathbf{S} := \mathbf{Z}^\top \mathbf{Z} \in \mathbb{R}^{n \times n}$ and $\mathbf{G}(\zeta) := (\mathbf{S} - \zeta \mathbf{I}_n)^{-1}$, $\zeta \in \mathbb{C} \setminus \mathbb{R}$. Assume $n/d \rightarrow \psi \in (0, \infty)$, and define MP edges $a := (1 - \sqrt{\psi})^2$ and $b := (1 + \sqrt{\psi})^2$. Fix $\kappa > 0$ and $\omega \in (0, 1)$, and consider the outside-spectrum domain*

$$\mathcal{S}_{\text{out}} := \left\{ \zeta = E + i\eta : \text{dist}(E, [a, b]) \geq \kappa, \quad n^{-1+\omega} \leq \eta \leq 1 \right\}.$$

Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ be Lipschitz with $f(1) = 0$ and polynomial growth, and define the random vector $\mathbf{m} \in \mathbb{R}^n$ such that its j -th entry is $m_j := f(\|\mathbf{z}_j\|^2)$ for $j \in [n]$, where $\mathbf{z}_j$ is the j -th column of $\mathbf{Z}$. Let $m_{\text{MP}, \psi}(\zeta)$ be the Stieltjes transform of ρ_{ψ}^{MP} . Then uniformly for $\zeta \in \mathcal{S}_{\text{out}}$,

$$\mathbf{m}^\top \mathbf{G}(\zeta) \mathbf{m} = m_{\text{MP}, \psi}(\zeta) \|\mathbf{m}\|^2 + O_{\prec}(d^{-1/2}). \quad (\text{E.32})$$

In particular, for any fixed real E with $\text{dist}(E, [a, b]) \geq \kappa$, $\mathbf{m}^\top (\mathbf{S} - E\mathbf{I})^{-1} \mathbf{m} = m_{\text{MP}, \psi}(E) \|\mathbf{m}\|^2 + o_{\mathbb{P}}(1)$.

Proof. Since $\|\mathbf{z}_j\|^2 = 1 + O_{\prec}(d^{-1/2})$ and f is Lipschitz with $f(1) = 0$,

$$m_j = O_{\prec}(d^{-1/2}) \quad \text{uniformly in } j. \quad (\text{E.33})$$

Hence $\|\mathbf{m}\|^2 = \sum_{j=1}^n m_j^2 = O_{\prec}(n/d) = O_{\prec}(1)$, and moreover $\|\mathbf{m}\| = O_{\prec}(1)$.

We first recall the isotropic local law of Bloemendal et al. (2014) for $\mathbf{S}$. By the entrywise local law for sample covariance matrices and its minor-stability (Bloemendal et al., 2014; Knowles and Yin, 2017), we have

$$\max_{i, j \in [n]} |\mathbf{G}_{ij}(\zeta) - \delta_{ij} m_{\text{MP}, \psi}(\zeta)| \prec \Psi(\zeta) := n^{-1/2}, \quad (\text{E.34})$$

and the same holds for all resolvent minors $\mathbf{G}^{(j)}(\zeta) := (\mathbf{S}^{(j)} - \zeta \mathbf{I})^{-1}$, uniformly in $j \in [n]$, where $\mathbf{S}^{(j)}$ denote the $(n-1) \times (n-1)$ minor of $\mathbf{S}$ obtained by removing j -th row and column.

We will also use the standard resolvent comparison identity between $\mathbf{G}$ and $\mathbf{G}^{(j)}$: for any $i, k \neq j$,

$$\mathbf{G}_{ik}(\zeta) = \mathbf{G}_{ik}^{(j)}(\zeta) + \frac{\mathbf{G}_{ij}(\zeta) \mathbf{G}_{jk}(\zeta)}{\mathbf{G}_{jj}(\zeta)}. \quad (\text{E.35})$$

Since $\mathbf{G}_{ij} = O_{\prec}(n^{-1/2})$ for $i \neq j$ and $\mathbf{G}_{jj} = O_{\prec}(1)$ on $\mathcal{S}_{\text{out}}$,

$$\max_{i, k \neq j} |\mathbf{G}_{ik}(\zeta) - \mathbf{G}_{ik}^{(j)}(\zeta)| \prec \Psi(\zeta)^2 = n^{-1}. \quad (\text{E.36})$$

Write the block form of $\mathbf{S} - \zeta \mathbf{I}$ w.r.t. index j :

$$\mathbf{S} - \zeta \mathbf{I} = \begin{pmatrix} q_j - \zeta & \mathbf{r}_j^\top \\ \mathbf{r}_j & \mathbf{S}^{(j)} - \zeta \mathbf{I} \end{pmatrix}, \quad q_j := \|\mathbf{z}_j\|^2, \quad \mathbf{r}_j := (\mathbf{z}_k^\top \mathbf{z}_j)_{k \neq j} \in \mathbb{R}^{n-1}.$$

Denote vector $\mathbf{G}_{j,-j}(\zeta) = (\mathbf{G}_{jk}(\zeta))_{k \neq j} \in \mathbb{R}^{1 \times (n-1)}$. The Schur complement identities give

$$\mathbf{G}_{jj}(\zeta) = \frac{1}{q_j - \zeta - \mathbf{r}_j^\top \mathbf{G}^{(j)}(\zeta) \mathbf{r}_j}, \quad (\text{E.37})$$

$$\mathbf{G}_{j,-j}(\zeta) = -\mathbf{G}_{jj}(\zeta) \mathbf{r}_j^\top \mathbf{G}^{(j)}(\zeta). \quad (\text{E.38})$$

Let $\mathbf{m}^{(j)}$ be $\mathbf{m}$ with the j -th component removed, and define $a_j(\zeta) := \mathbf{r}_j^\top \mathbf{G}^{(j)}(\zeta) \mathbf{m}^{(j)}$. Then $(\mathbf{G}\mathbf{m})_j = \mathbf{G}_{jj}(m_j - a_j(\zeta))$, so

$$\mathbf{m}^\top \mathbf{G}(\zeta) \mathbf{m} = \sum_{j=1}^n m_j^2 \mathbf{G}_{jj}(\zeta) - \sum_{j=1}^n m_j \mathbf{G}_{jj}(\zeta) a_j(\zeta) =: T_1(\zeta) - T_2(\zeta). \quad (\text{E.39})$$

Now we control $T_1(\zeta)$ and $T_2(\zeta)$ separately.

For $T_1(\zeta)$ term, by (E.34) with $i = j$, $\mathbf{G}_{jj}(\zeta) = m_{\text{MP},\psi}(\zeta) + O_{\prec}(n^{-1/2})$. Using $\|\mathbf{m}\|^2 = O_{\prec}(1)$,

$$T_1(\zeta) = m_{\text{MP},\psi}(\zeta) \|\mathbf{m}\|^2 + O_{\prec}(n^{-1/2}). \quad (\text{E.40})$$

Hence, it suffices to control the dependence term $T_2(\zeta)$ below. Write $X_j(\zeta) := m_j \mathbf{G}_{jj}(\zeta) a_j(\zeta)$, so $T_2(\zeta) = \sum_{j=1}^n X_j(\zeta)$. We follow the fluctuation average argument in (Wang et al., 2024) 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 $\mathbf{z}_j$ (i.e. conditioning on $\mathbf{Z}^{(j)}$). Then, we have that: (i) $m_j = f(\|\mathbf{z}_j\|^2)$ is even in $\mathbf{z}_j$; (ii) $\mathbf{G}_{jj}(\zeta)$ is even in $\mathbf{z}_j$ by (E.37); (iii) $a_j(\zeta)$ is odd in $\mathbf{z}_j$ because $\mathbf{r}_j = \mathbf{Z}^{(j)\top} \mathbf{z}_j$ is linear in $\mathbf{z}_j$, while $\mathbf{G}^{(j)}$ and $\mathbf{m}^{(j)}$ are $\mathbf{Z}^{(j)}$ -measurable, independent with $\mathbf{z}_j$. Hence $X_j(\zeta)$ is odd in $\mathbf{z}_j$, so

$$\mathbb{E}_j[X_j(\zeta)] = 0. \quad (\text{E.41})$$

Next, we control the size of a_j and X_j . Conditional on $\mathbf{Z}^{(j)}$, we have $\mathbf{r}_j = \mathbf{Z}^{(j)\top} \mathbf{z}_j$, with $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/d)$ independent of $\mathbf{Z}^{(j)}$. Thus $\mathbf{r}_j$ is centered Gaussian with $\text{Cov}_j(\mathbf{r}_j) = (1/d) \mathbf{S}^{(j)}$. Therefore $a_j(\zeta) = \mathbf{m}^{(j)\top} \mathbf{G}^{(j)}(\zeta) \mathbf{r}_j$ is centered Gaussian with conditional variance

$$\text{Var}_j(a_j(\zeta)) = \frac{1}{d} \mathbf{m}^{(j)\top} \mathbf{G}^{(j)}(\zeta) \mathbf{S}^{(j)} \mathbf{G}^{(j)}(\zeta) \mathbf{m}^{(j)} \leq \frac{C}{d} \|\mathbf{m}\|^2,$$

because on $\mathcal{S}_{\text{out}}$, $\|\mathbf{G}^{(j)}(\zeta)\| \leq C(\kappa)$ and $\|\mathbf{S}^{(j)}\| = O_{\prec}(1)$. Using $\|\mathbf{m}\| = O_{\prec}(1)$, we get

$$a_j(\zeta) = O_{\prec}(d^{-1/2}). \quad (\text{E.42})$$

Combining (E.33), $\mathbf{G}_{jj}(\zeta) = O_{\prec}(1)$, and (E.42) yields

$$X_j(\zeta) = O_{\prec}(d^{-1}). \quad (\text{E.43})$$

Finally, we take high-moment fluctuation averaging for T_2 . Fix an integer $p \geq 1$. We claim

$$\mathbb{E}|T_2(\zeta)|^{2p} \leq C_p \left(\frac{n}{d^2}\right)^p \asymp C_p d^{-p}, \quad \text{uniformly for } \zeta \in \mathcal{S}_{\text{out}}. \quad (\text{E.44})$$

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

$$\mathbb{E}|T_2|^{2p} = \sum_{j_1, \dots, j_{2p} \in [n]} \mathbb{E}[X_{j_1} \cdots X_{j_{2p}}].$$

For any fixed multi-index $\mathbf{j} = (j_1, \dots, j_{2p})$, let $\mathcal{I}(\mathbf{j})$ be the set of distinct values in $\mathbf{j}$, and write $r := |\mathcal{I}(\mathbf{j})|$. For each $t \in \mathcal{I}(\mathbf{j})$, let ν_t be its multiplicity in the list $\mathbf{j}$, so $\sum_{t \in \mathcal{I}} \nu_t = 2p$. Let $\mathcal{S}(\mathbf{j}) := \{t \in \mathcal{I}(\mathbf{j}) : \nu_t = 1\}$ be the set of singleton indices, and write $q := |\mathcal{S}(\mathbf{j})|$. Notice that

$$2p = \sum_{t \in \mathcal{I}} \nu_t \geq \sum_{t \in \mathcal{S}} 1 + \sum_{t \notin \mathcal{S}} 2 = q + 2(r - q) = 2r - q,$$

hence $q \geq 2r - 2p$ for any fixed multi-index $\mathbf{j}$.

Fix a monomial $X_{j_1} \cdots X_{j_{2p}}$ and pick a singleton index $t \in \mathcal{S}(\mathbf{j})$. 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 $\mathbf{G}$ (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 $\mathbf{Z}^{(t)}$ -measurable, hence independent of $\mathbf{z}_t$, and such that

$$F_t = F_t^{(t)} + \Delta_t, \quad |\Delta_t| \prec \Psi(\zeta)^2 \prod_{\alpha \neq t} |X_{j_\alpha}|. \quad (\text{E.45})$$

The estimate (E.45) follows by repeated use of the resolvent identity (E.35) and the replacement bound (E.36): each replacement introduces at least one factor $\mathbf{G}_{it}\mathbf{G}_{tk}/\mathbf{G}_{tt}$, hence a factor $O_{\prec}(\Psi^2)$, and the total number of replacements is $O_p(1)$ since p is fixed. Hence, (E.45) decouples each singleton index and gains a factor $\Psi(\zeta)^2$ per singleton.

Now take $\mathbb{E}_t[\cdot]$. Since $F_t^{(t)}$ is $\mathbf{Z}^{(t)}$ -measurable, by (E.41), we have $\mathbb{E}_t[X_t F_t^{(t)}] = F_t^{(t)} \mathbb{E}_t[X_t] = 0$. Hence $\mathbb{E}[X_t F_t] = \mathbb{E}[X_t \Delta_t]$. Using (E.43) and (E.45), we obtain that

$$|\mathbb{E}[X_t \Delta_t]| \prec \Psi(\zeta)^2 \mathbb{E}[|X_t| \prod_{\alpha \neq t} |X_{j_\alpha}|] \prec \Psi(\zeta)^2 d^{-2p}.$$

Then, we can repeat this decoupling for each singleton index in $\mathcal{S}(\mathbf{j})$, and each repetition contributes an additional factor $\Psi(\zeta)^2$. Therefore, for the monomial with q singleton indices,

$$|\mathbb{E}[X_{j_1} \cdots X_{j_{2p}}]| \prec d^{-2p} \Psi(\zeta)^{2q}. \quad (\text{E.46})$$

For fixed $r = |\mathcal{I}|$, the number of multi-indices with r distinct values is $O(n^r)$. Using (E.46) and $q \geq 2r - 2p$, we get

$$\sum_{j: |\mathcal{I}(j)|=r} |\mathbb{E}[X_{j_1} \cdots X_{j_{2p}}]| \prec n^r d^{-2p} \Psi(\zeta)^{2(2r-2p)}.$$

On $\mathcal{S}_{\text{out}}$, $\Psi(\zeta) = n^{-1/2}$, so $n^r d^{-2p} \Psi^{4r-4p} = n^{2p-r} d^{-2p} \asymp d^{-r}$. Summing over $r = 1, \dots, 2p$ 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 (E.44), proving $T_2(\zeta) = O_{\prec}(d^{-1/2})$. By (E.39), (E.40), and $T_2(\zeta) = O_{\prec}(d^{-1/2})$, we can conclude that

$$\mathbf{m}^\top \mathbf{G}(\zeta) \mathbf{m} = m_{\text{MP}, \psi}(\zeta) \|\mathbf{m}\|^2 + O_{\prec}(n^{-1/2}) + O_{\prec}(d^{-1/2}).$$

Since $n \asymp d$, this is exactly (E.32). ■

Lemma 41 (Right resolvent limits for quadratic forms) *Uniformly for $\lambda \in \mathcal{D}$,*

$$\mathbf{v}_k^\top \mathbf{Q}_R(\lambda) \mathbf{v}_\ell = \delta_{k\ell} m_\mu(\lambda) + o_{\mathbb{P}}(1), \quad (\text{E.47})$$

$$\mathbf{m}^\top \mathbf{Q}_R(\lambda) \mathbf{m} = \|\mathbf{m}\|^2 m_\mu(\lambda) + o_{\mathbb{P}}(1), \quad (\text{E.48})$$

$$\mathbf{v}_k^\top \mathbf{Q}_R(\lambda) \mathbf{m} = o_{\mathbb{P}}(1). \quad (\text{E.49})$$

Proof. Conditional on $\mathbf{Z}$, $\mathbf{K}_0 = \mathbf{G}^\top \mathbf{G}$ is a general sample covariance matrix with population covariance Σ and aspect ratio $\phi = n/N$. Similar as the proof of Lemma 32, we can verify that Assumption 5 of Wang et al. (2024) is satisfied for $\mathbf{G}$. Hence, we can apply Theorem 10 of Wang et al. (2024) to obtain a deterministic equivalence of quadratic form of the resolvent $\mathbf{Q}_R(\lambda)$. Precisely for any deterministic unit vectors $\mathbf{b}, \mathbf{b}'$ (possibly $\mathbf{Z}$ -measurable),

$$\mathbf{b}^\top (\mathbf{K}_0 - \lambda \mathbf{I})^{-1} \mathbf{b}' = \mathbf{b}^\top \left(-\lambda \tilde{m}_\mu(\lambda) \Sigma - \lambda \mathbf{I} \right)^{-1} \mathbf{b}' + O_{\prec}(N^{-1/2}), \quad (\text{E.50})$$

uniformly for $\lambda \in \mathcal{D}$.

By Corollary 27, $\Sigma = \Sigma_0 + \tau \mathbf{u} \mathbf{u}^\top + \mathbf{R}$ with $\|\mathbf{R}\| = o_{\mathbb{P}}(1)$. Let $\mathbf{A} := -\lambda s(\lambda) \Sigma_0 - \lambda \mathbf{I}$. Then

$$-\lambda s(\lambda) \Sigma - \lambda \mathbf{I} = \mathbf{A} - \lambda s(\lambda) \tau \mathbf{u} \mathbf{u}^\top + o_{\mathbb{P}}(1).$$

By Woodbury identity, we have

$$(\mathbf{A} - \lambda s \tau \mathbf{u} \mathbf{u}^\top)^{-1} = \mathbf{A}^{-1} + \frac{\lambda s \tau}{1 - \lambda s \tau \mathbf{u}^\top \mathbf{A}^{-1} \mathbf{u}} \mathbf{A}^{-1} \mathbf{u} \mathbf{u}^\top \mathbf{A}^{-1}.$$

If $\mathbf{b}^\top \mathbf{u} = 0$ and $\mathbf{b}'^\top \mathbf{u} = 0$, then $\mathbf{b}^\top (\mathbf{A}^{-1} \mathbf{u}) = O(1) \cdot (\mathbf{b}^\top \mathbf{u}) = 0$, hence the rank-one correction vanishes:

$$\mathbf{b}^\top (\mathbf{A} - \lambda s \tau \mathbf{u} \mathbf{u}^\top)^{-1} \mathbf{b}' = \mathbf{b}^\top \mathbf{A}^{-1} \mathbf{b}'.$$

Applying this with $\mathbf{b} = \mathbf{v}_k$, $\mathbf{b}' = \mathbf{v}_\ell$ (note $\mathbf{v}_k^\top \mathbf{u} = 0$ by (E.1)) yields that the τ -spike does not affect (E.47). Also, Lemma 37 implies that $\mathbf{b}^\top (\mathbf{A} - \lambda s \tau \mathbf{u} \mathbf{u}^\top)^{-1} \mathbf{b}' = \mathbf{b}^\top \mathbf{A}^{-1} \mathbf{b}' + o_{\mathbb{P}}(1)$ for $\mathbf{b} = \mathbf{b}' = \mathbf{m}/\|\mathbf{m}\|$.

For $\mathbf{v}_k^\top \mathbf{A}^{-1} \mathbf{v}_\ell$, since $\Sigma_0 = (1 - b_\sigma^2) \mathbf{I} + b_\sigma^2 \mathbf{S}$ is orthogonally invariant (shifted Wishart), the anisotropic local law from Knowles and Yin (2017) implies $\mathbf{v}_k^\top \mathbf{A}^{-1} \mathbf{v}_\ell = \delta_{k\ell} m_\mu(\lambda) + o_{\mathbb{P}}(1)$, giving (E.47). Here, we have removed the rank-one $\tau \mathbf{u} \mathbf{u}^\top$ term in Σ when applying (E.50).

For the quadratic form with respect to $\mathbf{m}$, we apply (E.50) with $\mathbf{b} = \mathbf{b}' = \mathbf{m}/\|\mathbf{m}\|$ and then show that $\mathbf{m}^\top \mathbf{A}^{-1} \mathbf{m} = \|\mathbf{m}\|^2 m_\mu(\lambda) + o_{\mathbb{P}}(1)$ even though $\mathbf{m}$ depends on $\mathbf{Z}$. Write

$$\mathbf{A}^{-1} = \left(-\lambda s(\lambda) ((1 - b_\sigma^2) \mathbf{I} + b_\sigma^2 \mathbf{S}) - \lambda \mathbf{I} \right)^{-1} = \alpha(\lambda) (\mathbf{S} - \zeta(\lambda) \mathbf{I})^{-1},$$

for explicit scalars $\alpha(\lambda)$ and $\zeta(\lambda)$ depending on λ (and bounded away from $\text{supp}(\rho_\psi^{\text{MP}})$ for $\lambda \in \mathcal{D}$). From Lemma 40, we can derive that

$$\mathbf{m}^\top (\mathbf{S} - \zeta \mathbf{I})^{-1} \mathbf{m} = m_{\text{MP}, \psi}(\zeta) \|\mathbf{m}\|^2 + o_{\mathbb{P}}(1), \quad (\text{E.51})$$

uniformly for $\zeta = \zeta(\lambda)$ in the relevant domain.

For the cross term $\mathbf{v}_k^\top \mathbf{Q}_R(\lambda) \mathbf{m}$, we use (E.50) with $\mathbf{b} = \mathbf{v}_k$, $\mathbf{b}' = \mathbf{m}$. The deterministic equivalent reduces to $\mathbf{v}_k^\top \mathbf{A}^{-1} \mathbf{m}$ since $\mathbf{v}_k \perp \mathbf{u}$. Because $\mathbf{v}_k^\top \mathbf{m} = o_{\mathbb{P}}(1)$ (Lemma 37) and $\|\mathbf{A}^{-1}\| = O(1)$ on $\mathcal{D}$, we get $\mathbf{v}_k^\top \mathbf{A}^{-1} \mathbf{m} = o_{\mathbb{P}}(1)$, yielding (E.49). $\blacksquare$

E.4.4. ANALYSES OF ADDITIVE SPIKES AND ALIGNMENTS

In this section, we define the rank-three spike matrix

$$\mathbf{P} := \frac{1}{\sqrt{N}} \mathbf{1}_N \mathbf{m}^\top + \frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}} (\mathbf{g}_1 \mathbf{v}_1^\top + \mathbf{g}_2 \mathbf{v}_2^\top) = \sum_{k=0}^2 \sqrt{\beta_k} \mathbf{a}_k \mathbf{b}_k^\top, \quad (\text{E.52})$$

where

$$\mathbf{a}_0 := \frac{\mathbf{1}_N}{\sqrt{N}}, \quad \mathbf{b}_0 := \frac{\mathbf{m}}{\|\mathbf{m}\|}, \quad \mathbf{a}_1 := \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|}, \quad \mathbf{b}_1 := \mathbf{v}_1, \quad \mathbf{a}_2 := \frac{\mathbf{g}_2}{\|\mathbf{g}_2\|}, \quad \mathbf{b}_2 := \mathbf{v}_2, \quad (\text{E.53})$$

and by Lemmas 37 and 38, asymptotically,

$$\beta_0 = \|\mathbf{m}\|^2 \rightarrow \beta_{\text{mean}} = \frac{c_\sigma^2}{2} \psi = \tau, \quad \beta_1 = \beta_2 = \theta_{\text{SNR}}^2 b_\sigma^2 \rightarrow \beta_{\text{lin}} = \frac{r^2 \psi}{2} b_\sigma^2. \quad (\text{E.54})$$

Then, we recall $\mathbf{Y}^\sharp = \mathbf{G} + \mathbf{P}$ and $\mathbf{K}^\sharp = \mathbf{Y}^{\sharp\top} \mathbf{Y}^\sharp$ defined by Proposition 34.

Lemma 42 Recall $\mathbf{G} \in \mathbb{R}^{N \times n}$ in (E.3) and $\mathbf{K}_0 = \mathbf{G}^\top \mathbf{G}$. Assume $\mathbf{K}_0$ has an isolated outlier eigenvalue $\hat{\lambda}_\tau^{(0)}$ with unit-norm eigenvector $\hat{\mathbf{u}}_\tau \in \mathbb{R}^n$ so that for any $\epsilon > 0$, $\mathbf{K}_0 \hat{\mathbf{u}}_\tau = \hat{\lambda}_\tau^{(0)} \hat{\mathbf{u}}_\tau$, and $\hat{\lambda}_\tau^{(0)} \notin \text{supp}(\mu) + (-\epsilon, \epsilon)$ for all large n . Define the associated unit left singular vector of $\mathbf{G}$ as

$$\hat{\mathbf{w}}_\tau := \frac{\mathbf{G} \hat{\mathbf{u}}_\tau}{\sqrt{\hat{\lambda}_\tau^{(0)}}} \in \mathbb{R}^N, \quad \|\hat{\mathbf{w}}_\tau\| = 1.$$

Consider the rank-3 additive spike matrix $\mathbf{P} = \sum_{k=0}^2 \sqrt{\beta_k} \mathbf{a}_k \mathbf{b}_k^\top$ with $\|\mathbf{a}_k\| = \|\mathbf{b}_k\| = 1$ and $\beta_k = O(1)$, $\mathbf{Y}^\sharp = \mathbf{G} + \mathbf{P}$, and $\mathbf{K}^\sharp = \mathbf{Y}^{\sharp\top} \mathbf{Y}^\sharp$. Assume that

$$\max_{k \in \{0,1,2\}} |\mathbf{b}_k^\top \hat{\mathbf{u}}_\tau| = o_{\mathbb{P}}(1), \quad \max_{k \in \{0,1,2\}} |\mathbf{a}_k^\top \hat{\mathbf{w}}_\tau| = o_{\mathbb{P}}(1). \quad (\text{E.55})$$

Then, we have

(i) **(Eigenvalue stability.)** There exists an eigenvalue $\hat{\lambda}_\tau^\sharp \in \text{spec}(\mathbf{K}^\sharp)$ such that $|\hat{\lambda}_\tau^\sharp - \hat{\lambda}_\tau^{(0)}| = o_{\mathbb{P}}(1)$. Hence, $\hat{\lambda}_\tau^\sharp$ is an outlier eigenvalue of $\mathbf{K}^\sharp$, i.e., for any $\epsilon > 0$, $\hat{\lambda}_\tau^\sharp \notin \text{supp}(\mu) + (-\epsilon, \epsilon)$ for all large n .

(ii) **(Eigenvector-alignment stability.)** Fix any constant $\delta > 0$. Let $\mathbf{\Pi}_\delta$ be the spectral projector of $\mathbf{K}^\sharp$ onto $I_\delta := (\hat{\lambda}_\tau^{(0)} - \delta, \hat{\lambda}_\tau^{(0)} + \delta)$. Then $\|(\mathbf{I} - \mathbf{\Pi}_\delta) \hat{\mathbf{u}}_\tau\| = o_{\mathbb{P}}(1)$. In particular, the normalized vector

$$\hat{\mathbf{u}}_\tau^\sharp := \frac{\mathbf{\Pi}_\delta \hat{\mathbf{u}}_\tau}{\|\mathbf{\Pi}_\delta \hat{\mathbf{u}}_\tau\|} \in \text{Range}(\mathbf{\Pi}_\delta)$$

satisfies $|\langle \hat{\mathbf{u}}_\tau^\sharp, \hat{\mathbf{u}}_\tau \rangle|^2 = 1 - o_{\mathbb{P}}(1)$. Thus $\hat{\mathbf{u}}_\tau^\sharp$ inherits the same eigenvector alignment limit as $\hat{\mathbf{u}}_\tau$ (e.g. alignment with $\mathbf{u} = \mathbf{1}_n / \sqrt{n}$ in Lemma 32).

(iii) **(Projected resolvents remove the singularity at $\hat{\lambda}_\tau^{(0)}$.)** Define the orthogonal projectors

$$\mathbf{P}_R^\perp := \mathbf{I}_n - \hat{\mathbf{u}}_\tau \hat{\mathbf{u}}_\tau^\top, \quad \mathbf{P}_L^\perp := \mathbf{I}_N - \hat{\mathbf{w}}_\tau \hat{\mathbf{w}}_\tau^\top,$$

and the spike-removed resolvents

$$\mathbf{Q}_R^\perp(\lambda) := \mathbf{P}_R^\perp (\mathbf{K}_0 - \lambda \mathbf{I})^{-1} \mathbf{P}_R^\perp, \quad \mathbf{Q}_L^\perp(\lambda) := \mathbf{P}_L^\perp (\mathbf{G} \mathbf{G}^\top - \lambda \mathbf{I})^{-1} \mathbf{P}_L^\perp, \quad (\text{E.56})$$

which admit a continuous extension to $\lambda = \hat{\lambda}_\tau^{(0)}$. For each fixed $k, \ell \in \{0, 1, 2\}$, $\mathbf{b}_k^\top \mathbf{Q}_R^\perp(\hat{\lambda}_\tau^{(0)}) \mathbf{b}_\ell$ and $\mathbf{a}_k^\top \mathbf{Q}_L^\perp(\hat{\lambda}_\tau^{(0)}) \mathbf{a}_\ell$ are well-defined and remain $O_{\mathbb{P}}(1)$. Hence, applying (E.55), we have

$$\mathbf{b}_k^\top \mathbf{Q}_R(\hat{\lambda}_\tau^{(0)}) \mathbf{b}_\ell = \mathbf{b}_k^\top \mathbf{Q}_R^\perp(\hat{\lambda}_\tau^{(0)}) \mathbf{b}_\ell + o_{\mathbb{P}}(1), \quad (\text{E.57})$$

$$\mathbf{a}_k^\top \mathbf{Q}_L(\hat{\lambda}_\tau^{(0)}) \mathbf{a}_\ell = \mathbf{a}_k^\top \mathbf{Q}_L^\perp(\hat{\lambda}_\tau^{(0)}) \mathbf{a}_\ell + o_{\mathbb{P}}(1). \quad (\text{E.58})$$

Proof. We can expand $\mathbf{K}^\sharp$ as a low-rank perturbation of $\mathbf{K}_0$. By definition,

$$\mathbf{K}^\sharp = (\mathbf{G} + \mathbf{P})^\top (\mathbf{G} + \mathbf{P}) = \mathbf{K}_0 + \underbrace{\mathbf{G}^\top \mathbf{P} + \mathbf{P}^\top \mathbf{G} + \mathbf{P}^\top \mathbf{P}}_{=: \mathbf{\Delta}}.$$

Thus $\mathbf{\Delta}$ has rank at most 9 and $\|\mathbf{\Delta}\| = O_{\mathbb{P}}(1)$ (since $\|\mathbf{G}\| = O_{\mathbb{P}}(1)$ and $\|\mathbf{P}\| = O_{\mathbb{P}}(1)$).

