

NCPRT 2026 · Fukuoka · October 7, 2026

Gaussian Equivalence for Self-Attention

Asymptotic spectral analysis of the attention matrix

Tomohiro Hayase^{1,2} Benoît Collins³ Ryo Karakida^{1,2}

¹AIST ²RIKEN AIP ³Kyoto University

PART 1 / 5

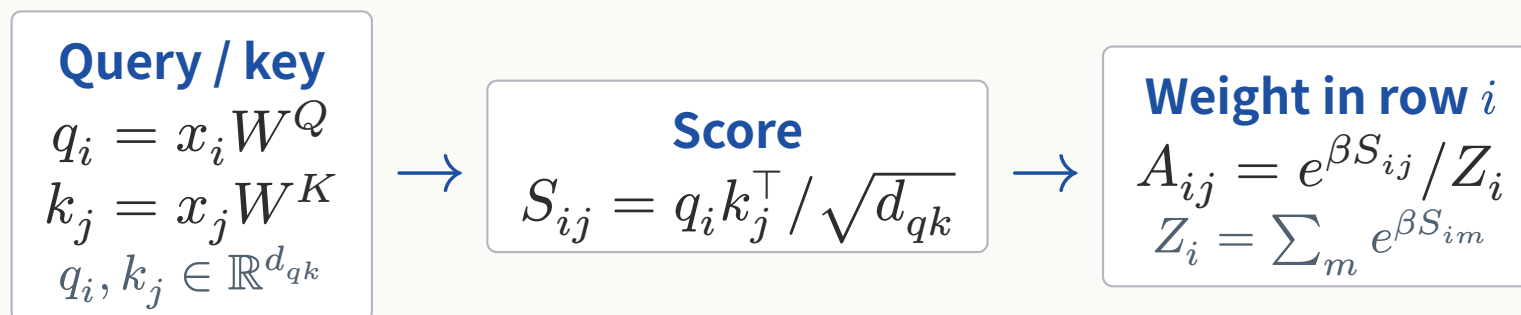
The attention matrix and the question

A random matrix from deep learning that is neither entrywise nor a functional calculus.

- **1. The attention matrix and the question**
- 2. The main theorem
- 3. The proof route
- 4. Why not Marchenko–Pastur?
- 5. Summary and open problems

Attention scores every pair of positions (i, j)

The **input**: ℓ vectors $x_1, \dots, x_\ell \in \mathbb{R}^d$, the rows of a matrix X .
Position i asks a **query**, position j offers a **key**; their inner product scores how much i attends to j .



$W^Q, W^K \in \mathbb{R}^{d \times d_{qk}}$ are the **weights** of the neural network; $\beta > 0$ is a fixed scalar.

Example, one row with $\ell = 4$ and $\beta = 1$:

row i of S	$\xrightarrow{\text{exp}}$	$e^{\beta S_{ij}}, \text{sum } Z_i = 5.48$	$\xrightarrow{\div Z_i}$	row i of A : sums to 1												
<table border="1"><tr><td>0.3</td><td>-1.2</td><td>1.0</td><td>0.1</td></tr></table>	0.3	-1.2	1.0	0.1		<table border="1"><tr><td>1.35</td><td>0.30</td><td>2.72</td><td>1.11</td></tr></table>	1.35	0.30	2.72	1.11		<table border="1"><tr><td>0.25</td><td>0.05</td><td>0.50</td><td>0.20</td></tr></table>	0.25	0.05	0.50	0.20
0.3	-1.2	1.0	0.1													
1.35	0.30	2.72	1.11													
0.25	0.05	0.50	0.20													

Softmax couples each row through its normalizer Z_i

Stacking all positions, the **score matrix** is

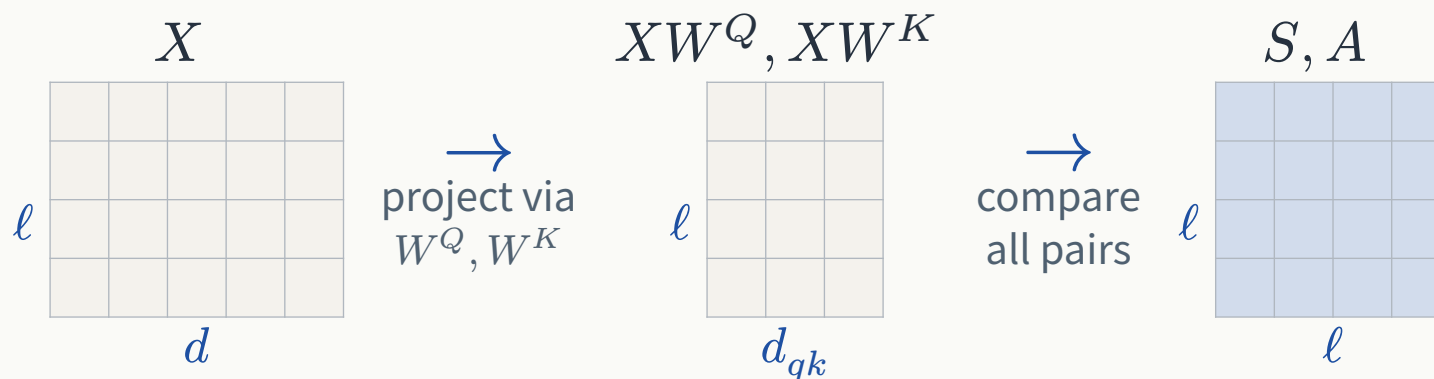
$$S = \frac{1}{\sqrt{d_{qk}}} XW^Q (W^K)^\top X^\top \in \mathbb{R}^{\ell \times \ell}.$$

The **attention matrix** is its **row-wise softmax**, with the **inverse temperature** $\beta > 0$:

$$A = \text{softmax}(\beta S), \quad A_{ij} = \frac{e^{\beta S_{ij}}}{\sum_{m=1}^{\ell} e^{\beta S_{im}}} = \frac{e^{\beta S_{ij}}}{Z_i}.$$

The exponential acts entrywise, but Z_i **couples all of row i** : A_{ij} depends on the whole i -th row of S . This row coupling makes A (**the object of this talk**) hard.

