# Gaussian Equivalence for Self-Attention: Asymptotic Spectral Analysis of Attention Matrix

**Tomohiro Hayase**
AIST and CoeFont, Co., Ltd.

**Benoît Collins**
Kyoto University

**Ryo Karakida**
AIST and RIKEN AIP

## Abstract

Self-attention layers have become fundamental building blocks of modern deep neural networks, yet their theoretical understanding remains limited, particularly from the perspective of random matrix theory. In this work, we provide a rigorous analysis of the singular value spectrum of the attention matrix and establish the first Gaussian equivalence result for attention. In a natural regime where the inverse temperature remains of constant order, we show that the singular value distribution of the attention matrix is asymptotically characterized by a tractable linear model. We further demonstrate that the distribution of squared singular values deviates from the Marchenko–Pastur law, which has been believed in previous work. Our proof relies on two key ingredients: precise control of fluctuations in the normalization term and a refined linearization that leverages favorable Taylor expansions of the exponential. This analysis also identifies a threshold for linearization and elucidates why attention, despite not being an entrywise operation, admits a rigorous Gaussian equivalence in this regime.

## 1 INTRODUCTION

Self-attention, popularized by the Transformer (Vaswani et al., 2017), has become a standard component across modalities. Yet, compared with feedforward layers, the asymptotic behavior of self-attention remains less understood from the viewpoint of random matrix theory (RMT). For standard feedforward architectures, RMT has already provided a powerful toolkit. Pennington and Worah (2017) initiated nonlinear random matrix theory for deep learning by deriving the spectral distribution of random-feature Gram matrices, opening the door to precise spectral analyses of neural kernels (Fan and Wang, 2020), universality results for high-dimensional learning (Hastie et al., 2022; Montanari and Saeed, 2022; Goldt et al., 2022), and spectral characterizations of networks updated by finitely many training steps (Wang et al., 2024). On the signal-propagation side, RMT and free probability have characterized Jacobian spectra and dynamical isometry at initialization (Pennington et al., 2017, 2018) and clarified Fisher-information spectra in wide random networks (Karakida et al., 2019; Amari et al., 2019; Hayase and Karakida, 2021). High-dimensional limits often show universality, enabling tractable deterministic equivalents that capture limiting spectra and performance.

Recent theory has analyzed attention under several complementary asymptotic regimes. Infinite-head limits recover GP/NTK descriptions for multi-head attention (Hron et al., 2020); tensor-program analyses identify the exact infinite-width law of a single standard attention layer (Sakai et al., 2025); and infinite-size analyses study either training dynamics or signal propagation in deep transformers (Bordelon et al., 2024; Giorlandino and Goldt, 2026). Our focus is complementary: we analyze a single standard softmax attention matrix at random initialization in a proportional high-dimensional regime where the context length grows with the embedding dimension and $\beta = O(1)$ remains fixed. In this regime, we obtain an exact baseline for the bulk non-Perron singular-value distribution. We do not study training dynamics here; rather, the present theorem is intended as a building block for later analyses of trainability, depth-wise signal propagation, and early feature learning.

A recurring principle in these previous works on usual feedforward architectures is Gaussian equivalence: high-dimensional nonlinear models often share their asymptotic behavior with Gaussian models determined by a few statistics (Pennington and Worah, 2017). The

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

attention structure, however, brings a distinct difficulty. The softmax over pairwise scores produces a full matrix with row-wise normalization, coupling all tokens and breaking i.i.d. structure. Standard tools for entrywise nonlinearities do not directly apply.

We prove the first Gaussian equivalence for self-attention at random initialization in the fixed softmax-temperature regime, yielding a tractable linear model that asymptotically characterizes the bulk singular-value distribution of the attention matrix (Theorem 3.2). Our proof sharply controls fluctuations of the softmax normalizers and approximates the exponential by Taylor polynomials of growing degree and applies a linearization argument that preserves the bilinear score structure. This suggests a candidate breakdown scale where Gaussian equivalence fails and explains why attention, despite not acting entrywise, admits a rigorous Gaussian approximation in this regime. The predicted limit for the squared singular values departs from the classical Marchenko-Pastur law (contrary to prior claims (Nait Saada et al., 2025)), and we confirm the prediction numerically at a large dimension (Figure 1 and Section 5.1). We expect this Gaussian equivalence to provide a foundation for the statistical analysis of self-attention layers.

The singular spectrum is relevant because recent theory connects spectral or signal-propagation properties of attention to concrete failure modes of transformers. In particular, the gap between leading singular values has been identified as a mechanism for rank collapse in width (Nait Saada et al., 2025), while signal-propagation analyses at initialization link the scale of attention to rank collapse, entropy collapse, gradient imbalance, and trainability in deep transformers (Noci et al., 2022; Giorlandino and Goldt, 2026). Our result should therefore be read as a precise baseline for the bulk spectrum of standard softmax attention before repeated depth composition or optimization modifies the layer. Outliers beyond the Perron mode and long-context scalings beyond $\beta = O(1)$ are not resolved here and are discussed only heuristically in Sections 6.1 and 6.3.

### 1.1 Related Work

**Complementary asymptotic theories of attention.** Prior theoretical analyses of attention mostly operate in regimes different from ours. Infinite-head limits recover GP/NTK descriptions for multi-head attention (Hron et al., 2020). For single standard attention layers, Sakai et al. (2025) identify the exact infinite-width law via tensor programs, again in a different limit from the proportional regime considered here. At the level of deep transformers, Bordelon et al. (2024) analyze infinite limits of training dynamics, whereas Giorlandino and Goldt (2026) study signal propagation at initialization and characterize rank-collapse and entropy-collapse regimes. Our proportional-limit result is complementary to these works: it gives an exact random-initialization baseline for the bulk non-Perron singular-value distribution of a single standard softmax attention matrix when the context length and embedding dimension grow together.

**Spectral analysis of attention matrices.** Prior work (Nait Saada et al., 2025) modeled first-layer key-query attention with orthonormal inputs by a random Markov surrogate generated from i.i.d. weights followed by row normalization; for this surrogate, the bulk singular-value law is quartercircular (Bordenave et al., 2012), equivalently yielding a Marchenko–Pastur law for squared singular values. Our result instead treats the exact bilinear softmax attention matrix in a proportional regime and shows that its limiting bulk law differs from that surrogate prediction. A recent analysis (Liao et al., 2025) studies an entrywise attention variant under signal-plus-noise inputs and a full-plus-rank-one structure in the attention weights ($W_K^\top W_Q = I + w_K w_Q^\top$), and remarks that the normalizer is asymptotically constant for truncated-exponential attention up to a scaling factor. Our control of fluctuations of the normalization term provides related support for analogous normalizer replacements and may strengthen such entrywise approximations. Our results, however, target the bulk singular-value distribution of standard softmax attention with general bilinear scores.

**Nonlinear RMT, Gaussian equivalence, and early feature learning.** Beyond GP/NTK limits, a substantial random-matrix literature studies nonlinear random-feature and kernel matrices directly. Pennington and Worah (2017) initiated nonlinear random matrix theory for deep learning, Louart et al. (2018) derived deterministic equivalents for Gram matrices of random neural feature maps, and Benigni and Péché (2021) established rigorous bulk laws for nonlinear models. Gaussian equivalence has since been extended to structured-data settings (Goldt et al., 2022). In a related direction, Wang et al. (2024) study nonlinear spiked covariance models and signal propagation through trained features, providing spectral characterizations that go beyond static kernels. Closest to the present motivation, recent one-step feature-learning analyses show that even a very small amount of training can create low-rank spikes or spiked random-feature equivalents, thereby changing both spectra and generalization (Ba et al., 2022; Cui et al., 2024; Moniri et al., 2024). Our contribution extends this nonlinear RMT / Gaussian-equivalence program to standard softmax attention, where the softmax normalizer couples entries

across each row and the score matrix itself is not i.i.d.

A related but distinct line uses free probability to analyze deep-network operators rather than random-feature Gram matrices. Collins and Hayase (2023) established asymptotic freeness of layerwise Jacobians in Haar-orthogonal MLPs, providing a rigorous basis for spectral propagation analyses. At a more matrix-theoretic level, entrywise nonlinear transforms have also been studied from free-probabilistic and combinatorial viewpoints (Dabo and Male, 2024; Speicher and Wendel, 2024). Unlike these settings, attention is not an entrywise transform; we show that a softmax-type normalization nonetheless admits a Gaussian equivalent.

**Scale of inverse temperature.** Most prior work has theoretically analyzed self-attention in infinite limits with a fixed context length, such as infinite embedding dimension or infinitely many heads (Hron et al., 2020; Bordelon et al., 2024; Sakai et al., 2025). In contrast, we consider a proportional limit where the context length grows with the embedding dimension, allowing for longer contexts, which is a more realistic setting. Empirical and practical work suggests that the effective attention scale matters in long contexts: Zhou et al. (2025) analyze length-induced embedding collapse and propose TempScale as mitigation, and Nakanishi (2025) empirically motivate a context-dependent scaling in SSMax. On the theoretical side, Giorlandino and Goldt (2026) analyze a REM-motivated $\Theta(\sqrt{\log \ell})$ scaling in a complementary infinite-sequence-length setting and identify a critical parameter separating rank-collapse and entropy-collapse regimes. At a more task-specific level, Chiang and Cholak (2022) study single-position focusing and length generalization under explicit temperature adjustment. Our analysis is restricted to $\beta = O(1)$; Section 6 provides heuristic evidence that $\sqrt{\log \ell}$ may mark the breakdown of the present linearization, but a rigorous treatment of this scale remains open. This long-context perspective is also consistent with practical proposals that explicitly modify or rescale attention in order to preserve selectivity at large context lengths, including LogN-Scaling as used in Qwen (Bai et al., 2023), YaRN (Peng et al., 2023), and SWAN-GPT (Puvvada et al., 2025). On the theoretical side, Chen et al. (2026) prove a sharp phase transition in a simplified long-context attention model under explicit attention scaling. Their regime and critical law are different from ours, but together these results support the view that the scale of attention is a first-order issue in long-context transformers.

## 2 PRELIMINARIES

Let $X \in \mathbb{R}^{\ell \times d}$ be an input sequence of context length $\ell \in \mathbb{N}$ with $d \in \mathbb{N}$. Let $W^Q, W^K$ be $d \times d_{\mathrm{qk}}$ random matrices with entries independently sampled from $\mathcal{N}(0,1)$ such that the collection of their entries forms an independent family. Given the input matrix $X$, we define the score matrix $S \in \mathbb{R}^{\ell \times \ell}$ by

$$S = \frac{1}{\sqrt{d_{\mathrm{qk}}}} X W^Q (W^K)^\top X^\top. \tag{2.1}$$

We consider the same assumption on $X$ as (Nait Saada et al., 2025): $X$ is deterministic with $\ell \le d$ and $XX^\top = I_\ell$.

**Remark 2.1.** The difference in scaling conventions is immaterial. Specifically, one may either adopt (1) $\|X_{i:}\|^2 = 1$ with $W_{ij} \sim \mathcal{N}(0,1)$, or (2) $\|X_{i:}\|^2/d = 1$ with $W_{ij} \sim \mathcal{N}(0,1/d)$. While (2) is the more practical choice, it obscures the order of individual entries in the proof. For clarity and consistency of exposition, we employ the convention in (1) throughout this work, which is compatible with $XX^\top = I_\ell$.