Part (i). Let $\mathbf{A}$ be symmetric, $\lambda \in \mathbb{R}$, and $\mathbf{v}$ a unit vector. Then

$$\text{dist}(\lambda, \text{spec}(\mathbf{A})) \leq \|(\mathbf{A} - \lambda \mathbf{I})\mathbf{v}\|. \quad (\text{E.59})$$

Indeed, considering eigen-decomposition $\mathbf{A} = \sum_i \lambda_i \mathbf{q}_i \mathbf{q}_i^\top$ and $\mathbf{v} = \sum_i c_i \mathbf{q}_i$, we have $\|(\mathbf{A} - \lambda \mathbf{I})\mathbf{v}\|^2 = \sum_i (\lambda_i - \lambda)^2 |c_i|^2 \geq \min_i (\lambda_i - \lambda)^2 \sum_i |c_i|^2 = \text{dist}(\lambda, \text{spec}(\mathbf{A}))^2$. Since $\mathbf{K}_0 \hat{\mathbf{u}}_\tau = \hat{\lambda}_\tau^{(0)} \hat{\mathbf{u}}_\tau$, we have

$$(\mathbf{K}^\sharp - \hat{\lambda}_\tau^{(0)} \mathbf{I}) \hat{\mathbf{u}}_\tau = (\mathbf{K}_0 - \hat{\lambda}_\tau^{(0)} \mathbf{I}) \hat{\mathbf{u}}_\tau + \Delta \hat{\mathbf{u}}_\tau = \Delta \hat{\mathbf{u}}_\tau.$$

Therefore, $\text{dist}(\hat{\lambda}_\tau^{(0)}, \text{spec}(\mathbf{K}^\sharp)) \leq \|\Delta \hat{\mathbf{u}}_\tau\|$. It remains to show $\|\Delta \hat{\mathbf{u}}_\tau\| = o_{\mathbb{P}}(1)$.

Write $\mathbf{P} = \mathbf{A} \Theta \mathbf{B}^\top$ with $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2]$, $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2]$ and $\Theta = \text{diag}(\sqrt{\beta_0}, \sqrt{\beta_1}, \sqrt{\beta_2})$. Then

$$\begin{aligned} \mathbf{G}^\top \mathbf{P} \hat{\mathbf{u}}_\tau &= \mathbf{G}^\top \mathbf{A} \Theta (\mathbf{B}^\top \hat{\mathbf{u}}_\tau), \\ \mathbf{P}^\top \mathbf{P} \hat{\mathbf{u}}_\tau &= \mathbf{B} \Theta (\mathbf{A}^\top \mathbf{A}) \Theta (\mathbf{B}^\top \hat{\mathbf{u}}_\tau), \\ \mathbf{P}^\top \mathbf{G} \hat{\mathbf{u}}_\tau &= \mathbf{B} \Theta \mathbf{A}^\top (\mathbf{G} \hat{\mathbf{u}}_\tau) = \sqrt{\hat{\lambda}_\tau^{(0)}} \mathbf{B} \Theta (\mathbf{A}^\top \hat{\mathbf{w}}_\tau), \end{aligned}$$

where we used $\mathbf{G} \hat{\mathbf{u}}_\tau = \sqrt{\hat{\lambda}_\tau^{(0)}} \hat{\mathbf{w}}_\tau$. Hence,

$$\|\Delta \hat{\mathbf{u}}_\tau\| \leq \|\mathbf{G}^\top \mathbf{A} \Theta\| \cdot \|\mathbf{B}^\top \hat{\mathbf{u}}_\tau\| + \sqrt{\hat{\lambda}_\tau^{(0)}} \|\mathbf{B} \Theta\| \cdot \|\mathbf{A}^\top \hat{\mathbf{w}}_\tau\| + \|\mathbf{B} \Theta (\mathbf{A}^\top \mathbf{A}) \Theta\| \cdot \|\mathbf{B}^\top \hat{\mathbf{u}}_\tau\|.$$

Now $\|\mathbf{G}^\top \mathbf{A} \Theta\| \leq \|\mathbf{G}\| \|\mathbf{A}\| \|\Theta\| = O_{\mathbb{P}}(1)$, $\|\mathbf{B} \Theta\| = O(1)$, and $\|\mathbf{B} \Theta (\mathbf{A}^\top \mathbf{A}) \Theta\| = O(1)$. By the decoupling assumptions (E.55), $\|\mathbf{B}^\top \hat{\mathbf{u}}_\tau\| = o_{\mathbb{P}}(1)$, and $\|\mathbf{A}^\top \hat{\mathbf{w}}_\tau\| = o_{\mathbb{P}}(1)$. Therefore $\|\Delta \hat{\mathbf{u}}_\tau\| = o_{\mathbb{P}}(1)$, and combining with (E.59) yields $\text{dist}(\hat{\lambda}_\tau^{(0)}, \text{spec}(\mathbf{K}^\sharp)) = o_{\mathbb{P}}(1)$, which proves (i).

Part (ii). Let $\mathbf{K}^\sharp = \sum_i \hat{\lambda}_i^\sharp \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^\top$ be an eigen-decomposition, and write $\hat{\mathbf{u}}_\tau = \sum_i c_i \hat{\mathbf{v}}_i$. Then

$$\|(\mathbf{K}^\sharp - \hat{\lambda}_\tau^{(0)} \mathbf{I}) \hat{\mathbf{u}}_\tau\|^2 = \sum_i (\hat{\lambda}_i^\sharp - \hat{\lambda}_\tau^{(0)})^2 |c_i|^2 = \|\Delta \hat{\mathbf{u}}_\tau\|^2 = o_{\mathbb{P}}(1).$$

Split the sum into indices with $\hat{\lambda}_i^\sharp \in I_\delta$ and those with $\hat{\lambda}_i^\sharp \notin I_\delta$. For $\hat{\lambda}_i^\sharp \notin I_\delta$ we have $|\hat{\lambda}_i^\sharp - \hat{\lambda}_\tau^{(0)}| \geq \delta$, hence

$$\delta^2 \sum_{\hat{\lambda}_i^\sharp \notin I_\delta} |c_i|^2 \leq \sum_{\hat{\lambda}_i^\sharp \notin I_\delta} (\hat{\lambda}_i^\sharp - \hat{\lambda}_\tau^{(0)})^2 |c_i|^2 \leq \|(\mathbf{K}^\sharp - \hat{\lambda}_\tau^{(0)} \mathbf{I}) \hat{\mathbf{u}}_\tau\|^2 = o_{\mathbb{P}}(1).$$

Hence $\sum_{\hat{\lambda}_i^\sharp \notin I_\delta} |c_i|^2 = o_{\mathbb{P}}(1)$, i.e. $\|(\mathbf{I} - \Pi_\delta) \hat{\mathbf{u}}_\tau\| = o_{\mathbb{P}}(1)$, proving (ii). This argument makes no assumption about how many eigenvalues lie inside I_δ , so it remains valid when there are more than one eigenvalues of $\mathbf{K}^\sharp$ in I_δ .

Part (iii). Let $\{\hat{\lambda}_i^{(0)}, \hat{\mathbf{u}}_i^{(0)}\}$ be an eigen-decomposition of $\mathbf{K}_0$ with $\hat{\lambda}_\tau^{(0)}$ corresponding to $\hat{\mathbf{u}}_\tau$. Then,

$$(\mathbf{K}_0 - \lambda \mathbf{I})^{-1} = \frac{1}{\hat{\lambda}_\tau^{(0)} - \lambda} \hat{\mathbf{u}}_\tau \hat{\mathbf{u}}_\tau^\top + \sum_{i \neq \tau} \frac{1}{\hat{\lambda}_i^{(0)} - \lambda} \hat{\mathbf{u}}_i^{(0)} \hat{\mathbf{u}}_i^{(0)\top}$$

for $\lambda \neq \hat{\lambda}_\tau^{(0)}$. Multiplying by $\mathbf{P}_R^\perp$ on both sides removes the pole term, giving $\mathbf{Q}_R^\perp(\lambda) = \sum_{i \neq \tau} \frac{1}{\hat{\lambda}_i^{(0)} - \lambda} \hat{\mathbf{u}}_i^{(0)} \hat{\mathbf{u}}_i^{(0)\top}$, which is continuous at $\lambda = \hat{\lambda}_\tau^{(0)}$ (since the sum is over $i \neq \tau$). The

same argument applies to $\mathbf{Q}_L^\perp(\lambda)$ using the eigen-decomposition of $\mathbf{G}\mathbf{G}^\top$. 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 $\lambda = \hat{\lambda}_\tau^{(0)}$. $\blacksquare$

Lemma 43 (Verification of the decoupling conditions (E.55) for $\mathbf{K}_0$) *Under the assumptions of Lemma 32, additionally suppose the separated regime, that is, $\tau > \tau_{\text{crit}}$ and $z'(-1/\Lambda_\tau) > 0$. Under these conditions, the matrix $\mathbf{K}_0 = \mathbf{G}^\top \mathbf{G}$ has a unique isolated outlier eigenvalue, denoted by λ_τ , associated with a unit-norm eigenvector $\hat{\mathbf{u}}_\tau$. Define the associated left singular vector $\hat{\mathbf{w}}_\tau := \mathbf{G}\hat{\mathbf{u}}_\tau/\sqrt{\lambda_\tau}$. Then the decoupling conditions in (E.55) hold where the spike directions $\mathbf{a}_k$ and $\mathbf{b}_k$ defined in (E.53).*

Proof. First we consider the right decoupling condition $\max_k |\mathbf{b}_k^\top \hat{\mathbf{u}}_\tau| = o_{\mathbb{P}}(1)$. By the XOR construction, $\mathbf{v}_1^\top \mathbf{u} = \mathbf{v}_2^\top \mathbf{u} = 0$ where $\mathbf{u} = \frac{1}{\sqrt{n}} \mathbf{1}_n$. Applying (E.19) in Lemma 32 with $\mathbf{v} = \mathbf{v}_k$ gives $\mathbf{v}_k^\top \hat{\mathbf{u}}_\tau = o_{\mathbb{P}}(1)$, for $k \in \{1, 2\}$. Moreover, Lemma 37 gives $\mathbf{u}^\top \mathbf{m} = o_{\mathbb{P}}(1)$ and $\|\mathbf{m}\| \rightarrow \sqrt{\beta_{\text{mean}}} > 0$. Therefore, we can obtain

$$\mathbf{b}_0^\top \mathbf{u} = \frac{\mathbf{m}^\top \mathbf{u}}{\|\mathbf{m}\|} = o_{\mathbb{P}}(1).$$

Applying Lemma 32 again with $\mathbf{v} = \mathbf{b}_0$ in (E.19) yields $\mathbf{b}_0^\top \hat{\mathbf{u}}_\tau = o_{\mathbb{P}}(1)$. Taking the maximum over $k \in \{0, 1, 2\}$ completes the first part.

Second, we consider the left decoupling condition $\max_k |\mathbf{a}_k^\top \hat{\mathbf{w}}_\tau| = o_{\mathbb{P}}(1)$. Let $\mathbf{G}^e$ be $\tilde{\mathbf{G}}$ defined in Lemma 36, so Lemma 36 indicates that $\|\mathbf{G} - \mathbf{G}^e\| \prec N^{-1/2}$ and $\mathbf{G}^e$ is independent of $(\mathbf{g}_1, \mathbf{g}_2)$. Define $\mathbf{K}_0^e := (\mathbf{G}^e)^\top \mathbf{G}^e$ and let $(\lambda_\tau^e, \mathbf{u}_\tau^e)$ denote the isolated eigenpair of $\mathbf{K}_0^e$ corresponding to the outlier of $\mathbf{K}_0$. In fact, Lemma 32 holds for $\mathbf{K}_0^e$ as well. Set $\mathbf{w}_\tau^e := \mathbf{G}^e \mathbf{u}_\tau^e / \sqrt{\lambda_\tau^e}$. Since

$$\|\mathbf{K}_0 - \mathbf{K}_0^e\| = \|\mathbf{G}^\top \mathbf{G} - (\mathbf{G}^e)^\top \mathbf{G}^e\| \leq (\|\mathbf{G}\| + \|\mathbf{G}^e\|) \|\mathbf{G} - \mathbf{G}^e\| = O_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(N^{-1/2}) = o_{\mathbb{P}}(1),$$

we can claim that

$$\|\hat{\mathbf{w}}_\tau - \mathbf{w}_\tau^e\| = o_{\mathbb{P}}(1). \quad (\text{E.60})$$

by Davis–Kahan Theorem. Also, we have $\|\hat{\mathbf{u}}_\tau - \mathbf{u}_\tau^e\| = o_{\mathbb{P}}(1)$ and $|\lambda_\tau - \lambda_\tau^e| = o_{\mathbb{P}}(1)$.

Fix $k \in \{1, 2\}$. Since $\mathbf{a}_k$ is unit,

$$|\mathbf{a}_k^\top \hat{\mathbf{w}}_\tau| \leq |\mathbf{a}_k^\top \mathbf{w}_\tau^e| + \|\hat{\mathbf{w}}_\tau - \mathbf{w}_\tau^e\| = |\mathbf{a}_k^\top \mathbf{w}_\tau^e| + o_{\mathbb{P}}(1)$$

by (E.60). Condition on $\mathbf{w}_\tau^e$ (which is measurable w.r.t. $\mathbf{G}^e$). By Lemmas 35 and 36, $\mathbf{g}_k \sim \mathcal{N}(0, \mathbf{I}_N)$ is independent of $\mathbf{G}^e$, hence independent of $\mathbf{w}_\tau^e$. Therefore $\mathbf{g}_k^\top \mathbf{w}_\tau^e \sim \mathcal{N}(0, 1)$ and $\|\mathbf{g}_k\| = \sqrt{N} + o_{\mathbb{P}}(\sqrt{N})$, so

$$|\mathbf{a}_k^\top \mathbf{w}_\tau^e| = \frac{|\mathbf{g}_k^\top \mathbf{w}_\tau^e|}{\|\mathbf{g}_k\|} = O_{\mathbb{P}}(1) \cdot \frac{1}{\sqrt{N} + o_{\mathbb{P}}(\sqrt{N})} = o_{\mathbb{P}}(1).$$

Hence $|\mathbf{a}_k^\top \hat{\mathbf{w}}_\tau| = o_{\mathbb{P}}(1)$ for $k \in \{1, 2\}$.

Lastly, we consider $\mathbf{a}_0 = \mathbf{1}_N/\sqrt{N}$. Let Π_δ be the spectral projector of $\mathbf{G}\mathbf{G}^\top$ onto interval $(\lambda_\tau - \delta, \lambda_\tau + \delta)$ for some fixed and small $\delta > 0$. On the event that the outlier is simple, $\Pi_\delta = \hat{\mathbf{w}}_\tau \hat{\mathbf{w}}_\tau^\top$, so

$$|\mathbf{a}_0^\top \hat{\mathbf{w}}_\tau|^2 = \mathbf{a}_0^\top \Pi_\delta \mathbf{a}_0 = -\frac{1}{2\pi i} \oint_{\Gamma} \mathbf{a}_0^\top \mathbf{Q}_L(z) \mathbf{a}_0 dz,$$

where $\mathbf{Q}_L(z) = (\mathbf{G}\mathbf{G}^\top - z\mathbf{I}_N)^{-1}$, $\Gamma := \{z \in \mathbb{C} : |z - \lambda_\tau| = \delta/2\}$, and $\mathbf{a}_0^\top \mathbf{Q}_L(z) \mathbf{a}_0$ is well-defined on Γ . Following the proof of Lemma 39, we can apply Theorem 2.10 in Fan et al. (2026) to get: uniformly for $z \in \Gamma$,

$$\mathbf{a}_0^\top \mathbf{Q}_L(z) \mathbf{a}_0 = s(z) + o_{\mathbb{P}}(1),$$

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 $|\mathbf{a}_0^\top \widehat{\mathbf{w}}_\tau|^2 = o_{\mathbb{P}}(1)$. $\blacksquare$

Lemma 44 (Master equation for additive outliers of $\mathbf{K}^\sharp$) Fix $\lambda \in \mathbb{C}$ such that

$$\lambda \notin \text{supp}(\mu) \cup \{0\}. \quad (\text{E.61})$$

Define $\mathbf{Q}_L(\lambda) = (\mathbf{G}\mathbf{G}^\top - \lambda\mathbf{I}_N)^{-1}$ and $\mathbf{Q}_R(\lambda) = (\mathbf{K}_0 - \lambda\mathbf{I}_n)^{-1}$. Let us define three 3×3 matrices as

$$\mathbf{A}(\lambda) := (\mathbf{a}_k^\top \mathbf{Q}_L(\lambda) \mathbf{a}_\ell)_{k,\ell=0,1,2} \quad \mathbf{B}(\lambda) := (\mathbf{b}_k^\top \mathbf{Q}_R(\lambda) \mathbf{b}_\ell)_{k,\ell=0,1,2} \quad (\text{E.62})$$

and $\mathbf{\Theta} := \text{diag}(\sqrt{\beta_0}, \sqrt{\beta_1}, \sqrt{\beta_2})$. Then λ is an outlier eigenvalue of $\mathbf{K}^\sharp$ if

$$\det(\mathbf{I}_3 - \lambda \mathbf{\Theta} \mathbf{A}(\lambda) \mathbf{\Theta} \cdot \mathbf{B}(\lambda)) = 0. \quad (\text{E.63})$$

Moreover, (E.63) characterizes all outlier eigenvalues of $\mathbf{K}^\sharp$, induced by $\{\mathbf{a}_k, \mathbf{b}_k\}_{k=0}^2$, that lie in any compact set $\mathcal{D} \subset \mathbb{C}$ satisfying (E.61).

Proof. Notice that from the last part of Lemma 42, we know all the entries of $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ are well-defined and finite, even if $\lambda = \widehat{\lambda}_\tau^{(0)} \notin \text{supp}(\mu) \cup \{0\}$ where $\widehat{\lambda}_\tau^{(0)}$ is characterized by Lemma 32.

Define

$$\mathbf{H}^\sharp := \begin{pmatrix} \mathbf{0} & \mathbf{Y}^\sharp \\ \mathbf{Y}^{\sharp\top} & \mathbf{0} \end{pmatrix}, \quad \mathbf{H}_0 := \begin{pmatrix} \mathbf{0} & \mathbf{G} \\ \mathbf{G}^\top & \mathbf{0} \end{pmatrix}.$$

If $\rho \neq 0$, then $\rho \in \text{spec}(\mathbf{H}^\sharp)$ iff $\rho^2 \in \text{spec}(\mathbf{K}^\sharp)$. Set $\lambda := \rho^2$. We now consider the low rank matrix $\mathbf{P} = \mathbf{A}_0 \mathbf{\Theta} \mathbf{B}_0^\top$ where $\mathbf{A}_0 := [\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2] \in \mathbb{R}^{N \times 3}$ and $\mathbf{B}_0 := [\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2] \in \mathbb{R}^{n \times 3}$. Then

$$\mathbf{H}^\sharp = \mathbf{H}_0 + \mathbf{\Delta}, \quad \mathbf{\Delta} := \begin{pmatrix} \mathbf{0} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{0} \end{pmatrix}.$$

Factor $\mathbf{\Delta} = \mathbf{U} \mathbf{C} \mathbf{U}^\top$ with

$$\mathbf{U} := \begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_0 \end{pmatrix} \in \mathbb{R}^{(N+n) \times 6}, \quad \mathbf{C} := \begin{pmatrix} \mathbf{0} & \mathbf{\Theta} \\ \mathbf{\Theta} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{6 \times 6}.$$

Under (E.61), $\rho \notin \text{spec}(\mathbf{H}_0)$ so $(\mathbf{H}_0 - \rho \mathbf{I})$ is invertible and

$$\det(\mathbf{H}^\sharp - \rho \mathbf{I}) = \det(\mathbf{H}_0 - \rho \mathbf{I}) \det(\mathbf{I}_6 + \mathbf{C} \mathbf{U}^\top (\mathbf{H}_0 - \rho \mathbf{I})^{-1} \mathbf{U}).$$

Hence $\rho \in \text{spec}(\mathbf{H}^\sharp) \setminus \text{spec}(\mathbf{H}_0)$ iff the 6×6 determinant above vanishes. For $\lambda = \rho^2 \notin \text{spec}(\mathbf{K}_0) \cup \text{spec}(\mathbf{G}\mathbf{G}^\top)$ we have

$$(\mathbf{H}_0 - \rho \mathbf{I})^{-1} = \begin{pmatrix} -\rho(\mathbf{G}\mathbf{G}^\top - \lambda \mathbf{I})^{-1} & \mathbf{G}(\mathbf{K}_0 - \lambda \mathbf{I})^{-1} \\ (\mathbf{K}_0 - \lambda \mathbf{I})^{-1} \mathbf{G}^\top & -\rho(\mathbf{K}_0 - \lambda \mathbf{I})^{-1} \end{pmatrix} = \begin{pmatrix} -\rho \mathbf{Q}_L(\lambda) & \mathbf{G} \mathbf{Q}_R(\lambda) \\ \mathbf{Q}_R(\lambda) \mathbf{G}^\top & -\rho \mathbf{Q}_R(\lambda) \end{pmatrix}.$$

Using the identity $\mathbf{Q}_L(\lambda)\mathbf{G} = \mathbf{G}\mathbf{Q}_R(\lambda)$, by the Schur complement, the 6×6 determinant reduces to the equivalent 3×3 condition

$$\det \left(\mathbf{I}_3 - \lambda \boldsymbol{\Theta} \mathbf{A}_0^\top \mathbf{Q}_L(\lambda) \mathbf{A}_0 \boldsymbol{\Theta} \cdot \mathbf{B}_0^\top \mathbf{Q}_R(\lambda) \mathbf{B}_0 \right) = 0,$$

which is exactly (E.63). The final uniform statement over $\mathcal{D}$ follows because all steps are analytic in λ on the domain (E.61). $\blacksquare$

Lemma 45 (Scalar additive outlier equations) *Let $\lambda \in \mathcal{D}$ and $s = s(\lambda)$ satisfy (E.26). Then, asymptotically, the determinant equation (E.63) reduces to three scalar equations:*

$$1 - \beta_k T(s) = o_{\mathbb{P}}(1), \quad k \in \{0, 1, 2\}. \quad (\text{E.64})$$

Equivalently, each additive spike of strength $\beta \in \{\beta_0, \beta_1, \beta_2\}$ produces an outlier eigenvalue at

$$\lambda = z(s) + o_{\mathbb{P}}(1) \quad \text{for any real } s \text{ with } z'(s) > 0 \text{ satisfying } \beta T(s) = 1.$$

Proof. Fix $\lambda \in \mathcal{D}$ and let $s = s(\lambda)$. By Lemma 39, using $\mathbf{a}_0 = \mathbf{1}/\sqrt{N}$, $\mathbf{a}_1 = \mathbf{g}_1/\|\mathbf{g}_1\|$, $\mathbf{a}_2 = \mathbf{g}_2/\|\mathbf{g}_2\|$, we have

$$\mathbf{A}_{k\ell}(\lambda) = \mathbf{a}_k^\top \mathbf{Q}_L(\lambda) \mathbf{a}_\ell = \delta_{k\ell} s + o_{\mathbb{P}}(1), \quad k, \ell \in \{0, 1, 2\}.$$

Thus $\mathbf{A}(\lambda) = s \mathbf{I}_3 + o_{\mathbb{P}}(1)$.

By Lemma 41 and orthogonality $\mathbf{v}_k^\top \mathbf{u} = 0$, $\mathbf{v}_k^\top \mathbf{m} = o_{\mathbb{P}}(1)$, we obtain

$$\mathbf{B}_{k\ell}(\lambda) = \mathbf{b}_k^\top \mathbf{Q}_R(\lambda) \mathbf{b}_\ell = \delta_{k\ell} m_\mu(\lambda) + o_{\mathbb{P}}(1), \quad k, \ell \in \{0, 1, 2\}.$$

Thus $\mathbf{B}(\lambda) = m_\mu(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1)$.

Plugging the above approximations of $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ into (E.63) gives

$$\det \left(\mathbf{I}_3 - \lambda s m_\mu(\lambda) \text{diag}(\beta_0, \beta_1, \beta_2) \right) = o_{\mathbb{P}}(1),$$

hence, 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 Lemmas 20 and 21, $\lambda s m_\mu(\lambda) = T(s)$, proving (E.64). $\blacksquare$

Lemma 46 (Eigenvector alignment for additive outliers) *Fix a distinct spike strength value $\beta_\star \in \{\beta_0, \beta_1, \beta_2\}$, and define $J_\star := \{k \in \{0, 1, 2\} : \beta_k = \beta_\star\}$, with $r_\star := |J_\star|$. Let*

$$\boldsymbol{\Pi}_{B,\star} := \sum_{k \in J_\star} \mathbf{b}_k \mathbf{b}_k^\top, \quad \boldsymbol{\Pi}_{A,\star} := \sum_{k \in J_\star} \mathbf{a}_k \mathbf{a}_k^\top.$$

Assume that $s_\star \in \mathbb{R}$ satisfies $\beta_\star T(s_\star) = 1$, $z'(s_\star) > 0$, and $\lambda_\star := z(s_\star) \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$. Then, there exists a sufficiently small $\delta > 0$ with interval $I_\delta = (\lambda_\star - \delta, \lambda_\star + \delta) \subset \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$

such that the spectral projector $\widehat{P}_\star$ of $K^\sharp$ onto interval I_δ and the corresponding spectral projector $\widehat{P}_\star^{(L)}$ of $Y^\sharp Y^{\sharp\top}$ onto I_δ satisfy:

$$\mathrm{Tr}(\widehat{P}_\star \Pi_{B,\star}) = r_\star \cdot \gamma_R(s_\star) + o_{\mathbb{P}}(1), \quad (\text{E.65})$$

$$\mathrm{Tr}(\widehat{P}_\star^{(L)} \Pi_{A,\star}) = r_\star \cdot \gamma_L(s_\star) + o_{\mathbb{P}}(1), \quad (\text{E.66})$$

where

$$\gamma_R(s) := \frac{m_\mu(z(s)) z'(s)}{\beta_\star T'(s)} = -\frac{T(s)}{\beta_\star s^2 T'(s)} \varphi(s) = -\frac{\varphi(s)}{\beta_\star^2 s^2 T'(s)}. \quad (\text{E.67})$$

and

$$\gamma_L(s) := \frac{s z'(s)}{\beta_\star T'(s)} = -\frac{z(s)}{\beta_\star T'(s)} \varphi(s). \quad (\text{E.68})$$

In particular, if $r_\star = 1$ and $J_\star = \{k\}$, then the unique outlier eigenvalue inside I_δ is simple, and if $\widehat{v}_k$ denotes its unit right eigenvector, then

$$|\langle \widehat{v}_k, \mathbf{b}_k \rangle|^2 = \gamma_R(s_\star) + o_{\mathbb{P}}(1). \quad (\text{E.69})$$

Proof. By Lemma 45, every additive outlier of $K^\sharp$ in $\mathcal{D}$ is asymptotically described by a real solution s of $\beta T(s) = 1$ such that $z'(s) > 0$ for some $\beta \in \{\beta_0, \beta_1, \beta_2\}$. Let $\lambda = z(s)$. Since $z'(s_\star) > 0$, Lemma 20 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

$$f'_\star(\lambda_\star) = -\beta_\star T'(s_\star) \frac{ds}{d\lambda}(\lambda_\star) = -\beta_\star \frac{T'(s_\star)}{z'(s_\star)}. \quad (\text{E.70})$$

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: $I_\delta \subset \mathcal{D}$; $\lambda_\star$ 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 45, the eigenvalues of $K^\sharp$ inside I_δ are precisely the outlier cluster associated with the repeated block $J_\star$; its total multiplicity is $r_\star$.

Let

$$\mathbf{R}^\sharp(\lambda) := (K^\sharp - \lambda I_n)^{-1}, \quad \mathbf{Q}_R(\lambda) := (K_0 - \lambda I_n)^{-1}, \quad \mathbf{Q}_L(\lambda) := (G G^\top - \lambda I_N)^{-1}.$$

Recall from Lemma 44 that

$$Y^\sharp = G + P, \quad P = A_0 \Theta B_0^\top, \quad A_0 = [a_0, a_1, a_2], \quad B_0 = [b_0, b_1, b_2].$$

Introduce the linearizations

$$H^\sharp := \begin{pmatrix} 0 & Y^\sharp \\ Y^{\sharp\top} & 0 \end{pmatrix}, \quad H_0 := \begin{pmatrix} 0 & G \\ G^\top & 0 \end{pmatrix},$$

and

$$U := \begin{pmatrix} A_0 & 0 \\ 0 & B_0 \end{pmatrix} \in \mathbb{R}^{(N+n) \times 6}, \quad C := \begin{pmatrix} 0 & \Theta \\ \Theta & 0 \end{pmatrix} \in \mathbb{R}^{6 \times 6},$$

so that $\mathbf{H}^\sharp = \mathbf{H}_0 + \mathbf{U}\mathbf{C}\mathbf{U}^\top$. Fix ρ with $\rho^2 = \lambda$, and set

$$\mathbf{R}_0(\rho) := (\mathbf{H}_0 - \rho \mathbf{I}_{N+n})^{-1}, \quad \mathbf{S}(\rho) := \mathbf{U}^\top \mathbf{R}_0(\rho) \mathbf{U}.$$

By the Woodbury identity,

$$(\mathbf{H}^\sharp - \rho \mathbf{I})^{-1} = \mathbf{R}_0(\rho) - \mathbf{R}_0(\rho) \mathbf{U} \mathbf{C} (\mathbf{I}_6 + \mathbf{S}(\rho) \mathbf{C})^{-1} \mathbf{U}^\top \mathbf{R}_0(\rho). \quad (\text{E.71})$$

Multiplying (E.71) on the left by $\mathbf{U}^\top$ and on the right by $\mathbf{U}$, we obtain the exact identity

$$\mathbf{U}^\top (\mathbf{H}^\sharp - \rho \mathbf{I})^{-1} \mathbf{U} = (\mathbf{I}_6 + \mathbf{S}(\rho) \mathbf{C})^{-1} \mathbf{S}(\rho), \quad (\text{E.72})$$

since $\mathbf{S} - \mathbf{S} \mathbf{C} (\mathbf{I}_6 + \mathbf{S} \mathbf{C})^{-1} \mathbf{S} = (\mathbf{I}_6 + \mathbf{S} \mathbf{C})^{-1} \mathbf{S}$. Using the block formula for $(\mathbf{H}_0 - \rho \mathbf{I})^{-1}$ from Lemma 44,

$$\mathbf{S}(\rho) = \begin{pmatrix} -\rho \mathbf{A}(\lambda) & \mathbf{X}(\lambda) \\ \mathbf{X}(\lambda)^\top & -\rho \mathbf{B}(\lambda) \end{pmatrix}, \quad \mathbf{X}(\lambda) := \mathbf{A}_0^\top \mathbf{G} \mathbf{Q}_R(\lambda) \mathbf{B}_0 = \mathbf{A}_0^\top \mathbf{Q}_L(\lambda) \mathbf{G} \mathbf{B}_0.$$

Here $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ are defined by (E.62). Also, the lower-right block of $(\mathbf{H}^\sharp - \rho \mathbf{I})^{-1}$ equals $-\rho (\mathbf{K}^\sharp - \lambda \mathbf{I}_n)^{-1} = -\rho \mathbf{R}^\sharp(\lambda)$. Therefore, taking the lower-right 3×3 block in (E.72), there exists a 3×3 matrix-valued function $\mathbf{H}_B(\lambda)$, analytic on a neighborhood of $\mathcal{D}$, such that

$$\mathbf{B}_0^\top \mathbf{R}^\sharp(\lambda) \mathbf{B}_0 = \mathbf{B}(\lambda) \mathbf{M}(\lambda)^{-1} + \mathbf{H}_B(\lambda), \quad \mathbf{M}(\lambda) := \mathbf{I}_3 - \lambda \boldsymbol{\Theta} \mathbf{A}(\lambda) \boldsymbol{\Theta} \mathbf{B}(\lambda). \quad (\text{E.73})$$

Hence every pole of $\mathbf{B}_0^\top \mathbf{R}^\sharp(\lambda) \mathbf{B}_0$ inside $\mathcal{D}$ comes from a zero of $\det \mathbf{M}(\lambda)$, i.e., from an additive outlier of $\mathbf{K}^\sharp$.