S and A are $\ell \times \ell$: one entry per pair of positions



ℓ number of vectors (the **context length**)

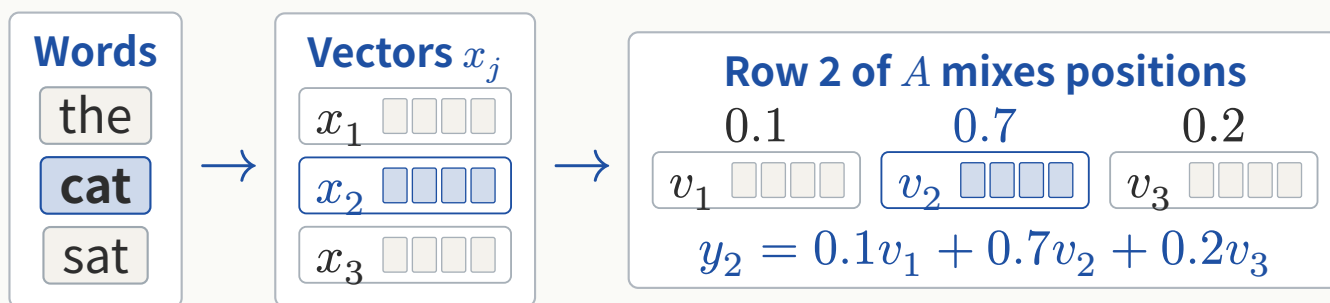
d coordinates per vector (the **embedding dimension**)

d_{qk} coordinates kept for comparison (**query/key dimension**)

Simplification. From here on $\ell = d = d_{qk}$. The general proportional regime is in the paper.

Self-attention averages the sequence with the weights in A

Self-attention returns a **new** sequence $Y = AV$: each vector becomes a weighted average, $y_i = \sum_j A_{ij} v_j$ (with $v_j = x_j W^V$).



Self-attention is **the characteristic layer of the Transformer**, the architecture behind large language models: an input-dependent operator $V \mapsto A(X)V$ that mixes the ℓ positions.

Each deep-learning module poses a spectral problem

Linear layers

Products of random matrices; singular-value laws.

Activations

Entrywise nonlinear maps of random features.

Jacobians

Signal propagation; spectra of layerwise derivatives.

Self-attention · this talk

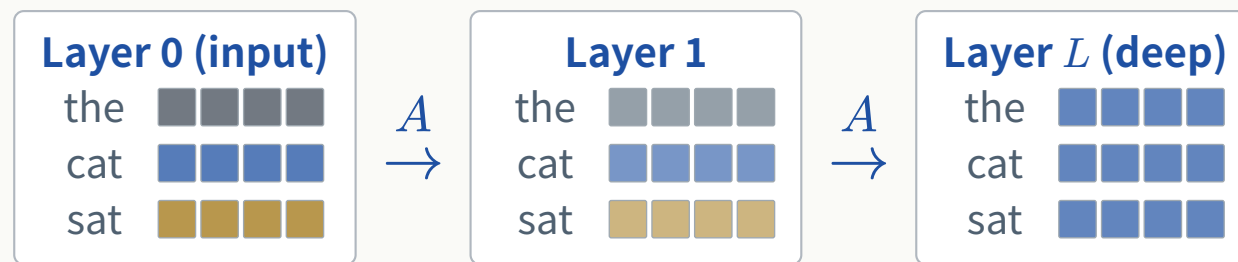
Row-wise softmax of random scores.

Tools: operator theory · spectral theory · free probability · random matrices

Deep learning is a **new source** of high-dimensional spectral problems, one per module. **Self-attention is one of them.**

With depth, attention can collapse the tokens to rank one

A Transformer **stacks** many attention layers. Across depth the token vectors can **collapse together**:



Rank collapse: the whole sequence is driven onto a single direction. This stalls signal propagation and training in deep Transformers, so the behavior of A **at initialization** is worth analyzing.

At initialization, A is an exactly specified random matrix

The weights are drawn at random; with the input X fixed, all randomness enters A through W^Q, W^K .

Fixed input

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

held fixed

Random weights

$$W^Q, W^K \text{ with entries } \sim \mathcal{N}(0, 1)$$

independent

Random score

$$S = XW^Q(W^K)^\top X^\top / \sqrt{d_{qk}}$$

Random A

$$A = \text{softmax}(\beta S), A\mathbf{1} = \mathbf{1}$$

Not a claim about a trained network: at initialization A **is** this random matrix. The only approximation in the talk is the Gaussian equivalence itself.

The common $\mathcal{N}(0, 1/d)$ initialization is equivalent after rescaling.

Row-stochastic and non-normal: we study $AA^\top$

Two structural facts hold for **every** choice of W^Q, W^K :

Rows sum to one. $A\mathbf{1} = \mathbf{1}$, so the all-ones **Perron direction** $u = \mathbf{1}/\sqrt{\ell}$ is a fixed, distinguished direction of A .

Not self-adjoint. A is generally not symmetric, so its own eigenvalues carry little information. We pass to the self-adjoint, positive **Gram** $AA^\top$ and study its eigenvalues, the squared singular values of A .

We track the **eigenvalues of $AA^\top$** and single out the direction u .

We measure the eigenvalue distribution of the Gram $BB^\top$

For $B \in \mathbb{R}^{d \times d}$, $BB^\top$ is self-adjoint and positive semidefinite, with eigenvalues $s_1(B)^2 \geq \dots \geq s_d(B)^2 \geq 0$. Its **empirical eigenvalue distribution** is

$$\nu_B := \mu_{BB^\top} = \frac{1}{d} \sum_{i=1}^d \delta_{s_i(B)^2} \quad \text{on } [0, \infty).$$

Moments are normalized traces:

$$\int x^q d\nu_B = \text{tr}[(BB^\top)^q],$$

with $\text{tr} = \frac{1}{d} \text{Tr}$.

$(M_d(\mathbb{R}), \text{tr})$ is a tracial $*$ -probability space; $\|B\|_\infty = s_1(B)$ is the operator norm.