Fix $\beta > 0$. We then define the $\ell \times \ell$ matrix $A$ by

$$A_{ij} = \frac{\exp\big(\beta S_{ij}\big)}{Z_i}, \tag{2.2}$$

where

$$Z_i = \sum_{j=1}^{\ell} \exp\big(\beta S_{ij}\big). \tag{2.3}$$

In particular, we write $A = \mathrm{softmax}(\beta S)$.

**Remark 2.2.** Let $a, b, x, y \in \mathbb{R}^\ell$ and write $\xi = a^\top W^Q (W^K)^\top b$, and $\zeta = x^\top W^Q (W^K)^\top y$. Then a direct computation yields $\mathrm{Cov}(\xi^2, \zeta^2)/2d = \|a\|_2^2 \|x\|_2^2 \, (b^\top y)^2 + (a^\top x)^2 \, \|b\|_2^2 \|y\|_2^2 + (d+1)\,(a^\top x)^2 (b^\top y)^2$. Thus, the entries of $XW^Q(W^K)^\top X^\top$ are **not** independent. Therefore, one can not directly apply the circular law theorem for Markov random matrices by (Bordenave et al., 2012).

**Definition 2.3.** For any $d \in \mathbb{N}$ and real matrix $Y \in \mathbb{R}^{d \times d}$, we denote its singular values (i.e. the square roots of the eigenvalues of $YY^\top$) by $s_1(Y) \ge s_2(Y) \ge \cdots \ge s_d(Y)$. The empirical **squared** singular value distribution $\nu_Y$ of $Y$ is defined as

$$\nu_Y = \mu_{YY^\top} = \frac{1}{d} \sum_{i=1}^{d} \delta_{s_i(Y)^2}, \tag{2.4}$$

where $\delta_s$ denotes the Dirac measure at $s \in \mathbb{R}$.

**Remark 2.4.** For each $q \in \mathbb{N}$, the $q$-th moment of $\nu_Y$ is given by

$$m_q(YY^\top) = \mathrm{tr}[(YY^\top)^q] = m_q(\nu_Y), \tag{2.5}$$

where $\mathrm{tr} = d^{-1}\,\mathrm{Tr}$ denotes the normalized trace.

## 3 GAUSSIAN EQUIVALENCE

Fix $\beta > 0$ and define

$$f(x) = \exp\big(\beta x - \tfrac{\beta^2}{2}\big) - 1, \tag{3.1}$$

so that $\mathbb{E}[f(\chi)] = 0$ when $\chi \sim \mathcal{N}(0,1)$. Throughout this paper, we denote by $\chi$ a random variable distributed as $\mathcal{N}(0,1)$, i.e. the standard normal distribution. For any function $g$, define

$$\theta_1(g) = \mathbb{E}[g(\chi)^2], \quad \theta_2(g) = \mathbb{E}[g'(\chi)]^2, \tag{3.2}$$

if the derivative and the expectations are well-defined. Set

$$\theta_1 = \theta_1(f) = e^{\beta^2} - 1, \quad \theta_2 = \theta_2(f) = \beta^2. \tag{3.3}$$

Note that $\theta_1 \geq \theta_2$ by $e^x \geq 1 + x$. Let $W$ be $\ell \times \ell$ random matrix of $\mathcal{N}(0,1)$ entries, with the family of all entries of $(W^Q, W^K, W)$ being independent. Define

$$Y^f_{\text{lin}} = \sqrt{\theta_2} S/\sqrt{\ell} + \sqrt{\theta_1 - \theta_2} W/\sqrt{\ell}, \tag{3.4}$$

where $S$ is given by (2.1).

We work in the proportional asymptotic regime

$$d, \ell, d_{\text{qk}} \to \infty, \tag{3.5}$$

such that

$$\frac{\ell}{d} \to \gamma, \qquad \frac{d_{\text{qk}}}{d} \to \psi, \tag{3.6}$$

for some constants $\gamma, \psi > 0$. Then, by the asymptotic freeness of $(W^Q, W^K, W)$, one has almost surely

$$\nu_{Y^f_{\text{lin}}} \xrightarrow{\text{moments}} \nu_\infty(\beta, \gamma, \psi) \tag{3.7}$$

as (3.5) with (3.6), where $\nu_\infty(\beta, \gamma, \psi)$ is a deterministic compactly supported probability measure on $\mathbb{R}_{\geq 0}$ depending only on $\beta, \gamma, \psi$.

**Remark 3.1.** Unless otherwise stated, we suppress the dependence on the parameters $\beta$, $\gamma$, and $\psi$ in the notation, and write $\nu_\infty(\beta, \gamma, \psi)$ simply as $\nu_\infty$. The limiting distribution $\nu_\infty$ is characterized by a fixed-point equation for its Cauchy–Stieltjes transform $G_{\nu_\infty}(z)$, which agrees with the formulation obtained in (Pennington and Worah, 2017). For a rigorous proof, we refer to (Benigni and Péché, 2021).

Now we have prepared to describe the main theorem.

**Theorem 3.2.** *Let $A \in \mathbb{R}^{\ell \times \ell}$ be the attention matrix from* (2.2)*. Then almost surely it holds that*

$$s_1(A) = \|AA^\top\|_\infty^{1/2} \to 1, \tag{3.8}$$

$$\nu_{\sqrt{\ell} A^\perp} \xrightarrow{\text{moments}} \nu_\infty, \tag{3.9}$$

*where*

$$A^\perp = A - u_\ell u_\ell^\top, \tag{3.10}$$

*with $u_\ell = (1, \ldots, 1)^\top/\sqrt{\ell}$, and,*

$$\nu_{\sqrt{\ell} A} \xrightarrow{\text{weakly}} \nu_\infty(\beta, \gamma, \psi), \tag{3.11}$$

*with* (3.5) *and* (3.6)*.*

The proof of Theorem 3.2 is deferred to Section 4.5.

In particular, Theorem 3.2 also fixes the bulk scale of the spectrum: after removing the Perron direction via the rank-one projection $u_\ell u_\ell^\top$, the empirical distribution of the singular values of $\sqrt{\ell}(A - u_\ell u_\ell^\top)$ converges almost surely to a nondegenerate limit $\nu_\infty$. Consequently, apart from the largest singular value (which tends to 1), almost all singular values of $A$ have magnitudes on the order of $d^{-1/2}$. This delineates a clean rank-one plus diffusive bulk structure for attention: one macroscopic mode at scale 1 and a sea of modes at scale $d^{-1/2}$.

**Remark 3.3** (Reductions)**.** By applying a $d \times d$ rank-$\ell$ orthogonal projection to both models $\sqrt{\ell} A$ and $Y^f_{\text{lin}}$, the case $\ell < d$ is obtained as a projected version of the case $\ell = d$. Hence, it suffices to prove the result for $\ell = d$. Next, by left-orthogonal invariance of the Ginibre ensemble and the independence of $(W^Q, W^K, W)$, the joint law of $(XW^Q, XW^K, W)$ is the same as that of $(W^Q, W^K, W)$. Hence, the distributional statements below are unchanged if one replaces $X$ with $I_d$. We, therefore, work without loss of generality with $X = I_d$. Lastly, since we take the proportional limit (3.6) with $\psi \in (0, +\infty)$, the rectangular shape of each $W^Q$ and $W^K$ does not cause a significant change in the proof. Thus, we also assume that $d_{\text{qk}} = d$ to avoid complicated notation. In summary, later in this section, we assume $\ell = d = d_{\text{qk}}$ and $X = I_d$.

## 4 EQUIVALENT RANDOM MATRIX MODELS

We next explain how Theorem 3.2 is established. The proof proceeds by a sequence of transformations, each of which preserves the asymptotic singular value distribution. Along the way, these transformations also suggest the underlying reason why the theorem holds. Table 1 summarizes the sequence of random matrices, all of which share the same limiting singular value distribution.

Intuitively, the argument relies on the representation

$$\sqrt{d} A = \left[\operatorname{diag}\left(\frac{Z}{\mathbb{E}[e^{\beta\chi}]d}\right)\right]^{-1} \frac{\exp(\beta S)}{\mathbb{E}[e^{\beta\chi}]\sqrt{d}}, \tag{4.1}$$

Table 1: Equivalent Random matrix models, in the sense of sharing the asymptotic singular value distributions.

| Notation | Definition | Key Parameters | Eq. |
|---|---|---|---|
| $\sqrt{d}A$ | $\sqrt{d}\,\mathrm{softmax}(\beta S)$ | $\beta > 0$ (constant). | (2.2) |
| $\sqrt{d}A^{\perp}$ | $\sqrt{d}(A - u_d u_d^{\top})$ | $u_d = (1, 1, \ldots, 1)/\sqrt{d}$ | (3.10) |
| $Y$ | $\exp(\beta S)/(e^{\beta^2/2}\sqrt{d})$ | - | (4.5) |
| $Y^f$ | $f(S)/\sqrt{d}$ | $f(x) = \exp(\beta x - \beta^2/2) - 1.$ | (4.11) |
| $Y^Q$ | $Q_{n_d}(S)/\sqrt{d}$ | $n_d = \lceil c \log d / \log\log d \rceil.$ | (4.14) |
| $Y^Q_{\mathrm{lin}}$ | $\sqrt{\theta_2^Q} S/\sqrt{d} + \sqrt{\theta_1^Q - \theta_2^Q} W/\sqrt{d}$ | $\theta_i^Q = \theta_i(Q_{n_d})$ $(i = 1, 2).$ | (4.21) |
| $Y^f_{\mathrm{lin}}$ | $\sqrt{\theta_2} S/\sqrt{d} + \sqrt{\theta_1 - \theta_2} W/\sqrt{d}$ | $\theta_1 = e^{\beta^2} - 1,\ \theta_2 = \beta^2.$ | (3.4) |

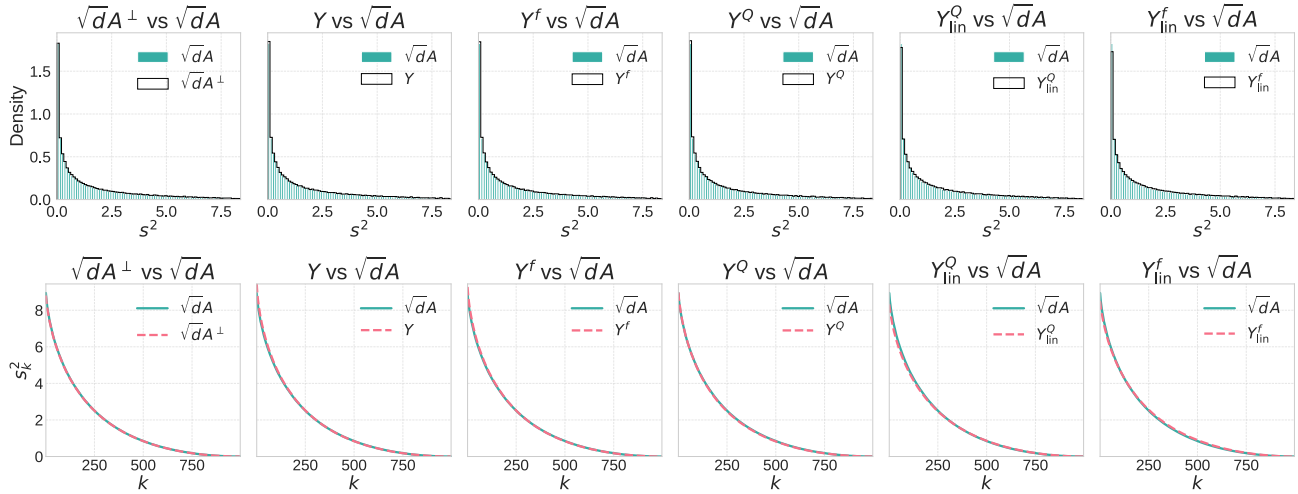


Figure 1: **Stepwise approximation of attention leading to the Gaussian Equivalence $Y^f_{\mathrm{lin}}$.** Top: histograms of empirical squared singular values; bottom: sorted $s_k^2$. Each panel compares the scaled attention matrix $\sqrt{d}A$ with a model introduced in the proof, shown left to right in the order $\sqrt{d}A^{\perp}, Y, Y^f, Y^Q, Y^Q_{\mathrm{lin}}, Y^f_{\mathrm{lin}}$. The bulk spectra (top three singular values removed; see Figure 2) are nearly indistinguishable, showing that the $\sqrt{d}A$ is accurately approximated throughout and culminates in $Y^f_{\mathrm{lin}}$. Settings: $d = d_{qk} = 1000$, $\beta = 1$, 10 draws.

where the normalizers $Z_i/d$ are first replaced by deterministic constants, after which the exponential function is expanded via a Taylor series to obtain a linear approximation. The proof consists of making each of these steps rigorous.