By Lemmas 39 and 41, uniformly for $\lambda \in \mathcal{D} \subset \mathbb{C}$,

$$\mathbf{A}(\lambda) = s(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1), \quad \mathbf{B}(\lambda) = m_\mu(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1).$$

Therefore, uniformly on $\mathcal{D}$, we have

$$\mathbf{M}(\lambda) = \text{diag}(1 - \beta_0 T(s(\lambda)), 1 - \beta_1 T(s(\lambda)), 1 - \beta_2 T(s(\lambda))) + o_{\mathbb{P}}(1), \quad (\text{E.74})$$

because of $\lambda s(\lambda) m_\mu(\lambda) = T(s(\lambda))$.

Restrict now to the block $J_\star$ and let us define $\mathbf{E}_\star := \sum_{k \in J_\star} \mathbf{e}_k \mathbf{e}_k^\top \in \mathbb{R}^{3 \times 3}$, 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 $k \in J_\star$,

$$\mathbf{E}_\star \mathbf{M}(\lambda) \mathbf{E}_\star = (1 - \beta_\star T(s(\lambda))) \mathbf{E}_\star + o_{\mathbb{P}}(1), \quad \lambda \in \mathcal{D}.$$

Hence,

$$\mathbf{E}_\star \mathbf{M}(\lambda)^{-1} \mathbf{E}_\star = \frac{1}{1 - \beta_\star T(s(\lambda))} \mathbf{E}_\star + o_{\mathbb{P}}(1), \quad \lambda \in \mathcal{D}. \quad (\text{E.75})$$

Combining (E.73), $\mathbf{B}(\lambda) = m_\mu(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1)$, and (E.75), we obtain

$$\mathbf{E}_\star \mathbf{B}_0^\top \mathbf{R}^\sharp(\lambda) \mathbf{B}_0 \mathbf{E}_\star = \frac{m_\mu(\lambda)}{1 - \beta_\star T(s(\lambda))} \mathbf{E}_\star + \mathbf{H}_\star(\lambda) + o_{\mathbb{P}}(1), \quad \lambda \in \mathcal{D}, \quad (\text{E.76})$$

where $\mathbf{H}_\star(\lambda) := \mathbf{E}_\star \mathbf{H}_B(\lambda) \mathbf{E}_\star$ is analytic on a neighborhood of $\mathcal{D}$.

By the contour formula for spectral projectors,

$$\widehat{\mathbf{P}}_\star = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \mathbf{R}^\sharp(\lambda) d\lambda,$$

for some circle $\Gamma_\star \subset \mathcal{D}$ with $\Gamma_\star \cap \mathbb{R} = \{\lambda_\star - \delta, \lambda_\star + \delta\}$. Therefore

$$\text{Tr}(\widehat{\mathbf{P}}_\star \mathbf{\Pi}_{B,\star}) = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \text{Tr}(\mathbf{\Pi}_{B,\star} \mathbf{R}^\sharp(\lambda)) d\lambda = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \text{Tr}(E_\star \mathbf{B}_0^\top \mathbf{R}^\sharp(\lambda) \mathbf{B}_0 E_\star) d\lambda. \quad (\text{E.77})$$

Insert (E.76). The analytic term $\mathbf{H}_\star(\lambda)$ has zero contour integral, and $\text{Tr}(E_\star) = r_\star$, so

$$\text{Tr}(\widehat{\mathbf{P}}_\star \mathbf{\Pi}_{B,\star}) = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \frac{r_\star m_\mu(\lambda)}{1 - \beta_\star T(s(\lambda))} d\lambda + o_{\mathbb{P}}(1). \quad (\text{E.78})$$

Now $1 - \beta_\star T(s(\lambda))$ has a simple zero at $\lambda_\star$, and by (E.70)

$$\left. \frac{d}{d\lambda} (1 - \beta_\star T(s(\lambda))) \right|_{\lambda=\lambda_\star} = -\beta_\star \frac{T'(s_\star)}{z'(s_\star)}.$$

Hence, $-\text{Res}_{\lambda=\lambda_\star} \frac{m_\mu(\lambda)}{1 - \beta_\star T(s(\lambda))} = \frac{m_\mu(\lambda_\star) z'(s_\star)}{\beta_\star T'(s_\star)}$ and Cauchy Residue Theorem implies that

$$\text{Tr}(\widehat{\mathbf{P}}_\star \mathbf{\Pi}_{B,\star}) = r_\star \cdot \frac{m_\mu(\lambda_\star) z'(s_\star)}{\beta_\star T'(s_\star)} + o_{\mathbb{P}}(1),$$

which is (E.65). By Lemma 21, $T(s) = z(s) s m_\mu(z(s))$ and $m_\mu(z(s)) = \frac{T(s)}{z(s)s}$. Substituting into (E.65),

$$\gamma_R(s) = \frac{T(s)}{\beta_\star z(s)s} \cdot \frac{z'(s)}{T'(s)} = -\frac{\varphi(s_\star)}{\beta_\star^2 s_\star^2 T'(s_\star)}$$

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 $\mathbf{K}^\sharp$ by $\mathbf{Y}^\sharp \mathbf{Y}^{\sharp\top}$, $\mathbf{B}_0$ by $\mathbf{A}_0$, and $\mathbf{B}(\lambda)$ by $\mathbf{A}(\lambda)$. Since $\mathbf{A}(\lambda) = s(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1)$, the numerator becomes $s(\lambda)$ instead of $m_\mu(\lambda)$, yielding

$$\text{Tr}(\widehat{\mathbf{P}}_\star^{(L)} \mathbf{\Pi}_{A,\star}) = r_\star \cdot \frac{s_\star z'(s_\star)}{\beta_\star T'(s_\star)} + o_{\mathbb{P}}(1).$$

If $r_\star = 1$, then $\widehat{\mathbf{P}}_\star = \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^\top$, and hence $|\langle \widehat{\mathbf{v}}_k, \mathbf{b}_k \rangle|^2 = \text{Tr}(\widehat{\mathbf{P}}_\star \mathbf{\Pi}_{B,\star}) = \gamma_R(s_\star) + o_{\mathbb{P}}(1)$, which proves (E.69). $\blacksquare$

Notice that from Lemmas 31 and 32, we observe that the mean spike generated by β_0 are asymptotically identified with the separated covariance outlier of $\mathbf{K}_0$ generated by τ if they exist. The following lemma shows that this mean spike is collision-safe and produces a stable additive outlier of $\mathbf{K}^\sharp$ with non-trivial overlap on the mean spike direction $\mathbf{b}_0$, even in the presence of the linear outlier when $\beta_{\text{lin}} > 0$.

Lemma 47 Assume $c_\sigma \neq 0$, $z'(s_{\text{un}}) > 0$, and, when $\beta_{\text{lin}} > 0$ and $z'(s_{\text{lin}}) > 0$, assume $s_{\text{un}} \neq s_{\text{lin}}$. Let $\lambda_{\text{un}} := z(s_{\text{un}})$. Then there exists a sufficiently small deterministic constant $\delta > 0$ such that, with probability tending to one,

$$I_{\text{un},\delta} := (\lambda_{\text{un}} - \delta, \lambda_{\text{un}} + \delta) \quad (\text{E.79})$$

is disjoint from $\text{supp}(\mu) \cup \{0\}$ and from the linear outlier location $z(s_{\text{lin}})$ (when present), and $\mathbf{K}^\sharp$ has exactly one additive mean-spike eigenvalue in $I_{\text{un},\delta}$.

More precisely, let Γ_{un} be the positively oriented circle of radius $\delta/2$ centered at λ_{un} and write, for $z \in \Gamma_{\text{un}}$,

$$\mathbf{Q}_R(z) = \frac{\hat{\mathbf{u}}_\tau \hat{\mathbf{u}}_\tau^\top}{\hat{\lambda}_\tau^{(0)} - z} + \mathbf{Q}_R^\perp(z), \quad \mathbf{Q}_L(z) = \frac{\hat{\mathbf{w}}_\tau \hat{\mathbf{w}}_\tau^\top}{\hat{\lambda}_\tau^{(0)} - z} + \mathbf{Q}_L^\perp(z),$$

where $\mathbf{Q}_R^\perp, \mathbf{Q}_L^\perp$ are the projected resolvents from Lemma 42. Then uniformly on Γ_{un} ,

$$\mathbf{A}^\perp(z) := (\mathbf{a}_k^\top \mathbf{Q}_L^\perp(z) \mathbf{a}_\ell)_{k,\ell=0}^2 = s(z) \mathbf{I}_3 + o_{\mathbb{P}}(1), \quad (\text{E.80})$$

$$\mathbf{B}^\perp(z) := (\mathbf{b}_k^\top \mathbf{Q}_R^\perp(z) \mathbf{b}_\ell)_{k,\ell=0}^2 = m_\mu(z) \mathbf{I}_3 + o_{\mathbb{P}}(1), \quad (\text{E.81})$$

and therefore, uniformly on Γ_{un} ,

$$\det(\mathbf{I}_3 - z \boldsymbol{\Theta} \mathbf{A}^\perp(z) \boldsymbol{\Theta} \mathbf{B}^\perp(z)) = (1 - \tau T(s(z)))(1 - \beta_{\text{lin}} T(s(z)))^2 + o_{\mathbb{P}}(1). \quad (\text{E.82})$$

Let $\hat{\mathbf{P}}_{\text{mean},\text{un}}$ be the spectral projector of $\mathbf{K}^\sharp$ onto the unique additive outlier in $I_{\text{un},\delta}$. Then,

$$\text{Tr}(\hat{\mathbf{P}}_{\text{mean},\text{un}} \mathbf{b}_0 \mathbf{b}_0^\top) = \gamma_R(s_{\text{un}}) + o_{\mathbb{P}}(1), \quad (\text{E.83})$$

$$\text{Tr}(\hat{\mathbf{P}}_{\text{mean},\text{un}} \mathbf{u} \mathbf{u}^\top) = o_{\mathbb{P}}(1). \quad (\text{E.84})$$

Proof. Since $z'(s_{\text{un}}) > 0$, Lemma 31 implies $\tau > \tau_{\text{crit}}$ and hence $\mathbf{K}_0$ has a separated covariance outlier $\hat{\lambda}_\tau^{(0)} = \lambda_{\text{un}} + o_{\mathbb{P}}(1)$ by Lemma 32. Choose $\delta > 0$ so small that the interval (E.79) is disjoint from the bulk and from the linear outlier location when the latter exists.

On the contour Γ_{un} , the raw resolvents are well-defined. By the spectral decomposition of $\mathbf{K}_0$ and Lemma 42,

$$\mathbf{Q}_R(z) = \frac{\hat{\mathbf{u}}_\tau \hat{\mathbf{u}}_\tau^\top}{\hat{\lambda}_\tau^{(0)} - z} + \mathbf{Q}_R^\perp(z), \quad \mathbf{Q}_L(z) = \frac{\hat{\mathbf{w}}_\tau \hat{\mathbf{w}}_\tau^\top}{\hat{\lambda}_\tau^{(0)} - z} + \mathbf{Q}_L^\perp(z),$$

and the pole parts contribute only $o_{\mathbb{P}}(1)$ to quadratic forms against the spike directions because

$$\max_{k \in \{0,1,2\}} |\mathbf{b}_k^\top \hat{\mathbf{u}}_\tau| = o_{\mathbb{P}}(1), \quad \max_{k \in \{0,1,2\}} |\mathbf{a}_k^\top \hat{\mathbf{w}}_\tau| = o_{\mathbb{P}}(1)$$

by Lemma 43, while $|\hat{\lambda}_\tau^{(0)} - z| \asymp 1$ uniformly on Γ_{un} . Therefore the same proofs as in Lemmas 39 and 41, now applied to the projected resolvents, yield (E.80) and (E.81). The factorization (E.82) then follows exactly as in Lemma 45.

Because s_{un} is the unique solution of $\tau T(s) = 1$ on the connected component containing $I_{\text{un},\delta}$, because $z'(s_{\text{un}}) > 0$, and because the linear factor is bounded away from zero on Γ_{un} by the separation assumption $s_{\text{un}} \neq s_{\text{lin}}$, Rouché's theorem implies that the projected determinant has exactly one zero inside Γ_{un} , hence $\mathbf{K}^\sharp$ has exactly one additive mean-spike eigenvalue in $I_{\text{un},\delta}$.

Finally, (E.83) is the same computation as in Lemma 46 with the projected resolvents replacing the raw ones. Also, (E.84) follows from the same contour representation with the test vector $\mathbf{u}$, because the only possible residue is proportional to quadratic forms of the type $\mathbf{u}^\top \mathbf{Q}_R^\perp(z) \mathbf{b}_0 = o_{\mathbb{P}}(1)$ and because $\mathbf{u}^\top \mathbf{b}_0 = o_{\mathbb{P}}(1)$ by Lemma 37. $\blacksquare$

Lemma 48 Assume $c_\sigma \neq 0$, $z'(s_{\text{un}}) > 0$, and, when $\beta_{\text{lin}} > 0$ and $z'(s_{\text{lin}}) > 0$, assume $s_{\text{un}} \neq s_{\text{lin}}$. Let $\lambda_{\text{un}} = z(s_{\text{un}})$ and let $I_{\text{un},\delta}$ be as in Lemma 47. Then, with probability tending to one, $\mathbf{K}^\sharp$ has exactly two eigenvalues in $I_{\text{un},\delta}$: one inherited from the covariance outlier of $\mathbf{K}_0$ and one produced by the additive mean spike. If $\hat{\mathbf{P}}_{\text{un}}^\sharp$ denotes the spectral projector onto this two-dimensional cluster, then

$$\|\hat{\mathbf{P}}_{\text{un}}^\sharp \mathbf{u}\|^2 = \varphi(s_{\text{un}}) + o_{\mathbb{P}}(1), \quad (\text{E.85})$$

$$\left\| \hat{\mathbf{P}}_{\text{un}}^\sharp \frac{\mathbf{m}}{\|\mathbf{m}\|} \right\|^2 = -\frac{\varphi(s_{\text{un}})}{\tau^2 s_{\text{un}}^2 T'(s_{\text{un}})} + o_{\mathbb{P}}(1), \quad (\text{E.86})$$

$$\mathbf{u}^\top \hat{\mathbf{P}}_{\text{un}}^\sharp \frac{\mathbf{m}}{\|\mathbf{m}\|} = o_{\mathbb{P}}(1). \quad (\text{E.87})$$

Proof. By Lemma 42, the separated covariance outlier of $\mathbf{K}_0$ persists under the rank-three additive perturbation and yields one eigenvalue of $\mathbf{K}^\sharp$ inside $I_{\text{un},\delta}$. By Lemma 47, the additive mean spike yields exactly one further eigenvalue in the same interval. Since the linear spikes are excluded from $I_{\text{un},\delta}$ by construction, there are exactly two eigenvalues in the interval. The $\mathbf{u}$ -alignment comes from the inherited covariance direction. Indeed, by Lemma 32,

$$|\hat{\mathbf{u}}_\tau^\top \mathbf{u}|^2 = \varphi(s_{\text{un}}) + o_{\mathbb{P}}(1).$$

Since $\hat{\mathbf{u}}_\tau$ lies asymptotically inside $\text{Range}(\hat{\mathbf{P}}_{\text{un}}^\sharp)$ by Lemma 42, we obtain (E.85); the additive mean-spike contribution to the $\mathbf{u}$ -mass is negligible by (E.84). Similarly, the $\mathbf{m}/\|\mathbf{m}\|$ -alignment comes entirely from the additive mean-spike component: the inherited covariance eigenvector contributes only $o_{\mathbb{P}}(1)$ in the $\mathbf{b}_0 = \mathbf{m}/\|\mathbf{m}\|$ direction by (E.19) in Lemma 32 together with Lemma 37, while Lemma 47 yields the main term (E.86). The cross term (E.87) follows by combining the two preceding decoupling statements. $\blacksquare$

Proposition 49 (No label alignment with any separated outliers) Under the assumptions of Theorem 3, let $\hat{\mathbf{V}}_{\text{out}}$ denote the orthogonal projector onto the direct sum of all separated outlier eigenvalues of $\mathbf{K}$. Then

$$\frac{1}{n} \|\hat{\mathbf{V}}_{\text{out}} \mathbf{y}\|^2 = \frac{1}{n} \mathbf{y}^\top \hat{\mathbf{V}}_{\text{out}} \mathbf{y} \xrightarrow{\mathbb{P}} 0. \quad (\text{E.88})$$

Proof. By Proposition 34, it suffices to prove the statement for $\mathbf{K}^\sharp$. In the separated regime covered by Theorem 3, the outlier projector of $\mathbf{K}^\sharp$ decomposes as the orthogonal sum of the uninformative cluster projector $\hat{\mathbf{P}}_{\text{un}}^\sharp$ from Lemma 48 and, when present, the linear cluster projector $\hat{\mathbf{P}}_{\text{lin}}^\sharp = \hat{\mathbf{P}}_\star$ from Lemma 46 by setting $\lambda_\star = z(s_{\text{lin}})$ and $r_\star = 2$ in the latter. Hence, $\hat{\mathbf{V}}_{\text{out}}^\sharp = \hat{\mathbf{P}}_{\text{un}}^\sharp + \hat{\mathbf{P}}_{\text{lin}}^\sharp$. It is therefore enough to show

$$\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_{\text{un}}^\sharp \mathbf{y} = o_{\mathbb{P}}(1), \quad \frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_{\text{lin}}^\sharp \mathbf{y} = o_{\mathbb{P}}(1).$$

Set $\bar{\mathbf{y}} := \mathbf{y}/\sqrt{n}$. By the XOR construction,

$$\bar{\mathbf{y}}^\top \mathbf{u} = 0, \quad \bar{\mathbf{y}}^\top \mathbf{v}_1 = 0, \quad \bar{\mathbf{y}}^\top \mathbf{v}_2 = 0.$$

Moreover Lemma 37 gives $\bar{\mathbf{y}}^\top \mathbf{m}/\|\mathbf{m}\| = o_{\mathbb{P}}(1)$. 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 $\lambda_3 = z(s_{\text{cov}})$. Lemma 31 implies $\lambda_0 = \lambda_3$.

For the linear outliers, consider $\lambda = \lambda_1 = \lambda_2$. Then, by Lemmas 42 and 45, 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 $\mathbf{K}^\sharp$ inside I_δ will converge to λ in probability, where $I_\delta := (\lambda - \delta, \lambda + \delta)$. Let $\widehat{\mathbf{V}}_\lambda^\sharp$ be the spectral projector of $\mathbf{K}^\sharp$ 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 $\text{spec}(\mathbf{K}^\sharp)$. Then

$$\widehat{\mathbf{V}}_\lambda^\sharp = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} (\mathbf{K}^\sharp - z\mathbf{I})^{-1} dz.$$

Hence

$$\bar{\mathbf{y}}^\top \widehat{\mathbf{V}}_\lambda^\sharp \bar{\mathbf{y}} = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\mathbf{y}}^\top (\mathbf{K}^\sharp - z\mathbf{I})^{-1} \bar{\mathbf{y}} dz. \quad (\text{E.89})$$

Now use the same linearization trick as in Lemma 46. Recall that $\mathbf{A}_0 = [\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2]$, $\mathbf{B}_0 = [\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2]$, and $\boldsymbol{\Theta} = \text{diag}(\sqrt{\beta_0}, \sqrt{\beta_1}, \sqrt{\beta_2})$. Then one obtains the exact projected-resolvent representation

$$(\mathbf{K}^\sharp - z\mathbf{I})^{-1} = \mathbf{Q}_R(z) + \mathbf{Q}_R(z)\mathbf{B}_0 \mathbf{S}(z) \mathbf{B}_0^\top \mathbf{Q}_R(z), \quad (\text{E.90})$$

where $\mathbf{S}(z) = z\boldsymbol{\Theta}\mathbf{A}(z)\boldsymbol{\Theta}(\mathbf{I}_3 - z\boldsymbol{\Theta}\mathbf{A}(z)\boldsymbol{\Theta}\mathbf{B}(z))^{-1}$ is a 3×3 matrix-valued meromorphic function, and all poles inside $\mathcal{D}$ come from the additive outlier equation described by Lemma 45. Hence, $\mathbf{S}(z)$ is analytic and bounded on Γ_λ . Insert (E.90) into (E.89). The possible pole term is

$$\bar{\mathbf{y}}^\top \widehat{\mathbf{V}}_\lambda^\sharp \bar{\mathbf{y}} = -\frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\mathbf{y}}^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} dz - \frac{1}{2\pi i} \oint_{\Gamma_\lambda} \bar{\mathbf{y}}^\top \mathbf{Q}_R(z) \mathbf{B}_0 \mathbf{S}(z) \mathbf{B}_0^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} dz.$$

By (E.47), $\bar{\mathbf{y}}\mathbf{Q}_R(z)\bar{\mathbf{y}} = m_\mu(z) + o_{\mathbb{P}}(1)$ and

$$\mathbf{B}_0^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} = \begin{bmatrix} \mathbf{b}_0^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} \\ \mathbf{b}_1^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} \\ \mathbf{b}_2^\top \mathbf{Q}_R(z) \bar{\mathbf{y}} \end{bmatrix} = o_{\mathbb{P}}(1)$$

uniformly on the contour $z \in \Gamma_\lambda$. Since $m_\mu(z)$ is analytic on Γ_λ , its contour integral vanishes. Hence $\bar{\mathbf{y}}^\top \widehat{\mathbf{P}}_{\text{lin}}^\sharp \bar{\mathbf{y}} = o_{\mathbb{P}}(1)$.

For the uninformative eigenvalues, let Γ_{un} be the contour from Lemma 47. Using the resolvent decomposition into the covariance pole plus the projected resolvent,

$$\mathbf{Q}_R(z) = \frac{\widehat{\mathbf{u}}_\tau \widehat{\mathbf{u}}_\tau^\top}{\widehat{\lambda}_\tau^{(0)} - z} + \mathbf{Q}_R^\perp(z),$$

we note first that $\bar{\mathbf{y}}^\top \widehat{\mathbf{u}}_\tau = o_{\mathbb{P}}(1)$ by (E.19) in Lemma 32, because $\bar{\mathbf{y}}^\top \mathbf{u} = 0$. Hence, the covariance pole contributes only $o_{\mathbb{P}}(1)$ to the contour integral for $\bar{\mathbf{y}}^\top \widehat{\mathbf{P}}_{\text{un}}^\sharp \bar{\mathbf{y}}$. Similarly as the linear outlier case, we have that uniformly on Γ_{un} , $\bar{\mathbf{y}}^\top \mathbf{Q}_R^\perp(z) \bar{\mathbf{y}} = m_\mu(z) + o_{\mathbb{P}}(1)$, and $\bar{\mathbf{y}}^\top \mathbf{Q}_R^\perp(z) \mathbf{b}_k = o_{\mathbb{P}}(1)$, for $k = 0, 1, 2$. Thus, the contour integral of the analytic main term vanishes and the remainder is $o_{\mathbb{P}}(1)$. Therefore $\bar{\mathbf{y}}^\top \widehat{\mathbf{P}}_{\text{un}}^\sharp \bar{\mathbf{y}} = o_{\mathbb{P}}(1)$. Combining the two clusters yields (E.88) for $\mathbf{K}^\sharp$, and hence also for $\mathbf{K}$. $\blacksquare$

E.5. Proof of Theorem 3

By Proposition 34, $\|K - K^\sharp\| = o_{\mathbb{P}}(1)$. Therefore it suffices to analyze spectrum of $K^\sharp$.

Item (i) follows from Lemma 32(i): the bulk ESD of K_0 is μ , and the finite-rank perturbation from K_0 to $K^\sharp$ does not change the limiting ESD.

For item (ii), Lemma 31 shows that the covariance parameter and the mean parameter are identical on the uninformative branch, i.e. $s_{\text{cov}} = s_{\text{mean}} = s_{\text{un}}$. Thus the covariance and mean mechanisms collide at the same first-order location $\lambda_{\text{un}} = z(s_{\text{un}})$. Lemma 48 proves that $K^\sharp$ has exactly two eigenvalues in a small interval around λ_{un} and that the corresponding spectral projector satisfies (4.2)–(4.4). By Proposition 34, the same statements hold for K .

Next, for item (iii), the linear spikes are away from the covariance outlier by $s_{\text{un}} \neq s_{\text{lin}}$. Therefore Lemmas 45 and 46 apply directly with $\beta_1 = \beta_2 = \beta_{\text{lin}}$, giving the two linear outlier eigenvalues at $z(s_{\text{lin}})$ and the projector-overlap formula (4.5).

Finally, item (iv) is proved by Proposition 49.

Appendix F. Kernel Spectral Clustering for XOR

In this section, we present the performance of the kernel spectral clustering for XOR in the finite-SNR proportional limit regime given by Theorem 6. Results in this section on kernel spectral clustering are applications of the main results of Couillet and Benaych-Georges (2016).

Recall the four-component XOR Gaussian mixture data $\mathbf{X} \in \mathbb{R}^{d \times n}$ defined in (3.4), with cluster means defined in (3.7) and common covariance $\mathbf{C} = \mathbf{I}_d$. Define the matrix of cluster means $\mathbf{M} := [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4] \in \mathbb{R}^{d \times 4}$ from (3.7), and cluster-indicator matrix $\mathbf{J}$

$$\mathbf{J} \in \{0, 1\}^{n \times 4}, \quad J_{i,a} = 1 \Leftrightarrow \mathbf{x}_i \text{ belongs to cluster } a.$$

Each cluster $a \in [4]$ has size $n_a = n/4$. For an affinity function $f : \mathbb{R} \rightarrow \mathbb{R}$, define the Euclidean distance kernel matrix $\mathbf{K}^{(f)} \in \mathbb{R}^{n \times n}$ as

$$(\mathbf{K}^{(f)})_{ij} := f(\|\mathbf{x}_i - \mathbf{x}_j\|^2), \quad i, j \in [n] \quad (\text{F.1})$$

We consider the normalized Laplacian of $\mathbf{K}^{(f)}$, which is defined as

$$\mathbf{L} := n \mathbf{D}^{-1/2} \mathbf{K}^{(f)} \mathbf{D}^{-1/2},$$

where $\mathbf{D} = \text{diag}(\mathbf{K}^{(f)} \mathbf{1}_n)$ is the degree matrix. Similar to the results on the spectra of the CK in Section 4.1, the analysis of the eigenvalues of $\mathbf{L}$ relies on a LE for $\mathbf{L}$. We refer to Theorem 1 of Couillet and Benaych-Georges (2016) for more details.

It is not difficult to prove that under Assumption 2, the pairwise distances between distinct samples $\mathbf{x}_i$ and $\mathbf{x}_j$ of the XOR dataset concentrate around $\tau = 2$ (see Lemma 13 for details). The core of the analysis of $\mathbf{L}$ then relies upon a careful entry-wise Taylor expansion of $\mathbf{K}^{(f)}$ around the limiting value τ . We therefore make the following assumption on the kernel function $f : \mathbb{R} \rightarrow \mathbb{R}$.

Assumption 50 *Let $\tau = 2$. The kernel function f is three-times differentiable in some neighborhood of τ . Furthermore, $f(\tau) > 0$ and $f'(\tau) \neq 0$.*

Theorem 51 (A detailed statement of Theorem 6) *Suppose $\text{SNR } r^2 = \Theta(1)$ and let Assumptions 2 and 50 hold. Then we have that:*

(i) **(BBP for two linear informative outliers.)** If $r^2 > 2\sqrt{\psi^{-1}}$, then there exists two isolated eigenvalues of $\mathbf{L}$, which are asymptotically approximated by

$$-\frac{2f'(\tau)}{f(\tau)}\rho + \frac{f(0) - f(\tau) + \tau f'(\tau)}{f(\tau)} \quad (\text{F.2})$$

in probability, where $\rho = \psi(1 + \frac{r^2}{2}) + \frac{2+r^2}{r^2}$.

(ii) **(BBP for additional uninformative 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 $\mathbf{L}$, which is approximated by

$$\lambda_+ := -\frac{2f'(\tau)}{f(\tau)}\rho_+ + \frac{f(0) - f(\tau) + \tau f'(\tau)}{f(\tau)}, \quad (\text{F.3})$$

in probability, where $\rho_+ = \ell_+(\psi - \frac{1}{1-\ell_+})$.

(iii) **(Eigenvector alignment.)** If ρ is such an isolated eigenvalue of $\mathbf{L}$ corresponding to (F.2), let Π_ρ denote the orthogonal projection onto the two-dimensional eigenspace corresponding to ρ . Then we have, in probability,

$$\frac{1}{d}\mathbf{J}^\top \Pi_\rho \mathbf{J} \rightarrow \frac{1}{1 + \frac{r^2}{2}} \left(\psi - \frac{4}{r^4} \right) \frac{r^2}{16} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2 \quad (\text{F.4})$$

where $\otimes$ is the Kronecker product. In particular, $\frac{1}{n}\mathbf{y}^\top \Pi_\rho \mathbf{y} \xrightarrow{\mathbb{P}} 0$, where $\mathbf{y}$ denotes the XOR-labels.

Remark 52 The trivial eigenvalue n , and the possible additional isolated eigenvalue of $\mathbf{L}$ corresponding to λ_+ carry no relevant clustering information. The other non-trivial outlier eigenvalues of $\mathbf{L}$ characterized in Theorem 51 (i) turn out to be the only spikes that carry relevant clustering information. However, similar to the eigenvectors of the CK matrix in Theorem 3, the last statement of Theorem 51 implies that the outlier eigenvectors of the Laplacian $\mathbf{L}$ are asymptotically orthogonal to the XOR-labels $\mathbf{y}$.

Proof of Theorem 51. The common covariance $\mathbf{C} = \mathbf{I}_d$ of the XOR dataset greatly simplifies the analysis of the spectrum of $\mathbf{L}$. We apply Corollary 1 from Couillet and Benaych-Georges (2016), which gives the isolating eigenvalues of $\mathbf{L}$ in terms of the isolated eigenvalues of $\mathbf{C} + \mathbf{M}\mathcal{D}(\mathbf{c})\mathbf{M}^\top$, where $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4]$, $\mathbf{c} = (\frac{1}{4})_{\alpha=1}^4$ is the vector of class proportions, and $\mathcal{D}(\cdot)$ is the diagonal operator. For the XOR dataset, this matrix becomes

$$\mathbf{C} + \mathbf{M}\mathcal{D}(\mathbf{c})\mathbf{M}^\top = \mathbf{I}_d + \frac{1}{4} \sum_{k=1}^4 \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top = \mathbf{I}_d + \frac{r^2}{d} \begin{bmatrix} \mathbf{1}_{\frac{d}{2}} \mathbf{1}_{\frac{d}{2}}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\frac{d}{2}} \mathbf{1}_{\frac{d}{2}}^\top \end{bmatrix}. \quad (\text{F.5})$$

The eigenvalues of (F.5) are $\ell = 1 + \frac{r^2}{2}$ with multiplicity two, and the rest are equal to 1. The phase transition in Corollary 1 of Couillet and Benaych-Georges (2016) can therefore be restated as $|\ell - 1| = \frac{r^2}{2} > \sqrt{\psi^{-1}}$. In this case, by Corollary 1 of Couillet and Benaych-Georges (2016) the isolating eigenvalues of $\mathbf{L}$ are asymptotically well approximated via (F.2) in probability, where

$\rho = \psi(1 + \frac{r^2}{2}) + \frac{2+r^2}{r^2}$. Similarly, define $\ell_+ = \frac{5f'(\tau)}{4f(\tau)} - \frac{f''(\tau)}{f'(\tau)}$. The additional isolating eigenvalue of $\mathbf{L}$ corresponding to $\rho_+ = \ell_+(\psi - \frac{1}{1-\ell_+})$ emerges if $(1 - \ell_+)^2 > \psi^{-1}$, in which case, the corresponding outlier eigenvalue of $\mathbf{L}$ converges to, in probability, λ_+ defined in (F.3).