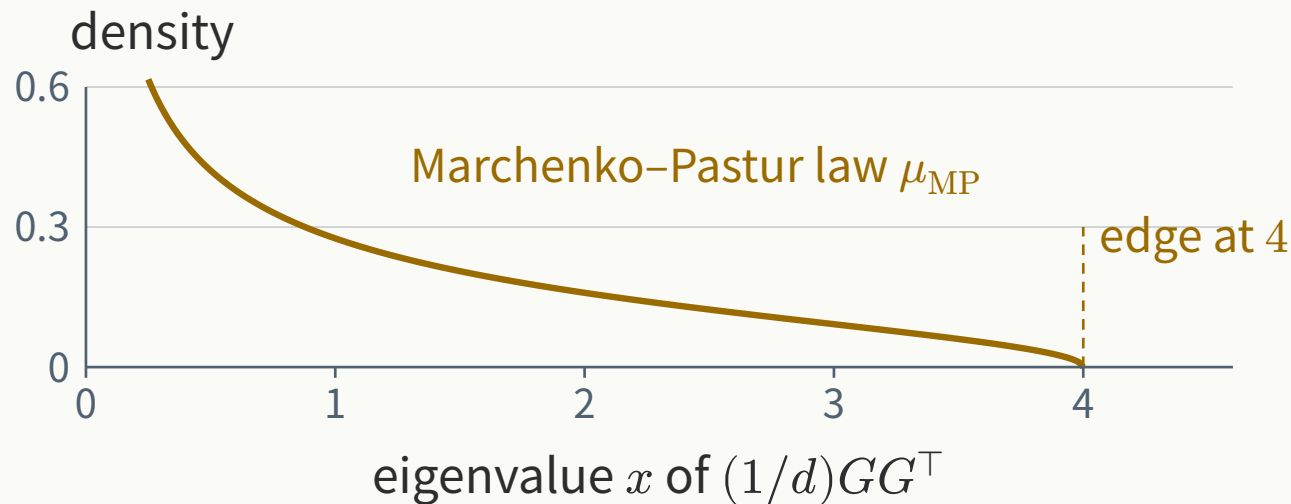
Convergence of all moments, with a uniform bound on the top eigenvalue, gives weak convergence. This is all the language we need.

Benchmark: one i.i.d. Gaussian matrix gives Marchenko–Pastur

Marchenko–Pastur law

For an **i.i.d. Gaussian matrix** $G \in \mathbb{R}^{d \times d}$ (independent $\mathcal{N}(0, 1)$ entries), $\nu_{G/\sqrt{d}} = \mu_{(1/d)GG^\top}$ converges almost surely to

$$d\mu_{\text{MP}}(x) = \frac{1}{2\pi} \sqrt{\frac{4-x}{x}} dx \quad \text{on } [0, 4].$$



The benchmark: the squared singular values of a **single** Gaussian matrix.

But the score is a product of two Gaussian matrices

Earlier work predicted a **Marchenko–Pastur** bulk for attention, in effect treating the score S as one i.i.d. Gaussian matrix.

Two features are missed:

- $S = XW^Q(W^K)^\top X^\top / \sqrt{d_{qk}}$ is a **product** of two Gaussian matrices, so its entries are **not independent**;
- the softmax is **row-wise**, not entrywise.

Does the Marchenko–Pastur prediction survive?

PART 2 / 5

The main theorem

A has the same limiting spectrum as a multiple of S plus an independent Gaussian matrix.

- 1. The attention matrix and the question
- **2. The main theorem**
- 3. The proof route
- 4. Why not Marchenko–Pastur?
- 5. Summary and open problems

As $d \rightarrow \infty$ the spectrum is deterministic; we may take $X = I_d$

Take $d \rightarrow \infty$ with $\ell = d = d_{qk}$ and $\beta > 0$ fixed. The empirical eigenvalue distribution of $AA^\top$ becomes **deterministic**.

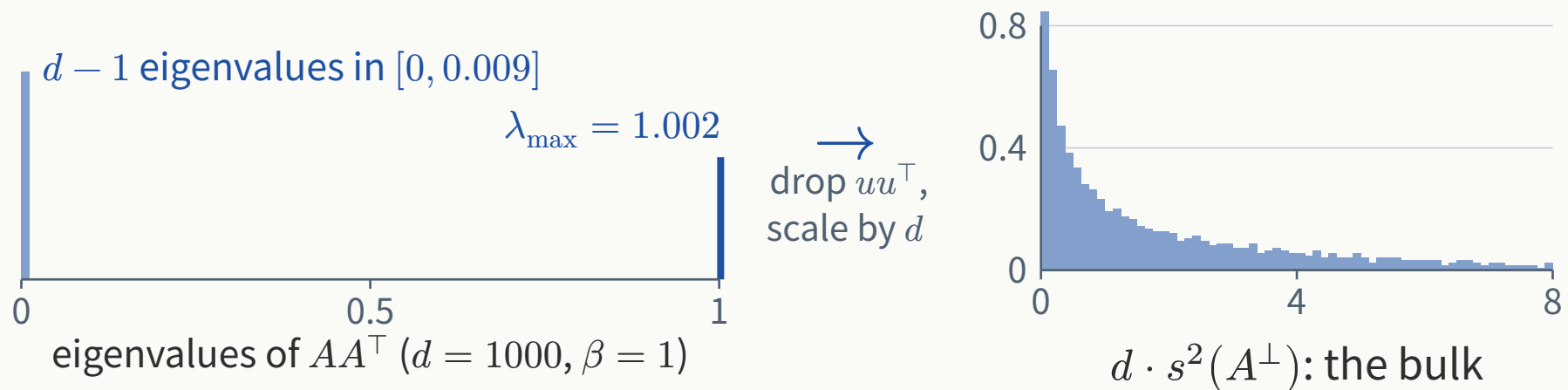
The input is orthonormal. By **left-orthogonal invariance** of the Gaussian weights we may take $X = I_d$, so

$$S = W^Q (W^K)^\top / \sqrt{d},$$

a **product of two independent i.i.d. Gaussian matrices**.

Everything below is about the random matrix $A = \text{softmax}\left(\beta W^Q (W^K)^\top / \sqrt{d}\right)$.

One eigenvalue near 1, the rest at scale $1/\sqrt{d}$



Why $1/\sqrt{d}$. $A\mathbf{1} = \mathbf{1}$ gives the Perron direction $u = \mathbf{1}/\sqrt{d}$ and $\lambda_{\max}(AA^\top) \rightarrow 1$. Every other entry is $\approx (1/d) e^{\beta S_{ij} - \beta^2/2}$ (Step 1): size $1/d$ times an $O(1)$ fluctuation, so the remainder has singular values of order $1/\sqrt{d}$.

$A \approx uu^\top + d^{-1/2} \cdot [\text{bulk}]$. So drop the Perron part and rescale: the theorem is about $\sqrt{d}A^\perp$, with $A^\perp := A - uu^\top$.

One draw (tools/gen_fa45_data.py). Interlacing: removing the rank-one $uu^\top$ does not change the bulk.

Main theorem: compare A with a multiple of S plus Gaussian

Both matrices are built from the **same** score S :

The original A (row-normalized softmax)

$$A_{ij} = \frac{\exp(\beta S_{ij})}{Z_i}$$

Y_{lin} , with W an independent i.i.d. Gaussian $d \times d$ matrix

$$Y_{\text{lin}} = \sqrt{\theta_2} \frac{S}{\sqrt{d}} + \sqrt{\theta_1 - \theta_2} \frac{W}{\sqrt{d}}$$

Theorem (Hayase–Collins–Karakida)

As $d \rightarrow \infty$ with β fixed, $\nu_{\sqrt{d}A^\perp}$ and $\nu_{Y_{\text{lin}}}$ converge to **one** limiting law ν_∞ , which depends only on β . Here

$$\theta_1 = e^{\beta^2} - 1, \quad \theta_2 = \beta^2.$$

“Equivalence” means equal limiting laws, not equal matrices.