The central task is to ensure the simultaneous validity of three bounds: (i) the tail estimate for $|S_{ij}|$ with truncation parameter $K = c \log d / \log\log d$ or $(\log d)^{1/2+\delta}$, (ii) the fluctuations of the normalizer, controlled at order $d^{-1/2+\delta}$, and (iii) the polynomial approximation degree $n_d = \lceil c \log d / \log\log d \rceil$. Showing that these bounds are compatible constitutes the core of the overall strategy.

### 4.1 Fluctuation of Normalizer

In this section, we replace the normalization term with a deterministic constant. Let $X_1, \ldots, X_d \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The *law of the iterated logarithm* (LIL; (Durrett, 2019, Theorem 8.5.1)) determines the fluctuations of the sum;

$$\limsup_d \frac{|X_1 + \cdots + X_d|}{\sqrt{2d \log\log d}} = 1, \quad \text{a.s.} \tag{4.2}$$

Since $d^{\delta} > \sqrt{2 \log\log d}$ for any $\delta > 0$ and sufficiently large $d$, the following holds:

$$\lim_d d^{1/2-\delta} \frac{X_1 + \cdots + X_d}{d} = 0, \quad \text{a.s.} \tag{4.3}$$

The same conclusion with $\delta$ holds for the non-independent arrays after replacing $X_j$ by $S_{ij} - \mathbb{E}[S_{ij}]$ from Lemma B.10 without LIL. Precisely, we obtain concentration of $Z_i$ around the constant mean as follows: for fixed $\delta \in (0, 1/2]$, it holds a.s. as $d \to \infty$,

$$d^{1/2-\delta} \max_{i=1,2,\ldots,d} \left| d^{-1} Z_i - \mathbb{E}[e^{\beta \chi}] \right| \to 0. \tag{4.4}$$

Therefore, we define the approximation obtained by applying concentration:

$$Y = \frac{\exp(\beta S)}{e^{\beta^2/2}\sqrt{d}}, \tag{4.5}$$

where we have used $\mathbb{E}[e^{\beta\chi}] = e^{\beta^2/2}$.

**Lemma 4.1** (Concentration)**.** *Fix $\delta \in (0, 1/2]$. Then, almost surely, there exists a sequence $(\varepsilon_d)_d$ with $\lim_d d^{1/2-\delta}\varepsilon_d = 0$ such that*

$$\lim_{d\to\infty} \max_{i=1,2,\dots,d} \big| s_i(\sqrt{d}A) - (1+\varepsilon_d)s_i(Y) \big| = 0. \tag{4.6}$$

*Proof.* Define $D = \mathrm{diag}(Z_1, \dots, Z_d)/(\mathbb{E}[e^{\beta\chi}]d)$. Then

$$\sqrt{d}\,A = D^{-1}Y. \tag{4.7}$$

By Lemma B.10, we obtain the concentration of $Z_i$ around the constant mean as (4.4). By the diagonality of $D$, it holds that

$$\frac{s_i(Y)}{\max D} \le s_i(\sqrt{d}A) \le \frac{s_i(Y)}{\min D}. \tag{4.8}$$

From (4.4), $|1/D_{ii} - 1| = |1 - D_{ii}|/|D_{ii}| = o(d^{-1/2+\delta})/(1 + o(d^{-1/2+\delta})) = o(d^{-1/2+\delta})$. Thus $1/\min D, 1/\max D = 1 + o(d^{-1/2+\delta})$ as $d \to \infty$ a.s., which proves the assertion. □

From Lemma 4.1, we have the estimation of $s_1(A)$ without the detail of $\varepsilon_d$ in Proposition 4.2. We use the order of $\varepsilon_d$ in Section 4.5.

**Proposition 4.2.** $\lim_d s_1(A) = \lim_d s_1(Y/\sqrt{d}) = 1$, *a.s.*

*Proof.* By the symmetry of $S$, the same bound holds in the column direction as (4.4) and thus

$$\max_{j=1,2,\dots,d} \Big| \frac{1}{\sqrt{d}} \sum_{i=1}^{d} Y_{ij} - 1 \Big| \to 0, \quad \text{a.s.} \tag{4.9}$$

Since

$$\|YY^\top\|_\infty \le \Big(\max_j \sum_i Y_{ij}\Big)\Big(\max_i \sum_j Y_{ij}\Big), \tag{4.10}$$

we obtain $\limsup_d \|Y/\sqrt{d}\|_\infty \le 1$.

On the other hand, as $(Y/d)u_d = (1 + o(1))u_d$, we also have $\liminf_d \|Y/\sqrt{d}\|_\infty \ge 1$. In summary, $\lim_{d\to\infty} s_1(Y/\sqrt{d}) = 1$ a.s. Finally, by Lemma 4.1, we conclude that $\lim_d s_1(A) = 1$ a.s. □

### 4.2 Rank-One Perturbation and Interlacing

To apply linearization, we consider the function $f$ defined in (3.1) by normalizing $x \mapsto \exp(\beta x)$ so that $\mathbb{E}[f(\chi)] = 0$. Define the corresponding matrix by

$$Y^f = \frac{1}{\sqrt{d}} f(S). \tag{4.11}$$

In this formulation, $Y^f$ serves as an approximation of $Y$. Subtracting the mean corresponds exactly to removing a rank-one projection from $Y$: the two matrices differ only by a perturbation aligned with the all-ones vector. Since such rank-one updates affect singular values only in a tightly controlled manner, the spectra of $Y$ and $Y^f$ must remain close, a fact formalized by the interlacing theorem below.

**Lemma 4.3** (Interlacing)**.** *For $i = 1, \dots, d$, we have*

$$s_{i+1}(Y^f) \le s_i(Y) \le s_{i-1}(Y^f),$$

*under the convention $s_0 = +\infty$ and $s_{d+1} = 0$.*

*Proof.* Observe that $Y^f = Y - \sqrt{d}\,u_d u_d^\top$, where $u_d = (1, 1, \dots, 1)^\top/\sqrt{d}$. Thus, $Y$ differs from $Y^f$ by a rank-one perturbation. By the interlacing theorem for rank-one perturbations (Thompson, 1976, Theorem 1), the claim follows. □

### 4.3 Taylor Expansion and Strong Convergence

We introduce the linearization $Y^f_{\mathrm{lin}}$ of $Y^f$ following Benigni and Péché (2021) and using $\mathbb{E}[f(\chi)] = 0$. The idea is to approximate $f$ by centered polynomials $Q_n$ with $\mathbb{E}[Q_n(\chi)] = 0$; since linearization is proved for polynomial nonlinearities, a sufficiently accurate polynomial approximation transfers the result to $f$.

The essential component of $f$ is $e^x$, which does not satisfy the assumption (Benigni and Péché, 2021, (2.4)). However, the assumption can be relaxed: it suffices to require the approximation only for *large* degrees. Specifically, by Lemma B.11, it holds that for any $K > 0$ and all $n > K$,

$$\max_{\beta|x|\le K} \big| e^{\beta x} - P_n(\beta x) \big| \le \Big(\frac{eK}{n}\Big)^n, \tag{4.12}$$

where $P_n(x) = \sum_{k=0}^{n-1} x^k/k!$. This yields the entrywise approximation of $f(S)$ in Lemma B.13, on which the proof crucially relies.

Set

$$Q_n(x) = e^{-\beta^2/2}\Big(P_n(x) - \mathbb{E}\big[P_n(\chi)\big]\Big), \tag{4.13}$$

and define

$$Y^Q = \frac{1}{\sqrt{d}} Q_{n_d}(S), \tag{4.14}$$

where

$$n_d = \lceil c \frac{\log d}{\log\log d} \rceil, \tag{4.15}$$

for a constant $c > 1/(1-2\delta)$ with $\delta \in (0, 1/2)$, and $\lceil \cdot \rceil$ denotes the ceiling function.

**Lemma 4.4.** $||Y^f - Y^Q||_\infty \to 0$ *as* $d \to \infty$, *a.s.*

*Proof.* Let $K > 0$, then choose a good set

$$G_K = \{\max_{i,j} |S_{ij}| \le K\}. \tag{4.16}$$

Then the concentration of $\chi^2$-distribution,

$$P(G_K^c) \le Cd^2 \exp(-K^2/2). \tag{4.17}$$

Thus, we set

$$K = (\log d)^{1/2+\delta}, \tag{4.18}$$

then $P(G_K^c)$ is super-polynomial decay. Thus, on the good set $G_K$,

$$||Y^f - Y^Q||_\infty \le \sqrt{d} \sup_{x\in[-K,K]} |f(x) - Q_n(x)| \tag{4.19}$$

By Lemma B.13, there exist constants $c_1, c_2 > 0$

$$\sup_{x\in[-K,K]} |f(x) - Q_n(x)| \le c_2(c_1 K/n)^n, \tag{4.20}$$

where $n \in \mathbb{N}$ with $n > \beta K$ and sufficiently large $d$. Then $n/K \ge c(\log d)^{1/2-\delta}/\log\log d > 1$. Now $n\log(n/K) = c((1/2-\delta) + o(1))\log d$. With $c > 1/(1-2\delta)$, $n\log(n/K) > (1/2+\varepsilon)\log d$ and $(K/n)^n = o(d^{-1/2})$. Then the RHS of (4.20) is $o(1/\sqrt{d})$ and it proves the assertion. □

### 4.4 Gaussian Equivalence for Polynomials

For $n \in \mathbb{N}$, define

$$Y_{\text{lin}}^Q = \left(\sqrt{\theta_2^Q} S + \sqrt{\theta_2^Q - \theta_1^Q} W\right)/\sqrt{d}, \tag{4.21}$$

where $\theta_1^Q = \theta_1(Q_n)$ and $\theta_2^Q = \theta_2(Q_n)$. For each fixed $n$, let $\nu_{n,\infty} = \lim_{d\to\infty} \nu_{Y_{\text{lin}}^Q}$ denote the limiting singular-value distribution.