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

$$\Upsilon_1 = \frac{1}{\sqrt{d/2}} \begin{pmatrix} \mathbf{1}_{d/2} \\ \mathbf{0}_{d/2} \end{pmatrix}, \quad \Upsilon_2 = \frac{1}{\sqrt{d/2}} \begin{pmatrix} \mathbf{0}_{d/2} \\ \mathbf{1}_{d/2} \end{pmatrix}. \quad (\text{F.6})$$

Let $\Upsilon_\rho = (\Upsilon_1, \Upsilon_2) \in \mathbb{R}^{d \times 2}$ be the column-concatenated eigenvectors, then we apply Corollary 2 of Couillet and Benaych-Georges (2016) to obtain

$$\frac{1}{d} \mathbf{J}^\top \mathbf{\Pi}_\rho \mathbf{J} = \frac{1}{\ell} \left(\psi - \frac{1}{(1-\ell)^2} \right) \mathcal{D}(c) \mathbf{M}^\top \Upsilon_\rho \Upsilon_\rho^\top \mathbf{M} \mathcal{D}(c) + o_{\mathbb{P}}(1). \quad (\text{F.7})$$

Or equivalently,

$$\frac{1}{d} \mathbf{J}^\top \mathbf{\Pi}_\rho \mathbf{J} \rightarrow \frac{r^2}{16 + 8r^2} \left(\psi - \frac{4}{r^4} \right) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2 \quad (\text{F.8})$$

in probability. Applying (F.8) to $\mathbf{y}$, we have

$$\frac{1}{d} \mathbf{y}^\top \mathbf{\Pi}_\rho \mathbf{y} \rightarrow \frac{r^2}{16 + 8r^2} \left(\psi - \frac{4}{r^4} \right) \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}^\top \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2 \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = 0$$

in probability. Hence, we complete the proof of this theorem. $\blacksquare$

Appendix G. Proof for Large-SNR Regime

This appendix presents the proof for Theorem 7. Throughout this section, we denote $r = r_0 n^{1/4}$, for some constant $r_0 \in (0, \infty)$, and we set $\theta_{\text{SNR}} := r \sqrt{\frac{n}{2d}}$. Denote that

$$\tau = \frac{c_\sigma^2}{2} \psi, \quad \kappa_0 = \lim_{n \rightarrow \infty} \frac{\theta_{\text{SNR}}^4}{n} = \frac{r_0^4 \psi^2}{4}, \quad \alpha := c_\sigma \sqrt{\kappa_0}, \quad \beta := \beta_{\text{quad}} := \alpha^2 = c_\sigma^2 \kappa_0. \quad (\text{G.1})$$

We write

$$\bar{\mathbf{y}} := \frac{\mathbf{y}}{\sqrt{n}}, \quad \mathbf{P}_{\text{lin}} := \mathbf{v}_1 \mathbf{v}_1^\top + \mathbf{v}_2 \mathbf{v}_2^\top, \quad \mathbf{P}_\perp := \mathbf{I}_n - \mathbf{P}_{\text{lin}}, \quad \mathbf{K}_\perp := \mathbf{P}_\perp \mathbf{K} \mathbf{P}_\perp. \quad (\text{G.2})$$

G.1. Projected QE Model

Lemma 53 Recall $\mathbf{Y}_{\text{QE}}$ defined by (A.1). Let $\mathbf{K}_{\text{QE}, \perp} := (\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp)^\top (\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp)$. Then

$$\|\mathbf{Y} \mathbf{P}_\perp - \mathbf{Y}_{\text{QE}} \mathbf{P}_\perp\| = O_{\prec}(n^{-1/4}), \quad \|\mathbf{K}_\perp - \mathbf{K}_{\text{QE}, \perp}\| = O_{\prec}(n^{-1/4}). \quad (\text{G.3})$$

Moreover,

$$\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp = \mathbf{Y}_0 \mathbf{P}_\perp + \frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N}} \left(\mathbf{g}_1^{\odot 2} \mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2} \mathbf{v}_2^{\odot 2\top} \right), \quad (\text{G.4})$$

because $\mathbf{P}_\perp \mathbf{v}_k = 0$ and, in the XOR basis, $\mathbf{P}_\perp \mathbf{v}_k^{\odot 2} = \mathbf{v}_k^{\odot 2}$.

Proof. The QE error in (G.3) follows from Proposition 10 after multiplication on the right by the contraction $\mathbf{P}_\perp$. The linear QE term vanishes after the projection because its right directions are $\mathbf{v}_1, \mathbf{v}_2$. The projected quadratic term has bounded operator norm since $\|\mathbf{g}_k^{\odot 2}\| = O_\prec(\sqrt{N})$, $\|\mathbf{v}_k^{\odot 2}\| = O(n^{-1/2})$, and $\theta_{\text{SNR}}^2 = O(n^{1/2})$. Also $\|\mathbf{Y}_0 \mathbf{P}_\perp\| = O_\prec(1)$ by the null CK bounds used in Appendix E. Hence $\|\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp\| = O_\prec(1)$, and

$$\|\mathbf{K}_\perp - \mathbf{K}_{\text{QE}, \perp}\| \leq (\|\mathbf{Y} \mathbf{P}_\perp\| + \|\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp\|) \|\mathbf{Y} \mathbf{P}_\perp - \mathbf{Y}_{\text{QE}} \mathbf{P}_\perp\| = O_\prec(n^{-1/4}).$$

This proves the lemma. $\blacksquare$

Lemma 54 Let $\mathbf{G} := \mathbf{Y}_0 - \frac{1}{\sqrt{N}} \mathbf{1}_N \mathbf{m}^\top$ and $\mathbf{b}_0 := \frac{\mathbf{P}_\perp \mathbf{m}}{\|\mathbf{P}_\perp \mathbf{m}\|}$ when $\|\mathbf{P}_\perp \mathbf{m}\| > 0$. Define the centered square vectors

$$\tilde{\mathbf{g}}_k := \mathbf{g}_k^{\odot 2} - \mathbf{1}_N, \quad \mathbf{h}_u := \tilde{\mathbf{g}}_1 + \tilde{\mathbf{g}}_2, \quad \mathbf{h}_y := \tilde{\mathbf{g}}_1 - \tilde{\mathbf{g}}_2,$$

and

$$\mathbf{a}_0 := \frac{\mathbf{1}_N}{\sqrt{N}}, \quad \mathbf{a}_u := \frac{\mathbf{h}_u}{\|\mathbf{h}_u\|}, \quad \mathbf{a}_y := \frac{\mathbf{h}_y}{\|\mathbf{h}_y\|}.$$

Then, in operator norm,

$$\mathbf{Y}_{\text{QE}} \mathbf{P}_\perp = \mathbf{G} \mathbf{P}_\perp + \mathbf{a}_0 (\sqrt{\tau} \mathbf{b}_0^\top + \alpha \mathbf{u}^\top) + \alpha \mathbf{a}_u \mathbf{u}^\top + \alpha \mathbf{a}_y \bar{\mathbf{y}}^\top + o_{\mathbb{P}}(1). \quad (\text{G.5})$$

Furthermore,

$$\|\mathbf{P}_\perp \mathbf{m}\|^2 \xrightarrow{\mathbb{P}} \tau, \quad \mathbf{b}_0^\top \mathbf{u} = o_{\mathbb{P}}(1), \quad \mathbf{b}_0^\top \bar{\mathbf{y}} = o_{\mathbb{P}}(1), \quad \mathbf{u}^\top \bar{\mathbf{y}} = 0,$$

and the left directions $\mathbf{a}_0, \mathbf{a}_u, \mathbf{a}_y$ are asymptotically orthonormal.

Proof. Recall $\mathbf{u} = \frac{1}{\sqrt{n}} \mathbf{1}_n$. Notice that

$$\mathbf{v}_1^{\odot 2} = \frac{1}{\sqrt{n}} (\mathbf{u} + \bar{\mathbf{y}}), \quad \mathbf{v}_2^{\odot 2} = \frac{1}{\sqrt{n}} (\mathbf{u} - \bar{\mathbf{y}}).$$

Therefore

$$\mathbf{g}_1^{\odot 2} \mathbf{v}_1^{\odot 2\top} + \mathbf{g}_2^{\odot 2} \mathbf{v}_2^{\odot 2\top} = \frac{1}{\sqrt{n}} \left[(\mathbf{h}_u + 2\mathbf{1}_N) \mathbf{u}^\top + \mathbf{h}_y \bar{\mathbf{y}}^\top \right]. \quad (\text{G.6})$$

The deterministic term in (G.6) gives

$$\frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N}} \cdot \frac{2}{\sqrt{n}} \mathbf{1}_N \mathbf{u}^\top = \frac{\theta_{\text{SNR}}^2 c_\sigma}{\sqrt{n}} \mathbf{a}_0 \mathbf{u}^\top = \alpha \mathbf{a}_0 \mathbf{u}^\top + o_{\mathbb{P}}(1),$$

because $\theta_{\text{SNR}}^2 c_\sigma / \sqrt{n} \rightarrow c_\sigma \sqrt{\kappa_0} = \alpha$ and $\alpha^2 = \beta$ where α and β are defined in (G.1). Similarly,

$$\frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N} \sqrt{n}} \mathbf{h}_u \mathbf{u}^\top = \alpha \mathbf{a}_u \mathbf{u}^\top + o_{\mathbb{P}}(1), \quad \frac{\theta_{\text{SNR}}^2 c_\sigma}{2\sqrt{N} \sqrt{n}} \mathbf{h}_y \bar{\mathbf{y}}^\top = \alpha \mathbf{a}_y \bar{\mathbf{y}}^\top + o_{\mathbb{P}}(1),$$

since

$$N^{-1} \|\mathbf{h}_u\|^2 = 4 + o_{\mathbb{P}}(1), \quad N^{-1} \|\mathbf{h}_y\|^2 = 4 + o_{\mathbb{P}}(1), \quad N^{-1} \mathbf{h}_u^\top \mathbf{h}_y = o_{\mathbb{P}}(1).$$

The null mean term is

$$\frac{1}{\sqrt{N}} \mathbf{1}_N (\mathbf{P}_\perp \mathbf{m})^\top = \mathbf{a}_0 \|\mathbf{P}_\perp \mathbf{m}\| \mathbf{b}_0^\top = \sqrt{\tau} \mathbf{a}_0 \mathbf{b}_0^\top + o_{\mathbb{P}}(1),$$

by Lemma 37; the same argument as in that lemma, with the deterministic vector $\bar{\mathbf{y}}$, gives $\bar{\mathbf{y}}^\top \mathbf{m} = o_{\mathbb{P}}(1)$. Since $\mathbf{m}$ is also asymptotically orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \mathbf{u}$, projection by $\mathbf{P}_\perp$ does not change its norm at first order and $\mathbf{b}_0^\top \mathbf{u} = \mathbf{b}_0^\top \bar{\mathbf{y}} = o_{\mathbb{P}}(1)$. Finally $\mathbf{a}_0^\top \mathbf{a}_u = \mathbf{a}_0^\top \mathbf{a}_y = o_{\mathbb{P}}(1)$ because $\mathbf{h}_u, \mathbf{h}_y$ are centered, and $\mathbf{a}_u^\top \mathbf{a}_y = o_{\mathbb{P}}(1)$ by the covariance estimate. $\blacksquare$

G.2. The Covariance Spike and the Projected Resolvents

Recall $\mathbf{P}_\perp$ defined in (G.2) and $\mathbf{G}$ defined in Lemma (54). Set

$$\mathbf{K}_0 := (\mathbf{G}\mathbf{P}_\perp)^\top (\mathbf{G}\mathbf{P}_\perp). \quad (\text{G.7})$$

The population covariance of the centered rows of $\mathbf{G}\mathbf{P}_\perp$ has the same rank-one spike $\tau \mathbf{u} \mathbf{u}^\top$ as in Appendix E, because $\mathbf{u} \perp \mathbf{v}_1, \mathbf{v}_2$. Thus, conditional on $\mathbf{Z}$,

$$\mathbf{P}_\perp \Sigma \mathbf{P}_\perp = \mathbf{P}_\perp \Sigma_0 \mathbf{P}_\perp + \tau \mathbf{u} \mathbf{u}^\top + o_{\mathbb{P}}(1).$$

The projection $\mathbf{P}_\perp$ does not change the limiting ESD of bulk covariance, hence we have the following lemma.

Lemma 55 (BBP for $\mathbf{K}_0$) *The ESD of $\mathbf{K}_0$ in (G.7) converges weakly in probability to μ . If $\tau > \tau_{\text{crit}}$, let Λ_τ be the population outlier from (E.13). If in addition $z'(-1/\Lambda_\tau) > 0$, then $\mathbf{K}_0$ has one separated outlier*

$$\hat{\lambda}_\tau^{(0)} = z\left(-\frac{1}{\Lambda_\tau}\right) + o_{\mathbb{P}}(1).$$

If $\hat{\mathbf{u}}_\tau$ is its unit right eigenvector, then $|\hat{\mathbf{u}}_\tau^\top \mathbf{u}|^2 \xrightarrow{\mathbb{P}} \varphi\left(-\frac{1}{\Lambda_\tau}\right)$. For every deterministic unit vector $\mathbf{b}$ with $\mathbf{b}^\top \mathbf{u} = o(1)$, we have $|\hat{\mathbf{u}}_\tau^\top \mathbf{b}| = o_{\mathbb{P}}(1)$. In particular the covariance outlier is asymptotically orthogonal to $\mathbf{b}_0$ and $\bar{\mathbf{y}}$, but it is generally not orthogonal to the additive quadratic right direction $\mathbf{u}$.

Proof. This is Lemma 32 applied to the projected matrix. The assumptions used there are unchanged by removing the two deterministic directions $\mathbf{v}_1, \mathbf{v}_2$, since $\mathbf{u}, \mathbf{b}_0, \bar{\mathbf{y}}$ are orthogonal to $\mathbf{v}_1, \mathbf{v}_2$ up to $o_{\mathbb{P}}(1)$, and the removed subspace has fixed dimension. The last statement follows from the general alignment formula (E.19) in Lemma 32. $\blacksquare$

Lemma 56 (Resolvent limits retaining the covariance pole) *Let $\mathcal{D} \subset \mathbb{C} \setminus (\text{supp}(\mu) \cup \{0\})$ be a compact spectral domain. For $\lambda \in \mathcal{D}$, set*

$$\mathbf{Q}_R(\lambda) := (\mathbf{K}_0 - \lambda \mathbf{I}_n)^{-1}, \quad \mathbf{Q}_L(\lambda) := ((\mathbf{G}\mathbf{P}_\perp)(\mathbf{G}\mathbf{P}_\perp)^\top - \lambda \mathbf{I}_N)^{-1}, \quad s = s(\lambda) := \tilde{m}_\mu(\lambda).$$

Let $\mathbf{A} := [\mathbf{a}_0, \mathbf{a}_u, \mathbf{a}_y]$ and $\mathbf{B} := [\mathbf{b}_0, \mathbf{u}, \bar{\mathbf{y}}]$. Uniformly on $\mathcal{D}$ such that $1 - \tau T(s(\lambda)) \neq 0$, we have

$$\mathbf{A}^\top \mathbf{Q}_L(\lambda) \mathbf{A} = s(\lambda) \mathbf{I}_3 + o_{\mathbb{P}}(1), \quad (\text{G.8})$$

and

$$\mathbf{B}^\top \mathbf{Q}_R(\lambda) \mathbf{B} = m_\mu(\lambda) \begin{pmatrix} 1 & 0 & 0 \\ 0 & (1 - \tau T(s(\lambda)))^{-1} & 0 \\ 0 & 0 & 1 \end{pmatrix} + o_{\mathbb{P}}(1). \quad (\text{G.9})$$

Proof. The left-resolvent estimate in (G.8) is the same argument as Lemma 39. The possible covariance outlier does not contribute to the quadratic forms because its left singular vector is asymptotically orthogonal to $\mathbf{a}_0, \mathbf{a}_u, \mathbf{a}_y$, as shown in Lemma 43. The random vectors $\mathbf{a}_u, \mathbf{a}_y$ are handled by conditioning on the independent copy of the centered bulk and applying the Hanson–Wright concentration inequality (Götze et al., 2021).

For the right resolvent in (G.9), the deterministic equivalent in Lemma 41 for a general sample covariance matrix gives

$$\mathbf{b}^\top \mathbf{Q}_R(\lambda) \mathbf{b}' = \mathbf{b}^\top (-\lambda s \mathbf{\Sigma} - \lambda \mathbf{I}_n)^{-1} \mathbf{b}' + o_{\mathbb{P}}(1).$$

Using $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \tau \mathbf{u} \mathbf{u}^\top + o_{\mathbb{P}}(1)$ and the Sherman–Morrison formula, with $\mathbf{A}_0(\lambda) := -\lambda(\mathbf{I}_n + s \mathbf{\Sigma}_0)$, we get

$$(\mathbf{A}_0 - \lambda s \tau \mathbf{u} \mathbf{u}^\top)^{-1} = \mathbf{A}_0^{-1} + \frac{\lambda s \tau}{1 - \lambda s \tau \mathbf{u}^\top \mathbf{A}_0^{-1} \mathbf{u}} \mathbf{A}_0^{-1} \mathbf{u} \mathbf{u}^\top \mathbf{A}_0^{-1}.$$

The isotropic law for $\mathbf{\Sigma}_0$ yields $\mathbf{u}^\top \mathbf{A}_0^{-1} \mathbf{u} = m_\mu(\lambda) + o_{\mathbb{P}}(1)$. Recalling $T(s) = \lambda s m_\mu(\lambda)$, we have

$$\mathbf{u}^\top \mathbf{Q}_R(\lambda) \mathbf{u} = \frac{m_\mu(\lambda)}{1 - \tau T(s(\lambda))} + o_{\mathbb{P}}(1).$$

For $\mathbf{b} \in \{\mathbf{b}_0, \bar{\mathbf{y}}\}$, $\mathbf{b}^\top \mathbf{u} = o_{\mathbb{P}}(1)$. Thus the rank-one Sherman–Morrison term is negligible and $\mathbf{b}^\top \mathbf{Q}_R(\lambda) \mathbf{b} = m_\mu(\lambda) + o_{\mathbb{P}}(1)$; and all cross terms are $o_{\mathbb{P}}(1)$ by the same argument. $\blacksquare$

G.3. Master Determinant Equation for Coupled Outliers

Define the deterministic 3×3 coefficient matrix

$$\mathbf{C} := \begin{pmatrix} \sqrt{\tau} & \alpha & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}, \quad (\text{G.10})$$

so that the finite-rank perturbation in (G.5) is

$$\mathbf{P} := \mathbf{A} \mathbf{C} \mathbf{B}^\top = \mathbf{a}_0 (\sqrt{\tau} \mathbf{b}_0^\top + \alpha \mathbf{u}^\top) + \alpha \mathbf{a}_u \mathbf{u}^\top + \alpha \mathbf{a}_y \bar{\mathbf{y}}^\top$$

where $\mathbf{A}$ and $\mathbf{B}$ are defined in Lemma 56. Let

$$\mathbf{Y}^\sharp := \mathbf{G} \mathbf{P}_\perp + \mathbf{P}, \quad \mathbf{K}^\sharp := \mathbf{Y}^{\sharp\top} \mathbf{Y}^\sharp.$$

By Lemmas 53 and 54, we know that

$$\|\mathbf{K}_\perp - \mathbf{K}^\sharp\| = o_{\mathbb{P}}(1). \quad (\text{G.11})$$

Lemma 57 (Coupled outlier determinant) *Let $\lambda \in \mathbb{C} \setminus (\text{supp}(\mu) \cup \{0\})$ and $s = s(\lambda) = \tilde{m}_\mu(\lambda)$. Away from the possible covariance pole of $\mathbf{K}_0$, the relative determinant for additive outliers of $\mathbf{K}^\sharp$ satisfies*

$$D_{\text{rel}}(\lambda) := \det \left(\mathbf{I}_3 - \lambda \mathbf{C}^\top (\mathbf{A}^\top \mathbf{Q}_L(\lambda) \mathbf{A}) \mathbf{C} (\mathbf{B}^\top \mathbf{Q}_R(\lambda) \mathbf{B}) \right) \quad (\text{G.12})$$

$$= (1 - \beta T(s)) \frac{1 - 2(\tau + \beta)T(s) + \tau(\tau + \beta)T(s)^2}{1 - \tau T(s)} + o_{\mathbb{P}}(1). \quad (\text{G.13})$$

Consequently, after multiplying by the covariance-spike factor of $\mathbf{K}_0$, the full outlier master equation for $\mathbf{K}^\sharp$ outside $\text{supp}(\mu)$ is

$$(1 - \beta T(s)) [1 - 2(\tau + \beta)T(s) + \tau(\tau + \beta)T(s)^2] = 0. \quad (\text{G.14})$$

Proof. Similarly as Lemma 44, the finite-rank additive determinant obtained from the linearization $\begin{pmatrix} 0 & \mathbf{Y}^\sharp \\ \mathbf{Y}^{\sharp\top} & 0 \end{pmatrix}$ gives (G.12). Substituting (G.8)–(G.9) and $T(s) = \lambda s m_\mu(\lambda)$ gives

$$D_{\text{rel}}(\lambda) = \det \left(\mathbf{I}_3 - T(s) \mathbf{C}^\top \mathbf{C} \begin{pmatrix} 1 & 0 & 0 \\ 0 & (1 - \tau T(s))^{-1} & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) + o_{\mathbb{P}}(1).$$

From (G.10), we have

$$\mathbf{C}^\top \mathbf{C} = \begin{pmatrix} \tau & \alpha\sqrt{\tau} & 0 \\ \alpha\sqrt{\tau} & 2\beta & 0 \\ 0 & 0 & \beta \end{pmatrix}.$$

Hence we get (G.13).

The denominator $1 - \tau T(s)$ is not an extra root of $\mathbf{K}^\sharp$; it is the pole of the resolvent of the null model $\mathbf{K}_0$ at the covariance outlier. In the characteristic determinant of $\mathbf{K}^\sharp$, the zero of $\det(\mathbf{K}_0 - \lambda \mathbf{I})$ at this covariance location is multiplied by the pole of D_{rel} . The product is governed by the numerator in (G.13). Equivalently, an argument-principle count on a contour around any separated cluster counts zeros of (G.14), with algebraic multiplicity. $\blacksquare$

The outlier master equation (G.14) captures the asymptotic locations of the eigenvalues of $\mathbf{K}^\sharp$ outside the bulk spectrum. Recall the functions z, φ, T from (3.11)–(3.12). For $\beta > 0$, define the label factor and the non-label coupled factor, respectively, as

$$F_y(s) := 1 - \beta T(s), \quad F_{\text{nl}}(s) := 1 - 2(\tau + \beta)T(s) + \tau(\tau + \beta)T(s)^2. \quad (\text{G.15})$$

Then (G.14) is equivalent to $F_y(s)F_{\text{nl}}(s) = 0$. The label factor $F_y = 0$ has a root at $s_y := T^{-1}(1/\beta)$, and the non-label factor $F_{\text{nl}} = 0$ has two possible roots at $s_\pm$: when $c_\sigma \neq 0$, equivalently $\tau > 0$ and $\beta > 0$,

$$t_\pm := \frac{1 \pm \sqrt{\beta/(\tau + \beta)}}{\tau}, \quad s_\pm := T^{-1}(t_\pm). \quad (\text{G.16})$$

The factor F_{nl} is the joint contribution of three effects: the population-covariance spike of the centered null matrix, the mean spike from $\mathbf{m}$, and the large-SNR quadratic $\mathbf{u}$ -channel. The original covariance outlier studied in Lemma 32 for finite SNR case with location $s_\tau = T^{-1}(1/\tau) = -1/\Lambda_\tau$ is shifted into the two roots $s_\pm$ by β_{quad} if they exist.

Lemma 58 (Cluster count and no-label property) *Let $\lambda_\star \in \mathbb{R} \setminus (\text{supp}(\mu) \cup \{0\})$ be a limiting location generated by the roots of (G.14), and let $m_\star$ be its multiplicity as in (4.12). Then $\mathbf{K}^\sharp$ has exactly $m_\star$ eigenvalues in a sufficiently small interval around $\lambda_\star$. If the interval contains no label root s_y , then the associated spectral projector $\hat{\mathbf{P}}_\star^\sharp$ satisfies $\bar{\mathbf{y}}^\top \hat{\mathbf{P}}_\star^\sharp \bar{\mathbf{y}} = o_{\mathbb{P}}(1)$. If the interval contains the label root s_y , then*

$$\bar{\mathbf{y}}^\top \hat{\mathbf{P}}_\star^\sharp \bar{\mathbf{y}} \xrightarrow{\mathbb{P}} \frac{m_\mu(z(s_y))z'(s_y)}{\beta T'(s_y)}.$$

Proof. Choose a small contour $\Gamma_\star$ enclosing $\lambda_\star$, disjoint from $\text{supp}(\mu)$, and containing no other limiting candidate. On this contour the deterministic factors in (G.14) are bounded away from zero except for the roots inside the contour. Lemma 57 and Rouché's theorem give exactly $m_\star$ zeros of the finite-dimensional outlier determinant, after the covariance-pole cancellation described above. The linearization determinant is equivalent to the eigenvalue equation of $\mathbf{K}^\sharp$ outside the bulk, so $\mathbf{K}^\sharp$ has exactly $m_\star$ eigenvalues in the interval.

For the projector statements, use the contour representation

$$\bar{\mathbf{y}}^\top \hat{\mathbf{P}}_\star^\sharp \bar{\mathbf{y}} = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \bar{\mathbf{y}}^\top (\mathbf{K}^\sharp - \lambda \mathbf{I})^{-1} \bar{\mathbf{y}} d\lambda.$$

The same Woodbury-linearization calculation used for additive outlier overlaps shows that, uniformly on $\Gamma_\star$,

$$\bar{\mathbf{y}}^\top (\mathbf{K}^\sharp - \lambda \mathbf{I})^{-1} \bar{\mathbf{y}} = \frac{m_\mu(\lambda)}{1 - \beta T(s(\lambda))} + H_y(\lambda) + o_{\mathbb{P}}(1), \quad (\text{G.17})$$

where H_y is analytic in a neighborhood of the contour. The non-label block does not appear in the leading singular term because the left direction $\mathbf{a}_y$ and the right direction $\bar{\mathbf{y}}$ are asymptotically orthogonal to the non-label directions $\mathbf{a}_0, \mathbf{a}_u$ and $\mathbf{b}_0, \mathbf{u}$.

If the contour contains no label root, the leading term in (G.17) is analytic and its contour integral is zero, giving the no-label statement. If the contour contains s_y , then $1 - \beta T(s(\lambda))$ has a simple zero at $\lambda_y = z(s_y)$, and

$$\frac{d}{d\lambda} \{1 - \beta T(s(\lambda))\} \Big|_{\lambda=\lambda_y} = -\beta \frac{T'(s_y)}{z'(s_y)}.$$

Thus

$$-\text{Res}_{\lambda=\lambda_y} \frac{m_\mu(\lambda)}{1 - \beta T(s(\lambda))} = \frac{m_\mu(z(s_y)) z'(s_y)}{\beta T'(s_y)}.$$

This proves the label projector formula. Using $T(s) = z(s) s m_\mu(z(s))$, $\beta T(s_y) = 1$, and $\varphi(s) = -s z'(s)/z(s)$, the same quantity equals $-\varphi(s_y)/(\beta^2 s_y^2 T'(s_y))$. $\blacksquare$

G.4. Proof of Theorem 7

The bulk statement follows by comparing the full CK matrix with its QE model in normalized trace norm. Proposition 10 gives $\|\mathbf{Y} - \mathbf{Y}_{\text{QE}}\| = O_{\prec}(n^{-1/4})$, hence $\|\mathbf{Y} - \mathbf{Y}_{\text{QE}}\|_F = O_{\prec}(n^{1/4})$. Since $n^{-1}\|\mathbf{Y}\|_F^2$ and $n^{-1}\|\mathbf{Y}_{\text{QE}}\|_F^2$ are tight, the normalized Schatten-1 norm difference of the Gram matrices satisfies

$$\frac{1}{n} \|\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}}\|_1 \leq \frac{\|\mathbf{Y}\|_F + \|\mathbf{Y}_{\text{QE}}\|_F}{n} \|\mathbf{Y} - \mathbf{Y}_{\text{QE}}\|_F = o_{\mathbb{P}}(1).$$

Thus $\mathbf{K}$ and $\mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}}$ have the same ESD limit. Moreover, $\mathbf{Y}_{\text{QE}}^\top \mathbf{Y}_{\text{QE}}$ differs from $\mathbf{G}^\top \mathbf{G}$ by a finite-rank matrix, and the limiting ESD is the same as $\mathbf{G}^\top \mathbf{G}$.

For the diverging linear outliers, recall that the linear QE term is $\mathbf{T}_1 = \frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}} (\mathbf{g}_1 \mathbf{v}_1^\top + \mathbf{g}_2 \mathbf{v}_2^\top)$. Since $N^{-1}\|\mathbf{g}_k\|^2 \rightarrow 1$ and $N^{-1}\mathbf{g}_1^\top \mathbf{g}_2 \rightarrow 0$ in probability, we have

$$\mathbf{T}_1^\top \mathbf{T}_1 = b_\sigma^2 \theta_{\text{SNR}}^2 (\mathbf{v}_1 \mathbf{v}_1^\top + \mathbf{v}_2 \mathbf{v}_2^\top) + o_{\mathbb{P}}(\theta_{\text{SNR}}^2).$$

All other terms have operator norm $O_{\prec}(1)$ after separating the linear part, whereas $\theta_{\text{SNR}}^2 = (r_0^2 \psi/2)n^{1/2}(1 + o(1))$. Weyl's inequality gives the top two eigenvalue asymptotics, and Davis–Kahan gives convergence of the two-dimensional projector to $\mathbf{P}_{\text{lin}}$. Since $\bar{\mathbf{y}} \perp \mathbf{v}_1, \mathbf{v}_2$, the linear outlier space has no asymptotic label alignment.