Top eigenvalue $\rightarrow 1$; the bulk converges in moments

Three statements, in decreasing strength:

- 1 Top eigenvalue.** $\lambda_{\max}(AA^\top) = s_1(A)^2 \rightarrow 1$ a.s.; the upper bound uses concentration of the row and column sums.
- 2 Bulk, in moments.** $\nu_{\sqrt{d}A^\perp} \rightarrow \nu_\infty$ a.s.
- 3 Whole matrix, weakly.** $\nu_{\sqrt{d}A} \rightarrow \nu_\infty$; the top eigenvalue is an atom of mass $1/d$ escaping to ∞ , so only weak convergence survives.

ν_∞ is identified through Y_{lin} ; its Cauchy–Stieltjes transform satisfies a fixed-point equation.

PART 3 / 5

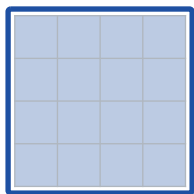
The proof route

Concentrate the normalizer, replace $\exp$ by a polynomial, then apply a known Gaussian equivalence.

- 1. The attention matrix and the question
- 2. The main theorem
- **3. The proof route**
- 4. Why not Marchenko–Pastur?
- 5. Summary and open problems

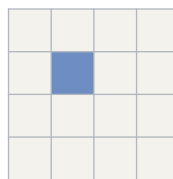
Softmax is row-wise: neither functional calculus nor entrywise

Functional calculus



$$B \mapsto f(B)$$
$$\sigma(f(B)) = f(\sigma(B))$$

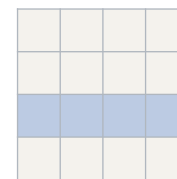
Entrywise



$$M_{ij} \mapsto f(M_{ij})$$

Gaussian equivalence, known

Row-wise (softmax)

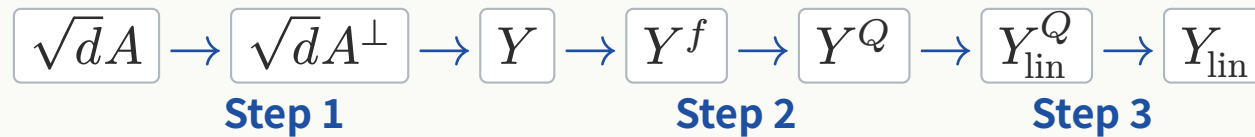


$$S_{ij} \mapsto e^{\beta S_{ij}} / Z_i$$

the new case

The novelty is controlling the row coupling Z_i . **Step 1** concentrates Z_i and reduces the row-wise problem to an **entrywise** one, where the standard tool applies.

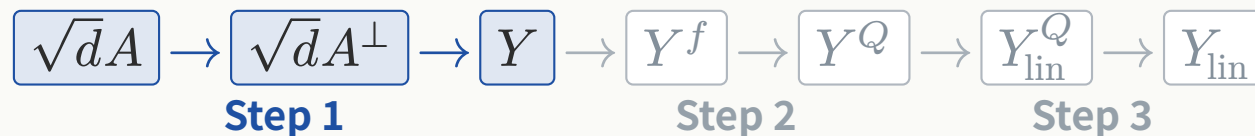
Proof: three steps, each preserving the limiting law



Step	What it does	Tool
1	remove Z_i : row-wise $\rightarrow$ entrywise	concentration of Z_i
2	replace $\exp$ by a polynomial	Taylor truncation
3	polynomial $\rightarrow$ multiple of S plus independent Gaussian	Benigni–Péché (2021)

Each arrow changes the matrix but preserves the limiting eigenvalue distribution of the Gram. All of them are compared numerically at the end of this part.

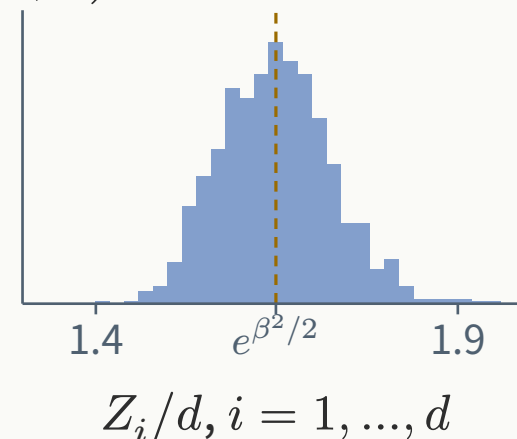
Step 1: the row normalizer concentrates



Uniformly in i , $Z_i/d \rightarrow \mathbb{E}[e^{\beta\chi}] = e^{\beta^2/2}$ with $\chi \sim \mathcal{N}(0, 1)$:

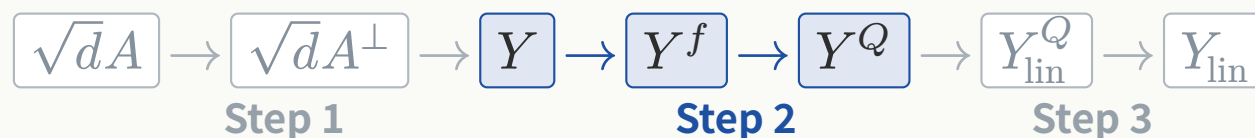
$$d^{1/2-\delta} \max_i |d^{-1} Z_i - e^{\beta^2/2}| \rightarrow 0 \quad \text{a.s.}$$

Hence $\sqrt{d}A = D^{-1}Y$ with $D = \text{diag}(Z_i)/(e^{\beta^2/2}d) \rightarrow I$ and $Y = \exp(\beta S)/(e^{\beta^2/2}\sqrt{d})$ entrywise.



Difficulty. The S_{ij} are not independent. Condition on one Gaussian factor, then use χ^2 - and Hoeffding-type bounds; the fluctuation scale is that of the law of the iterated logarithm.