**Lemma 4.5** (Gaussian Equivalence)**.** *For any* $q \in \mathbb{N}$, *we have the self-averaging*

$$m_q(\nu_{Y^Q}) \xrightarrow[d\to\infty]{a.s.} \mathbb{E}[m_q(\nu_{Y^Q})]. \tag{4.22}$$

*Moreover, as* $d \to \infty$,

$$\mathbb{E}[m_q(\nu_{Y^Q})] = \big(1 + o(1)\big) m_q(\nu_\infty). \tag{4.23}$$

*Proof.* The almost-sure convergence (4.22) follows from (Benigni and Péché, 2021, Lemma 3.10). Take $n_d = c\log d/\log\log d$ as (4.15), Benigni and Péché (2021, Theorem 3.5) yields as $d \to \infty$,

$$\mathbb{E}[m_q(\nu_{Y^Q})] = \big(1 + o(1)\big) m_q(\nu_{n_d,\infty}). \tag{4.24}$$

Finally, by Lemma B.14, $m_q(\nu_{n_d,\infty}) \to m_q(\nu_\infty)$, which proves the assertion. □

### 4.5 Proof of Theorem

We are now ready to prove the main theorem. Here we use the order of $\varepsilon_d$ in Lemma 4.1. The argument proceeds by tracing back from $Y_{\text{lin}}^f$, for which the limiting distribution $\nu_\infty$ is known to exist.

*Proof of Theorem 3.2.* Fix $q \in \mathbb{N}$. We need to show that almost surely as $d \to \infty$,

$$m_q\big(\nu_{\sqrt{d}(A - u_d u_d^\top)}\big) \to m_q(\nu_\infty). \tag{4.25}$$

By Lemmas 4.5 and B.14, we have almost surely $m_q(\nu_{Y^Q})$ has the same limit as $m_q(\nu_{Y_{\text{lin}}^Q})$, i.e. $m_q(\nu_\infty)$. Since $\nu_\infty$ has finite moments of all orders, Lemmas 4.4 and D.1 further yields a.s.$m_q(\nu_{Y^f})$ has the same limit as $m_q(\nu_{Y^Q})$.

Next, by Lemma 4.1, writing $\mathcal{E} = D^{-1} - I_d$,

$$\sqrt{d}(A - u_d u_d^\top) = (I_d + \mathcal{E})Y^f + \sqrt{d}\mathcal{E} u_d u_d^\top, \tag{4.26}$$

where $\mathcal{E}$ is diagonal with $\max_i |\mathcal{E}_{ii}| \le \varepsilon_d = o(d^{-1/2+\delta})$. We bound the rank-one term: $m_q(d\mathcal{E}^2 u_d u_d^\top)^{1/q} \le d m_q(u_d u_d^\top)^{1/q} \varepsilon_d^2$. Since $m_q(u_d u_d^\top) = \frac{1}{d}\|u_d\|_2^2 = \frac{1}{d}$, we obtain

$$m_q\big(d\,\mathcal{E}^2 u_d u_d^\top\big)^{1/q} \le d^{\,1-1/q}\,\varepsilon_d^2. \tag{4.27}$$

Choosing $\delta = 1/(2q)$ and using $\varepsilon_d = o(d^{-1/2+\delta})$ gives $d^{\,1-1/q}\varepsilon_d^2 = o(1)$, hence $m_q\big(d\,\mathcal{E}^2 u_d u_d^\top\big)^{1/q} \to 0$. Therefore, by the triangle inequality applied to $L^q$-norm, a.s., $m_q\big(\nu_{\sqrt{d}A^\perp}\big) \to \lim_{d\to\infty} m_q\big(\nu_{(I_d+\mathcal{E})Y^f}\big) = m_q(\nu_\infty)$.

Finally, the weak convergence of $\nu_{\sqrt{d}A}$ follows from interlacing for rank-one perturbations together with the compactness of the support of $\nu_\infty$. □

### 4.6 Outliers and the Gap

Let us first check $Y_{\text{lin}}^f$ has no outliers, that is, $\|Y_{\text{lin}}^f\|^2 = s_1(Y_{\text{lin}}^f)^2$ converges to the right edge of $\nu_\infty$. Set $W_1 = W^Q/\sqrt{d}, W_2 = (W^K)^\top/\sqrt{d}, W_3 = W/\sqrt{d}$, then they are i.i.d. standard real Ginibre ensembles. We apply free probability theory (see Appendix C.1 for related notations) to handle the operator norm $\|Y_{\text{lin}}^f\|$. We only need to show $\|\alpha W_1 W_2 + \beta W_3\|$ converges to $\|\alpha c_1 c_2 + \beta c_3\|$, where $c_1, c_2, c_3$ are $*$-free circular elements.

To achieve this, it is enough to prove that $(W_1, W_2, W_3)$ converges strongly to $(c_1, c_2, c_3)$ in the sense of (Collins and Male, 2014), namely, for any non-commuting polynomial $P$ in three abstract non-commuting variables and its transposes, $\|P(W_1, W_2, W_3)\|$ converges to $\|P(c_1, c_2, c_3)\|$ in the large $d$-limit.

Note that in the case of complex Ginibre, this is already known, as the real part and imaginary part of a complex Ginibre are iid GUE and the strong convergence of GUE is established. In the orthogonal case, we can not rely on this trick and we are not available of any direct proof, although we believe that the most recent results of (van Handel, 2025) suffice to prove this result.

The easiest way to overcome this problem is to revisit the proof of Collins and Male (2014), starting with the fact that one knows the strong convergence of random orthogonal i.i.d Haar distributed matrices (real) $V_i$. Then, we make the following observation: There exists $c > 0$ such that for any $\varepsilon > 0$, there exists a real polynomial $P_\varepsilon$ and i.i.d orthogonal matrices $V_1, V_2, V_3$ such that $P(\big|\|V_1 P_\varepsilon(V_2 + V_2^\top)V_3\| - 2\big| > \varepsilon) = O(\exp(-c\varepsilon^2 d))$.

The existence of such a polynomial follows closely the proof of Collins and Male (2014). Namely, one can find a continuous increasing function $P$ such that the pushforward under $\mu$ of the arcsine distribution (the asymptotic eigenvalue counting distribution of $V + V^\top$) is the quarter-circle distribution. It follows from the strong convergence of a single real Ginibre matrix that this function $P$ satisfies $P(\big|\|V_1 P(V_2 + V_2^\top)V_3\| - 2\big| > \varepsilon) = O(\exp(-c\varepsilon^2 d))$. (Note that $\|W_i\| \to 2$, the right edge of the quarter-circular law.) The next step consists of replacing the continuous function $P$ by an increasing real polynomial $P_\varepsilon$ that is uniformly close to $P$ on an open set containing the closed interval $[-1, 1]$

We repeat this argument for $W_2, W_3$ and observe that all the Haar orthogonal variables involved $V_1, \ldots, V_9$ can be chosen to be i.i.d. Using a $3\varepsilon$ argument like in Collins and Male (2014), we get strong convergence and therefore the expected result.

We conclude that the potential outlier of $A$ is caused by the rank-one perturbation and the nonlinearity, because of the fact that $Y^f_{\text{lin}}$ has no outlier asymptotically. Further, by Proposition C.1, it holds that

$$\lim_{d\to\infty} \|Y^f_{\text{lin}}\|^2 = \max \operatorname{supp}(\nu_\infty) > 4(e^{\beta^2} - 1), \quad (4.28)$$

From this, the right edge of the bulk $\nu_\infty$ of $dAA^\top$ is strictly larger than that of the corresponding Marchenko-Pastur law ($= 4\theta_1$) with the i.i.d. assumption in Nait Saada et al. (2025), showing an effect of non-independent term $S$ in the attention matrix: it enlarges the bulk and decreases the gap between the bulk and $s_1(A)$.

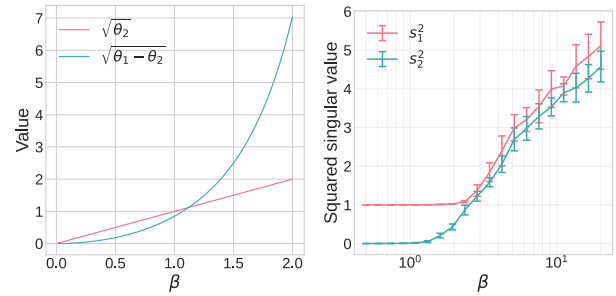


Figure 3: (Left): Growth of coefficients around $\beta = 1$. (Right): Growth of $s_1(A)^2$ and $s_2(A)^2$ with $d = 1000$ including $\beta \gg 1$.

## 5 NUMERICAL SIMULATIONS

### 5.1 Empirical Spectra

To verify, at finite $d$, that each approximation step in the proof chain is accurate, we compared six random matrix models against $\sqrt{d}\,A$ by examining the squared singular values. We set $n_T = d = d_{qk} = 1000$, used $n_s = 10$ samples, and fixed $\beta = 1$. For the polynomial approximation, we chose $c = 2$ and $n_d = 8$. Appendix E summarizes the settings. Figure 1 shows histograms and sorted curves of the squared singular values after removing the top three from each sample (the top three are shown separately in Figure 2). Across the six models, the bulks agree closely, consistent with the predicted limiting law. The bulk right edge exceeds $4\theta_1 = 4(e^{\beta^2} - 1)$ at $\beta = 1$, providing empirical support for (4.28). Figure 2 reveals a very large leading value $s_1^2 \sim d$ for $\sqrt{d}\,A$ and $Y$, as predicted. Even after removing this dominant outlier, the nonlinear models $\sqrt{d}\,A$, $\sqrt{d}(A - u_d u_d^\top)$, $Y$, $Y^f$, $Y^Q$ show additional outliers relative to the linearized models $Y^f_{\text{lin}}$, $Y^Q_{\text{lin}}$. A mechanism-level discussion is deferred to Section 6.3.

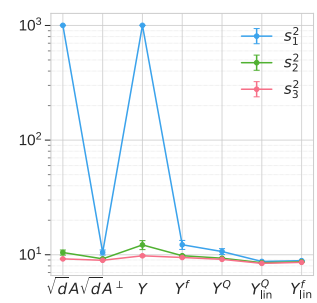


Figure 2: Plots of $s_1^2, s_2^2, s_3^2$.

### 5.2 Signal–Noise Balance

We quantify how the two Gaussian-equivalence components shape $\nu_\infty$ across $\beta$; the balance between $\sqrt{\theta_2}$ and $\sqrt{\theta_1 - \theta_2}$ dictates whether the model behaves more like a single Ginibre matrix or like $S/\sqrt{d}$. We examined how the comparison of the two scales varies with the inverse temperature $\beta$. In our setting, $\sqrt{\theta_2} = \beta$ and $\sqrt{\theta_1 - \theta_2} = \sqrt{e^{\beta^2} - 1 - \beta^2}$. Figure 3 (left) shows that these curves intersect at $\beta = 1.121 \pm 10^{-3}$.

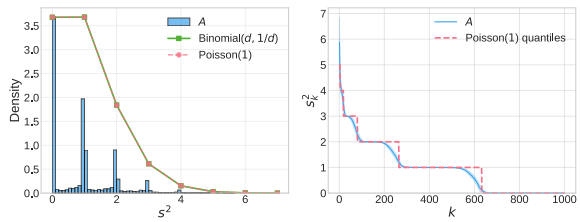


Figure 4: $\nu_A$ vs. Poisson(1). Left: histogram ($d = 1000$, $\beta = 50$, $n_s = 10$, bin width 0.1); right: sorted value and quantile plot using $F^{-1}(1 - k/d)$ for Poisson(1). Histogram heights differ because Poisson is discrete, but the quantiles match well.