It remains to analyze the order-one spectrum of $\mathbf{K}_{\perp}$. By (G.11), it suffices to analyze $\mathbf{K}^{\sharp}$. Lemma 57 gives the full outlier master equation (G.14). The roots of the label factor are exactly $s_y = T^{-1}(1/\beta)$; the roots of the non-label factor are exactly $s_{\pm} = T^{-1}(t_{\pm})$ with $t_{\pm}$ in (G.16). By the proof of Lemma 45, a real root produces a separated eigenvalue if and only if $z'(s) > 0$. The multiplicity and collision-safe statement follows from Lemma 58. The same lemma gives the label projector formula and the no-label statement for non-label clusters. Finally, (G.11) transfers the locations and spectral projectors from $\mathbf{K}^{\sharp}$ to $\mathbf{K}_{\perp}$ by standard stability of isolated spectral projectors under an $o_{\mathbb{P}}(1)$ operator-norm perturbation.

Appendix H. Proof for Pretrained Weight Regime

In this section we prove Theorem 8. The main point is that the spiked weight creates two different effects. The linear Hermite component creates one diverging rank-one spike in the sample direction $\mathbf{s} = \mathbf{X}^{\top} \mathbf{1}_d$. After this direction is removed, the quadratic Hermite component leaves an order-one rank-one feature proportional to $\mathbf{q} = \mathbf{s}^{\odot 2}$. The master equation of the outliers is more complicated than the scalar equation in Lemma 45, since $\mathbf{q}$ has a non-negligible component along $\mathbf{u} = \mathbf{1}_n/\sqrt{n}$ and the left vector $\mathbf{a}^{\odot 2}$ is not orthogonal to $\mathbf{1}_N$.

Throughout this section, we use the following notation for the sample-direction spike and its projection:

$$\mathbf{s} := \mathbf{X}^{\top} \mathbf{1}_d, \quad \mathbf{q} := \mathbf{s}^{\odot 2}, \quad \widehat{\mathbf{s}} := \frac{\mathbf{s}}{\|\mathbf{s}\|}, \quad \Pi_s := \mathbf{I}_n - \widehat{\mathbf{s}}\widehat{\mathbf{s}}^{\top}, \quad \mathbf{K}_s := \Pi_s \mathbf{K} \Pi_s.$$

H.1. Moment Estimates and Deterministic Directions

Lemma 59 (Moments of $\mathbf{s}$ and $\mathbf{q}$) *Under the assumptions of Theorem 8, we have*

$$\frac{1}{n} \|\mathbf{s}\|^2 \xrightarrow{\mathbb{P}} \kappa, \tag{H.1}$$

$$\frac{1}{n} \|\mathbf{q}\|^2 \xrightarrow{\mathbb{P}} \eta, \tag{H.2}$$

$$\frac{1}{\sqrt{n}} \mathbf{u}^{\top} \mathbf{q} \xrightarrow{\mathbb{P}} \kappa, \quad \frac{1}{n} \mathbf{y}^{\top} \mathbf{q} \xrightarrow{\mathbb{P}} -\frac{r^2}{2}, \tag{H.3}$$

$$\frac{1}{\sqrt{n}} \mathbf{v}_2^{\top} \mathbf{s} \xrightarrow{\mathbb{P}} \frac{r}{\sqrt{2}}, \quad \mathbf{v}_1^{\top} \widehat{\mathbf{s}} = o_{\mathbb{P}}(1), \quad \frac{1}{\sqrt{n}} \mathbf{y}^{\top} \widehat{\mathbf{s}} = o_{\mathbb{P}}(1), \tag{H.4}$$

$$\mathbf{q}^{\top} \widehat{\mathbf{s}} = O_{\mathbb{P}}(1) \tag{H.5}$$

where $\kappa := 1 + \frac{r^2}{2}$ and $\eta := 3 + 3r^2 + \frac{r^4}{2}$. Consequently,

$$\|\Pi_s \mathbf{v}_1\|^2 \xrightarrow{\mathbb{P}} 1, \quad \|\Pi_s \mathbf{v}_2\|^2 \xrightarrow{\mathbb{P}} \frac{1}{\kappa}, \tag{H.6}$$

$$\frac{1}{n} \|\Pi_s \mathbf{q}\|^2 \xrightarrow{\mathbb{P}} \eta, \quad \left\langle \frac{\Pi_s \mathbf{q}}{\|\Pi_s \mathbf{q}\|}, \mathbf{u} \right\rangle^2 \xrightarrow{\mathbb{P}} \frac{\kappa^2}{\eta}, \tag{H.7}$$

$$\left\langle \frac{\mathbf{\Pi}_s \mathbf{q}}{\|\mathbf{\Pi}_s \mathbf{q}\|}, \frac{\mathbf{y}}{\sqrt{n}} \right\rangle^2 \xrightarrow{\mathbb{P}} \chi_y. \quad (\text{H.8})$$

Moreover $\mathbf{\Pi}_s \mathbf{q} / \|\mathbf{\Pi}_s \mathbf{q}\|$ is asymptotically orthogonal to $\mathbf{v}_1$, to $\mathbf{\Pi}_s \mathbf{v}_2 / \|\mathbf{\Pi}_s \mathbf{v}_2\|$, and to the mean-spike direction $\mathbf{m} / \|\mathbf{m}\|$ from Appendix E, in the sense of the right-resolvent quadratic forms used below.

Proof. Since $\mathbf{b} = \mathbf{1}_d = \sqrt{d} \mathbf{u}_2$, the deterministic shifts of $s_j = \mathbf{1}_d^\top \mathbf{x}_j$ on the four balanced XOR blocks are $0, 0, r, -r$ and the noise part is i.i.d. $\mathcal{N}(0, 1)$. Hence

$$\frac{1}{n} \sum_{j=1}^n \mathbb{E} s_j^2 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (1 + r^2) = \kappa,$$

and

$$\frac{1}{n} \sum_{j=1}^n \mathbb{E} s_j^4 = \frac{1}{2} \cdot 3 + \frac{1}{2} \cdot (3 + 6r^2 + r^4) = 3 + 3r^2 + \frac{r^4}{2} = \eta.$$

The law of large numbers gives (H.1)–(H.2). The mean of s_j^2 is 1 on the first two blocks and $1 + r^2$ on the last two blocks. Since $\mathbf{u} = \mathbf{1}_n / \sqrt{n}$ and $\mathbf{y} = (\mathbf{1}_{n/2}, -\mathbf{1}_{n/2})$, this gives (H.3). Also $\mathbf{v}_2^\top \mathbf{s} = r\sqrt{n/2} + o_{\mathbb{P}}(1)$, while $\mathbf{v}_1^\top \mathbf{s} = o_{\mathbb{P}}(1)$ and $\mathbf{y}^\top \mathbf{s} = o_{\mathbb{P}}(\sqrt{n})$, proving (H.4).

Finally, $\mathbf{q}^\top \mathbf{s} = \sum_j s_j^3$. The third moments in the third and fourth XOR blocks are $r^3 + 3r$ and $-r^3 - 3r$, so the average third moment is zero. Therefore $\sum_j s_j^3 = O_{\mathbb{P}}(\sqrt{n})$, and after division by $\|\mathbf{s}\| \asymp \sqrt{n}$ we get (H.5). The projection statements (H.6)–(H.8) follow by subtracting the $\hat{\mathbf{s}}$ -component and using (H.1)–(H.5). For example,

$$\|\mathbf{\Pi}_s \mathbf{v}_2\|^2 = 1 - |\mathbf{v}_2^\top \hat{\mathbf{s}}|^2 \rightarrow 1 - \frac{r^2}{2\kappa} = \frac{1}{\kappa},$$

and the projection changes $\mathbf{q}$ by an $O_{\mathbb{P}}(1)$ -norm vector.

The last statement follows from the same leave-one-out argument as Lemma 41. The block means of $\mathbf{q}$ have no component along $\mathbf{v}_1$ or $\mathbf{v}_2$, and the additional projection by $\mathbf{s}$ only changes normalized inner products by $o_{\mathbb{P}}(1)$. For the mean direction, $m_j = F(\|\mathbf{z}_j\|^2)$ is of order $d^{-1/2}$, while the covariance of s_j^2 with $\|\mathbf{z}_j\|^2 - 1$ is $O(d^{-1})$ per sample. Hence $\mathbf{q}^\top \mathbf{m} = O_{\mathbb{P}}(1)$, whereas $\|\mathbf{q}\| \asymp \sqrt{n}$ and $\|\mathbf{m}\| \asymp 1$. The normalized overlap is therefore $o_{\mathbb{P}}(1)$, and the resolvent version is obtained by the same deterministic-equivalent estimate. $\blacksquare$

Lemma 60 Recall $\mathbf{a}$ defined in Theorem 8. Let $\mathbf{a}_0 := \frac{\mathbf{1}_N}{\sqrt{N}}$ and $\mathbf{h} := \frac{\mathbf{a}^{\odot 2}}{\|\mathbf{a}^{\odot 2}\|}$. Then

$$\|\mathbf{a}^{\odot 2}\|^2 = \frac{3}{N}(1 + o_{\mathbb{P}}(1)), \quad \mathbf{a}_0^\top \mathbf{h} \xrightarrow{\mathbb{P}} \rho := \frac{1}{\sqrt{3}}. \quad (\text{H.9})$$

Moreover $\mathbf{h}$ is asymptotically orthogonal to the Gaussian left spike directions $\mathbf{g}_1 / \|\mathbf{g}_1\|$ and $\mathbf{g}_2 / \|\mathbf{g}_2\|$ defined in (A.1), and the left-resolvent quadratic forms satisfy, uniformly for λ in compact subsets of $\mathbb{C} \setminus (\text{supp}(\mu) \cup \{0\})$,

$$\mathbf{A}^\top \mathbf{Q}_L(\lambda) \mathbf{A} = s(\lambda) \begin{pmatrix} 1 & 0 & 0 & \rho \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \rho & 0 & 0 & 1 \end{pmatrix} + o_{\mathbb{P}}(1), \quad (\text{H.10})$$

where

$$\mathbf{A} := \left[\mathbf{a}_0, \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|}, \frac{\mathbf{g}_2}{\|\mathbf{g}_2\|}, \mathbf{h} \right], \quad \mathbf{Q}_L(\lambda) := (\mathbf{G}\mathbf{G}^\top - \lambda \mathbf{I}_N)^{-1}, \quad s(\lambda) := \tilde{m}_\mu(\lambda).$$

Proof. The first two limits are immediate from the Gaussian moments $\mathbb{E}a_i^2 = N^{-1}$ and $\mathbb{E}a_i^4 = 3N^{-2}$:

$$\sum_i a_i^4 = \frac{3}{N}(1 + o_{\mathbb{P}}(1)), \quad \mathbf{a}_0^\top \mathbf{h} = \frac{N^{-1/2} \sum_i a_i^2}{(\sum_i a_i^4)^{1/2}} \rightarrow \frac{1}{\sqrt{3}}.$$

The vector $\mathbf{a}$ is independent of $(\mathbf{W}, \mathbf{X})$, hence $\mathbf{h}$ is independent of $\mathbf{G}, \mathbf{g}_1, \mathbf{g}_2$. Its centered component $\mathbf{h} - \rho \mathbf{a}_0$ has independent sub-exponential coordinates with variance $1 - \rho^2 + o(1)$, after normalization. The same Hanson–Wright inequality and anisotropic-law argument used in Lemma 39 gives (H.10). The off-diagonal entry with $\mathbf{a}_0$ retains the deterministic overlap ρ , while all other off-diagonal entries vanish by independence and orthogonality. $\blacksquare$

H.2. QE for Spiked Weight

Recall the finite-SNR surrogate model from Proposition 34,

$$\mathbf{Y}_0^\sharp := \mathbf{G} + \frac{1}{\sqrt{N}} \mathbf{1}_N \mathbf{m}^\top + \frac{\theta_{\text{SNR}} b_\sigma}{\sqrt{N}} (\mathbf{g}_1 \mathbf{v}_1^\top + \mathbf{g}_2 \mathbf{v}_2^\top), \quad \theta_{\text{SNR}} = r \sqrt{\frac{n}{2d}}. \quad (\text{H.11})$$

Lemma 61 *Define*

$$\mathbf{Y}_{\text{sw}}^\sharp := \mathbf{Y}_0^\sharp + \frac{b_\sigma \theta}{\sqrt{N}} \mathbf{a} \mathbf{s}^\top + \frac{c_\sigma \theta^2}{2\sqrt{N}} \mathbf{a}^{\odot 2} \mathbf{q}^\top. \quad (\text{H.12})$$

Then

$$\|\mathbf{Y} - \mathbf{Y}_{\text{sw}}^\sharp\| = O_{\prec}(n^{-1/4}). \quad (\text{H.13})$$

Consequently, with

$$\mathbf{Y}_s^\sharp := (\mathbf{Y}_0^\sharp + \frac{c_\sigma \theta^2}{2} \mathbf{a}^{\odot 2} \mathbf{q}^\top) \mathbf{\Pi}_s, \quad \mathbf{K}_s^\sharp := (\mathbf{Y}_s^\sharp)^\top \mathbf{Y}_s^\sharp, \quad (\text{H.14})$$

we have

$$\|\mathbf{Y} \mathbf{\Pi}_s - \mathbf{Y}_s^\sharp\| = O_{\prec}(n^{-1/4}), \quad \|\mathbf{K}_s - \mathbf{K}_s^\sharp\| = o_{\mathbb{P}}(1). \quad (\text{H.15})$$

Proof. Write

$$\mathbf{W}_1 \mathbf{X} = \mathbf{W} \mathbf{Z} + \Delta_{\text{lin}} + \Delta_{\text{sp}}, \quad \Delta_{\text{lin}} := \mathbf{W} \mathbf{M}, \quad \Delta_{\text{sp}} := \theta \mathbf{a} \mathbf{s}^\top.$$

The entries of Δ_{lin} are $O_{\prec}(n^{-1/2})$, while $\|\Delta_{\text{sp}}\|_\infty = O_{\prec}(N^{-1/4})$. Taylor expansion at $\mathbf{W} \mathbf{Z}$, using the boundedness of $\sigma^{(3)}$, gives

$$\frac{1}{\sqrt{N}} \sigma(\mathbf{W}_1 \mathbf{X}) = \frac{1}{\sqrt{N}} \sigma(\mathbf{W} \mathbf{Z}) + \frac{1}{\sqrt{N}} \sigma'(\mathbf{W} \mathbf{Z}) \odot \Delta_{\text{lin}} + \frac{1}{\sqrt{N}} \sigma'(\mathbf{W} \mathbf{Z}) \odot \Delta_{\text{sp}} \quad (\text{H.16})$$

$$+ \frac{1}{2\sqrt{N}} \sigma''(\mathbf{W} \mathbf{Z}) \odot \Delta_{\text{sp}}^{\odot 2} + \mathbf{R}, \quad (\text{H.17})$$

where the omitted terms $(2\sqrt{N})^{-1}\sigma''(\mathbf{W}\mathbf{Z}) \odot \Delta_{\text{lin}}^{\odot 2}$, $\sqrt{N}^{-1}\sigma''(\mathbf{W}\mathbf{Z}) \odot (\Delta_{\text{lin}} \odot \Delta_{\text{sp}})$, and the third-order remainder all have operator norm $O_{\prec}(n^{-1/4})$. Indeed, the mixed term has entries of size $O_{\prec}(n^{-1/2}N^{-1/4})$, and after multiplication by $1/\sqrt{N}$ its operator norm is bounded by $\sqrt{Nn}/\sqrt{N} \cdot O_{\prec}(n^{-1/2}N^{-1/4}) = O_{\prec}(N^{-1/4})$; the remainder is analogous.

The first two terms in (H.17) are reduced to $\mathbf{Y}_0^{\sharp}$ by Proposition 34. For the spiked linear term in (H.17),

$$\frac{1}{\sqrt{N}}\sigma'(\mathbf{W}\mathbf{Z}) \odot \Delta_{\text{sp}} = \frac{b_{\sigma}\theta}{\sqrt{N}}\mathbf{a}\mathbf{s}^{\top} + \mathbf{E}_1,$$

and $\|\mathbf{E}_1\| = O_{\prec}(n^{-1/4})$. This follows by writing $\mathbf{E}_1 = (\theta/\sqrt{N})\text{diag}(\mathbf{a})(\sigma'(\mathbf{W}\mathbf{Z}) - b_{\sigma}\mathbf{1}_N\mathbf{1}_n^{\top})\text{diag}(\mathbf{s})$, using $\|\text{diag}(\mathbf{a})\| = O_{\prec}(N^{-1/2})$, $\|\sigma'(\mathbf{W}\mathbf{Z}) - b_{\sigma}\mathbf{1}\mathbf{1}^{\top}\| = O_{\prec}(\sqrt{N} + \sqrt{n})$, and $\|\text{diag}(\mathbf{s})\| = O_{\prec}(1)$. The quadratic spiked term in (H.17) satisfies

$$\frac{1}{2\sqrt{N}}\sigma''(\mathbf{W}\mathbf{Z}) \odot \Delta_{\text{sp}}^{\odot 2} = \frac{c_{\sigma}\theta^2}{2\sqrt{N}}\mathbf{a}^{\odot 2}\mathbf{q}^{\top} + \mathbf{E}_2, \quad \|\mathbf{E}_2\| = O_{\prec}(n^{-1/4}),$$

by the same Hermite-coefficient concentration: after factoring $\text{diag}(\mathbf{a}^{\odot 2})$ and $\text{diag}(\mathbf{q})$, the non-constant part of $\sigma''(\mathbf{W}\mathbf{Z})$ has operator norm $O_{\prec}(\sqrt{N} + \sqrt{n})$, while $\|\text{diag}(\mathbf{a}^{\odot 2})\| = O_{\prec}(N^{-1})$. Combining these estimates proves (H.13).

Right multiplication by $\mathbf{\Pi}_s$ kills the diverging linear spike exactly, $(b_{\sigma}\theta/\sqrt{N})\mathbf{a}\mathbf{s}^{\top}\mathbf{\Pi}_s = 0$, and $\theta^2/\sqrt{N} = \theta_0^2$. Hence the first estimate in (H.15) follows. The matrices $\mathbf{Y}\mathbf{\Pi}_s$ and $\mathbf{Y}_s^{\sharp}$ have operator norm $O_{\mathbb{P}}(1)$: $\mathbf{Y}_0^{\sharp}\mathbf{\Pi}_s$ is bounded by Proposition 34, and the quadratic term has norm $\frac{|c_{\sigma}|\theta_0^2}{2}\|\mathbf{a}^{\odot 2}\|\|\mathbf{\Pi}_s\mathbf{q}\| = O_{\mathbb{P}}(1)$ by Lemmas 59 and 60. Therefore, we have

$$\|\mathbf{K}_s - \mathbf{K}_s^{\sharp}\| \leq (\|\mathbf{Y}\mathbf{\Pi}_s\| + \|\mathbf{Y}_s^{\sharp}\|)\|\mathbf{Y}\mathbf{\Pi}_s - \mathbf{Y}_s^{\sharp}\| = o_{\mathbb{P}}(1).$$

■

H.3. The Compressed Covariance Spike

We analyze the projected covariance spike below similarly to the analysis of the covariance spike present in the centered null matrix $\mathbf{G}$ in Lemma 32. Let $\mathbf{\Sigma}$ denote the conditional population covariance of the unscaled centered row underlying $\mathbf{G}$, as in Appendix E. Thus

$$\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \tau\mathbf{u}\mathbf{u}^{\top} + o_{\mathbb{P}}(1), \quad \mathbf{\Sigma}_0 = (1 - b_{\sigma}^2)\mathbf{I}_n + b_{\sigma}^2\mathbf{Z}^{\top}\mathbf{Z}, \quad \tau = \frac{c_{\sigma}^2}{2}\psi.$$

Define

$$\hat{\mathbf{u}}_s := \frac{\mathbf{\Pi}_s\mathbf{u}}{\|\mathbf{\Pi}_s\mathbf{u}\|}, \quad \tau_s := \tau\|\mathbf{\Pi}_s\mathbf{u}\|^2, \quad \mathbf{K}_{0,s} := (\mathbf{G}\mathbf{\Pi}_s)^{\top}(\mathbf{G}\mathbf{\Pi}_s).$$

Lemma 62 (Covariance spike after s -deflation) *Under the assumptions of Theorem 8,*

$$\mathbf{\Pi}_s\mathbf{\Sigma}\mathbf{\Pi}_s = \mathbf{\Pi}_s\mathbf{\Sigma}_0\mathbf{\Pi}_s + \tau_s\hat{\mathbf{u}}_s\hat{\mathbf{u}}_s^{\top} + o_{\mathbb{P}}(1), \quad \tau_s = \tau + o_{\mathbb{P}}(1), \quad \|\hat{\mathbf{u}}_s - \mathbf{u}\| = o_{\mathbb{P}}(1). \quad (\text{H.18})$$

The ESD of $\mathbf{K}_{0,s}$ converges weakly in probability to μ . Its possible separated covariance outlier is governed by the same equation as in Lemma 32, namely

$$1 - \tau T(s) = 0, \quad z'(s) > 0. \quad (\text{H.19})$$

Moreover, uniformly for λ in compact subsets of $\mathbb{C} \setminus (\text{supp}(\mu) \cup \{0\})$ avoiding the possible pole $1 - \tau T(s(\lambda)) = 0$, with $\mathbf{Q}_{R,0}^{(s)}(\lambda) := (\mathbf{K}_{0,s} - \lambda \mathbf{I})^{-1}$ on $\text{Range } \mathbf{\Pi}_s$,

$$\mathbf{b}^\top \mathbf{Q}_{R,0}^{(s)}(\lambda) \mathbf{b} = m_\mu(\lambda) + o_{\mathbb{P}}(1), \quad (\text{H.20})$$

$$\widehat{\mathbf{q}}_s^\top \mathbf{Q}_{R,0}^{(s)}(\lambda) \widehat{\mathbf{q}}_s = m_\mu(\lambda) R_q(T(s(\lambda))) + o_{\mathbb{P}}(1), \quad (\text{H.21})$$

$$\mathbf{b}^\top \mathbf{Q}_{R,0}^{(s)}(\lambda) \widehat{\mathbf{q}}_s = o_{\mathbb{P}}(1), \quad (\text{H.22})$$

for every unit vector $\mathbf{b}$ among $\mathbf{b}_0 := \mathbf{m}/\|\mathbf{m}\|$, $\mathbf{v}_1$, and $\tilde{\mathbf{v}}_2 := \mathbf{\Pi}_s \mathbf{v}_2 / \|\mathbf{\Pi}_s \mathbf{v}_2\|$. Here $\widehat{\mathbf{q}}_s := \mathbf{\Pi}_s \mathbf{q} / \|\mathbf{\Pi}_s \mathbf{q}\|$ and

$$R_q(t) = 1 + \frac{\kappa^2}{\eta} \frac{\tau t}{1 - \tau t} = \frac{1 - \tau \omega_q t}{1 - \tau t}. \quad (\text{H.23})$$

Proof. By Lemma 59, $\mathbf{u}^\top \widehat{\mathbf{s}} = O_{\mathbb{P}}(n^{-1/2})$. Hence $\|\mathbf{\Pi}_s \mathbf{u}\|^2 = 1 - |\mathbf{u}^\top \widehat{\mathbf{s}}|^2 = 1 + o_{\mathbb{P}}(1)$, which proves $\tau_s = \tau + o_{\mathbb{P}}(1)$ and $\|\widehat{\mathbf{u}}_s - \mathbf{u}\| = o_{\mathbb{P}}(1)$. Multiplying $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \tau \mathbf{u} \mathbf{u}^\top + o_{\mathbb{P}}(1)$ on both sides by $\mathbf{\Pi}_s$ gives (H.18). Since $\mathbf{\Pi}_s \mathbf{\Sigma}_0 \mathbf{\Pi}_s$ differs from $\mathbf{\Sigma}_0$ by a matrix of rank at most two plus one additional zero direction, its limiting covariance ESD is still ν ; consequently the limiting ESD of $\mathbf{K}_{0,s}$ is still μ . The multiplicative BBP equation for the rank-one covariance spike in (H.18) is therefore the same as in Lemma 32, with τ_s in place of τ ; since $\tau_s \rightarrow \tau$, this gives (H.19).

It remains to record the anisotropic resolvent limits needed below. The right-resolvent deterministic equivalent for the base covariance $\mathbf{\Pi}_s \mathbf{\Sigma}_0 \mathbf{\Pi}_s$ is unchanged by the rank-one compression, because $\widehat{\mathbf{s}}$ is asymptotically orthogonal to the deterministic finite-dimensional directions considered here and the removed subspace has fixed rank. Adding the rank-one covariance spike $\tau_s \widehat{\mathbf{u}}_s \widehat{\mathbf{u}}_s^\top$ and applying the Sherman–Morrison formula yields

$$\widehat{\mathbf{u}}_s^\top \mathbf{Q}_{R,0}^{(s)}(\lambda) \widehat{\mathbf{u}}_s = \frac{m_\mu(\lambda)}{1 - \tau T(s(\lambda))} + o_{\mathbb{P}}(1), \quad (\text{H.24})$$

whereas every unit vector asymptotically orthogonal to $\widehat{\mathbf{u}}_s$ has quadratic form $m_\mu(\lambda) + o_{\mathbb{P}}(1)$, and cross forms with $\widehat{\mathbf{u}}_s$ are negligible. The vectors $\mathbf{b}_0$, $\mathbf{v}_1$, and $\tilde{\mathbf{v}}_2$ are asymptotically orthogonal to $\widehat{\mathbf{u}}_s$, giving (H.20). Furthermore, Lemma 59 gives the decomposition

$$\widehat{\mathbf{q}}_s = \frac{\kappa}{\sqrt{\eta}} \widehat{\mathbf{u}}_s + \sqrt{\omega_q} \mathbf{e}_q + o_{\mathbb{P}}(1), \quad \mathbf{e}_q \perp \widehat{\mathbf{u}}_s, \quad \|\mathbf{e}_q\| = 1,$$

with $\mathbf{e}_q$ asymptotically isotropic for the base resolvent and asymptotically orthogonal to $\mathbf{b}_0$, $\mathbf{v}_1$, $\tilde{\mathbf{v}}_2$. Combining this decomposition with (H.24) gives

$$\begin{aligned} \widehat{\mathbf{q}}_s^\top \mathbf{Q}_{R,0}^{(s)}(\lambda) \widehat{\mathbf{q}}_s &= \frac{\kappa^2}{\eta} \frac{m_\mu(\lambda)}{1 - \tau T(s(\lambda))} + \omega_q m_\mu(\lambda) + o_{\mathbb{P}}(1) \\ &= m_\mu(\lambda) \left(1 + \frac{\kappa^2}{\eta} \frac{\tau T(s(\lambda))}{1 - \tau T(s(\lambda))} \right) + o_{\mathbb{P}}(1), \end{aligned}$$

which is (H.21)–(H.23). The same orthogonal decomposition gives the cross estimate (H.22). $\blacksquare$

H.4. Proof of Giant Spike and Bulk Limit

Proof of Theorem 8(i)–(ii). The rank-one matrix $\mathbf{P}_s := \frac{b_\sigma \theta}{\sqrt{N}} \mathbf{a} \mathbf{s}^\top$ has squared singular value

$$\|\mathbf{P}_s\|^2 = \frac{b_\sigma^2 \theta^2}{N} (\mathbf{a}^\top \mathbf{a}) \|\mathbf{s}\|^2 = b_\sigma^2 \theta_0^2 \phi \kappa \sqrt{N} (1 + o_{\mathbb{P}}(1)),$$

using $\mathbf{a}^\top \mathbf{a} \rightarrow 1$, $\|\mathbf{s}\|^2/n \rightarrow \kappa$, and $n/N \rightarrow \phi$. All terms in $\mathbf{Y}_{\text{sw}}^\# - \mathbf{P}_s$ have operator norm $O_{\mathbb{P}}(1)$. Thus the top eigenvalue of $(\mathbf{Y}_{\text{sw}}^\#)^\top \mathbf{Y}_{\text{sw}}^\#$ is $\|\mathbf{P}_s\|^2 + o_{\mathbb{P}}(\sqrt{N})$, and the corresponding eigenvector has overlap $1 - o_{\mathbb{P}}(1)$ with $\hat{\mathbf{s}}$. The feature error in Lemma 61 changes this eigenvalue by $o_{\mathbb{P}}(\sqrt{N})$, and Davis–Kahan Theorem gives (4.17). Since $\mathbf{y}^\top \hat{\mathbf{s}} = O_{\mathbb{P}}(1)$, the normalized label overlap tends to zero. If $b_\sigma = 0$, $\mathbf{P}_s$ is absent and no diverging spike is produced by this mechanism.