Step 2: a polynomial of degree $\sim \log d / \log \log d$ replaces $\exp$



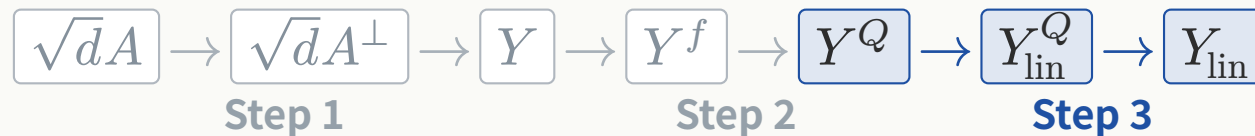
Center the nonlinearity, $f(x) = \exp(\beta x - \beta^2/2) - 1$, so that $\mathbb{E}[f(\chi)] = 0$. Approximate f by a polynomial Q_n of degree

$$n_d = \lceil c \log d / \log \log d \rceil.$$

Truncate the scores at $K = (\log d)^{1/2+\delta}$; exceeding K is super-polynomially rare. On $|x| \leq K$ the Taylor remainder is $\leq (e\beta K/n)^n = o(1/\sqrt{d})$, so $\|Y^f - Y^Q\|_\infty \rightarrow 0$.

This degree makes the truncation, the normalizer fluctuation, and the polynomial error compatible at once.

Step 3: for a polynomial, the Gaussian equivalence is known



Gaussian equivalence (Benigni–Péché 2021)

For a fixed polynomial Q , the entrywise $Q(S)/\sqrt{d}$ and

$$\left(\sqrt{\theta_2(Q)} S + \sqrt{\theta_1(Q) - \theta_2(Q)} W \right) / \sqrt{d}$$

have the same limiting eigenvalue distribution of the Gram, where

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

Letting $n_d \rightarrow \infty$: $\theta_1(f) = e^{\beta^2} - 1$ and $\theta_2(f) = \beta^2$, which is Y_{lin} of the main theorem.

Step 3 intuition: only θ_1 and θ_2 survive the graph counting

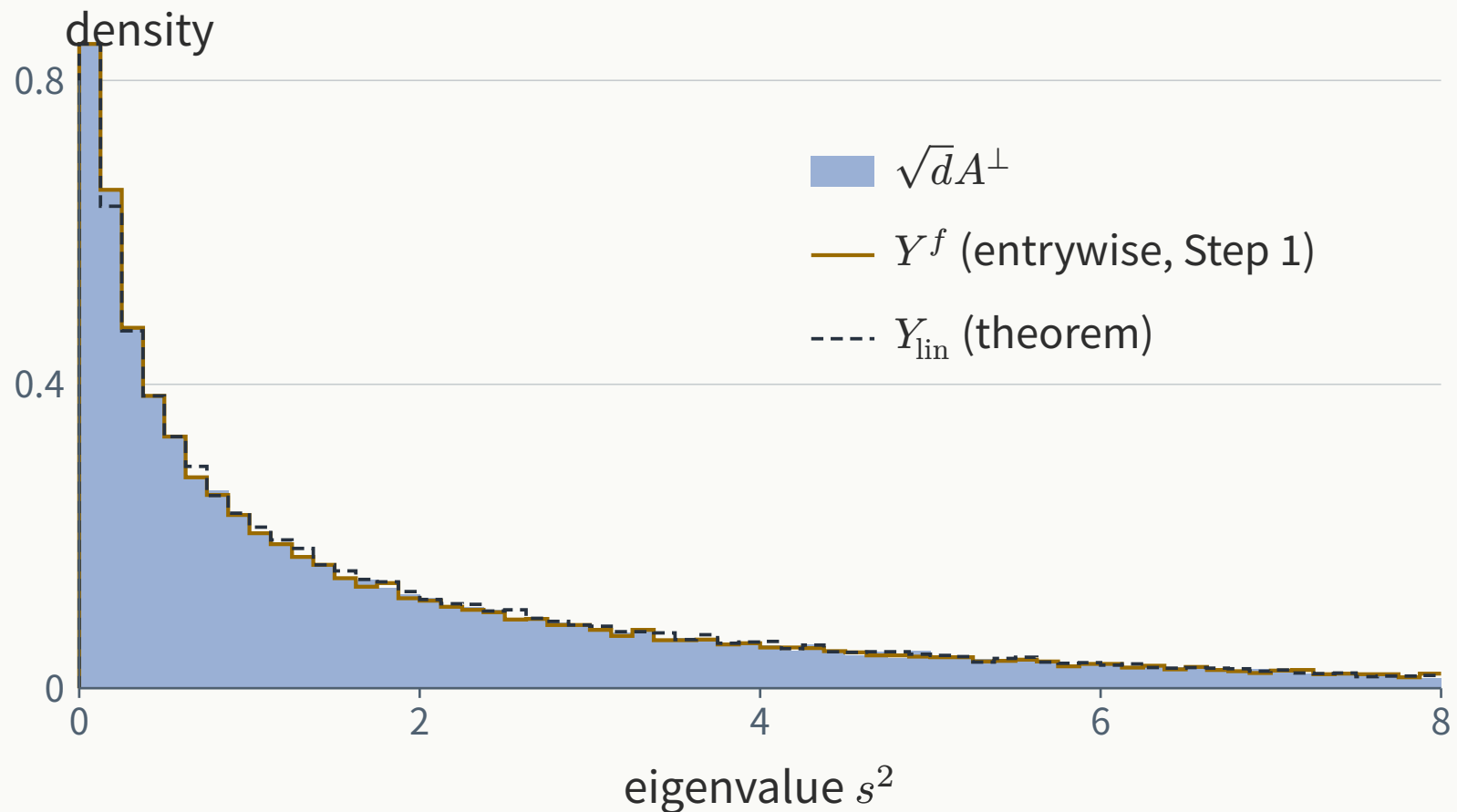
As usual in random matrix theory, the limiting moments of the Gram become a **counting problem on graphs**:

1. By Gaussianity (Wick's theorem), each moment is a sum over pairings, one graph per pairing.
2. As $d \rightarrow \infty$ only the leading graphs survive; the higher-order terms vanish.

The surviving graphs see Q only through $\theta_1(Q)$ and $\theta_2(Q)$: exactly the data kept by **a multiple of S plus an independent Gaussian matrix**.

Numerically, the chain's endpoints have the same spectrum

Bulk eigenvalues of the Gram, at the start, after Step 1, and at the end of the chain:



Simulation: $d = 1000$, $\beta = 1$, 10 draws. Intermediate matrices: paper, Fig. 1.

Y_{lin} = structured part + independent Gaussian part

$$Y_{\text{lin}} = \underbrace{\sqrt{\theta_2} S / \sqrt{d}}_{\text{structured}} + \underbrace{\sqrt{\theta_1 - \theta_2} W / \sqrt{d}}_{\text{independent Gaussian}}$$

Structured part, size $\sqrt{\theta_2} = \beta$:
keeps the score S ; the linear
response, from f' .

Independent Gaussian, size
 $\sqrt{\theta_1 - \theta_2} = \sqrt{e^{\beta^2} - 1 - \beta^2}$:
produced by the nonlinearity of
exp.