### 5.3 Large Inverse Temperature

To probe $\beta$ well beyond the valid range of our bounds and demonstrate why those bounds matter, we increased $\beta$ at fixed $d = 1000$ with $n_s = 10$. Figure 3 (right) shows that $s_2(A)^2$ rises near $\beta \gtrsim 1$, while $s_1(A)^2$ grows for $\beta \gtrsim 2$, indicating that the softmax linearization breaks before the normalizer dominates. At $\beta = 50$, Figure 4 exhibits discrete atoms in $\nu_A$ aligned with Poisson(1) quantiles, consistent with an argmax-like limit. Theoretical thresholds and the Poisson heuristic are discussed in Section 6.1.

## 6 DISCUSSION

### 6.1 Threshold in The Inverse Temperature

Our rigorous analysis establishes rank collapse for $\beta = O(1)$. We argue for a threshold $\beta = \Theta(\sqrt{\log \ell}) = \Theta(\sqrt{\log d})$ that separates regimes. (Recall we take a proportional limit of $d$ and $\ell$.) From the fluctuation scale $d^{-1/2+\delta}$ of the row normalizer, the truncation level $K = (\log d)^{1/2+\delta}$, and the observation that the bulk right edge scales like $O(e^{\beta^2})$ (4.28), the evidence points to a threshold at $\beta = \Theta(\sqrt{\log d})$. Around this scale, the softmax linearization fails and the bulk right edge inflates, so additional singular directions can rise to $O(1)$. This conclusion is consistent with Figure 3(right) (growth of $s_2(A)^2$ near $\beta \gtrsim 1$).

In the further large $\beta$ regime, Figure 4 illustrates that $s_1(A)$ departs from 1 and $\nu_A$ develops Poisson-like atoms, consistent with an argmax limit at very large $\beta$. For $\beta \gg 1$, the softmax is approximated by the *argmax*. The distribution of squared singular values associated with a random argmax coincides with that of drawing $d$ balls independently with replacement into $d$ bins, which is Binomial$(d, 1/d)$; this converges to a Poisson distribution with rate $\lambda = 1$ as $d \to \infty$. Thus, we hypothesize that for very large $\beta$, $\nu_A$ is well approximated by Poisson(1).

### 6.2 General Input Matrix

We outline how to extend beyond orthonormal $X$ to arbitrary $X$. Generalizing the input matrix $X$ should be attainable by considering a four-factor random-matrix model $XW^Q(XW^K)^\top$ combined with an entrywise nonlinearity. In particular, an extension of (Speicher and Wendel, 2024) should accommodate this setting.

### 6.3 Outliers

We explain outlier mechanisms, the bulk misses, and sketch their modeling. Interlacing alone does not control extreme eigenvalues; therefore, so edge behavior requires separate analysis. Nonlinearity is likely to generate outliers. Following the suggestion of Benigni and Péché (2022) under strong assumptions on the nonlinearity, introducing a rank-one term with weight $\sqrt{\theta_3}$, where $\theta_3 = \mathbb{E}\big[\frac{1}{2}f''(\chi)\big]^2$, may capture such outliers. A rigorous treatment would need uniform high-moment bounds $q = q(d)$ together with uniform control of normalizer fluctuations, with a slowly growing function dictated by the LIL (4.2).

## 7 CONCLUSION

This work establishes a Gaussian equivalent model for the bulk singular value distribution of $\sqrt{d}\,A$ in the $\beta = O(1)$ regime. The proof relies on (i) the concentration of normalizers at scale $d^{-1/2+\delta}$, (ii) a tight polynomial approximation of $\exp(\beta x)$ with degree $n_d = \lceil c \log d / \log\log d \rceil$, and (iii) a polynomial linearization transferring limits to $f(S)$ and then to $A$. The resulting bulk diverges from the Marčenko–Pastur law due to the non-independence in $S$, and our simulations (Figure 1) corroborate the enlarged right edge. Open directions include the large-$\beta$ regime, extending to general $X$, multi-head attention, and a precise theory of outliers which will require uniform high-moment bounds.

## Acknowledgements

We thank the anonymous reviewers for their constructive feedback, which helped us improve the introduction and related work sections. T.H. was supported by JST BOOST, Japan, Grant Number JPMJBY24G4. R.K. was supported by JST FOREST, Japan, Grant Number JPMJFR226Q, and JSPS KAKENHI Grant Numbers 22H05116 and 23K16965.

## References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory (COLT)*, volume 178, pages 4310–4312. PMLR, 2022.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference (MSML)*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471, 2022.

Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Conference on Learning Theory (COLT)*, volume 247, pages 4891–4957. PMLR, 2024.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 1924–1932. PMLR, 2018.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 1032–1041. PMLR, 2019.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher information and natural gradient learning in random deep networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 694–702. PMLR, 2019.

Tomohiro Hayase and Ryo Karakida. The spectrum of Fisher information of deep networks achieving dynamical isometry. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 334–342. PMLR, 2021.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In *International Conference on Machine Learning (ICML)*, volume 119, pages 4376–4386. PMLR, 2020.

Mana Sakai, Ryo Karakida, and Masaaki Imaizumi. Infinite-width limit of a single attention layer: Analysis via tensor programs. *arXiv preprint*, 2025, arXiv:2506.00846.

Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=utSqpxQHXq`.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 45561–45587. PMLR, 2025. URL `https://proceedings.mlr.press/v267/nait-saada25a.html`.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in Transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Charles Bordenave, Pietro Caputo, and Djalil Chafaï. Circular law theorem for random Markov matrices. *Probability Theory and Related Fields*, 152(3-4):751–779, 2012.

Zhenyu Liao, Jiaqing Liu, TianQi Hou, Difan Zou, and Zenan Ling. A random matrix analysis of in-context memorization for nonlinear attention. *arXiv preprint*, 2025, arXiv:2506.18656v1 [stat.ML].

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *Annals of Applied Probability*, 28(2):1190–1248, 2018. URL `https://doi.org/10.1214/17-AAP1328`.

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 9662–9695. PMLR, 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. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 36106–36159. PMLR, 2024.

Benoît Collins and Tomohiro Hayase. Asymptotic freeness of layerwise Jacobians caused by invariance of multilayer perceptron: The Haar orthogonal case. *Communications in Mathematical Physics*, 397(1): 85–109, 2023.

Issa Dabo and Camille Male. A traffic approach for profiled Pennington-Worah matrices. *arXiv preprint*, 2024, arXiv:2409.13433.

Roland Speicher and Alexander Wendel. Entrywise application of non-linear functions on orthogonally invariant matrices. *arXiv preprint*, 2024, arXiv:2412.06943.

Yuqi Zhou, Sunhao Dai, Zhanshuo Cao, Xiao Zhang, and Jun Xu. Length-induced embedding collapse in PLM-based models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 28767–28791, 2025.

Ken M Nakanishi. Scalable-Softmax is superior for attention. *arXiv preprint*, 2025, arXiv:2501.19399.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of Self-Attention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7654–7664, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint*, 2023, arXiv:2309.16609.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. *arXiv preprint*, 2023, arXiv:2309.00071.

Krishna C. Puvvada, Faisal Ladhak, Santiago Akle Serrano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, and Boris Ginsburg. SWAN-GPT: An efficient and scalable approach for long-context language modeling. *arXiv preprint*, 2025, arXiv:2504.08719.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=7SLtElfqCW`.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge, UK, 5th edition, 2019.

Robert C Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1-2):69–78, 1976.

Benoît Collins and Camille Male. The strong asymptotic freeness of Haar and deterministic matrices. *Annales scientifiques de l'École Normale Supérieure*, 47(1):147–163, 2014.

Ramon van Handel. The strong convergence phenomenon. *arXiv preprint*, 2025, arXiv:2507.00346.

Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint*, 2022, arXiv:2201.04753 [math.PR].

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*, volume 335 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 2006.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2009.

Uffe Haagerup and Flemming Larsen. Brown's spectral distribution measure for R-diagonal elements in

finite von Neumann algebras. *Journal of Functional Analysis*, 176(2):331–367, 2000.

Florent Benaych-Georges. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144:471–515, 2009.

Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1): 201–220, 1991.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Materials

## A NOTATIONS

Throughout this paper, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say an event $E \in \mathcal{F}$ occurs *almost surely* if $\mathbb{P}(E) = 1$.

**Definition A.1.** For a sequence of probability distributions $\nu_d (d \in \mathbb{N})$ and $\nu$, We say $\nu_d$ converges to $\nu$ *weakly* if

$$\lim_{d\to\infty} \int g(x)\nu_d(dx) = \int g(x)\nu(dx), \tag{A.1}$$

for any bounded continuous function $g$ on $\mathbb{R}$.

We say $\nu_d$ $(d \in \mathbb{N})$ converges to $\nu$ *in moments* if

$$\lim_{d\to\infty} \int x^q \nu_d(dx) = \int x^q \nu(dx), \tag{A.2}$$

for any $q \in \mathbb{N}$.

For $x \in \mathbb{R}^d$, we write $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$. If there is no confusion, we omit the index 2. For matrix $A \in \mathbb{R}^{d\times d}$, we define the operator norm as follows:

$$\|A\|_\infty = \sup_{x\in\mathbb{R}^d, \|x\|_2=1} \|Ax\|_2. \tag{A.3}$$

In fact, $\|AA^\top\|_\infty = s_1(A)^2$, where $A^\top$ is the transposed matrix of $A$. Since in this paper we only consider the operator norm for matrices, we omit the index $\infty$ if there is no confusion.

**Definition A.2** (Asymptotic notation)**.** Let $f : \mathbb{N} \to \mathbb{R}_+$ and $g : \mathbb{N} \to \mathbb{R}_+$ be two nonnegative functions, and assume $f(d) > 0$ for all sufficiently large $d$. We write

$$g(d) = \Theta(f(d)) \quad \text{as } d \to \infty \tag{A.4}$$

if there exist constants $c_1, c_2 > 0$ and $d_0 \in \mathbb{N}$ such that $c_1 f(d) \le g(d) \le c_2 f(d), \forall d \ge d_0$. We write

$$g(d) = O(f(d)) \quad \text{as } d \to \infty \tag{A.5}$$

if there exists a constant $C > 0$ and $d_0 \in \mathbb{N}$ such that $0 \le g(d) \le C\, f(d), \forall d \ge d_0$. We write

$$g(d) = o(f(d)) \quad \text{as } d \to \infty \tag{A.6}$$

if for every $\varepsilon > 0$ there exists $d_0 \in \mathbb{N}$ such that $0 \le g(d) \le \varepsilon\, f(d), \forall d \ge d_0$, equivalently, $\lim_{d\to\infty} \frac{g(d)}{f(d)} = 0$.

## B LEMMAS

### B.1 Concentration of Normalizer

#### B.1.1 A Single Row

Let $v, w$ be i.i.d. standard Gaussian vectors. Write $\xi = \langle v, w\rangle/\sqrt{d}$. Further, write

$$\sigma = \|v\|/\sqrt{d}. \tag{B.1}$$

Define $\Omega_K = \{|\xi| \le K\}$.