For the bulk, $\mathbf{Y}_{\text{sw}}^\#$ differs from the finite-SNR random-weight surrogate $\mathbf{Y}_0^\#$ by two rank-one feature perturbations in (H.12). Hence the corresponding Gram matrices differ by a finite-rank matrix. The error $\mathbf{Y} - \mathbf{Y}_{\text{sw}}^\#$ contributes negligibly to the limiting ESD: its product with the bounded part has normalized trace norm $o_{\mathbb{P}}(1)$, and its product with the single giant rank-one term has rank at most two. Since the random-weight surrogate has limiting ESD μ by Theorem 3, so does $\mathbf{K}$. Compressing by Π_s changes at most one eigenvalue, so the same bulk limit holds for $\mathbf{K}_s$. $\blacksquare$

H.5. Proof of Order-one Outliers and Label Alignment

Let

$$\hat{\mathbf{q}}_s := \frac{\Pi_s \mathbf{q}}{\|\Pi_s \mathbf{q}\|}, \quad \tilde{\mathbf{v}}_2 := \frac{\Pi_s \mathbf{v}_2}{\|\Pi_s \mathbf{v}_2\|}.$$

By Lemmas 59 and 60, the finite-rank part of $\mathbf{Y}_s^\#$ can be written, up to an $o_{\mathbb{P}}(1)$ operator-norm error, as

$$\mathbf{Y}_s^\# = \mathbf{G} \Pi_s + \sqrt{\tau} \mathbf{a}_0 \mathbf{b}_0^\top + \sqrt{\beta_{\text{lin}}} \mathbf{a}_1 \mathbf{v}_1^\top + \sqrt{\frac{\beta_{\text{lin}}}{\kappa}} \mathbf{a}_2 \tilde{\mathbf{v}}_2^\top + \sqrt{\beta_q} \mathbf{h} \hat{\mathbf{q}}_s^\top + o_{\mathbb{P}}(1), \quad (\text{H.25})$$

where

$$\mathbf{b}_0 := \frac{\mathbf{m}}{\|\mathbf{m}\|}, \quad \mathbf{a}_1 := \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|}, \quad \mathbf{a}_2 := \frac{\mathbf{g}_2}{\|\mathbf{g}_2\|}.$$

The factor $\beta_{\text{lin}}/\kappa$ in the third term is the loss of the $\mathbf{v}_2$ -component caused by deflating the giant $\mathbf{s}$ -direction. Set

$$\Theta := \text{diag}(\sqrt{\tau}, \sqrt{\beta_{\text{lin}}}, \sqrt{\beta_{\text{lin}}/\kappa}, \sqrt{\beta_q}), \quad \mathbf{A} := [\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2, \mathbf{h}], \quad \mathbf{B} := [\mathbf{b}_0, \mathbf{v}_1, \tilde{\mathbf{v}}_2, \hat{\mathbf{q}}_s].$$

Let $\mathbf{K}_0^{(s)} := (\mathbf{G} \Pi_s)^\top (\mathbf{G} \Pi_s)$, and let $\mathbf{Q}_R^{(s)}(\lambda) := (\mathbf{K}_0^{(s)} - \lambda \mathbf{I})^{-1}$ on the range of Π_s . By Lemma 62, the compressed null matrix keeps the covariance spike, now in the direction $\hat{\mathbf{u}}_s = \Pi_s \mathbf{u} / \|\Pi_s \mathbf{u}\|$ and with strength $\tau + o_{\mathbb{P}}(1)$. Thus, with $t = T(s(\lambda)) = \lambda s(\lambda) m_\mu(\lambda)$,

$$\mathbf{B}^\top \mathbf{Q}_R^{(s)}(\lambda) \mathbf{B} = m_\mu(\lambda) \text{diag}(1, 1, 1, R_q(t)) + o_{\mathbb{P}}(1), \quad (\text{H.26})$$

where $R_q(t)$ is given by (H.23). From the proof of Lemma 62, we know that $\hat{\mathbf{q}}_s$ has squared overlap κ^2/η with the projected covariance-spike direction $\hat{\mathbf{u}}_s$, while $\mathbf{b}_0$, $\mathbf{v}_1$, and $\tilde{\mathbf{v}}_2$ are asymptotically

orthogonal to it. Hence only the quadratic spiked-weight direction sees the covariance pole. Together with (H.10), the finite-rank determinant for order-one outliers of $\mathbf{K}_s^\sharp$ is

$$D_{\text{rel}}(\lambda) = \det \left(\mathbf{I}_4 - t \Theta \begin{pmatrix} 1 & 0 & 0 & \rho \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \rho & 0 & 0 & 1 \end{pmatrix} \Theta \text{diag}(1, 1, 1, R_q(t)) \right) + o_{\mathbb{P}}(1). \quad (\text{H.27})$$

Expanding the determinant gives

$$D_{\text{rel}}(\lambda) = (1 - \beta_{\text{lin}} t) \left(1 - \frac{\beta_{\text{lin}}}{\kappa} t \right) [(1 - \tau t)(1 - \beta_q R_q(t)t) - \rho^2 \tau \beta_q R_q(t)t^2] + o_{\mathbb{P}}(1). \quad (\text{H.28})$$

The denominator $1 - \tau t$ in $R_q(t)$ is the pole of the covariance outlier already present in the base matrix $\mathbf{K}_0^{(s)}$, not an additional root. As in the collision-safe argument for the uninformative cluster in Appendix E, the characteristic determinant is obtained by multiplying by the covariance factor $1 - \tau t$. Using $\rho^2 = 1/3$,

$$(1 - \tau t) D_{\text{rel}}(\lambda) = (1 - \beta_{\text{lin}} t) \left(1 - \frac{\beta_{\text{lin}}}{\kappa} t \right) F_q(t) + o_{\mathbb{P}}(1), \quad (\text{H.29})$$

where F_q is exactly (4.16). This is the desired simplification of the augmented determinant.

Proof of Theorem 8(iii). Choose a contour in $\mathbb{C}$ enclosing one candidate location λ_* and disjoint from $\text{supp}(\mu) \cup \{0\}$ and from all other candidate locations from (4.18). On this contour, (H.29) is a uniform approximation to the finite-dimensional determinant. Rouché's theorem gives the same number of zeros as the deterministic product in (H.29), counted with algebraic multiplicity. These zeros are equivalent to eigenvalues of the linearized matrix, hence to eigenvalues of $\mathbf{K}_s^\sharp$, outside the bulk. Lemma 61 transfers the conclusion to $\mathbf{K}_s$. The admissibility condition $z'(s) > 0$ is the usual BBP threshold from Lemma 45; roots on branches with $z'(s) \leq 0$ do not separate from the limiting support μ . ■

Proof of Theorem 8(iv). We now compute the resolvent in the label direction. Since $\mathbf{y}^\top \hat{\mathbf{s}} = O_{\mathbb{P}}(1)$, replacing $\mathbf{y}/\sqrt{n}$ by $\Pi_s \mathbf{y}/\|\Pi_s \mathbf{y}\|$ changes normalized overlaps by $o_{\mathbb{P}}(1)$. The linear spike directions $\mathbf{v}_1$ and $\tilde{\mathbf{v}}_2$ are asymptotically orthogonal to $\mathbf{y}/\sqrt{n}$, and the same is true of the mean direction $\mathbf{b}_0$. More precisely, uniformly for λ on any fixed contour away from $\text{supp}(\mu) \cup \{0\}$,

$$\frac{1}{\sqrt{n}} \mathbf{B}^\top \mathbf{Q}_R^{(s)}(\lambda) \mathbf{y} = m_\mu(\lambda) \sqrt{\chi_y} \mathbf{e}_4 + o_{\mathbb{P}}(1). \quad (\text{H.30})$$

The possible covariance pole in $\mathbf{Q}_R^{(s)}$ does not change this vector, because $\mathbf{y} \perp \mathbf{u}$. Thus the first three right coordinates $\mathbf{b}_0, \mathbf{v}_1, \tilde{\mathbf{v}}_2$ cannot create a pole in the label resolvent. In particular, the roots of the two linear factors have zero label residue, and a pure mean-spike cluster has zero label residue. Once $\beta_q > 0$, however, the mean/covariance cluster is no longer a separate pure $\mathbf{b}_0$ -cluster; it is absorbed into the coupled factor F_q . Any label mass at such a shifted location is caused by its mixing with the quadratic direction $\hat{\mathbf{q}}_s$, not by $\mathbf{b}_0$ itself.

The Woodbury formula for the finite-rank model (H.25), combined with (H.10), (H.26), and (H.30), gives uniformly away from the bulk

$$\frac{1}{n} \mathbf{y}^\top (\mathbf{K}_s^\sharp - \lambda \mathbf{I})^{-1} \mathbf{y} = H_y(\lambda) + m_\mu(\lambda) \frac{\beta_q \chi_y t (1 - \tau t) (1 - \frac{2}{3} \tau t)}{F_q(t)} + o_{\mathbb{P}}(1), \quad (\text{H.31})$$

where H_y is analytic in any neighborhood not intersecting $\text{supp}(\mu) \cup \{0\}$, and $t = T(s(\lambda))$. The absence of the factors $1 - \beta_{\text{lin}} t$ and $1 - \beta_{\text{lin}} t / \kappa$ from the denominator in (H.31) is the desired zero-alignment statement for the $\mathbf{v}_1$ - and $\tilde{\mathbf{v}}_2$ -outliers. The absence of a separate $1 - \tau t$ denominator is the corresponding zero-alignment statement for a pure mean/covariance cluster. The only non-analytic denominator relevant to the label is $F_q(t)$.

To see the numerator in (H.31), restrict the Woodbury correction to the coupled coordinates $(\mathbf{a}_0, \mathbf{h})$. The relevant 2×2 inverse has determinant $F_q(t)/(1 - \tau t)$, and its $(2, 2)$ -entry contributes $\beta_q[1 - (2/3)\tau t](1 - \tau t)/F_q(t)$. Multiplication by the squared right overlap χ_y gives the singular term.

Let $\Gamma_\star$ be a positively oriented contour around $I_\star$. The spectral projector satisfies

$$\frac{1}{n} \mathbf{y}^\top \hat{\mathbf{P}}_\star^{(s)} \mathbf{y} = -\frac{1}{2\pi i} \oint_{\Gamma_\star} \frac{1}{n} \mathbf{y}^\top (\mathbf{K}_s - \lambda \mathbf{I})^{-1} \mathbf{y} d\lambda.$$

The approximation $\|\mathbf{K}_s - \mathbf{K}_s^\sharp\| = o_{\mathbb{P}}(1)$ allows us to replace $\mathbf{K}_s$ by $\mathbf{K}_s^\sharp$. The analytic term H_y integrates to zero. Hence, if the contour contains no root of $F_q(T(s))$, the limit is zero. If the contour contains a single simple root $s_\star$ of $F_q(T(s))$, then

$$\left. \frac{d}{d\lambda} F_q(T(s(\lambda))) \right|_{\lambda=z(s_\star)} = F_q'(T(s_\star)) \frac{T'(s_\star)}{z'(s_\star)}.$$

Taking minus the residue of (H.31) gives

$$\Gamma_{\text{sw}}(s) := -\frac{m_\mu(z(s))z'(s)}{T'(s)} \frac{\beta_q \chi_y T(s) (1 - \tau T(s)) (1 - \frac{2}{3}\tau T(s))}{F_q'(T(s))}. \quad (\text{H.32})$$

If a contour encloses several simple F_q -roots, the same calculation sums their residues. This completes the proof of Theorem 8. $\blacksquare$

Appendix I. Proof for Quadratic Sample-size Regime

In this section we prove Theorem 9 for the CK matrix in the quadratic sample-size $n \asymp d^2$ and linear-width $N \asymp n$ regime. The CK matrix is a sample covariance matrix whose population covariance is a quadratic polynomial kernel. This produces two successive BBP transitions: one in the quadratic kernel of the data governed by the aspect ratio γ , and one from the linear-width sample covariance model governed by ϕ . Throughout this section we consider quadratic asymptotic limit of the sample size:

$$p := \frac{d(d+1)}{2}, \quad \frac{n}{p} \rightarrow \gamma \in (0, \infty), \quad r \in [0, \infty). \quad (\text{I.1})$$

Set

$$\delta := \frac{r^2}{2}, \quad \ell := \delta^2 = \frac{r^4}{4}, \quad \lambda_+(\gamma) := (1 + \sqrt{\gamma})^2. \quad (\text{I.2})$$

Let h_k be the orthonormal Hermite polynomials and define

$$a_k := \mathbb{E}[\sigma(\xi) h_k(\xi)], \quad \xi \sim \mathcal{N}(0, 1).$$

Thus $a_1 = b_\sigma$, $a_2 = \frac{c_\sigma}{\sqrt{2}}$, $\alpha_2 := a_2^2 = \frac{c_\sigma^2}{2}$, and $\alpha_0 := 1 - a_1^2 - a_2^2$.

Quadratic lift. Let $\text{svec} : \mathbb{S}^d \rightarrow \mathbb{R}^p$ denote the Frobenius-isometric symmetric vectorization. Define

$$\mathbf{q}_i := \text{svec}(\mathbf{x}_i \mathbf{x}_i^\top), \quad \mathbf{Q} := [\mathbf{q}_1, \dots, \mathbf{q}_n] \in \mathbb{R}^{p \times n}. \quad (\text{I.3})$$

Then

$$\mathbf{H} := (\mathbf{X}^\top \mathbf{X})^{\odot 2} = \mathbf{Q}^\top \mathbf{Q}. \quad (\text{I.4})$$

For the XOR model, write $\mathbf{x}_i = \mathbf{z}_i + \mathbf{m}_i$, where $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I}_d/d)$. Define the second- and first-chaos lifted matrices $\mathbf{E}_2, \mathbf{E}_1 \in \mathbb{R}^{p \times n}$ by their i -th columns, for $i \in [n]$,

$$(\mathbf{E}_2)_i := \text{svec}\left(\mathbf{z}_i \mathbf{z}_i^\top - \frac{1}{d} \mathbf{I}_d\right), \quad (\mathbf{E}_1)_i := \text{svec}\left(\mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top\right). \quad (\text{I.5})$$

Finally set

$$\mathbf{s} := \frac{1}{\sqrt{2}} \text{svec}\left(\mathbf{u}_1 \mathbf{u}_1^\top - \mathbf{u}_2 \mathbf{u}_2^\top\right), \quad \|\mathbf{s}\| = 1, \quad \mathbf{v} := \frac{\mathbf{y}}{\sqrt{n}}. \quad (\text{I.6})$$

The unprojected rank-one second-chaos spike is

$$\mathbf{Y}_0 := \mathbf{E}_2 + \frac{\delta}{\sqrt{p}} \mathbf{s} \mathbf{y}^\top = \mathbf{E}_2 + \theta_n \mathbf{s} \mathbf{v}^\top, \quad \theta_n := \delta \sqrt{\frac{n}{p}} \rightarrow \theta := \delta \sqrt{\gamma}. \quad (\text{I.7})$$

Projected nuisance space. For $s_i = \|\mathbf{x}_i\|$, define

$$\zeta_k^{(i)} := \mathbb{E}[\sigma(s_i \xi) h_k(\xi)], \quad \mathbf{D}_k := \text{diag}(\zeta_k^{(1)}, \dots, \zeta_k^{(n)}), \quad \mathbf{D} := \text{diag}(s_1, \dots, s_n), \quad (\text{I.8})$$

$$\boldsymbol{\mu}_\sigma := (\zeta_0^{(1)}, \dots, \zeta_0^{(n)})^\top, \quad \mathbf{A}_\sigma := \text{diag}\left(\frac{\zeta_1^{(i)}}{s_i}\right)_{i=1}^n, \quad \mathbf{B}_\sigma := \text{diag}\left(\frac{\zeta_3^{(i)}}{s_i}\right)_{i=1}^n. \quad (\text{I.9})$$

We define the BBP-projected CK nuisance space by

$$\mathcal{U}_\sigma^+ := \text{span}\{\mathbf{1}, \boldsymbol{\mu}_\sigma\} + \text{Range}(\mathbf{A}_\sigma \mathbf{X}^\top) + \text{Range}(\mathbf{B}_\sigma \mathbf{X}^\top) + \text{Range}(\mathbf{E}_1^\top). \quad (\text{I.10})$$

Finally, we define the projection by

$$\mathbf{P}_\sigma := \text{Proj}((\mathcal{U}_\sigma^+)^{\perp}). \quad (\text{I.11})$$

Limiting spectra and transforms. Define the limiting population law

$$\nu_q := \alpha_0 + \alpha_2 \rho_\gamma^{\text{MP}}, \quad (\text{I.12})$$

that is, the push-forward of the MP law ρ_γ^{MP} by $x \mapsto \alpha_0 + \alpha_2 x$. Let

$$\mu_q := \rho_\phi^{\text{MP}} \boxtimes \nu_q. \quad (\text{I.13})$$

Slightly different from (3.11), for this pair (ν_q, ϕ) we define the z -transform and φ -transform as

$$z(s) := -\frac{1}{s} + \phi \int \frac{t}{1+ts} \nu_q(dt), \quad \varphi(s) := -\frac{sz'(s)}{z(s)}. \quad (\text{I.14})$$

These are the transforms in (3.11) with ν replaced by ν_q in (I.12). The upper edge of ν_q is $\tau_+^\nu = \alpha_0 + \alpha_2 \lambda_+(\gamma)$.

Recall parameters in (I.1) and (I.2). For later use, define the BBP location and two maps

$$\Lambda_y := \alpha_0 + \alpha_2 \lambda_{\text{out}}(\gamma, \ell), \quad (\text{I.15})$$

$$\lambda_{\text{out}}(\gamma, \ell) := 1 + \gamma + \gamma \ell + \frac{1}{\ell}, \quad \text{Align}(\gamma, \ell) := \frac{\gamma \ell^2 - 1}{\gamma \ell (\ell + 1)}. \quad (\text{I.16})$$

I.1. The Projection Preserves the XOR Label

Lemma 63 (The projection preserves the XOR label) *Under the assumptions of Theorem 9, let P_σ be defined in (I.11). Then*

$$\text{rank}(\mathbf{I}_n - P_\sigma) = o(n), \quad \frac{1}{n} \mathbf{y}^\top P_\sigma \mathbf{y} \xrightarrow{\mathbb{P}} 1. \quad (\text{I.17})$$

Proof. We prove the rank bound first. Clearly, $\dim \text{span}\{\mathbf{1}, \boldsymbol{\mu}_\sigma\} \leq 2$, and $\text{rank}(\mathbf{A}_\sigma \mathbf{X}^\top) \leq d$, $\text{rank}(\mathbf{B}_{3,\sigma} \mathbf{X}^\top) \leq d$. It remains to control $\text{rank}(\mathbf{E}_1^\top)$. Recall that $(\mathbf{E}_1)_i = \text{svec}(\mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top)$, and in the XOR model $\mathbf{m}_i = \eta_i \frac{r}{\sqrt{d}} \mathbf{u}_{\kappa_i}$, $\eta_i \in \{\pm 1\}$, and $\kappa_i \in \{1, 2\}$. For $a \in \{1, 2\}$, define the $n \times d$ matrix $\mathbf{Z}_a^{\text{row}}$ by

$$(\mathbf{Z}_a^{\text{row}})_{i \cdot} := \mathbf{1}_{\{\kappa_i=a\}} \eta_i \frac{r}{\sqrt{d}} \mathbf{z}_i^\top.$$

Let $\mathbf{C}_E := [\mathbf{Z}_1^{\text{row}} \quad \mathbf{Z}_2^{\text{row}}] \in \mathbb{R}^{n \times 2d}$. We claim that $\text{Range}(\mathbf{E}_1^\top) \subseteq \text{Range}(\mathbf{C}_E)$. For any $\mathbf{w} \in \mathbb{R}^p$, let $\mathbf{W}$ be the symmetric matrix such that $\mathbf{w} = \text{svec}(\mathbf{W})$. Then the i -th entry of $\mathbf{E}_1^\top \mathbf{w}$ is

$$\begin{aligned} [\mathbf{E}_1^\top \mathbf{w}]_i &= \left\langle \text{svec}(\mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top), \text{svec}(\mathbf{W}) \right\rangle \\ &= \left\langle \mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top, \mathbf{W} \right\rangle_F \\ &= 2 \mathbf{z}_i^\top \mathbf{W} \mathbf{m}_i = \mathbf{1}_{\{\kappa_i=1\}} \eta_i \frac{r}{\sqrt{d}} (2 \mathbf{W} \mathbf{u}_1)^\top \mathbf{z}_i + \mathbf{1}_{\{\kappa_i=2\}} \eta_i \frac{r}{\sqrt{d}} (2 \mathbf{W} \mathbf{u}_2)^\top \mathbf{z}_i. \end{aligned}$$

Thus $\mathbf{E}_1^\top \mathbf{w}$ belongs to the column span of $\mathbf{C}_E$. Therefore $\text{rank}(\mathbf{E}_1^\top) \leq \text{rank}(\mathbf{C}_E) \leq 2d$. Then, because $n \asymp d^2$, we have

$$\text{rank}(\mathbf{I}_n - P_\sigma) = \dim(\mathcal{U}_\sigma^+) \leq 2 + d + d + 2d = 4d + 2 = o(n).$$

We now prove the label-preservation statement. It is convenient to work with a slightly larger nuisance space. Define the $n \times q$ matrix

$$\mathbf{F} := [\boldsymbol{\mu}_\sigma \quad \mathbf{A}_\sigma \mathbf{X}^\top \quad \mathbf{B}_{3,\sigma} \mathbf{X}^\top \quad \mathbf{C}_E], \quad q \leq 1 + d + d + 2d = 4d + 1.$$

Then $\mathcal{U}_\sigma^+ \subseteq \text{span}\{\mathbf{1}\} + \text{Range}(\mathbf{F})$. Let $\mathbf{P} := \mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$. Since $\mathbf{y}^\top \mathbf{1} = 0$, we have

$$\|\Pi_{\mathcal{U}_\sigma^+} \mathbf{y}\|_2^2 \leq \|\Pi_{\text{span}\{\mathbf{1}\} + \text{Range}(\mathbf{F})} \mathbf{y}\|_2^2 = \|\Pi_{\text{Range}(\mathbf{P}\mathbf{F})} \mathbf{y}\|_2^2.$$

Therefore

$$\|\Pi_{\mathcal{U}_\sigma^+} \mathbf{y}\|_2^2 \leq (\mathbf{F}^\top \mathbf{y})^\top (\mathbf{F}^\top \mathbf{P} \mathbf{F})^\dagger (\mathbf{F}^\top \mathbf{y}). \quad (\text{I.18})$$

Let the i -th row of $\mathbf{F}$ be $\mathbf{f}_i^\top \in \mathbb{R}^q$. Define $\bar{\mathbf{f}} := \frac{1}{n} \sum_{i=1}^n \mathbb{E} \mathbf{f}_i$, and $\boldsymbol{\Sigma}_F := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})^\top]$.

We first prove that $\sum_{i=1}^n y_i \mathbb{E} \mathbf{f}_i = \mathbf{0}$. Firstly, the first coordinate of $\mathbf{f}_i$ is $\zeta_0^{(i)} = \mathbb{E}[\sigma(s_i \xi)]$, whose expectation depends only on $\|\mathbf{m}_i\| = r/\sqrt{d}$, the same for all clusters. Since $\sum_i y_i = 0$, its label-weighted expectation vanishes. For the two linear blocks in $\mathbf{f}_i$, write $a(s) := \frac{\zeta_1(s)}{s}$, and $b(s) := \frac{\zeta_3(s)}{s}$. By rotational symmetry of the Gaussian noise, for any radial scalar function h , $\mathbb{E}[h(\|\mathbf{z} + \mathbf{m}\|)(\mathbf{z} + \mathbf{m})] = \kappa_h \mathbf{m}$ for some scalar κ_h depending only on $\|\mathbf{m}\|$. Therefore

$$\sum_i y_i \mathbb{E}[a(s_i) \mathbf{x}_i] = \kappa_a \sum_i y_i \mathbf{m}_i = \mathbf{0}, \quad (\text{I.19})$$

$$\sum_i y_i \mathbb{E}[b(s_i) \mathbf{x}_i] = \kappa_b \sum_i y_i \mathbf{m}_i = \mathbf{0}. \quad (\text{I.20})$$

Here we used the XOR symmetry: the two signs on each axis are balanced, so $\sum_i y_i \mathbf{m}_i = \mathbf{0}$. Finally, the C_E block has zero mean because it is linear in $\mathbf{z}_i$, hence $\mathbb{E}\left[\mathbf{1}_{\{\kappa_i=a\}} \eta_i \frac{r}{\sqrt{d}} \mathbf{z}_i\right] = \mathbf{0}$. Thus $\sum_i y_i \mathbb{E} \mathbf{f}_i = \mathbf{0}$. Since $\mathbf{y}^\top \mathbf{1} = 0$, we have $\mathbf{F}^\top \mathbf{y} = \sum_{i=1}^n y_i (\mathbf{f}_i - \bar{\mathbf{f}})$. The previous population orthogonality gives

$$\mathbb{E}[\mathbf{F}^\top \mathbf{y}] = \mathbf{0}. \quad (\text{I.21})$$

We next use sample-covariance concentration bound for independent non-identically distributed rows in dimension $q = o(n)$. The rows $\mathbf{f}_i$ are smooth finite-dimensional functions of the Gaussian vectors $\mathbf{z}_i$, with uniformly bounded derivatives on the event $\max_i |s_i - 1| = o(1)$, and the remaining tails are Gaussian. We now prove that

$$\left\| \Sigma_F^{\dagger/2} \left(\frac{1}{n} \mathbf{F}^\top \mathbf{P} \mathbf{F} - \Sigma_F \right) \Sigma_F^{\dagger/2} \right\| = o_{\mathbb{P}}(1). \quad (\text{I.22})$$

Consider an orthonormal basis $\mathbf{U} \in \mathbb{R}^{n \times r_F}$ of $\text{supp}(\Sigma_F)$ where $r_F = \text{rank}(\Sigma_F)$, and define

$$\tilde{\Sigma}_F := \mathbf{U}^\top \Sigma_F \mathbf{U}, \quad \mathbf{g}_i := \tilde{\Sigma}_F^{-1/2} \mathbf{U}^\top (\mathbf{f}_i - \bar{\mathbf{f}}),$$

then $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbf{g}_i \mathbf{g}_i^\top] = \mathbf{I}$. Let $\mathbf{G} \in \mathbb{R}^{n \times r_F}$ be the matrix whose i -th row is $\mathbf{g}_i$. Then

$$\left\| \Sigma_F^{\dagger/2} \left(\frac{1}{n} \mathbf{F}^\top \mathbf{P} \mathbf{F} - \Sigma_F \right) \Sigma_F^{\dagger/2} \right\| = \left\| \frac{1}{n} \mathbf{G}^\top \mathbf{P} \mathbf{G} - \mathbf{I}_{r_F} \right\|, \quad \text{and} \quad \frac{1}{n} \mathbf{P} \mathbf{G}^\top \mathbf{P} \mathbf{G} = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i^\top - \bar{\mathbf{g}} \bar{\mathbf{g}}^\top$$

where $\bar{\mathbf{g}} = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i$. Hence to prove (I.22), we only need to show that

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i^\top - \mathbf{I}_{r_F} \right\| = o_{\mathbb{P}}(1) \quad (\text{I.23})$$

and $\|\bar{\mathbf{g}}\| = o_{\mathbb{P}}(1)$. The second one is derict: since $\sum_{i=1}^n \mathbb{E}[\mathbf{g}_i]$, $\mathbb{E}[\|\bar{\mathbf{g}}\|^2] \leq \frac{1}{n^2} \sum_{i=1}^n \text{Tr} \mathbb{E}[\mathbf{g}_i \mathbf{g}_i^\top] = O(1/d)$, and then we can apply Markov's inequality to conclude that $\|\bar{\mathbf{g}}\| = o_{\mathbb{P}}(1)$. For the first part, we can apply Proposition 2.6 of [Nourdin et al. \(2010\)](#) for the Gaussian-chaos tail bound of F and Theorem 6.1 of [Vershynin \(2012\)](#) for sample covariance concentration inequality to conclude (I.23).

Therefore, with probability tending to one, (I.22) implies that $\mathbf{F}^\top \mathbf{P} \mathbf{F} \succeq \frac{n}{2} \Sigma_F$ on $\text{supp}(\Sigma_F)$ and hence $(\mathbf{F}^\top \mathbf{P} \mathbf{F})^\dagger \preceq \frac{2}{n} \Sigma_F^\dagger$. From (I.18), it follows that, with probability tending to one,

$$\|\Pi_{\mathcal{U}_\sigma^+} \mathbf{y}\|_2^2 \leq (\mathbf{F}^\top \mathbf{y})^\top (\mathbf{F}^\top \mathbf{P} \mathbf{F})^\dagger (\mathbf{F}^\top \mathbf{y}) \leq \frac{2}{n} (\mathbf{F}^\top \mathbf{y})^\top \Sigma_F^\dagger (\mathbf{F}^\top \mathbf{y}). \quad (\text{I.24})$$

It remains to bound the whitened score. Since the rows are independent and (I.21), we have

$$\mathbb{E}\left[(\mathbf{F}^\top \mathbf{y})^\top \Sigma_F^\dagger (\mathbf{F}^\top \mathbf{y})\right] = \text{Tr}\left[\Sigma_F^\dagger \text{Cov}(\mathbf{F}^\top \mathbf{y})\right] = \text{Tr}\left[\Sigma_F^\dagger \sum_{i=1}^n \text{Cov}(\mathbf{f}_i)\right]. \quad (\text{I.25})$$

Moreover, since $\mathbb{E}[(\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})^\top] = \text{Cov}(\mathbf{f}_i) + (\mathbb{E}[\mathbf{f}_i] - \bar{\mathbf{f}})(\mathbb{E}[\mathbf{f}_i] - \bar{\mathbf{f}})^\top$, we have that

$$\sum_{i=1}^n \text{Cov}(\mathbf{f}_i) \preceq \sum_{i=1}^n \mathbb{E}[(\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})^\top] = n\mathbf{\Sigma}_F.$$

Therefore, from (I.25), we have $\mathbb{E}[(\mathbf{F}^\top \mathbf{y})^\top \mathbf{\Sigma}_F^\dagger (\mathbf{F}^\top \mathbf{y})] \leq n \text{rank}(\mathbf{\Sigma}_F) \leq n(4d+1)$. By Markov's inequality, $(\mathbf{F}^\top \mathbf{y})^\top \mathbf{\Sigma}_F^\dagger (\mathbf{F}^\top \mathbf{y}) = O_{\mathbb{P}}(nd)$. Consequently, by (I.24) $\|\Pi_{\mathcal{U}_\sigma^\perp} \mathbf{y}\|_2^2 = O_{\mathbb{P}}(d)$. Finally, $\mathbf{y}^\top \mathbf{P}_\sigma \mathbf{y} = \|\mathbf{y}\|_2^2 - \|\Pi_{\mathcal{U}_\sigma^\perp} \mathbf{y}\|_2^2 = n - O_{\mathbb{P}}(d)$, and hence we complete the proof. $\blacksquare$

I.2. QE of the Population Covariance Matrix

Let $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and define the population CK matrix

$$\Phi(\mathbf{X}) := \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{X})^\top \sigma(\mathbf{w}^\top \mathbf{X}) \mid \mathbf{X}] \in \mathbb{R}^{n \times n}, \quad \Phi(\mathbf{X})_{ij} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{x}_i) \sigma(\mathbf{w}^\top \mathbf{x}_j) \mid \mathbf{X}].$$

The following lemmas derive the QE of the population kernel $\Phi(\mathbf{X})$ after the projection $\mathbf{P}_\sigma$ defined in (I.11).

Lemma 64 (Expansion of $\Phi(\mathbf{X})$) Recall $\mathbf{D}_k$ defined in (I.8). Let $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$. Then

$$\Phi(\mathbf{X}) = \sum_{k \geq 0} \mathbf{D}_k \mathbf{G}^{\odot k} \mathbf{D}_k, \quad (\mathbf{G}^{\odot k})_{ij} := (\mathbf{G}_{ij})^k. \quad (\text{I.26})$$

Proof. The pair $(\mathbf{w}^\top \mathbf{x}_i, \mathbf{w}^\top \mathbf{x}_j)$ is jointly Gaussian with variances (s_i^2, s_j^2) and correlation $\mathbf{R}_{ij} := \mathbf{G}_{ij}/(s_i s_j)$. Writing $\mathbf{w}^\top \mathbf{x}_i = s_i g_i$ with (g_i, g_j) standard normal, $\mathbb{E}[g_i g_j] = \mathbf{R}_{ij}$, expanding $\sigma(s_i g_i) = \sum_k \zeta_k^{(i)} h_k(g_i)$ in L^2 and using $\mathbb{E}[h_k(g_i) h_\ell(g_j)] = \delta_{k\ell} \mathbf{R}_{ij}^k$ yields (I.26). See the proof of Lemma 5.2 in Wang and Zhu (2024) for more details. $\blacksquare$

Lemma 65 We have $\mathbf{D}_0 \mathbf{G}^{\odot 0} \mathbf{D}_0 = \mu_\sigma \mu_\sigma^\top$ and $\mathbf{D}_1 \mathbf{G} \mathbf{D}_1 = \mathbf{A}_\sigma \mathbf{G} \mathbf{A}_\sigma$.

