Gaussian Equivalence for Self-Attention

Asymptotic Spectral Analysis of Attention Matrix

Tomohiro Hayase, Benoit Collins, Ryo Karakida

AISTATS 2026

Why Analyze the Spectrum of Attention?

① Signal Propagation

At random initialization, the spectrum determines gradient propagation and training stability. Well established for FFN, but unknown for attention.

② Understanding Rank Collapse

The collapse of attention output to a low-rank subspace can be quantitatively understood from the spectral structure.

Scaling Limits of Self-Attention

Each prior approach scales a different combination of model and context dimensions

(Embedding dim. d, # of heads H, context length ℓ)

(d, H, ℓ)	**Approach**
$(\infty, \infty, \Theta(1))$	NNGP & NTK [Hron+ ‘20]
$(\infty, \Theta(1), \Theta(1))$	Tensor program [Sakai, RK & Imaizumi ‘25]
$(\Theta(1), \infty, \Theta(1))$	Dynamical mean field [Bordelon+ ‘24]
$\boldsymbol{(\infty, \Theta(1), \infty)}$	**This work [Hayase, Collins & Karakida ‘26]**

$$f(X) = \mathrm{softmax}\left(\beta \frac{XW^Q (W^K)^\top X^\top}{\sqrt{d}}\right) XW^V$$

Self-Attention: Setup

Score matrix:

$$S = \frac{1}{\sqrt{d}} X W^Q (W^K)^\top X^\top$$

Attention matrix:

$$A_{ij} = \frac{\exp(\beta S_{ij})}{Z_i}, \quad Z_i = \sum_{j=1}^{\ell} \exp(\beta S_{ij})$$

Simplified key assumption

$X \in \mathbb{R}^{\ell \times d}$, $XX^\top = I_\ell$

$W^Q, W^K \in \mathbb{R}^{d \times d_{qk}}$: iid $\mathcal{N}(0,1)$ entries

$\beta > 0$: inverse temperature (constant)

Limit: $d = \ell = d_{qk} \to \infty$

The paper treats the general proportional limit $\ell/d \to \gamma$, $d_{qk}/d \to \psi$; we set $\gamma = \psi = 1$ here for clarity.

Prior Work: The IID Simplification

Nait Saada et al. (2025)

Models attention as an i.i.d. random Markov matrix → Apply circular law → Conclude that squared singular values follow MP law

This Work

Entries of S are NOT independent → The i.i.d. assumption is invalid → Asymptotic spectrum differs from MP law

$\mathrm{Cov}(\xi^2, \zeta^2)$ contains a cross term $(a^\top x)^2(b^\top y)^2$ → entries have non-trivial dependencies.

Main Theorem: Gaussian Equivalence

Theorem 3.2

① Top singular value: $s_1(A) \to 1$ (a.s.)

② Bulk: $\nu_{\sqrt{\ell}A^\perp} \overset{\text{moments}}{\to} \nu_\infty$, where $A^\perp = A - u_\ell u_\ell^\top$ (Perron direction removed)

③ ν_∞ has the same limit as the linear model Y^f_{lin}:

$$Y^f_{\text{lin}} = \frac{\sqrt{\theta_2}S}{\sqrt{\ell}} + \frac{\sqrt{\theta_1 - \theta_2}W}{\sqrt{\ell}}, \quad \theta_1 = e^{\beta^2} - 1, \quad \theta_2 = \beta^2$$

In the constant-β regime, the spectrum of the attention matrix is fully described by the linear model “score matrix + independent noise”.

Structure of Attention at Initialization

Rank-1 Component

$s_1(A) \to 1$

From the row-sum=1 constraint of softmax. Energy concentrates on the Perron direction $u = (1, \dots, 1)^\top/\sqrt{\ell}$.

Diffuse Bulk

Remaining singular values $\sim O(\ell^{-1/2})$

After scaling, $\sqrt{\ell}(A - uu^\top)$ converges to a non-degenerate continuous distribution ν_∞.

$$A \approx uu^\top + \frac{1}{\sqrt{\ell}} \times [\text{Gaussian Equivalent bulk}]$$

One macroscopic mode (scale 1) + a sea of $O(\ell^{-1/2})$ modes

Proof Strategy: Chain of Approximations

(Reminder: $\ell = d = d_{qk}$ throughout the proof.)

$$\sqrt{\ell}A \to \sqrt{\ell}A^{\perp} \to Y \to Y^{f} \to Y^{Q} \to Y^{Q}_{\mathrm{lin}} \to Y^{f}_{\mathrm{lin}}$$

Three Technical Ingredients

	Ingredient	Content
(i)	Normalizer fluctuation control	Concentration of Z_i/ℓ (scale $\ell^{-1/2+\delta}$)
(ii)	Taylor approximation	Polynomial approx. of $\exp(\beta x)$ (degree $n_\ell = \lceil c \log \ell / \log\log \ell \rceil$)
(iii)	Linearization of polynomials	Apply Benigni–Péché (2021) GE to polynomials