Spectrally, the nonlinear row-normalized A is a **multiple of S** plus an **independent Gaussian matrix**.

At $\beta \approx 1.121$ the Gaussian part takes over

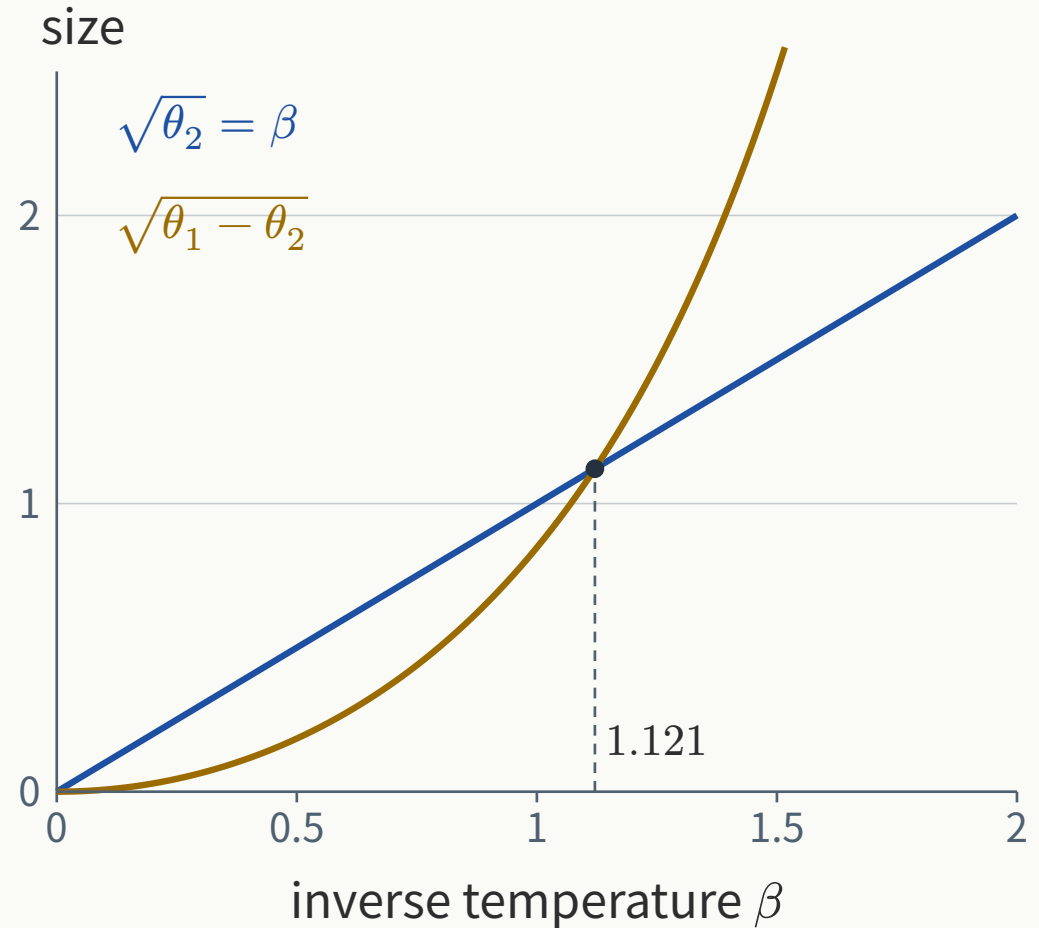
The two parts of Y_{lin} have sizes

$$\sqrt{\theta_2} = \beta,$$

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

$\beta < 1.121$ score part dominates

$\beta > 1.121$ Gaussian part dominates



A transition already visible **inside** the theorem: no large- β analysis is needed.

PART 4 / 5

Why not Marchenko–Pastur?

The structured part is a product of two Gaussian matrices, so the bulk is wider than Marchenko–Pastur.

- 1. The attention matrix and the question
- 2. The main theorem
- 3. The proof route
- **4. Why not Marchenko–Pastur?**
- 5. Summary and open problems

The limit is not Marchenko–Pastur

The structured part of Y_{lin} carries $S = W^Q (W^K)^\top / \sqrt{d}$: a **product of two** independent Gaussian matrices, not a single one.