Proof. Immediate from the definitions of $\mu_\sigma, \mathbf{A}_\sigma$ in (I.9). $\blacksquare$

Let $\mathbf{T}_k := \mathbf{D}_k \mathbf{G}^{\odot k} \mathbf{D}_k$. Then from Lemmas 64 and 65, we have

$$\mathbf{P}_\sigma \Phi(\mathbf{X}) \mathbf{P}_\sigma = \sum_{k=2}^{\infty} \mathbf{P}_\sigma \mathbf{T}_k \mathbf{P}_\sigma. \quad (\text{I.27})$$

In the following two subsections, we control the first three terms in the expansion of (I.27)

I.2.1. THE QUADRATIC TERM

Lemma 66 (Radial cancellation by the nuisance projector) Recall $s_i := \|\mathbf{x}_i\|$ for $i = 1, \dots, n$. Let

$$m(s) := \zeta_0(s) = \mathbb{E}[\sigma(s\xi)], \quad q(s) := \zeta_2(s) = \mathbb{E}[\sigma(s\xi)h_2(\xi)].$$

Assume $a_2 = q(1) \neq 0$. Let g be any C^2 function in a neighborhood of 1. For $g_i := g(s_i)$, set $\mathbf{g} := (g_1, \dots, g_n)^\top$. Then $\frac{1}{d} \|\mathbf{P}_\sigma \mathbf{g}\|_2^2 = o_{\mathbb{P}}(1)$. In particular,

$$\frac{1}{d} \|\mathbf{P}_\sigma \mathbf{v}\|_2^2 = o_{\mathbb{P}}(1), \quad \frac{1}{d} \|\mathbf{P}_\sigma \mathbf{q}\|_2^2 = o_{\mathbb{P}}(1), \quad (\text{I.28})$$

where $\mathbf{v} := (s_1^2, \dots, s_n^2)^\top$, and $\mathbf{q} := (\zeta_2(s_1), \dots, \zeta_2(s_n))^\top$.

Proof. By Stein's identity and the definition of a_2 , $m'(1) = \sqrt{2} a_2$. Since $a_2 \neq 0$, we have $m'(1) \neq 0$. For a fixed C^2 function g , define $\beta_g := \frac{g'(1)}{m'(1)}$ and $r_g(s) := g(s) - g(1) - \beta_g \{m(s) - m(1)\}$. Then $r_g(1) = 0$, and $r'_g(1) = g'(1) - \beta_g m'(1) = 0$. Therefore, by Taylor's theorem, there exists a constant $C_g < \infty$ such that for all s sufficiently close to 1, $|r_g(s)| \leq C_g |s - 1|^2$. Let $\mathbf{r}_g := (r_g(s_1), \dots, r_g(s_n))^\top$. Since

$$\mathbf{g} = \{g(1) - \beta_g m(1)\} \mathbf{1} + \beta_g \boldsymbol{\mu}_\sigma + \mathbf{r}_g,$$

and since $\mathbf{P}_\sigma \mathbf{1} = \mathbf{0}$ and $\mathbf{P}_\sigma \boldsymbol{\mu}_\sigma = \mathbf{0}$, we get $\mathbf{P}_\sigma \mathbf{g} = \mathbf{P}_\sigma \mathbf{r}_g$. Hence by Lemma 13

$$\|\mathbf{P}_\sigma \mathbf{g}\|_2^2 \leq \|\mathbf{r}_g\|_2^2 \leq C_g^2 \sum_{i=1}^n |s_i - 1|^4 \lesssim \left(\max_i |s_i - 1|^2 \right) \sum_{i=1}^n |s_i - 1|^2 \leq \tau_n^2 B^2.$$

where $\tau_n^2 = O\left(\frac{\log n}{d}\right)$, and $B^2 = O(d)$. Thus $\frac{1}{d} \|\mathbf{P}_\sigma \mathbf{g}\|_2^2 = o_{\mathbb{P}}(1)$. Taking $g(s) = s^2$ and $g(s) = q(s) = \zeta_2(s)$ implies (I.28). $\blacksquare$

Lemma 67 Recall $\mathbf{H}$ defined in (I.4). We have that

$$\|\mathbf{P}_\sigma \mathbf{T}_2 \mathbf{P}_\sigma - a_2^2 \mathbf{P}_\sigma \mathbf{H} \mathbf{P}_\sigma\| = o_{\mathbb{P}}(1). \quad (\text{I.29})$$

Proof. Recall $s_i = \|\mathbf{x}_i\|$ and $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$. Define the normalized data $\mathbf{u}_i := \frac{\mathbf{x}_i}{s_i}$, and $\mathbf{R}_{ij} := \mathbf{u}_i^\top \mathbf{u}_j$. Thus $\mathbf{X}^\top \mathbf{X} = \mathbf{D} \mathbf{R} \mathbf{D}$, and $\mathbf{H} = (\mathbf{X}^\top \mathbf{X})^{\odot 2} = \mathbf{D}^2 \mathbf{R}^{\odot 2} \mathbf{D}^2$. Set $q_i := \zeta_2(s_i)$, and $v_i := s_i^2$, and define

$$\mathbf{q} := (q_1, \dots, q_n)^\top, \quad \mathbf{v} := (v_1, \dots, v_n)^\top, \quad \mathbf{D}_q := \text{diag}(q_1, \dots, q_n), \quad \mathbf{D}_v := \text{diag}(v_1, \dots, v_n).$$

Thus, $\mathbf{H} = \mathbf{D}_v \mathbf{R}^{\odot 2} \mathbf{D}_v$ and $\mathbf{T}_2 = \mathbf{D}_q \mathbf{R}^{\odot 2} \mathbf{D}_q$. We now separate the large constant component of $\mathbf{R}^{\odot 2}$. Define $\boldsymbol{\Psi} := \mathbf{R}^{\odot 2} - \frac{1}{d} \mathbf{1} \mathbf{1}^\top$. This is the degree-two spherical harmonic Gram matrix, since for any $i, j \in [n]$,

$$\Psi_{ij} = \left\langle \mathbf{u}_i \mathbf{u}_i^\top - \frac{1}{d} \mathbf{I}_d, \mathbf{u}_j \mathbf{u}_j^\top - \frac{1}{d} \mathbf{I}_d \right\rangle_F.$$

Hence $\mathbf{R}^{\odot 2} = \frac{1}{d} \mathbf{1} \mathbf{1}^\top + \boldsymbol{\Psi}$. Consequently, if we define $\widetilde{\mathbf{H}} := \mathbf{D}_v \boldsymbol{\Psi} \mathbf{D}_v$ and $\mathbf{C}_\sigma := \mathbf{D}_v^{-1} \mathbf{D}_q$, then we have

$$\mathbf{H} = \frac{1}{d} \mathbf{v} \mathbf{v}^\top + \widetilde{\mathbf{H}}, \quad \mathbf{T}_2 = \frac{1}{d} \mathbf{q} \mathbf{q}^\top + \mathbf{C}_\sigma \widetilde{\mathbf{H}} \mathbf{C}_\sigma.$$

Therefore

$$\mathbf{P}_\sigma \mathbf{T}_2 \mathbf{P}_\sigma - a_2^2 \mathbf{P}_\sigma \mathbf{H} \mathbf{P}_\sigma = \frac{1}{d} \mathbf{P}_\sigma (\mathbf{q} \mathbf{q}^\top - a_2^2 \mathbf{v} \mathbf{v}^\top) \mathbf{P}_\sigma + \mathbf{P}_\sigma (\mathbf{C}_\sigma \widetilde{\mathbf{H}} \mathbf{C}_\sigma - a_2^2 \widetilde{\mathbf{H}}) \mathbf{P}_\sigma. \quad (\text{I.30})$$

By Lemma 66, we can control the first term at the right-hand side of (I.30) as

$$\left\| \frac{1}{d} \mathbf{P}_\sigma (\mathbf{q} \mathbf{q}^\top - a_2^2 \mathbf{v} \mathbf{v}^\top) \mathbf{P}_\sigma \right\| \leq \frac{1}{d} \|\mathbf{P}_\sigma \mathbf{q}\|_2^2 + \frac{a_2^2}{d} \|\mathbf{P}_\sigma \mathbf{v}\|_2^2 = o_{\mathbb{P}}(1).$$

Next, we use the centered quadratic-lift bound

$$\|\widetilde{\mathbf{H}}\| = O_{\mathbb{P}}(1). \quad (\text{I.31})$$

We defer the proof of (I.31) to Lemma 68 below. Now define $c(s) := \frac{\zeta_2(s)}{s^2}$. Since c is continuous near 1, $c(1) = a_2$, and $\max_i |s_i - 1| = o_{\mathbb{P}}(1)$, we have

$$\|C_\sigma - a_2 \mathbf{I}_n\| = \max_i |c(s_i) - a_2| = o_{\mathbb{P}}(1).$$

Also $\|C_\sigma\| = O_{\mathbb{P}}(1)$. Therefore

$$\left\| P_\sigma \left(C_\sigma \widetilde{H} C_\sigma - a_2^2 \widetilde{H} \right) P_\sigma \right\| = o_{\mathbb{P}}(1).$$

Combining (I.30), we prove (I.29). ■

Lemma 68 (Operator norm bound for the centered quadratic lift) *Assume that X is XOR model with $\mathbf{x}_i = \mathbf{z}_i + \mathbf{m}_i$, where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, d^{-1} \mathbf{I}_d)$ and $\|\mathbf{m}_i\| = r/\sqrt{d}$ for $i \in [n]$ where $r = O(1)$. Let $s_i := \|\mathbf{x}_i\|$, and $\mathbf{u}_i := \frac{\mathbf{x}_i}{s_i}$, and define $\tilde{\mathbf{q}}_i := s_i^2 \text{svec}(\mathbf{u}_i \mathbf{u}_i^\top - \frac{1}{d} \mathbf{I}_d)$. Let $\tilde{\mathbf{Q}} := [\tilde{\mathbf{q}}_1, \dots, \tilde{\mathbf{q}}_n] \in \mathbb{R}^{p \times n}$, and $\tilde{\mathbf{H}} := \tilde{\mathbf{Q}}^\top \tilde{\mathbf{Q}}$. If $\frac{n}{p} \rightarrow \gamma \in (0, \infty)$, then $\|\tilde{\mathbf{H}}\| = O_{\mathbb{P}}(1)$.*

Proof. Since

$$s_i^2 \left(\mathbf{u}_i \mathbf{u}_i^\top - \frac{1}{d} \mathbf{I}_d \right) = \mathbf{x}_i \mathbf{x}_i^\top - \frac{\|\mathbf{x}_i\|^2}{d} \mathbf{I}_d,$$

we have the exact identity

$$\tilde{\mathbf{q}}_i = \text{svec} \left(\mathbf{x}_i \mathbf{x}_i^\top - \frac{\|\mathbf{x}_i\|^2}{d} \mathbf{I}_d \right).$$

Using $\mathbf{x}_i = \mathbf{z}_i + \mathbf{m}_i$, expand

$$\mathbf{x}_i \mathbf{x}_i^\top - \frac{\|\mathbf{x}_i\|^2}{d} \mathbf{I}_d = \left(\mathbf{z}_i \mathbf{z}_i^\top - \frac{\|\mathbf{z}_i\|^2}{d} \mathbf{I}_d \right) + \left(\mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top - \frac{2\mathbf{z}_i^\top \mathbf{m}_i}{d} \mathbf{I}_d \right) + \left(\mathbf{m}_i \mathbf{m}_i^\top - \frac{\|\mathbf{m}_i\|^2}{d} \mathbf{I}_d \right).$$

Therefore we have decomposed $\tilde{\mathbf{Q}}$ as $\tilde{\mathbf{Q}} = \mathbf{A} + \mathbf{B} + \mathbf{C}$, where the columns of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are, respectively,

$$\mathbf{a}_i = \text{svec} \left(\mathbf{z}_i \mathbf{z}_i^\top - \frac{\|\mathbf{z}_i\|^2}{d} \mathbf{I}_d \right), \tag{I.32}$$

$$\mathbf{b}_i = \text{svec} \left(\mathbf{z}_i \mathbf{m}_i^\top + \mathbf{m}_i \mathbf{z}_i^\top - \frac{2\mathbf{z}_i^\top \mathbf{m}_i}{d} \mathbf{I}_d \right), \tag{I.33}$$

$$\mathbf{c}_i = \text{svec} \left(\mathbf{m}_i \mathbf{m}_i^\top - \frac{\|\mathbf{m}_i\|^2}{d} \mathbf{I}_d \right). \tag{I.34}$$

Thus, to prove the lemma, it suffices to prove that We prove $\|\mathbf{A}\| = O_{\mathbb{P}}(1)$, $\|\mathbf{B}\| = O_{\mathbb{P}}(1)$, and $\|\mathbf{C}\| = O(1)$.

Step 1: the pure quadratic Gaussian part $\mathbf{A}$. Write $\mathbf{z}_i = \rho_i \mathbf{g}_i$, where $\rho_i := \|\mathbf{z}_i\|$ and $\mathbf{g}_i := \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}$. Then $\mathbf{g}_i$ is uniform on the sphere $\mathbb{S}^{d-1}$, independent of ρ_i , and $\rho_i^2 \sim \frac{1}{d} \chi_d^2$. Moreover,

$$\mathbf{a}_i = \rho_i^2 \text{svec} \left(\mathbf{g}_i \mathbf{g}_i^\top - \frac{1}{d} \mathbf{I}_d \right).$$

Let $\phi_i := \text{svec}(\mathbf{g}_i \mathbf{g}_i^\top - \frac{1}{d} \mathbf{I}_d)$, and $\Phi := [\phi_1, \dots, \phi_n]$. Then $\mathbf{A} = \Phi \mathbf{D}_\rho$, where $\mathbf{D}_\rho := \text{diag}(\rho_1^2, \dots, \rho_n^2)$. By the standard chi-square concentration bound and $n = O(d^2)$, we have that $\max_{1 \leq i \leq n} \rho_i^2 = O_{\mathbb{P}}(1)$. Therefore

$$\|\mathbf{A}\| \leq \|\Phi\| \|\mathbf{D}_\rho\| = O_{\mathbb{P}}(1) \|\Phi\|,$$

so it remains to control $\|\Phi\|$. We use the following Lemma 70 below to get $\|\Phi\| = O_{\mathbb{P}}(1)$.

Step 2: the linear cross part B. For $\mathbf{m} \in \mathbb{R}^d$, define the linear map $\mathcal{L}_{\mathbf{m}} : \mathbb{R}^d \rightarrow \mathbb{S}_0^d$ by

$$\mathcal{L}_{\mathbf{m}}(\mathbf{v}) := \mathbf{m} \mathbf{v}^\top + \mathbf{v} \mathbf{m}^\top - \frac{2 \mathbf{v}^\top \mathbf{m}}{d} \mathbf{I}_d.$$

Here $\mathbb{S}_0^d$ denotes the space of $d \times d$ symmetric matrices with zero trace. Then $\mathbf{b}_i = \text{svec}(\mathcal{L}_{\mathbf{m}_i}(\mathbf{z}_i))$. Moreover,

$$\|\mathcal{L}_{\mathbf{m}}(\mathbf{v})\|_F \leq 2\|\mathbf{m}\| \|\mathbf{v}\| + \frac{2|\mathbf{v}^\top \mathbf{m}|}{\sqrt{d}} \leq 4\|\mathbf{m}\| \|\mathbf{v}\|.$$

Hence $\|\mathcal{L}_{\mathbf{m}}\|_{\text{op}} \leq 4\|\mathbf{m}\|$. In the XOR model, $\mathbf{m}_i = \eta_i \frac{r}{\sqrt{d}} \mathbf{u}_{\kappa_i}$, where $\eta_i \in \{\pm 1\}$, and $\kappa_i \in \{1, 2\}$. Let $I_a := \{i : \kappa_i = a\}$. On I_a , $\mathbf{b}_i = \eta_i r \text{svec}(\mathcal{L}_{\mathbf{u}_a}(\mathbf{z}_i)) / \sqrt{d}$. Let $\mathbf{Z}_a := [\mathbf{z}_i]_{i \in I_a} \in \mathbb{R}^{d \times |I_a|}$. Since the entries of $\mathbf{Z}_a$ are Gaussian with variance $1/d$, standard Gaussian matrix norm bounds give

$$\|\mathbf{Z}_a\| = O_{\mathbb{P}}\left(1 + \sqrt{\frac{|I_a|}{d}}\right) = O_{\mathbb{P}}(\sqrt{d}),$$

because $|I_a| \asymp n \asymp d^2$. Therefore

$$\|\mathbf{B}_{I_a}\| \leq \frac{r}{\sqrt{d}} \|\mathcal{L}_{\mathbf{u}_a}\|_{\text{op}} \|\mathbf{Z}_a\| = O_{\mathbb{P}}(1).$$

Since there are only two axes $a = 1, 2$, we can conclude that $\|\mathbf{B}\| = O_{\mathbb{P}}(1)$.

Step 3: the deterministic mean part C. For each $i \in [n]$, we have $\|\mathbf{m}_i\| = \frac{r}{\sqrt{d}}$, and hence

$$\begin{aligned} \|\mathbf{c}_i\|_2 &= \left\| \mathbf{m}_i \mathbf{m}_i^\top - \frac{\|\mathbf{m}_i\|^2}{d} \mathbf{I}_d \right\|_F \\ &\leq \|\mathbf{m}_i \mathbf{m}_i^\top\|_F + \frac{\|\mathbf{m}_i\|^2}{d} \|\mathbf{I}_d\|_F = \|\mathbf{m}_i\|^2 + \frac{\|\mathbf{m}_i\|^2}{\sqrt{d}} \leq \frac{Cr^2}{d}. \end{aligned}$$

Therefore $\|\mathbf{C}\| \leq \|\mathbf{C}\|_F = O(1)$. ■

Lemma 69 (Uniform-spherical harmonic Gram bound) *Let $\mathbf{g}_1, \dots, \mathbf{g}_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$, and assume $n \asymp d^2$. For fixed $m \geq 2$, let $q_m^{(d)}$ be the normalized Gegenbauer polynomial of degree m defined in [Lu and Yau \(2025\)](#), and define*

$$(\mathbf{A}_m)_{ij} := \frac{1}{\sqrt{n}} q_m^{(d)}(\sqrt{d} \mathbf{g}_i^\top \mathbf{g}_j) \mathbf{1}_{\{i \neq j\}}. \quad (\text{I.35})$$

Then $\|\mathbf{A}_m\|_{\text{op}} \prec d^{\max\{2-m, 0\}/2}$. Equivalently, if $\psi_m^{(d)}$ denotes the degree- m zonal spherical harmonic kernel normalized so that, for fixed m ,

$$\psi_m^{(d)}(t) = c_{m,d} d^{-m/2} q_m^{(d)}(\sqrt{d} t), \quad c_{m,d} \asymp 1, \quad (\text{I.36})$$

then

$$\left\| \left(\psi_m^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \right)_{i,j=1}^n - \psi_m^{(d)}(1) \mathbf{I}_n \right\| = O_{\mathbb{P}} \left(\frac{\sqrt{n}}{d^{m/2}} d^{\max\{2-m, 0\}/2} \right). \quad (\text{I.37})$$

Thus, for $n \asymp d^2$,

$$\left\| \left(\psi_2^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \right)_{i,j} - \psi_2^{(d)}(1) \mathbf{I}_n \right\| = O_{\mathbb{P}}(1), \quad (\text{I.38})$$

$$\left\| \left(\psi_3^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \right)_{i,j} - \psi_3^{(d)}(1) \mathbf{I}_n \right\| = O_{\mathbb{P}}(d^{-1/2}) = o_{\mathbb{P}}(1), \quad (\text{I.39})$$

$$\left\| \left(\psi_4^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \right)_{i,j} - \psi_4^{(d)}(1) \mathbf{I}_n \right\| = O_{\mathbb{P}}(d^{-1}) = o_{\mathbb{P}}(1). \quad (\text{I.40})$$

Proof. This is a direct consequence of Proposition 8 of [Lu and Yau \(2025\)](#). In their notation, if $n = \alpha d^\ell + o(d^\ell)$, then

$$\|\mathbf{A}_m\|_{\text{op}} \prec d^{\max\{\ell-m, 0\}/2}.$$

Taking $\ell = 2$ gives the first result of this lemma. For the equivalent harmonic-kernel form, note that

$$\left(\psi_m^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \mathbf{1}_{\{i \neq j\}} \right)_{i,j} = c_{m,d} d^{-m/2} \sqrt{n} \mathbf{A}_m.$$

The diagonal entries are $\psi_m^{(d)}(1)$. Notice that $c_{m,d} = \Theta(1)$. Therefore, (I.37) is due to

$$\left(\psi_m^{(d)}(\mathbf{g}_i^\top \mathbf{g}_j) \right)_{i,j} - \psi_m^{(d)}(1) \mathbf{I}_n = c_{m,d} d^{-m/2} \sqrt{n} \mathbf{A}_m. \quad \blacksquare$$

Lemma 70 (Degree-two spherical Gram bound) *Let $\mathbf{g}_1, \dots, \mathbf{g}_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$, and define*

$$\phi_i := \text{svec} \left(\mathbf{g}_i \mathbf{g}_i^\top - \frac{1}{d} \mathbf{I}_d \right), \quad \Phi := [\phi_1, \dots, \phi_n].$$

If $n/d^2 = O(1)$, then $\|\Phi\| = O_{\mathbb{P}}(1)$.

Proof. For $i, j \in [n]$,

$$(\Phi^\top \Phi)_{ij} = \left\langle \mathbf{g}_i \mathbf{g}_i^\top - \frac{1}{d} \mathbf{I}_d, \mathbf{g}_j \mathbf{g}_j^\top - \frac{1}{d} \mathbf{I}_d \right\rangle_F = (\mathbf{g}_i^\top \mathbf{g}_j)^2 - \frac{1}{d}.$$

Thus $\Phi^\top \Phi = (1 - \frac{1}{d}) \mathbf{I}_n + \mathbf{K}_2$, where for $i \neq j \in [n]$, $(\mathbf{K}_2)_{ij} = (\mathbf{g}_i^\top \mathbf{g}_j)^2 - \frac{1}{d}$. From [Lu and Yau \(2025\)](#), we know that the normalized Gegenbauer polynomial $q_2^{(d)}$ is

$$q_2^{(d)}(x) = \frac{1}{\sqrt{2}} \sqrt{\frac{d+2}{d-1}} (x^2 - 1). \quad (\text{I.41})$$

Therefore, for $i \neq j$,

$$(\mathbf{g}_i^\top \mathbf{g}_j)^2 - \frac{1}{d} = \frac{\sqrt{2}}{d} \sqrt{\frac{d-1}{d+2}} q_2^{(d)}(\sqrt{d} \mathbf{g}_i^\top \mathbf{g}_j).$$

Recall the definition of $\mathbf{A}_2$ in (I.35). Then by Lemma 69, we have that

$$\mathbf{K}_2 = \frac{\sqrt{2n}}{d} \sqrt{\frac{d-1}{d+2}} \mathbf{A}_2,$$

and $\|\mathbf{A}_2\|_{\text{op}} = O_{\mathbb{P}}(1)$. Since $\sqrt{n}/d = O(1)$, it follows that $\|\mathbf{K}_2\| = O_{\mathbb{P}}(1)$. Consequently, $\|\Phi^\top \Phi\| = O_{\mathbb{P}}(1)$. $\blacksquare$

I.2.2. THE CUBIC AND QUARTIC TERMS

We next analyze the terms $T_k = D_k G^{\odot k} D_k$, for $k = 3, 4$, from (I.26). Let

$$S := \text{diag}(s_1, \dots, s_n), \quad R := S^{-1} G S^{-1}, \quad Z_k := \text{diag}(\zeta_k^{(1)}, \dots, \zeta_k^{(n)}). \quad (\text{I.42})$$

Thus $R_{ij} = \langle x_i/s_i, x_j/s_j \rangle$ is the normalized Gram matrix. Therefore $T_k = Z_k R^{\odot k} Z_k$. We shall use the following fixed-degree harmonic Gram estimate. It is a standard consequence of the trace-moment method for random inner-product kernel matrices; see, for instance, [Cheng and Singer \(2013\)](#), [Lu and Yau \(2025\)](#), and the quadratic-regime estimates in [Pandit et al. \(2024\)](#). We record the statement in the form needed here.

Lemma 71 (Harmonic Gram bounds for normalized XOR directions) *Recall the normalized Gram matrix R defined in (I.42). For fixed $m \geq 2$, let $\psi_m^{(d)}$ denote the degree- m zonal spherical harmonic kernel on $\mathbb{S}^{d-1}$, normalized so that*

$$\psi_m^{(d)}(u^\top v) = c_{m,d} d^{-m/2} q_m^{(d)}(\sqrt{d} u^\top v), \quad c_{m,d} \asymp 1,$$

where $q_m^{(d)}$ is the L^2 -normalized Gegenbauer polynomial used in [Lu and Yau \(2025\)](#). Define $\Psi_m := (\psi_m^{(d)}(R_{ij}))_{i,j=1}^n$ on XOR dataset. If $n \asymp d^2$, then for each fixed $m \geq 2$,

$$\|\Psi_m - \psi_m^{(d)}(1) I_n\| = O_{\mathbb{P}}\left(\sqrt{\frac{n}{d^m}} + \frac{n}{d^m}\right).$$

Proof. The proof has two parts. First we recall Lemma 69 for standard uniform-sphere distributions. Then we show that the same estimate remains valid for the normalized XOR directions.

Let $\theta_1, \dots, \theta_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$ and recall A_m defined in (I.35). Define the corresponding harmonic matrix by

$$K_m^{\text{sph}} := (\psi_m^{(d)}(\theta_i^\top \theta_j))_{i,j=1}^n - \psi_m^{(d)}(1) I_n = (\psi_m^{(d)}(\theta_i^\top \theta_j) \mathbf{1}_{\{i \neq j\}})_{i,j=1}^n.$$

Then Lemma 69 implies that $\|K_3^{\text{sph}}\| = o_{\mathbb{P}}(1)$, and $\|K_4^{\text{sph}}\| = o_{\mathbb{P}}(1)$.

For the XOR model, conditional on the cluster of i , we have $x_i = \frac{g_i + \mu_i}{\sqrt{d}}$, where $g_i \sim \mathcal{N}(\mathbf{0}, I_d)$ and $\mu_i \in \{\pm r u_1, \pm r u_2\}$, and therefore $u_i = \frac{g_i + \mu_i}{\|g_i + \mu_i\|_2}$ defined in Lemma 68. Thus u_i is not uniform on the sphere. However, we can claim that its angular law is absolutely continuous with respect to the uniform measure, with uniformly bounded L^p density for every fixed p . We now verify this claim. Let σ_{d-1} denote the uniform probability measure on $\mathbb{S}^{d-1}$. For fixed μ with $\|\mu\| \leq r$, let

$$\mathbb{P}_\mu := \text{Law}\left(\frac{g + \mu}{\|g + \mu\|_2}\right), \quad g \sim \mathcal{N}(\mathbf{0}, I_d).$$

Let $L_\mu := \frac{d\mathbb{P}_\mu}{d\sigma_{d-1}}$. Using polar coordinates for a standard Gaussian vector, $g = \rho\theta$, where $\theta \sim \sigma_{d-1}$, and $\rho \geq 0$ is independent of θ with $\rho^2 \sim \chi_d^2$. One can verify that the density L_μ is given by

$$L_\mu(\theta) = \mathbb{E}_\rho \exp\left(\rho \mu^\top \theta - \frac{\|\mu\|^2}{2}\right).$$

We prove that for every fixed $p < \infty$,

$$\sup_{d \geq 1} \sup_{\|\boldsymbol{\mu}\| \leq r} \|L_{\boldsymbol{\mu}}\|_{L^p(\sigma_{d-1})} \leq C_{p,r} < \infty. \quad (\text{I.43})$$

To prove this, Jensen's inequality gives

$$L_{\boldsymbol{\mu}}(\boldsymbol{\theta})^p \leq \mathbb{E}_{\rho} \exp\left(p\rho \boldsymbol{\mu}^{\top} \boldsymbol{\theta} - \frac{p\|\boldsymbol{\mu}\|^2}{2}\right).$$

Integrating over $\boldsymbol{\theta}$ and using the standard spherical sub-Gaussian bound

$$\mathbb{E}_{\boldsymbol{\theta} \sim \sigma_{d-1}} \exp(t \boldsymbol{\mu}^{\top} \boldsymbol{\theta}) \leq \exp\left(\frac{Ct^2 \|\boldsymbol{\mu}\|^2}{d}\right),$$

we get

$$\|L_{\boldsymbol{\mu}}\|_{L^p(\sigma_{d-1})}^p \leq e^{-p\|\boldsymbol{\mu}\|^2/2} \mathbb{E}_{\rho} \exp\left(\frac{Cp^2 \rho^2 \|\boldsymbol{\mu}\|^2}{d}\right).$$

Since $\rho^2 \sim \chi_d^2$, the last expectation is bounded uniformly in d for every p and $\|\boldsymbol{\mu}\| \leq r$. This proves (I.43).

Now fix $m \geq 2$. Let $\mathbf{K}_m^{\text{xor}} := (\psi_m^{(d)}(\mathbf{u}_i^{\top} \mathbf{u}_j) \mathbf{1}_{\{i \neq j\}})_{i,j=1}^n$. For an even integer $2q$,

$$\mathbb{E}_{\text{xor}} \text{Tr}((\mathbf{K}_m^{\text{xor}})^{2q}) = \sum_{i_1, \dots, i_{2q}} \mathbb{E}_{\text{xor}} \prod_{\ell=1}^{2q} \psi_m^{(d)}(\mathbf{u}_{i_{\ell}}^{\top} \mathbf{u}_{i_{\ell+1}}) \mathbf{1}_{\{i_{\ell} \neq i_{\ell+1}\}},$$

where $i_{2q+1} = i_1$. For a fixed index sequence $(i_1, \dots, i_{2q})$, let $V(i)$ be the set of distinct vertices appearing in the sequence. Conditional on the cluster assignments, the variables $(\mathbf{u}_j)_{j \in V(i)}$ are independent shifted angular variables. Hence their joint law has density $\prod_{j \in V(i)} L_{\boldsymbol{\mu}_j}(\boldsymbol{\theta}_j)$ with respect to the corresponding product of uniform spherical measures.