Step 1: Normalizer Concentration

Key Insight: $Z_i/\ell \to \mathbb{E}\left[e^{\beta\chi}\right] = e^{\beta^2/2}$ with sufficiently fast concentration

$$\ell^{1/2-\delta} \max_i \left| \ell^{-1} Z_i - e^{\beta^2/2} \right| \to 0 \quad \text{a.s.}$$

→ fluctuation scale $o\left(\ell^{-1/2+\delta}\right)$ for any $\delta > 0$

Consequently, the diagonal matrix $D = \mathrm{diag}(Z_i)/\left(\mathbb{E}\left[e^{\beta\chi}\right]\ell\right)$ approaches the identity:

$$\sqrt{\ell}A = D^{-1}Y \quad \text{where} \quad Y = \frac{\exp(\beta S)}{e^{\beta^2/2}\sqrt{\ell}}$$

Difficulty: S_{ij} are not independent, so standard concentration inequalities do not apply directly. → A combination of conditional independence + χ^2 concentration + Hoeffding + truncation is required.

Step 2: Polynomial Approximation

Centering:

$$f(x) = \exp\left(\beta x - \frac{\beta^2}{2}\right) - 1 \quad \Rightarrow \quad \mathbb{E}[f(\chi)] = 0$$

Taylor approximation:

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

Choosing the truncation level and approximation order

Truncation: $K = (\log \ell)^{1/2+\delta} \to P\big(\max|S_{ij}| > K\big)$ decays super-polynomially

Approx. degree: $n_\ell = \lceil c \log \ell / \log\log \ell \rceil$

$\to |f(x) - Q_n(x)| \leq c_2 (c_1 K/n)^n = o\big(1/\sqrt{\ell}\big)$ on $|x| \leq K$

Step 3: Gaussian Equivalence for Polynomials

Benigni–Péché (2021) framework:

$$\frac{Q_n(S)}{\sqrt{\ell}} \quad \text{and} \quad \frac{\sqrt{\theta_2(Q_n)}S}{\sqrt{\ell}} + \frac{\sqrt{\theta_1(Q_n) - \theta_2(Q_n)}W}{\sqrt{\ell}}$$

share the same asymptotic singular value distribution, where

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

Intuition — decompose the nonlinear function $g(S)$ into two components:

- $\sqrt{\theta_2} \cdot S$: signal correlated with input (contribution from g')
- $\sqrt{\theta_1 - \theta_2} \cdot W$: noise component independent of input

→ Asymptotically, a nonlinear transform is equivalent to a linear mixture of "signal + noise".

The Spectrum Is NOT Marchenko–Pastur

Y^f_{lin} contains the score matrix S.

$S = \left(1/\sqrt{\ell}\right) W^Q \left(W^K\right)^\top$ (reduced to $X = I_\ell$ under $\ell = d = d_{qk}$)

→ Product of two independent Ginibre matrices.
→ Its singular value distribution follows the Fuss–Catalan family, NOT the MP law.

Right Edge of Bulk

$$\max \operatorname{supp}(\nu_\infty) > 4\left(e^{\beta^2} - 1\right)$$

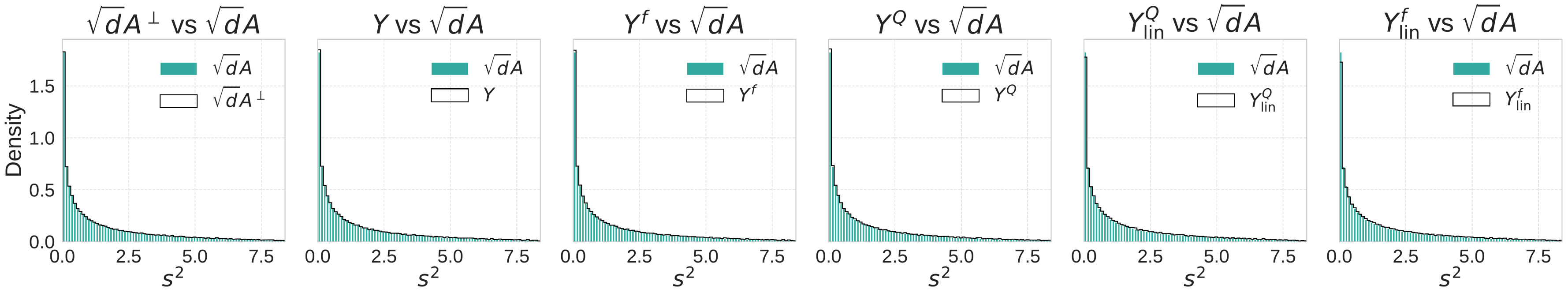
$= 4\theta_1$ = right edge of MP

Bulk widens because entries of S are dependent.

Practical implication: bulk right edge widens → gap between $s_1(A)$ and the bulk shrinks → the "spectral gap → vanishing" claim is overestimated under the i.i.d. model.