**Lemma B.1.** $\mathbb{P}(\Omega_K^c) \le 2\exp(-K^2/2)$ *for* $K < \sqrt{d}$.

*Proof.* By Chernoff, $\mathbb{P}(\Omega_K^c) \le \inf_{0<\lambda<\sqrt{d}} \exp\left(-\lambda K - \frac{d}{2}\log(1-\lambda^2/d)\right)$. Now $-d\log(1-\lambda^2/d) \le \lambda^2$ since $a + \log(1-a) \ge 0$ for $0 < a < 1$. By $K < \sqrt{d}$, we can set $\lambda = K$ and get $\mathbb{P}(\Omega_K^c) \le 2\exp(-K^2 + K^2/2)$. □

**Lemma B.2.**

$$\mathbb{E}[e^{\xi} \mid \Omega_K, v] = \mathcal{L}_K(\sigma) := e^{\sigma^2/2}\,\frac{\Phi\left(\frac{K-\sigma^2}{\sigma}\right) - \Phi\left(\frac{-K-\sigma^2}{\sigma}\right)}{\Phi\left(\frac{K}{\sigma}\right) - \Phi\left(\frac{-K}{\sigma}\right)} \tag{B.2}$$

*where* $\Phi$ *is the standard normal CDF.*

*Proof.* Recall $\sigma^2 = \|v\|^2/d$. Then

$$\mathbb{E}[e^{\xi} \mid \Omega_K, v] = \frac{\displaystyle\int_{-K}^{K} e^{x}\,\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/(2\sigma^2)}\,dx}{\displaystyle\int_{-K}^{K} \frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/(2\sigma^2)}\,dx}. \tag{B.3}$$

Thus, we complete the square in the numerator:

$$e^{x-\frac{x^2}{2\sigma^2}} = e^{\sigma^2/2}\,\exp\left(-\frac{(x-\sigma^2)^2}{2\sigma^2}\right). \tag{B.4}$$

Thus, the assertion holds. □

Therefore, the conditional expectation only depends on $K$ and $\sigma = \|v\|/\sqrt{d}$. As $K \to \infty$, the ratio $\to 1$, so it converges to $\mathbb{E}[e^{\xi} \mid v] = e^{\sigma^2/2}$.

**Lemma B.3.** *For* $\beta > 0$, *we have* $\mathbb{E}[e^{\beta\xi} \mid \Omega_{K/\beta}, v] = \mathcal{L}_K(\beta\sigma)$ *with* $\sigma = \|v\|_2/\sqrt{d}$.

*Proof.* By replacing $v$ with $\beta v$, the assertion holds. □

Recall $\mathbb{E}[e^{\beta\xi}] = e^{\beta^2/2}$.

**Lemma B.4.** *Fix* $\delta \in (0,1)$. *Then*

$$\mathbb{P}\left(\left|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}\right| > \varepsilon\right) \le 2\exp\left(-\frac{d}{2}\left(\frac{(\varepsilon - \Delta^{\beta}(K))_+}{L_{\delta}^{\beta}(K)}\right)^2\right) + 2e^{-d\delta^2/2}.$$

*where* $(x)_+ = \max(x, 0)$,

$$\Delta^{\beta}(K) = |\mathcal{L}_K(\beta) - e^{\beta^2/2}|, \tag{B.5}$$

$$L_{\delta}^{\beta}(K) = \beta \sup_{|\sigma-1|\le\delta} |\mathcal{L}_K'(\beta\sigma)|. \tag{B.6}$$

*Proof.* Consider $G_\delta = \{|\sigma - 1| < \delta\}$. On $G_\delta$, by the mean-value theorem around $\sigma = 1$,

$$|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}| \le \Delta^{\beta}(K) + L_{\delta}^{\beta}(K)|\sigma - 1|. \tag{B.7}$$

Hence, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}\right| > \varepsilon\right) \le \mathbb{P}\left(|\sigma - 1| > \frac{(\varepsilon - \Delta^{\beta}(K))_+}{L_{\delta}^{\beta}(K)}\right) + \mathbb{P}(G_\delta^c). \tag{B.8}$$

Now use the $\chi$-concentration for $\sigma = \|v\|/\sqrt{d}$:

$$\mathbb{P}(|\sigma - 1| > u) \le 2\exp\left(-\frac{d\,u^2}{2}\right) \qquad (0 < u \le 1/2). \tag{B.9}$$

Therefore, we have the claim. □

**Lemma B.5** (Bias bound for truncated normalizer)**.** *Assume $K/\beta \geq 3/2$. Let*

$$\Delta^\beta(K) := \left|\mathcal{L}_K(\beta) - e^{\beta^2/2}\right|$$

*be the bias of the truncated normalizer at $\sigma = 1$. Then*

$$\Delta^\beta(K) \leq C_{\text{bias}}(\beta)\, \exp\Big( - \frac{(K-\beta^2)^2}{2\beta^2}\Big),$$

*where $C_{\text{bias}}(\beta) = \frac{\beta\, e^{\beta^2}}{\sqrt{2\pi}\,\left(2\Phi(3/2)-1\right)} = O(e^{\beta^2})$. In particular, if*

$$K \geq \beta^2 + \beta\sqrt{2\log\frac{2\,C_{\text{bias}}(\beta)}{\varepsilon}}\ ,$$

*then $\Delta^\beta(K) \leq \varepsilon/2$.*

*Proof.* Write $\rho := K/\beta$. Recalling that $\Phi$ is the standard normal CDF, we can express the bias as

$$\frac{\Delta^\beta(K)}{e^{\beta^2/2}} = \frac{2\Phi(\rho)\ -\ \Phi(\rho-\beta)\ -\ \Phi(\rho+\beta)}{2\Phi(\rho)-1}\,.$$

Using the inequality $e^{-t^2/2}\sinh(\rho t) \leq (\rho t)\, e^{-t^2/2}$ for $0 \leq t \leq \beta$, one finds that

$$2\Phi(\rho)\ -\ \Phi(\rho-\beta)\ -\ \Phi(\rho+\beta) \leq \frac{1}{\sqrt{2\pi}}\,\frac{\beta}{K}\,\exp\Big( - \frac{(K-\beta^2)^2}{2\beta^2} + \beta^2\Big).$$

Moreover, $2\Phi(\rho) - 1 \geq 2\Phi(3/2) - 1$ under the assumption $\rho \geq 3/2$. Combining these facts, we obtain

$$\Delta^\beta(K) \leq \frac{\beta\, e^{\beta^2}}{\sqrt{2\pi}\,\left(2\Phi(3/2)-1\right)K}\,\exp\Big( - \frac{(K-\beta^2)^2}{2\beta^2}\Big) \leq C_{\text{bias}}(\beta)\,\exp\Big( - \frac{(K-\beta^2)^2}{2\beta^2}\Big),$$

as claimed. In particular, solving $\Delta^\beta(K) \leq \varepsilon/2$ for $K$ yields the stated threshold on $K$, above which $\Delta^\beta(K) \leq \varepsilon/2$. □

**Lemma B.6** (Local Lipschitz continuity of $\mathcal{L}_K$)**.** *Assume $K/\beta \geq 3/2$. Then for all $\sigma$ with $|\sigma - 1| \leq 1/2$, it holds that*

$$\beta\left|\mathcal{L}'_K(\beta\sigma)\right| \leq K C_L(\beta), \tag{B.10}$$

*where*

$$C_L(\beta) = e^{9\beta^2/8}(c_0/\beta + c_1 + c_2\beta), \tag{B.11}$$

*and $c_1, c_2, c_3 > 0$ are constants. In particular, $L^\beta_{1/2}(K) \leq K C_L(\beta)$.*

*Proof.* We focus on $\mathcal{L}'_K(u)$ with $u \in [\beta/2,\, 3\beta/2]$, which corresponds to $|\sigma - 1| \leq 1/2$. In this range one checks that $D(u) := \Phi(K/u) - \Phi(-K/u) \geq 2\Phi(1) - 1$ (since $K/u \geq K/(3\beta/2) \geq 1$ by assumption) and $|N(u)| \leq 1$, where $N(u) := \Phi\big(\frac{K-u^2}{u}\big) - \Phi\big(\frac{-K-u^2}{u}\big)$. Writing $\mathcal{L}_K(u) = e^{u^2/2}\frac{N(u)}{D(u)}$ and differentiating, one can bound the derivative as

$$\left|\mathcal{L}'_K(u)\right| \leq e^{u^2/2}\left[\frac{|u|}{2\Phi(1)-1} + \frac{2}{\sqrt{2\pi}\,\left(2\Phi(1)-1\right)}\Big(1 + \frac{K}{u^2}\Big) + \frac{2}{\sqrt{2\pi}\,\left(2\Phi(1)-1\right)^2}\,\frac{K}{u^2}\right].$$

For $u \in [\beta/2, 3\beta/2]$, we have $e^{u^2/2} \leq e^{9\beta^2/8}$, $|u| \leq 3\beta/2$, and $K/u^2 \leq 4K/\beta^2$. Substituting these bounds above with $\beta/K$ and simplifying, we find $\beta\mathcal{L}'_K(u)$ is bounded by $K C_L(\beta)$ with (B.11), where $c_0, c_1, c_2 > 0$ are constants, which proves the assertion. □

**Lemma B.7.** *Fix $\beta > 0$ and $\varepsilon \in (0,1)$. Assume the* minimal natural truncation ratio $K \geq \frac{3}{2}\beta$. *There exists a threshold $K_0(\beta,\varepsilon) = \Theta\Big(\beta^2(1+\sqrt{\log\frac{1}{\varepsilon}})\Big)$ such that if $K \geq K_0(\beta,\varepsilon)$, then for all $d \in \mathbb{N}$,*

$$\mathbb{P}\Big(\big|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}\big| > \varepsilon\Big) \leq 2\exp\Big(-\,c(\beta)\,\frac{\varepsilon^2\,d}{K^2}\Big) + 2e^{-d/8},$$

*where $c(\beta) > 0$ is a constant that decays exponentially in $\beta^2$ (specifically, $c(\beta) = \Theta(e^{-\frac{9}{4}\beta^2})$). In particular, the deviation probability above is exponentially small in $d$ for fixed $\beta$, and $c(\beta)$ captures the $\beta$-dependence of the concentration rate.*

*Proof.* Apply Lemma B.4 with $\delta = \frac{1}{2}$. This gives the general tail bound

$$\mathbb{P}\Big(\big|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}\big| > \varepsilon\Big) \leq 2\exp\left(-\frac{d}{2}\left(\frac{\big(\varepsilon - \Delta^\beta(K)\big)_+}{L^\beta_{1/2}(K)}\right)^2\right) + 2e^{-d/8},$$

where $\Delta^\beta(K)$ and $L^\beta_{1/2}(K)$ are as defined in that lemma. Now, by Lemma B.5 we choose $K \geq K_0(\beta,\varepsilon)$ so that $\Delta^\beta(K) \leq \varepsilon/2$. Hence $(\varepsilon - \Delta^\beta(K))_+ \geq \varepsilon/2$. Moreover, by Lemma B.6 we have $L^\beta_{1/2}(K) \leq C_L(\beta)K$. Substituting these bounds into the expression above, we obtain

$$\exp\Big(-\,\frac{d}{2}\Big(\frac{\varepsilon/2}{L^\beta_{1/2}(K)}\Big)^2\Big) \leq \exp\Big(-\,\frac{\varepsilon^2\,d}{8\,C_L(\beta)^2K^2}\Big).$$

Thus, by writing $c(\beta) = 1/(8C_L(\beta)^2)$, we have the assertion. □

### B.1.2 Evaluation of Normalizer

**Lemma B.8.** *Let $v, w_1, \ldots, w_d$ be i.i.d. standard normal $d$-dimensional vectors. Write $\Omega_{d,K/\beta} = \cap_{j=1}^d\{|\langle v, w_j\rangle/\sqrt{d}| \leq K\}$ for $K > 0$. Write $\sigma = \|v\|_2/\sqrt{d}$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{j=1}^d \exp\left(\frac{\beta\langle v, w_j\rangle}{\sqrt{d}}\right) - \mathcal{L}_K(\beta\sigma)\right| > \varepsilon \mid \Omega_{d,K/\beta}, v\right) \leq \exp\left(-\frac{2\varepsilon^2}{(e^K - e^{-K})^2}d\right). \tag{B.12}$$