By Hölder's inequality and (I.43), for every fixed q there is a constant $C_{q,r}$ such that

$$\left| \mathbb{E}_{\text{xor}} \prod_{\ell=1}^{2q} \psi_m^{(d)}(\mathbf{u}_{i_{\ell}}^{\top} \mathbf{u}_{i_{\ell+1}}) \mathbf{1}_{\{i_{\ell} \neq i_{\ell+1}\}} \right| \leq C_{q,r} \left(\mathbb{E}_{\text{sph}} \left| \prod_{\ell=1}^{2q} \psi_m^{(d)}(\boldsymbol{\theta}_{i_{\ell}}^{\top} \boldsymbol{\theta}_{i_{\ell+1}}) \mathbf{1}_{\{i_{\ell} \neq i_{\ell+1}\}} \right|^2 \right)^{1/2}.$$

The right-hand side is controlled by the same moment calculation used for the uniform-sphere Gegenbauer matrices in (Lu and Yau, 2025, Appendix E); the only change is that the moment order is doubled, which only changes the constant because q is fixed. Therefore, for every fixed q ,

$$\mathbb{E}_{\text{xor}} \text{Tr}((\mathbf{K}_m^{\text{xor}})^{2q}) \leq C_{q,r} n \left(\sqrt{\frac{n}{d^m}} + \frac{n}{d^m} \right)^{2q} + C_{q,r}. \quad (\text{I.44})$$

Consequently, by Markov's inequality,

$$\|\mathbf{K}_m^{\text{xor}}\| = O_{\mathbb{P}} \left(\sqrt{\frac{n}{d^m}} + \frac{n}{d^m} \right).$$

Since $\boldsymbol{\Psi}_m - \psi_m^{(d)}(1) \mathbf{I}_n = \mathbf{K}_m^{\text{xor}}$, the claimed estimates follow. ■

We shall also use the following elementary bound on the normalized Gram matrix.

Lemma 72 (Operator norm of the normalized Gram matrix) *Under the XOR model and $n \asymp d^2$,*

$$\|\mathbf{R}\| = O_{\mathbb{P}}\left(\frac{n}{d}\right) = O_{\mathbb{P}}(d). \quad (\text{I.45})$$

Moreover, deterministically, $\|\mathbf{R}\| \geq \frac{n}{d}$. Thus $\|\mathbf{R}\| = \Theta_{\mathbb{P}}(d)$.

Proof. Let $\mathbf{U} := [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{d \times n}$. Then $\mathbf{R} = \mathbf{U}^{\top} \mathbf{U}$, so $\text{rank}(\mathbf{R}) \leq d$ and $\text{Tr } \mathbf{R} = \sum_{i=1}^n \|\mathbf{u}_i\|_2^2 = n$. Therefore

$$\|\mathbf{R}\| \geq \frac{\text{Tr } \mathbf{R}}{\text{rank}(\mathbf{R})} \geq \frac{n}{d}.$$

For the upper bound, write $\mathbf{R} = \mathbf{S}^{-1} \mathbf{X}^{\top} \mathbf{X} \mathbf{S}^{-1}$. By Lemma 13, there is a constant $c > 0$ such that $\min_{i \in [n]} s_i \geq c$ and $\|\mathbf{X}\| \leq c\sqrt{n/d}$, with probability tending to one. Hence we conclude (I.45). ■

Lemma 73 (The cubic term) *Recall $a_3 := \mathbb{E}[\sigma(\xi)h_3(\xi)]$. Then*

$$\mathbf{P}_{\sigma} \mathbf{T}_3 \mathbf{P}_{\sigma} = a_3^2 \mathbf{P}_{\sigma} + o_{\mathbb{P}}(1). \quad (\text{I.46})$$

Proof. The degree-three zonal spherical harmonic decomposition is

$$t^3 = \frac{3}{d+2}t + \psi_3^{(d)}(t), \quad (\text{I.47})$$

where $\psi_3^{(d)}(t) := t^3 - \frac{3}{d+2}t$ defined in (I.36). Applying (I.47) entrywise to $\mathbf{R}$, we get

$$\mathbf{R}^{\odot 3} = \frac{3}{d+2} \mathbf{R} + \mathbf{\Psi}_3, \quad (\mathbf{\Psi}_3)_{ij} := \psi_3^{(d)}(\mathbf{R}_{ij}).$$

By Lemma 71, $\mathbf{\Psi}_3 = \psi_3^{(d)}(1)\mathbf{I}_n + \mathbf{E}_3$, where $\|\mathbf{E}_3\| = o_{\mathbb{P}}(1)$. Since $\psi_3^{(d)}(1) = 1 + O(d^{-1})$, and $\|\mathbf{Z}_3 - a_3 \mathbf{I}_n\| = o_{\mathbb{P}}(1)$ from Lemma 13, we have

$$\mathbf{P}_{\sigma} \mathbf{Z}_3 \mathbf{\Psi}_3 \mathbf{Z}_3 \mathbf{P}_{\sigma} = \psi_3^{(d)}(1) \mathbf{P}_{\sigma} \mathbf{Z}_3^2 \mathbf{P}_{\sigma} + \mathbf{P}_{\sigma} \mathbf{Z}_3 \mathbf{E}_3 \mathbf{Z}_3 \mathbf{P}_{\sigma} = a_3^2 \mathbf{P}_{\sigma} + o_{\mathbb{P}}(1).$$

Therefore,

$$\mathbf{P}_{\sigma} \mathbf{T}_3 \mathbf{P}_{\sigma} = \mathbf{P}_{\sigma} \mathbf{Z}_3 \mathbf{R}^{\odot 3} \mathbf{Z}_3 \mathbf{P}_{\sigma} = a_3^2 \mathbf{P}_{\sigma} + \frac{3}{d+2} \mathbf{P}_{\sigma} \mathbf{Z}_3 \mathbf{R} \mathbf{Z}_3 \mathbf{P}_{\sigma} + o_{\mathbb{P}}(1).$$

Finally, (I.46) is derived from definition of $\mathbf{P}_{\sigma}$ in (I.11) and $\mathbf{Z}_3 \mathbf{R} \mathbf{Z}_3 = \mathbf{Z}_3 \mathbf{S}^{-1} \mathbf{G} \mathbf{S}^{-1} \mathbf{Z}_3 = \mathbf{B}_{3,\sigma} \mathbf{G} \mathbf{B}_{3,\sigma}$. ■

Lemma 74 (The quartic term) *Recall $a_4 := \mathbb{E}[\sigma(\xi)h_4(\xi)]$. Then $\mathbf{P}_{\sigma} \mathbf{T}_4 \mathbf{P}_{\sigma} = a_4^2 \mathbf{P}_{\sigma} + o_{\mathbb{P}}(1)$.*

Proof. The degree-four zonal spherical harmonic decomposition is

$$t^4 = \frac{3}{d(d+2)} + \frac{6}{d+4} \left(t^2 - \frac{1}{d} \right) + \psi_4^{(d)}(t), \quad (\text{I.48})$$

where $\psi_4^{(d)}(t)$ is defined in (I.36). Applying (I.48) entrywise to $\mathbf{R}$, define $\mathbf{Q}_2 := \mathbf{R}^{\odot 2} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ and $(\Psi_4)_{ij} := \psi_4^{(d)}(\mathbf{R}_{ij})$. Then

$$\mathbf{R}^{\odot 4} = \frac{3}{d(d+2)}\mathbf{1}\mathbf{1}^\top + \frac{6}{d+4}\mathbf{Q}_2 + \Psi_4.$$

Hence $\mathbf{P}_\sigma \mathbf{T}_4 \mathbf{P}_\sigma = \mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3$, where $\mathbf{I}_1 := \frac{3}{d(d+2)}\mathbf{P}_\sigma \mathbf{Z}_4 \mathbf{1}\mathbf{1}^\top \mathbf{Z}_4 \mathbf{P}_\sigma$, $\mathbf{I}_2 := \frac{6}{d+4}\mathbf{P}_\sigma \mathbf{Z}_4 \mathbf{Q}_2 \mathbf{Z}_4 \mathbf{P}_\sigma$, and $\mathbf{I}_3 := \mathbf{P}_\sigma \mathbf{Z}_4 \Psi_4 \mathbf{Z}_4 \mathbf{P}_\sigma$. We bound these three terms separately.

First, since $\mathbf{P}_\sigma \mathbf{1} = 0$, $\mathbf{P}_\sigma \mathbf{Z}_4 \mathbf{1} = \mathbf{P}_\sigma (\mathbf{Z}_4 - a_4 \mathbf{I}_n) \mathbf{1}$. By Lemma 13, we have $\|\mathbf{Z}_4 - a_4 \mathbf{I}_n\| \prec \sqrt{\frac{\log n}{d}}$ implying

$$\|\mathbf{P}_\sigma \mathbf{Z}_4 \mathbf{1}\|^2 \leq \|(\mathbf{Z}_4 - a_4 \mathbf{I}_n) \mathbf{1}\|^2 \prec n \frac{\log n}{d}.$$

Then we can conclude that $\mathbf{I}_1 = o_{\mathbb{P}}(1)$ since $n \asymp d^2$.

Second, by Lemma 71 with $m = 2$ and second-order Gegenbauer polynomial formula in (I.41), we know that $\|\mathbf{Q}_2\| = O_{\mathbb{P}}(1)$. Also $\|\mathbf{Z}_4\| = O_{\mathbb{P}}(1)$. Therefore $\|\mathbf{I}_2\| \leq \frac{C}{d}\|\mathbf{Q}_2\| = o_{\mathbb{P}}(1)$.

Third, again by Lemma 71 with $m = 4$, we have $\Psi_4 = \psi_4^{(d)}(1)\mathbf{I}_n + \mathbf{E}_4$, where $\|\mathbf{E}_4\| = o_{\mathbb{P}}(1)$. Moreover,

$$\psi_4^{(d)}(1) = 1 - \frac{6}{d+4} + \frac{3}{(d+2)(d+4)} = 1 + O(d^{-1}).$$

Therefore $\mathbf{I}_3 = \psi_4^{(d)}(1)\mathbf{P}_\sigma \mathbf{Z}_4^2 \mathbf{P}_\sigma + \mathbf{P}_\sigma \mathbf{Z}_4 \mathbf{E}_4 \mathbf{Z}_4 \mathbf{P}_\sigma$. The second term is $o_{\mathbb{P}}(1)$. Since $\|\mathbf{Z}_4 - a_4 \mathbf{I}_n\| = o_{\mathbb{P}}(1)$, we also have $\mathbf{P}_\sigma \mathbf{Z}_4^2 \mathbf{P}_\sigma = a_4^2 \mathbf{P}_\sigma + o_{\mathbb{P}}(1)$. Consequently, $\mathbf{I}_3 = a_4^2 \mathbf{P}_\sigma + o_{\mathbb{P}}(1)$.

Combining the estimates for $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3$ proves Lemma 74. $\blacksquare$

Lemma 75 (Population CK after projection) *Under the assumptions of Theorem 9,*

$$\|\mathbf{P}_\sigma \Phi(\mathbf{X}) \mathbf{P}_\sigma - (\alpha_0 \mathbf{P}_\sigma + \alpha_2 \mathbf{P}_\sigma \mathbf{H} \mathbf{P}_\sigma)\| \xrightarrow{\mathbb{P}} 0. \quad (\text{I.49})$$

Proof. In (I.26), the $k = 0$ term equals $\mu_\sigma \mu_\sigma^\top$ and the $k = 1$ term equals $\mathbf{A}_\sigma \mathbf{G} \mathbf{A}_\sigma$ by Lemma 65; both are killed by $\mathbf{P}_\sigma$. Following the proof of Lemma 5.2 of Wang and Zhu (2024), by Lemma 13, we can derive that with high probability

$$\left\| \mathbf{P}_\sigma \Phi(\mathbf{X}) \mathbf{P}_\sigma - \left((\alpha_0 - a_3^2 - a_4^2) \mathbf{P}_\sigma + \mathbf{P}_\sigma \left(\sum_{k=2}^4 \mathbf{T}_k \right) \mathbf{P}_\sigma \right) \right\| \lesssim \sqrt{\frac{\log n}{d}}. \quad (\text{I.50})$$

Then by applying Lemmas 67, 73 and 74 for $\mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4$, respectively, we can derive (I.49). $\blacksquare$

I.3. BBP Transition for the Projected Population CK

Lemma 76 (Population CK reduction under the projection) *Under the assumptions of Theorem 9, with $\alpha_2 = a_2^2 = \frac{c_\sigma^2}{2}$ and $\alpha_0 = \sum_{k \geq 3} a_k^2 = 1 - b_\sigma^2 - \frac{c_\sigma^2}{2}$, we have*

$$\|\mathbf{P}_\sigma \Phi(\mathbf{X}) \mathbf{P}_\sigma - (\alpha_0 \mathbf{P}_\sigma + \alpha_2 \mathbf{P}_\sigma \mathbf{H}_0 \mathbf{P}_\sigma)\| \xrightarrow{\mathbb{P}} 0, \quad (\text{I.51})$$

where $\mathbf{H}_0 = \mathbf{Y}_0^\top \mathbf{Y}_0$ and $\mathbf{Y}_0$ is defined in (I.7).

Proof. Lemma 75 shows that it suffices to control the difference between $P_\sigma H P_\sigma$ and $P_\sigma H_0 P_\sigma$. We split the proof into four steps. Notice that $H = Q^\top Q$ where Q is defined by (I.3). For each data point $x_i = z_i + m_i$, since $x_i x_i^\top = z_i z_i^\top + z_i m_i^\top + m_i z_i^\top + m_i m_i^\top$, we have that

$$q_i = \text{svec}\left(\frac{1}{d}I_d\right) + (E_2)_i + (E_1)_i + \text{svec}(m_i m_i^\top) \quad (\text{I.52})$$

where $(E_2)_i$ and $(E_1)_i$ are the i -th columns of E_2 and E_1 , respectively, in (I.5). Let $\bar{q} := \text{svec}(\frac{1}{d}I_d + \frac{r^2}{2d}(u_1 u_1^\top + u_2 u_2^\top))$. Hence we can decompose Q into

$$Q = \bar{q} \mathbf{1}^\top + E_2 + E_1 + \frac{\delta_d}{\sqrt{p}} s y^\top, \quad (\text{I.53})$$

where s is defined in (I.6), $\delta_d = \delta \sqrt{1 + \frac{1}{d}}$, and $\delta = \frac{r^2}{2}$. Thus $\left\| \frac{\delta_d - \delta}{\sqrt{p}} s y^\top \right\| = |\delta_d - \delta| \sqrt{\frac{n}{p}} = o(1)$.

By the definition of P_σ , we have $Q P_\sigma = Y_0 P_\sigma + o_{\mathbb{P}}(1)$. Moreover, we can claim that $\|Y_0 P_\sigma\| = O_{\mathbb{P}}(1)$. The traceless second-chaos component in E_2 has bounded operator norm in the quadratic regime, while the radial trace component is $o(1)$. The signal part has norm $\left\| \frac{\delta}{\sqrt{p}} s y^\top P_\sigma \right\| \leq \delta \sqrt{\frac{n}{p}} = O(1)$. Consequently, with Lemma 68,

$$\|P_\sigma H P_\sigma - P_\sigma H_0 P_\sigma\| = \|(Q P_\sigma)^\top (Q P_\sigma) - (Y_0 P_\sigma)^\top (Y_0 P_\sigma)\| = o_{\mathbb{P}}(1)$$

which proves the lemma. ■

Proposition 77 (Projected population CK: quadratic BBP transition) *Assume the XOR model and Assumption 1. Suppose $n/p \rightarrow \gamma \in (0, \infty)$, $r \in [0, \infty)$ is fixed, and $\alpha_2 = c_\sigma^2/2 > 0$. Let P_σ be defined by (I.10)–(I.11). Let $\Pi_\Phi^+(I)$ denote the spectral projector of $P_\sigma \Phi(X) P_\sigma$ associated with the interval I .*

- (a) *The ESD of $P_\sigma \Phi(X) P_\sigma$ converges weakly to ν_q in probability where ν_q is defined in (I.12).*
- (b) *If $\ell < 1/\sqrt{\gamma}$, then for every fixed $\epsilon > 0$,*

$$\left\| \Pi_\Phi^+([\alpha_0 + \alpha_2(\lambda_+(\gamma) + \epsilon), \infty)) \frac{\mathbf{y}}{\sqrt{n}} \right\|^2 \xrightarrow{\mathbb{P}} 0 \quad (\text{I.54})$$

where $\lambda_+(\gamma)$ is defined by (I.2). Thus there is no label-aligned outlier above the affine MP edge.

- (c) *If $\ell > 1/\sqrt{\gamma}$, then for every fixed sufficiently small $\epsilon > 0$,*

$$\left\| \Pi_\Phi^+([\Lambda_y - \epsilon, \Lambda_y + \epsilon]) \frac{\mathbf{y}}{\sqrt{n}} \right\|^2 \xrightarrow{\mathbb{P}} \text{Align}(\gamma, \ell). \quad (\text{I.55})$$

If, in addition, this interval contains a single eigenvalue with probability tending to one, then this eigenvalue λ_{lab} and its unit eigenvector $\hat{v}_{\text{lab}}$ satisfy $\lambda_{\text{lab}} \xrightarrow{\mathbb{P}} \Lambda_y$, and $\left| \left\langle \hat{v}_{\text{lab}}, \frac{\mathbf{y}}{\sqrt{n}} \right\rangle \right|^2 \xrightarrow{\mathbb{P}} \text{Align}(\gamma, \ell)$ where Λ_y and $\text{Align}(\gamma, \ell)$ are defined by (I.15) and (I.16) respectively.

Proof. Lemma 76 shows that it suffices to study the spectrum of $P_\sigma H_0 P_\sigma$. Notice that $H_0 = Y_0^\top Y_0$ and

$$Y_0 = E_2 + \theta_n s v^\top, \quad v = y/\sqrt{n}, \quad \theta_n \rightarrow \theta = \delta\sqrt{\gamma}.$$

The columns of E_2 are independent, centered Gaussian quadratic-chaos vectors with population covariance

$$\mathbb{E}[(E_2)_i (E_2)_i^\top] = \frac{2}{d^2} I_p = \frac{1 + o(1)}{p} I_p. \quad (\text{I.56})$$

The anisotropic MP local law Knowles and Yin (2017); Fan et al. (2026) gives

$$v^\top (\lambda I - E_2^\top E_2)^{-1} v \rightarrow -m_\gamma(\lambda), \quad (\text{I.57})$$

$$s^\top (\lambda I - E_2 E_2^\top)^{-1} s \rightarrow -\tilde{m}_\gamma(\lambda), \quad (\text{I.58})$$

$$s^\top E_2 (\lambda I - E_2^\top E_2)^{-1} v \rightarrow 0 \quad (\text{I.59})$$

uniformly for λ in compact subsets of $(\lambda_+(\gamma), \infty)$, where m_γ is the Stieltjes transform of the MP law ρ_γ^{MP} and $\tilde{m}_\gamma(\lambda) = (1 - \gamma)/\lambda + \gamma m_\gamma(\lambda)$ is the companion transform. Let $x = \sqrt{\lambda} > \|E_2\|$. Linearizing the rectangular matrix and using the matrix determinant lemma, an outlying singular value of Y_0 must satisfy

$$1 - \theta^2 \lambda m_\gamma(\lambda) \tilde{m}_\gamma(\lambda) = 0. \quad (\text{I.60})$$

This is the rectangular BBP equation of Benaych-Georges and Nadakuditi (2012). Solving it gives $\lambda = \lambda_{\text{out}}(\gamma, \ell)$ and the solution is outside $\text{supp}(\rho_\gamma^{\text{MP}})$ if and only if $\theta^2 > \sqrt{\gamma}$, i.e., $\ell > \frac{1}{\sqrt{\gamma}}$. Moreover, when $\ell > 1/\sqrt{\gamma}$, if $\hat{v}_0$ is the right singular vector of Y_0 corresponding to the separated singular value, then the singular-vector formula of Benaych-Georges and Nadakuditi (2012) yields

$$|\langle \hat{v}_0, v \rangle|^2 \xrightarrow{\mathbb{P}} \frac{\gamma \ell^2 - 1}{\gamma \ell (\ell + 1)}. \quad (\text{I.61})$$

Equivalently, $H_0 = Y_0^\top Y_0$ has a separated eigenvalue at $\lambda_{\text{out}}(\gamma, \ell)$ with the overlap in (I.61). In the subcritical regime, there is no separated label-aligned eigenvalue above $\lambda_+(\gamma)$. Also, with Lemma 63 and Benaych-Georges and Nadakuditi (2012), in the supercritical regime, we can get $\|(I - P_\sigma) \hat{v}_0\| \xrightarrow{\mathbb{P}} 0$.

We now use a deterministic compression lemma. If a symmetric matrix A has an isolated eigenpair (λ_*, u) and an orthogonal projector P satisfies $\|(I - P)u\| \rightarrow 0$, then the spectral projector of PAP in any fixed small interval around λ_* maps $Pu/\|Pu\|$ to itself up to $o(1)$. This follows immediately from

$$\left\| (PAP - \lambda_*) \frac{Pu}{\|Pu\|} \right\| \leq 2\|A\| \|(I - P)u\| + o(1)$$

and the spectral theorem. Applying this lemma to $A = H_0$, $P = P_\sigma$, and $u = \hat{v}_0$ shows that the BBP spectral mass of $P_\sigma H_0 P_\sigma$ near $\lambda_{\text{out}}(\gamma, \ell)$ has the same label overlap as in (I.61). Since $\text{rank}(I - P_\sigma) = o(n)$, the compression also leaves the MP bulk unchanged.

By Lemma 76, on $\text{Im}(P_\sigma)$, $P_\sigma \Phi(X) P_\sigma = \alpha_0 P_\sigma + \alpha_2 P_\sigma H_0 P_\sigma + o_{\mathbb{P}}(1)$. Since $\alpha_2 > 0$, this is an affine spectral map $\lambda \mapsto \alpha_0 + \alpha_2 \lambda$, up to an $o_{\mathbb{P}}(1)$ perturbation. Weyl's inequality gives the eigenvalue convergence, and the spectral-projector formulation follows from the Davis–Kahan theorem. The MP bulk is mapped to ν_q . The subcritical and supercritical claims (I.54)–(I.55) follow from the corresponding claims for H_0 . $\blacksquare$

I.4. BBP Transition for the Projected CK

Proposition 78 (Second BBP transition for linear-width projected CK) *Let*

$$\mathbf{K}_\sigma = \frac{1}{N} \mathbf{P}_\sigma \sigma(\mathbf{W}\mathbf{X})^\top \sigma(\mathbf{W}\mathbf{X}) \mathbf{P}_\sigma, \quad (\text{I.62})$$

and assume (I.1) holds. Assume that $\ell = \frac{r^4}{4} > \gamma^{-1/2}$. Recall Λ_y in (I.15), z and φ transforms defined in (I.14). Then the following hold.

- (a) The ESD of $\mathbf{K}_\sigma$ converges weakly in probability to $\mu_q = \rho_\phi^{\text{MP}} \boxtimes \nu_q$.
- (b) If $z'(-\frac{1}{\Lambda_y}) > 0$, then $\mathbf{K}_\sigma$ has exactly one label-aligned separated eigenvalue, and for every sufficiently small deterministic interval I_y around λ_y defined in (4.20), which is disjoint from $\text{supp}(\mu_q)$, $\mathbf{K}_\sigma$ has exactly one eigenvalue in I_y , and this eigenvalue converges in probability to λ_y . If $\hat{\mathbf{v}}_y$ is the associated unit eigenvector, then the alignment with the label $\langle \hat{\mathbf{v}}_y, \mathbf{y} \rangle / \sqrt{n}$ satisfies (4.21).
- (c) If either $\ell \leq \gamma^{-1/2}$ or $z'(-\frac{1}{\Lambda_y}) \leq 0$, then no such separated outlier exists.

Proof. We condition on $\mathbf{X}$ throughout the proof. Let $r_n := \text{rank}(\mathbf{P}_\sigma)$. By Lemma 63, $n - r_n = \text{rank}(\mathbf{I}_n - \mathbf{P}_\sigma)$ which is $o(n)$, and $\frac{r_n}{N} \rightarrow \phi$. Let $\mathbf{U}_+ \in \mathbb{R}^{n \times r_n}$ be an orthonormal basis of $\text{Range}(\mathbf{P}_\sigma)$, so that $\mathbf{P}_\sigma = \mathbf{U}_+ \mathbf{U}_+^\top$. Define the projected feature row

$$\mathbf{g}_a := \mathbf{U}_+^\top \sigma(\mathbf{w}_a^\top \mathbf{X})^\top \in \mathbb{R}^{r_n}, \quad a = 1, \dots, N.$$

Conditional on $\mathbf{X}$, the rows $\mathbf{g}_1, \dots, \mathbf{g}_N$ are independent and identically distributed. Since

$$\mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top \mathbf{X})^\top \mid \mathbf{X}] = \boldsymbol{\mu}_\sigma$$

and $\mathbf{P}_\sigma \boldsymbol{\mu}_\sigma = \mathbf{0}$, we have $\mathbb{E}[\mathbf{g}_a \mid \mathbf{X}] = \mathbf{0}$. Their conditional covariance is

$$\tilde{\boldsymbol{\Sigma}}_\sigma^+ := \mathbb{E}[\mathbf{g}_a \mathbf{g}_a^\top \mid \mathbf{X}] = \mathbf{U}_+^\top \boldsymbol{\Phi}(\mathbf{X}) \mathbf{U}_+. \quad (\text{I.63})$$

The nonzero eigenvalues of $\mathbf{P}_\sigma \boldsymbol{\Phi}(\mathbf{X}) \mathbf{P}_\sigma$ are exactly the eigenvalues of $\tilde{\boldsymbol{\Sigma}}_\sigma^+$. Moreover,

$$\mathbf{K}_\sigma = \mathbf{U}_+ \left(\frac{1}{N} \sum_{a=1}^N \mathbf{g}_a \mathbf{g}_a^\top \right) \mathbf{U}_+^\top.$$

Therefore the nonzero eigenvalues of $\mathbf{K}_\sigma$ are exactly those of the $r_n \times r_n$ sample covariance matrix $\tilde{\mathbf{K}}_\sigma := \frac{1}{N} \sum_{a=1}^N \mathbf{g}_a \mathbf{g}_a^\top$. We now verify the inputs of the nonlinear spiked covariance theorem from Wang et al. (2024).

First, by Lemma 63 and Proposition 77, the ESD of $\tilde{\boldsymbol{\Sigma}}_\sigma^+$ converges to ν_q and, in the supercritical regime ($\ell > \gamma^{-1/2}$) there is a label-aligned population covariance outlier $\hat{\Lambda}_y = \Lambda_y + o_{\mathbb{P}}(1)$, where $\Lambda_y = \alpha_0 + \alpha_2 \lambda_{\text{out}}(\gamma, \ell)$. If $\mathbf{v}_y^{\text{pop}}$ denotes the associated unit eigenvector, then the first BBP phase transition gives

$$\left| \left\langle \mathbf{v}_y^{\text{pop}}, \frac{\mathbf{y}}{\sqrt{n}} \right\rangle \right|^2 \xrightarrow{\mathbb{P}} \frac{\gamma \ell^2 - 1}{\gamma \ell (\ell + 1)}.$$

If $\ell \leq \gamma^{-1/2}$, then there is no separated label-aligned population spike.

Second, conditional on $\mathbf{X}$, we show that the row law of $\mathbf{g}_a$ satisfies the quadratic-form concentration assumptions required for the nonlinear spiked covariance theorem. The rows are projected nonlinear Gaussian feature vectors, and the projection is deterministic after conditioning on $\mathbf{X}$. For globally Lipschitz σ , the nonlinear Hanson–Wright inequality gives for deterministic matrices $\mathbf{A}$

$$\mathbf{g}_a^\top \mathbf{A} \mathbf{g}_a - \text{Tr}(\mathbf{A} \tilde{\Sigma}_\sigma^+) \prec \|\mathbf{A}\|_F,$$

after the standard covariance normalization.

Thus, on a high-probability event in $\mathbf{X}$, we may apply Theorem 13 Wang et al. (2024) to $\tilde{\mathbf{K}}_\sigma = \frac{1}{N} \sum_{a=1}^N \mathbf{g}_a \mathbf{g}_a^\top$ with population covariance $\tilde{\Sigma}_\sigma^+$, aspect ratio $\frac{r_n}{N} \rightarrow \phi$, bulk law ν_q , and population spike Λ_y . We can conclude that a population spike λ produces a separated sample outlier if and only if $z'(-\frac{1}{\lambda}) > 0$, and the sample outlier converges to $z(-\frac{1}{\lambda})$. Additionally, the corresponding sample eigenvector has squared overlap with deterministic test directions multiplied by the factor $\varphi(-\frac{1}{\lambda})$. Applying this with $\lambda = \Lambda_y$, if $z'(-\frac{1}{\Lambda_y}) > 0$, then $\tilde{\mathbf{K}}_\sigma$, and hence $\mathbf{K}_\sigma$, has exactly one sample eigenvalue converging to λ_y .

It remains to compute the overlap with the original label vector $\mathbf{v} := \frac{\mathbf{y}}{\sqrt{n}}$. Let $\tilde{\mathbf{v}} := \frac{\mathbf{U}_+^\top \mathbf{v}}{\|\mathbf{U}_+^\top \mathbf{v}\|_2}$. By the enlarged label-preservation lemma, $\|\mathbf{U}_+^\top \mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{P}_\sigma \mathbf{v} \xrightarrow{\mathbb{P}} 1$. Hence $\tilde{\mathbf{v}}$ is asymptotically the same direction as the projected label. Since the population eigenvector belongs to $\text{Range}(\mathbf{P}_\sigma)$, the first BBP phase gives

$$|\langle \mathbf{v}_y^{\text{pop}}, \mathbf{v} \rangle|^2 \xrightarrow{\mathbb{P}} \frac{\gamma \ell^2 - 1}{\gamma \ell(\ell + 1)}.$$

The eigenvector overlap statement of the nonlinear sample-covariance BBP theorem therefore yields

$$|\langle \hat{\mathbf{v}}_y, \mathbf{v} \rangle|^2 \xrightarrow{\mathbb{P}} \varphi\left(-\frac{1}{\Lambda_y}\right) \frac{\gamma \ell^2 - 1}{\gamma \ell(\ell + 1)}.$$

Finally, we expand the shrinkage factor. Since $z'(s) = \frac{1}{s^2} - \phi \int \frac{t^2}{(1+ts)^2} \nu_q(dt)$, we have

$$z'\left(-\frac{1}{\Lambda_y}\right) = \Lambda_y^2 \left(1 - \phi \int \frac{t^2}{(\Lambda_y - t)^2} \nu_q(dt)\right).$$

Therefore

$$\varphi\left(-\frac{1}{\Lambda_y}\right) = -\frac{(-1/\Lambda_y) z'(-1/\Lambda_y)}{z(-1/\Lambda_y)} = \frac{1 - \phi \int \frac{t^2}{(\Lambda_y - t)^2} \nu_q(dt)}{1 + \phi \int \frac{t}{\Lambda_y - t} \nu_q(dt)}. \quad (\text{I.64})$$

If $z'(-\frac{1}{\Lambda_y}) \leq 0$, then the population spike is not in the outlier-producing index set $I = \{i : z'(-1/\lambda_i) > 0\}$, and it sticks to the sample bulk and produces no outlier eigenvalue. If $\ell \leq \gamma^{-1/2}$, then the first population BBP transition does not occur, so there is no label-aligned population spike to feed into the linear-width sample-covariance BBP theorem. Hence in either case the quadratic label channel produces no separated linear-width CK outlier. $\blacksquare$

Proof of Theorem 9. Item (i) is due to the first part of Proposition 78. Item (ii) is proved by Proposition 77 since $\mathbb{E}[\mathbf{K}|\mathbf{X}] = \Phi(\mathbf{X})$. Item (iii) is based on the second and the third parts of Proposition 78, where φ transform is given by (I.64). $\blacksquare$