Numerical Verification

$\ell = d = d_{qk} = 1000$, $\beta = 1$, 10 samples



After removing the top-3 singular values, the bulk is close to the theoretical prediction. The right edge exceeds $4(e^1 - 1) \approx 6.87$ → confirms departure from the MP law.

Signal–Noise Decomposition

$$Y_{\mathrm{lin}}^{f} = \frac{\sqrt{\theta_2} S}{\sqrt{\ell}} + \frac{\sqrt{\theta_1 - \theta_2} W}{\sqrt{\ell}}$$

Signal: $\sqrt{\theta_2} = \beta$

Reflects the structure of the score matrix S. Preserves the Query–Key interaction pattern.

Noise: $\sqrt{\theta_1 - \theta_2} = \sqrt{e^{\beta^2} - 1 - \beta^2}$

Random noise from the independent Ginibre matrix W, generated by the nonlinearity of $\exp$.

Crossover Point: $\beta \approx \mathbf{1.121}$ — $\beta < 1.121$: Signal dominates → Ginibre product type / $\beta > 1.121$: Noise dominates → Ginibre type (MP-like)

The β Threshold: When Does Linearization Break?

$\beta = O(1)$

Regime of this work

Gaussian Equivalence holds Rank-1 + diffuse bulk Fully described by linear model

$\beta = \Theta\left(\sqrt{\log \ell}\right)$

Critical scale

Threshold where linearization breaks Bulk right edge blows up Extra singular directions reach $O(1)$

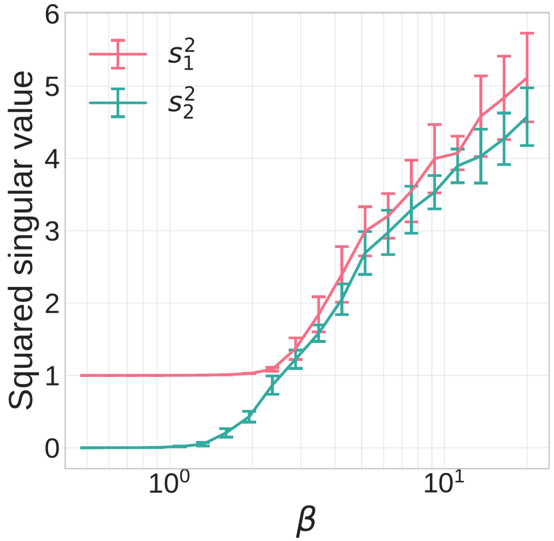
$\beta \gg \sqrt{\log \ell}$

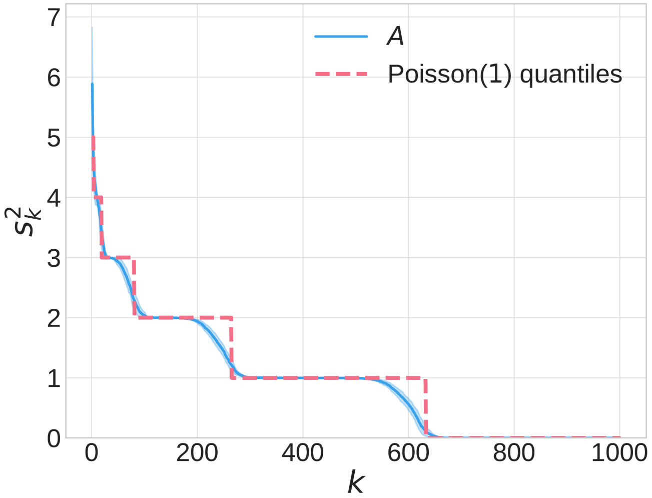
argmax regime

Softmax $\to$ argmax $\nu_A \to \mathrm{Poisson}(1)$ (ℓ balls into ℓ boxes independently)

First quantitative RMT-side support, in the form $\Theta\left(\sqrt{\log \ell}\right)$, for the conjecture $\beta^* = O\left(\sqrt{\log \ell}\right)$ of Giorlandino & Goldt (2025). (Recall $\ell = d$.)

Numerical Evidence: Large-β Transition





Left: s_2^2 starts increasing for $\beta \geq 1$, s_1^2 grows for $\beta \geq 2$. For $\beta \geq 5$ the rank-1 structure collapses. **Right**: $\beta = 50$, $\ell = d = 1000$. Matches Poisson(1) quantiles → confirms argmax regime.

Summary & Open Problems

Contributions

- First rigorous Gaussian Equivalence for self-attention
- Proved that the asymptotic spectrum differs from the MP law
- Critical scale $\beta = \Theta\left(\sqrt{\log \ell}\right)$ partially supported via RMT and experiments
- Identified clear rank-1 + diffuse bulk structure

Open Problems

Problem	Description
Large-β regime	Rigorous theory for $\beta \gg 1$ and the critical scale
General X	Extension to general input matrices (4-factor random matrix model)
Multi-head	Extension to multi-head attention
Outlier theory	Refined theory of outliers induced by nonlinearity

Poster 169. (Another related poster of ours will also be presented tomorrow at **Poster 49**.)