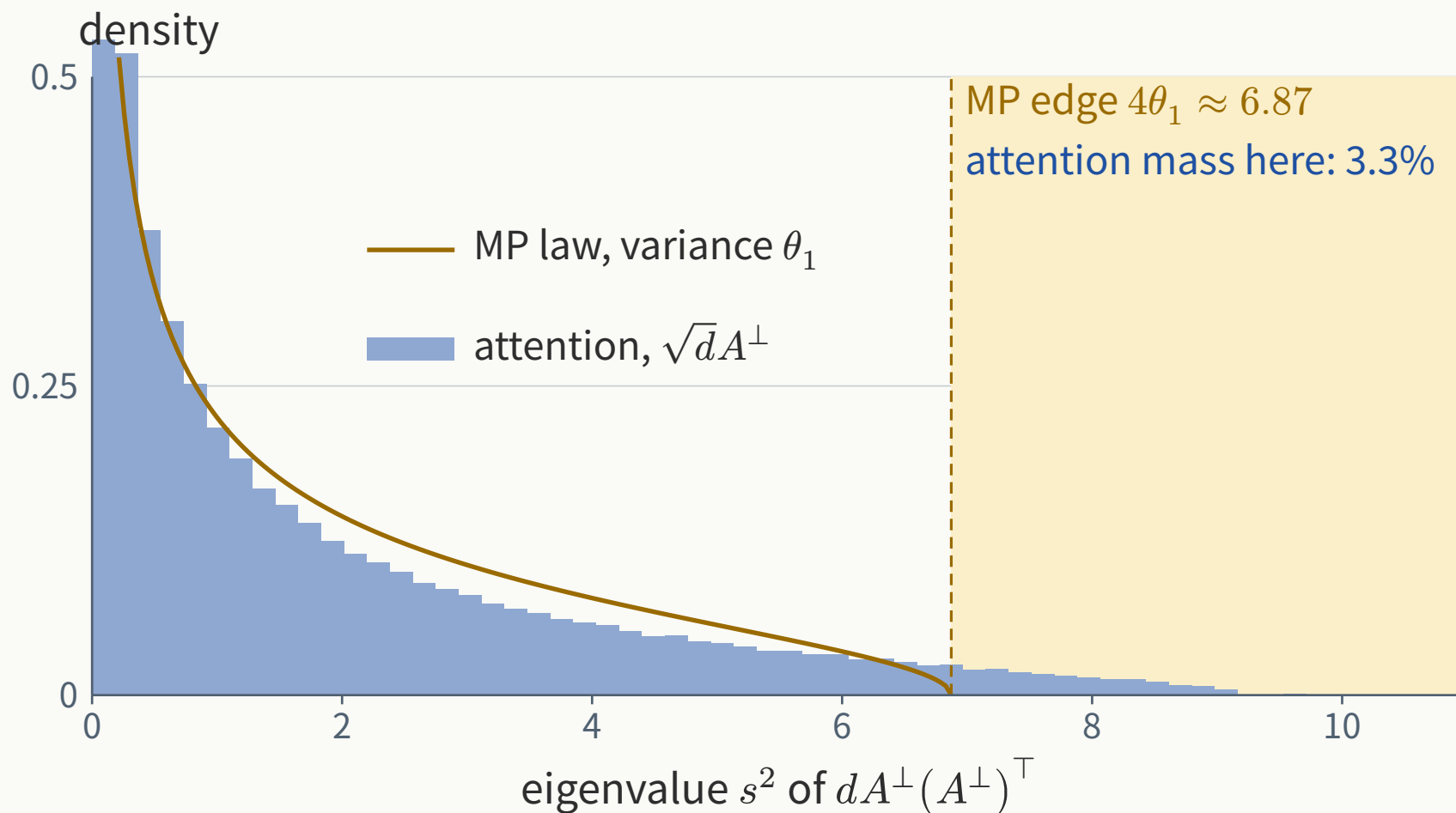
Corollary

The right edge of the bulk strictly exceeds the Marchenko–Pastur edge of the i.i.d. benchmark with the same entry variance θ_1 :

$$\max \text{supp}(\nu_\infty) > 4\theta_1 = 4(e^{\beta^2} - 1).$$

The strict inequality is proved with free probability (Appendix A1).

Numerically, the bulk spills past the MP edge



The i.i.d. benchmark has no mass beyond $4\theta_1$; the attention bulk does.

Simulation: $d = 1600$, 12 draws, $\beta = 1$ (tools/gen_fa45_data.py).

A wider bulk narrows the gap to the top eigenvalue



The top eigenvalue stays at $\lambda_{\max}(AA^\top) \rightarrow 1$. The bulk lives at scale $1/d$, with edge $\approx \max \text{supp}(\nu_\infty)/d$.

A **wider** ν_∞ raises that edge, so the bulk creeps toward the top mode, unlike the i.i.d. benchmark.

PART 5 / 5

Summary and open problems

What is proved, and what is open beyond $\beta = O(1)$, general inputs, and many heads.

- 1. The attention matrix and the question
- 2. The main theorem
- 3. The proof route
- 4. Why not Marchenko–Pastur?
- **5. Summary and open problems**

Summary

- Spectrally, A is a **multiple of the score S plus an independent Gaussian matrix** (Gaussian equivalence, $\beta = O(1)$).
- Structure: one Perron eigenvalue ($\lambda_{\max}(AA^T) \rightarrow 1$) plus a bulk with limiting law ν_{∞} .
- The bulk is **not** Marchenko–Pastur: the score is a product of two Gaussian matrices.

Method: concentrate the row normalizer, replace $\exp$ by a polynomial, apply Benigni–Péché. Self-attention joins the spectral map of deep-learning modules.

Open problem 1: the large- β threshold

Our theorem is for $\beta = O(1)$. What happens as β grows?

$$\beta = O(1)$$

This work: Gaussian equivalence; rank one + bulk.

$$\beta = \Theta(\sqrt{\log d})$$

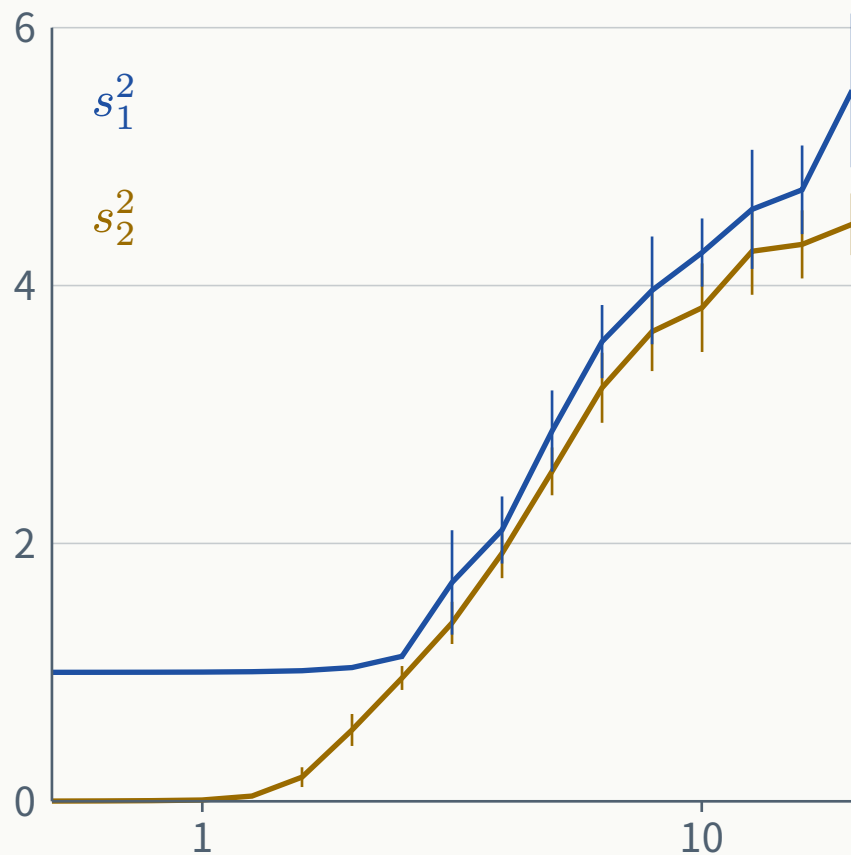
Conjectured threshold where linearization breaks.

$$\beta \rightarrow \infty$$

Softmax $\rightarrow$ argmax;
 $\nu_A \rightarrow \text{Poisson}(1)$ (d balls in d bins).