*In particular, the bound is uniform on $v$.*

*Proof.* Let $\xi_j = \exp(\beta\langle v, w_j\rangle/\sqrt{d}) \mid \Omega_{d,K/\beta}, v$. Then under the conditioning, $\xi_1, \ldots, \xi_d$ is IID family and bounded: $e^{-K} \leq \xi_j \leq e^K$. Thus, by Hoeffding's inequality (Hoeffding, 1963, Theorem 2), we have the claim. □

**Proposition B.9.** *Let $v, w_1, \ldots, w_d$ be i.i.d. $\mathcal{N}(0, I_d)$ in $\mathbb{R}^d$ and fix $\beta > 0$. For any $n \in \mathbb{N}$, $\delta \in (0, 1/2]$, and $\varepsilon > 0$, set $\varepsilon_d := d^{-1/2+\delta}\varepsilon$. Then*

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{j=1}^d \exp\Big(\frac{\beta\langle v, w_j\rangle}{\sqrt{d}}\Big) - e^{\beta^2/2}\right| > \varepsilon_d\right) = o(d^{-n}) \qquad (d \to \infty).$$

*Proof sketch.* Write $Z = \sum_j e^{\beta\langle v, w_j\rangle/\sqrt{d}}$ and $E = \{|Z/d - e^{\beta^2/2}| > \varepsilon_d\}$. Set $K := c\,\frac{\log d}{\log\log d}$ with any constant $c > 0$. Split $\mathbb{P}(E) \leq \mathbb{P}(E \cap \Omega_{d,K/\beta}) + \mathbb{P}(\Omega^c_{d,K/\beta})$. By Lemma B.1 and $K/\beta < \sqrt{d}$, $\mathbb{P}(\Omega^c_{d,K/\beta}) \leq 2d\exp(-\Theta(K^2/\beta^2))$. Conditioned on $\Omega_{d,K/\beta}$ and $v$, Lemma B.8 yields $\exp\big(-\Theta(d^{2\delta-\eta(d)})\big)$ (some $\eta(d) \to 0$). For the bias, Lemma B.7 implies $\mathbb{P}(|\mathcal{L}_K(\beta\sigma) - e^{\beta^2/2}| > \varepsilon_d/2) \leq 2\exp\big(-\Theta(\varepsilon^2 d^{2\delta}/K^2)\big) + 2e^{-d/8}$. Each term is $o(d^{-n})$ as $d \to \infty$ since $K = (\log d)/(\log\log d)$. □

**Lemma B.10.** *Fix $\delta \in (0, 1/2]$. Then $\max_{i=1,2,\ldots,d} d^{1/2-\delta}|\frac{1}{d}\sum_{j=1}^d \exp(\beta S_{ij}) - \mathbb{E}[e^{\beta\chi}]| \to 0$, almost surely.*

*Proof.* Let $\varepsilon > 0$ and $\varepsilon_d = d^{-1/2+\delta}\varepsilon$. Write $T_i = \frac{1}{d}\sum_j \exp(\beta S_{ij}) - \mathbb{E}[e^{\beta\chi}]$. Since $\mathbb{P}(d^{1/2-\delta}\max_{i=1}^d |T_i| > \varepsilon) = \mathbb{P}(\max_{i=1}^d |T_i| > \varepsilon_d)$, we only need to show $\sum_d \mathbb{P}(\max_{i=1}^d |T_i| > \varepsilon_d) < \infty$ for the almost-sure convergence. Now $T_1, \ldots, T_d$ is identically distributed; thus, $\mathbb{P}(\max_{i=1}^d |T_i| > \varepsilon_d) \le d\mathbb{P}(|T_1| > \varepsilon_d)$. By Proposition B.9, we have $\mathbb{P}(|T_1| > \varepsilon_d) = o(d^{-n})$ for any $n \in \mathbb{N}$. Choose $n > 2$. Then

$$\sum_d \mathbb{P}(\max_{i=1}^d |T_i| > \varepsilon_d) \le \sum_d Cd/d^n < \infty, \tag{B.13}$$

with a constant $C > 0$. It has proven the claim. □

### B.2 Polynomial Approximation

Fix $\beta > 0$. For $n \in \mathbb{N}$ define the $n$-term Taylor polynomial and the remainder by

$$P_n(\beta x) := \sum_{k=0}^{n-1} \frac{(\beta x)^k}{k!}, \qquad R_n(\beta x) := e^{\beta x} - P_n(\beta x). \tag{B.14}$$

Throughout, $\chi \sim \mathcal{N}(0,1)$ denotes a standard normal random variable, and $\mathrm{Poi}(\lambda)$ a Poisson variable with mean $\lambda > 0$.

**Lemma B.11.** *Let $K > 0$ and $n \in \mathbb{N}$ with $n > \beta K$. Then*

$$\max_{|x|\le K} |R_n(\beta x)| \le \Big(\frac{e\,\beta K}{n}\Big)^n. \tag{B.15}$$

*Proof.* Put $\mu := \beta|x|$. Then

$$R_n(\beta|x|) = \sum_{k=n}^{\infty} \frac{\mu^k}{k!} = e^{\mu}\,\mathbb{P}\big(N_\mu \ge n\big), \qquad N_\mu \sim \mathrm{Poi}(\mu). \tag{B.16}$$

By the Chernoff bound for Poisson tails, for $n > \mu$,

$$\mathbb{P}\big(N_\mu \ge n\big) \le \exp\big\{-n\log(n/\mu) + n - \mu\big\}. \tag{B.17}$$

Hence

$$|R_n(\beta x)| \le R_n(\beta|x|) \le \Big(\frac{e\,\mu}{n}\Big)^n \le \Big(\frac{e\,\beta K}{n}\Big)^n, \tag{B.18}$$

which completes the proof. □

**Lemma B.12.** *Let $m := \lceil n/2 \rceil$ and $\lambda := \beta^2/2$. Then*

$$\big|\mathbb{E}[e^{\beta\chi}] - \mathbb{E}[P_n(\beta\chi)]\big| = \sum_{r=m}^{\infty} \frac{\lambda^r}{r!} = e^{\lambda}\,\mathbb{P}\big(\mathrm{Poi}(\lambda) \ge m\big). \tag{B.19}$$

*In particular, if $n > \beta^2$ (equivalently $m > \lambda$), then*

$$\big|\mathbb{E}[e^{\beta\chi}] - \mathbb{E}[P_n(\beta\chi)]\big| \le \Big(\frac{e\,\lambda}{m}\Big)^m = \Big(\frac{e\,\beta^2}{2m}\Big)^m \le \Big(\frac{e\,\beta^2}{n}\Big)^{n/2}. \tag{B.20}$$

*Proof.* Using $\mathbb{E}[\chi^{2r}] = (2r)!/(2^r r!)$ and $\mathbb{E}[\chi^{2r+1}] = 0$,

$$\mathbb{E}[e^{\beta\chi}] - \mathbb{E}[P_n(\beta\chi)] = \sum_{k=n}^{\infty} \frac{\beta^k}{k!}\,\mathbb{E}[\chi^k] = \sum_{r=\lceil n/2\rceil}^{\infty} \frac{\beta^{2r}}{(2r)!}\,\mathbb{E}[\chi^{2r}] = \sum_{r=m}^{\infty} \frac{(\beta^2/2)^r}{r!}. \tag{B.21}$$

The series on the right equals $e^{\lambda}\mathbb{P}(\mathrm{Poi}(\lambda) \ge m)$. For $m > \lambda$, the Poisson Chernoff bound

$$\mathbb{P}\big(\mathrm{Poi}(\lambda) \ge m\big) \le \exp\big\{-m\log(m/\lambda) + m - \lambda\big\} \tag{B.22}$$

yields the stated inequality after multiplication by $e^{\lambda}$. □

**Lemma B.13.** *Define*

$$f(x) := e^{\beta x}/\mathbb{E}[e^{\beta\chi}] - 1, \tag{B.23}$$

$$Q_n(\beta x) := (P_n(\beta x) - \mathbb{E}[P_n(\beta\chi)])/\mathbb{E}[e^{\beta\chi}]. \tag{B.24}$$

*For any $K > 0$ and $n \in \mathbb{N}$ with $n > \max\{\beta K, \beta^2\}$,*

$$\max_{x\in[-K,K]} \big|f(x) - Q_n(\beta x)\big| \le e^{-\beta^2/2}\left[\Big(\frac{e\,\beta K}{n}\Big)^n + \Big(\frac{e\,\beta^2}{n}\Big)^{n/2}\right]. \tag{B.25}$$

*Proof.* By the triangle inequality,

$$e^{\beta^2/2} \max_{|x|\le K} \big|f(x) - Q_n(\beta x)\big| \le \max_{|x|\le K} |R_n(\beta x)| + \big|\mathbb{E}[e^{\beta\chi}] - \mathbb{E}[P_n(\beta\chi)]\big|. \tag{B.26}$$

Apply Lemma B.11 to the first term and Lemma B.12 (with $n > \beta^2$) to the second to obtain the claim. □

**Lemma B.14.** $\lim_{n\to\infty} m_q(\nu_{n,\infty}) = m_q(\nu_\infty)$

*Proof.* We only need to show $\theta_i(Q_n) \to \theta_i(f)$ $(i = 1, 2)$ as $d \to \infty$.

$$\theta_1(f) - \theta_1(Q_n) = \mathbb{E}[f(\chi)^2 - Q_n(\chi)^2] \tag{B.27}$$

Let $R_n(x) = e^{-\beta^2/2}[e^x - P_n(x)]$. Since $R_n' = R_{n-1}$, by CS,

$$|\theta_1(f) - \theta_1(Q_n)|^2 \le \mathbb{E}[R_n(\chi)^2]\mathbb{E}[(2f(\chi) - R_{n-1}(\chi))^2]. \tag{B.28}$$

$$\theta_2(f) - \theta_2(Q_n) = \mathbb{E}[f'(\chi)]^2 - \mathbb{E}[Q_n'(\chi)]^2 = \mathbb{E}[R_{n-1}(\chi)](\mathbb{E}[2e^\chi - R_{n-1}(\chi)]) \tag{B.29}$$

Take arbitrary $K > 1$. For $q = 1, 2$ and any $n > \max(\beta K, \beta^2)$, we have

$$|\mathbb{E}[R_n(\chi)^q]| \le \mathbb{E}[|R_n(\chi)|^q \mid |\chi| < K] + \mathbb{P}(|\chi| > K) \tag{B.30}$$

$$\le [e^{-\beta^2/2}(\Big(\frac{e\,\beta K}{n}\Big)^n + \Big(\frac{e\,\beta^2}{n}\Big)^{n/2})]^q + 2\exp(-K^2/2). \tag{B.31}$$

Since $K$ is arbitrary, $\lim_n \mathbb{E}[R_n(\chi)] = \lim_n \mathbb{E}[R_n(\chi)^2] = 0$. □

# C A LOWER BOUND FOR THE OPERATOR NORM

We compute the lower bound of $Y^f_{\text{lin}} = \alpha Z/\sqrt{d} + \beta W^Q(W^K)^\top/d$ with $\alpha = \sqrt{\theta_1 - \theta_2}$ and $\beta = \sqrt{\theta_2}$. From the strong convergence, we only need to find the maximum of the support of the limit distribution $\nu_\infty$.

We use tools of free probability. Background on circular and $R$-diagonal elements, transforms, and free convolutions can be found in Nica and Speicher (2006). The edge characterization via $K'(w) = 0$ is treated in Anderson et al. (2009). The $R$-diagonal singular-value calculus underlying $d = c_2c_3$ goes back to Haagerup and Larsen (2000); the Fuss–Catalan relation for $|d|^2$ is documented in Benaych-Georges (2009).

Let $c_1, c_2, c_3$ be $*$-free standard circular elements in a tracial C*-probability space and let $\alpha, \beta > 0$. Using hermitization and $R$-transform calculus, we show the strict lower bound.

$$\lim_{d\to\infty} \|Y^f_{\text{lin}}\|_\infty = \big\|\alpha c_1 + \beta c_2c_3\big\| > 2\sqrt{\alpha^2 + \beta^2}.$$

We also record the exact stationary equation determining the norm and derive asymptotic expansions in the regimes $\beta/\alpha \to 0$ and $\beta/\alpha \to \infty$.

### C.1 Definitions and Background

We work in a tracial C*-probability space $(\mathfrak{A}, \|\cdot\|, \mathrm{tr})$. Subalgebras are said to be free (resp. $*$-free) if all alternating products (resp. alternating products with choice of adjoint) of centered elements have zero trace (Voiculescu, 1991; Nica and Speicher, 2006). A *standard semicircular* element has density $(2\pi)^{-1}\sqrt{4-x^2}$ on $[-2,2]$. A *standard circular* element is $c = (s_1 + is_2)/\sqrt{2}$ with $s_1, s_2$ free standard semicirculars (Nica and Speicher, 2006); its operator norm equals 2.

For any $a \in \mathfrak{A}$, the *hermitization* is the self-adjoint element

$$\widetilde{a} = \begin{pmatrix} 0 & a \\ a^* & 0 \end{pmatrix} \in M_2(\mathfrak{A}), \qquad \|a\| = \|\widetilde{a}\|.$$

For self-adjoint $x$, write the Cauchy transform $G_x(z) = \mathrm{tr}\big[(z-x)^{-1}\big]$, the $K$-transform $K_x = G_x^{\langle -1\rangle}$ (functional inverse in a neighborhood of 0), and the $R$-transform $R_x(w) = K_x(w) - 1/w$. If $x, y$ are free and self-adjoint then $R_{x+y} = R_x + R_y$ (Nica and Speicher, 2006). For compactly supported laws, the right endpoint of the spectrum of $x$ is $K_x(w_*)$, where $w_* > 0$ is the unique real solution of $K_x'(w_*) = 0$; this is the standard "edge equation" characterization (see, e.g., Anderson et al., 2009, Ch. 2).

Two basic inputs will be used repeatedly (standard facts; see references cited inline):

1. If $c$ is standard circular, then $\widetilde{c}$ is semicircular. Hence $R_{\alpha\widetilde{c}}(w) = \alpha^2 w$ and $\|\alpha c\| = 2\alpha$ (Nica and Speicher, 2006).
2. If $d = c_2c_3$ is the product of two free standard circulars, then $d$ is $R$-diagonal and $|d|^2$ has the (order 2) Fuss–Catalan law, i.e. the free multiplicative square of the Marchenko–Pastur law. Its moment series $M(z)$ satisfies $M(z) = 1 + zM(z)^3$ (Benaych-Georges, 2009). Using the block resolvent identity $G_{\widetilde{d}}(z) = z\, G_{|d|^2}(z^2)$ (spectral symmetrization of singular values), one obtains

$$R_{\beta\widetilde{d}}(w) = \frac{\beta^2 w}{1-\beta^2 w^2}, \qquad 0 < w < 1/\beta, \tag{C.1}$$

   see also Haagerup and Larsen (2000) for $R$-diagonal calculus and singular-value relations.

### C.2 Hermitization

Let $X = \alpha c_1 + \beta c_2 c_3$ with $c_1, c_2, c_3$ $*$-free standard circulars and $\alpha, \beta > 0$. By freeness,

$$\widetilde{X} = \alpha\, \widetilde{c_1} \;\boxplus\; \beta\, \widetilde{(c_2c_3)}.$$

Using the two inputs above, its $K$-transform is

$$K(w) = \frac{1}{w} + \alpha^2 w + \frac{\beta^2 w}{1-\beta^2 w^2}, \qquad 0 < w < 1/\beta. \tag{C.2}$$

Let $w_* \in (0, 1/\beta)$ be the unique solution to $K'(w_*) = 0$. Then

$$\|X\| = \|\widetilde{X}\| = K(w_*).$$

It is convenient to also record the dimensionless stationary equation: writing $y = \beta^2 w^2$ and $\gamma = (\alpha/\beta)^2$, the equation $K'(w) = 0$ is equivalent to

$$\gamma y^3 - 2\gamma y^2 + (\gamma+3)y - 1 = 0, \qquad y \in (0,1), \tag{C.3}$$

which has a unique solution in $(0,1)$.

### C.3 A Strict Lower Bound

**Proposition C.1.** *For all $\alpha, \beta > 0$,*

$$\big\|\alpha c_1 + \beta c_2 c_3\big\| > 2\sqrt{\alpha^2+\beta^2}.$$

*Proof.* By (C.2) and $K'(w_*) = 0$,

$$\|X\| = K(w_*) = \frac{1}{w_*} + \alpha^2 w_* + \frac{\beta^2 w_*}{1-\beta^2 w_*^2} = \frac{1}{w_*} + (\alpha^2+\beta^2)w_* + \frac{\beta^4 w_*^3}{1-\beta^2 w_*^2}.$$

Since $w_* \in (0, 1/\beta)$, the last term is strictly positive. Therefore

$$\|X\| > \frac{1}{w_*} + (\alpha^2+\beta^2)w_*.$$

For any $A > 0$ and $w_* > 0$, the AM–GM inequality gives $\frac{1}{w_*} + Aw_* \geq 2\sqrt{A}$. Taking $A = \alpha^2 + \beta^2$ yields the strict inequality

$$\|X\| > 2\sqrt{\alpha^2+\beta^2}.$$

□

**Remark C.2.** The function $K$ in (C.2) is strictly convex on $(0, 1/\beta)$, since $K''(w) = 2w^{-3} + 2\beta^4 w(1-\beta^2 w^2)^{-2} + 4\beta^6 w^3(1-\beta^2 w^2)^{-3} > 0$. Hence $w_*$ is the unique global minimizer of $K$, and $K(w) \to \infty$ at both endpoints, so indeed $\|X\| = K(w_*)$ as used above (cf. Anderson et al., 2009).

# D TECHNICAL LEMMAS

**Lemma D.1.** *Let $(X_d)_{d\geq 1}$ and $(Y_d)_{d\geq 1}$ be sequences of real $d \times d$ matrices. Let $\nu$ be a compactly supported probability distribution on $[0, +\infty)$. Assume that $\|X_d - Y_d\|_\infty \to 0$, $m_q(YY^\top) \to m_q(\nu)(q \in \mathbb{N})$ then $m_q(XX^\top) \to m_q(\nu)(q \in \mathbb{N})$.*

*Proof.* It holds that $\max_i |s_i(X) - s_i(Y)| \leq \|X - Y\|_\infty$ and thus $C_q := \sup_d m_q(XX^\top) < \infty$. Thus, by the Markov inequality, $\frac{1}{d}\sum_{i:s_i\geq R} |s_i(XX^\top)^q| \leq C_{2q}/R^2$. Let $K_R(t) = \min(t^{2q}, R^{2q})$. By the Lipschitz bound of $K_R$, the moments converge to the desired value on the good set. Since $R > 0$ is arbitrary, we have proven the claim. □

**Lemma D.2.** *Let $(X_d)$ and $(Y_d)$ be $d \times d$ real matrices and write $s_i(A)$ for the singular values of $A$. Assume*

$$\|X_d - Y_d\|_\infty \to 0, \qquad m_q(Y_d Y_d^\top) := \frac{1}{d}\sum_{i=1}^{d} s_i(Y_d)^{2q} \to m_q(\nu) \in [0, \infty) \tag{D.1}$$

*for every $q \in \mathbb{N}$. Then for every $q \in \mathbb{N}$,*

$$m_q(X_d X_d^\top) \to m_q(\nu). \tag{D.2}$$

*Proof.* **Step 1 (uniform moment bound for $X_d$).** By Weyl's inequality for singular values,

$$\max_{1\leq i\leq d} \big|s_i(X_d) - s_i(Y_d)\big| \leq \|X_d - Y_d\|_\infty. \tag{D.3}$$

Hence, for any $q \in \mathbb{N}$ and all $i$,

$$s_i(X_d)^{2q} \leq 2^{2q-1}\big(s_i(Y_d)^{2q} + \|X_d - Y_d\|_\infty^{2q}\big). \tag{D.4}$$

Averaging,

$$m_q(X_d X_d^\top) \leq 2^{2q-1}\Big(m_q(Y_d Y_d^\top) + \|X_d - Y_d\|_\infty^{2q}\Big). \tag{D.5}$$

Since $m_q(Y_d Y_d^\top)$ converges and $\|X_d - Y_d\|_\infty \to 0$, we obtain

$$C_q := \sup_d m_q(X_d X_d^\top) < \infty. \tag{D.6}$$

**Step 2 (truncation and Lipschitz control on the "good" set).** Fix $q \in \mathbb{N}$ and $R > 0$. Let $K_R(t) = \min\{t^{2q}, R^{2q}\}$ on $[0, \infty)$. Then $K_R$ is Lipschitz with constant $L_R = 2q\, R^{2q-1}$. By the bound on singular value deviations,

$$\frac{1}{d}\sum_{i=1}^{d}\Big|K_R\big(s_i(X_d)\big) - K_R\big(s_i(Y_d)\big)\Big| \le L_R\, \|X_d - Y_d\|_\infty \to 0. \tag{D.7}$$

**Step 3 (tail bound).** By Markov's inequality and Step 1,

$$\frac{1}{d}\sum_{i:\, s_i(X_d)\ge R} s_i(X_d)^{2q} \le \frac{1}{R^2}\cdot\frac{1}{d}\sum_{i=1}^{d} s_i(X_d)^{2q+2} = \frac{m_{q+1}(X_d X_d^\top)}{R^2} \le \frac{C_{q+1}}{R^2}. \tag{D.8}$$

The same bound holds with $X_d$ replaced by $Y_d$ since $m_{q+1}(Y_d Y_d^\top)$ converges. Therefore,

$$\limsup_{d\to\infty}\Big|m_q(X_d X_d^\top) - m_q(Y_d Y_d^\top)\Big| \le \limsup_{d\to\infty}\frac{1}{d}\sum_{i=1}^{d}\Big|K_R\big(s_i(X_d)\big) - K_R\big(s_i(Y_d)\big)\Big| + \frac{2C_{q+1}}{R^2}. \tag{D.9}$$

The first term tends to 0 by Step 2. Letting $R \to \infty$ yields

$$m_q(X_d X_d^\top) - m_q(Y_d Y_d^\top) \to 0. \tag{D.10}$$

Since $m_q(Y_d Y_d^\top) \to m_q(\nu)$, we conclude $m_q(X_d X_d^\top) \to m_q(\nu)$. □

# E SETTINGS FOR NUMERICAL SIMULATIONS

All experiments were conducted in Python 3.10.12 on a Linux system. Unless otherwise noted, the plotted curves report the mean over 10 independent random trials, and error bars indicate one standard deviation; in several figures, the error bars are visually negligible due to their small magnitude.