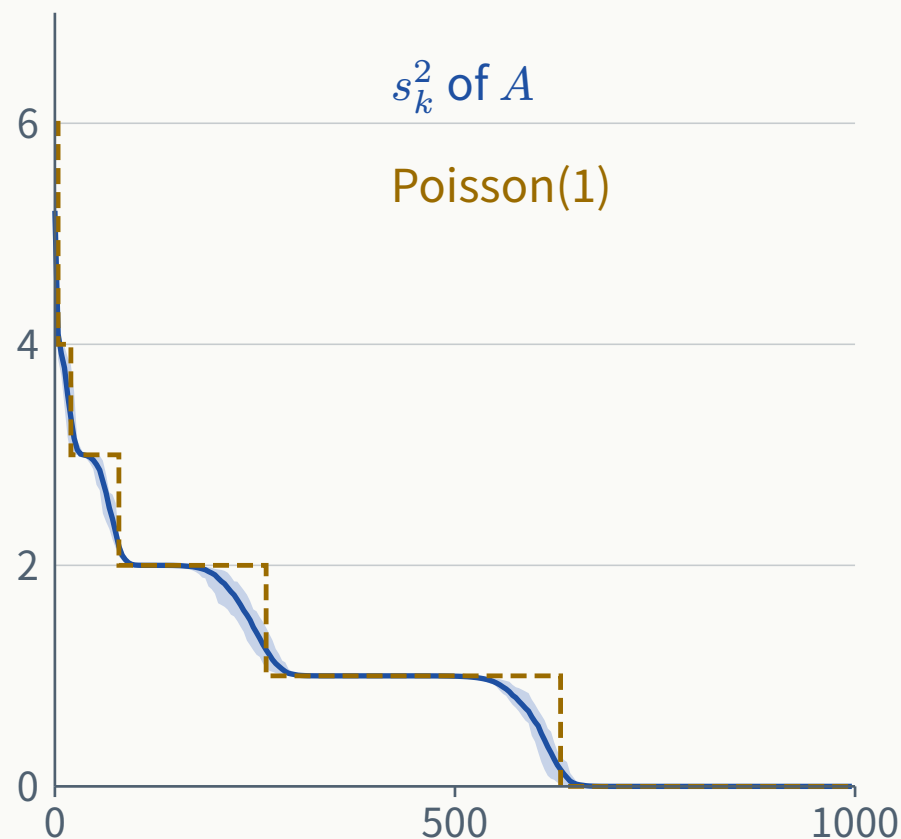
At $\beta = \Theta(\sqrt{\log d})$ the normalizer fluctuation, the score truncation, and the edge $\sim e^{\beta^2}$ collide. A rigorous theory **at and beyond** this threshold is open.

Numerically, the top eigenvalues rise and the law tends to Poisson



β (log scale)

Top eigenvalues s_1^2, s_2^2 of $AA^\top$ climb as β grows.



index k

$\beta = 50$ (near argmax): ν_A matches the Poisson(1) quantiles.

Simulation: $d = 1000$, 10 draws per point (tools/gen_fa45_data.py); cf. paper, § 6.

Open problem 2: general input matrix X

We assumed $XX^\top = I_d$. The goal is **arbitrary** X .

Proposed route. Treat the four-factor matrix

$$\frac{1}{\sqrt{d_{qk}}} XW^Q (XW^K)^\top$$

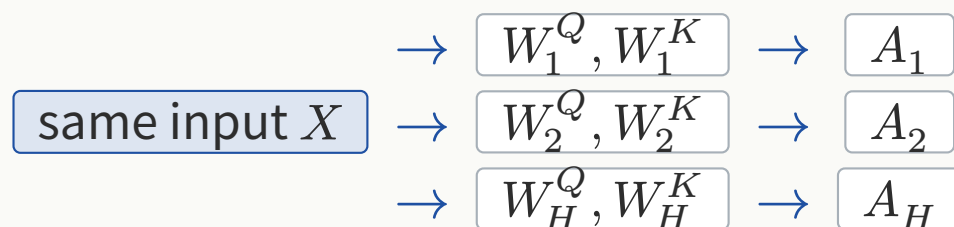
with the entrywise exponential, extending **Speicher–Wendel (2024)** (entrywise functions of orthogonally invariant matrices).

Why it is hard. The reduction to $X = I_d$ used orthonormality. A general X brings back anisotropy and ties the input geometry to the weights.

Open problem 3: multi-head attention

Real self-attention runs H **heads** in parallel: independent pairs (W_h^Q, W_h^K) give $A_1, \dots, A_H$ from the **same** X , combined as

$$X \mapsto (A_1 X W_1^V, \dots, A_H X W_H^V) W^O.$$



Our theorem is $H = 1$. The gap to real self-attention is the **joint** behavior of $A_1, \dots, A_H$.

Question. Independent weights but a **shared** input: are $A_1, \dots, A_H$ **asymptotically free**, so that their combined spectrum is a free convolution? Or does the shared X couple them?

Open problem 4: outliers

Removing the Perron part controls the **bulk**, not the extreme eigenvalues of $AA^\top$. Outliers beyond the Perron mode need a separate analysis.

Proposed route. Following **Benigni–Péché (2022)**, add a rank-one term of weight $\sqrt{\theta_3}$,

$$\theta_3 = \mathbb{E} \left[\frac{1}{2} f''(\chi) \right]^2.$$

Missing. Uniform high-moment bounds with $q = q(d)$, and uniform control of the normalizer at the law-of-the-iterated-logarithm scale.

Thank you

T. Hayase, B. Collins, R. Karakida.
Gaussian Equivalence for Self-Attention:
Asymptotic Spectral Analysis of Attention Matrix.
AISTATS 2026 (Oral) · [arXiv:2510.06685](https://arxiv.org/abs/2510.06685)

Appendix: the spectral edge via free probability

As $d \rightarrow \infty$, $Y_{\text{lin}} \rightarrow a := \alpha c_1 + \beta c_2 c_3$ in a tracial C^* -probability space, with c_1, c_2, c_3 *-free **circular elements** (limits of independent i.i.d. Gaussian matrices), $\alpha = \sqrt{\theta_1 - \theta_2}$ and $\sqrt{\theta_2} = \beta$.

$c_2 c_3$ is a product of two free circulars: its squared singular values follow the **Fuss–Catalan law**. Hermitize a and use the **K-transform** $K = G^{-1} = 1/w + R$:

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

The edge is $\|a\|_\infty = K(w_*)$ with $K'(w_*) = 0$. AM–GM gives $\|a\|_\infty > 2\sqrt{\alpha^2 + \beta^2} = 2\sqrt{\theta_1}$, so $\max \text{supp}(\nu_\infty) > 4\theta_1$.

Only the strict inequality is used in the talk. Haagerup–Larsen (2000); Benaych-Georges (2009); Benigni–Péché (2021).

References

- T. Hayase, B. Collins, R. Karakida. Gaussian equivalence for self-attention: asymptotic spectral analysis of attention matrix. **AISTATS 2026**. arXiv:2510.06685.
- L. Benigni, S. Péché. Eigenvalue distribution of some nonlinear models of random matrices. **Electron. J. Probab.** 26 (2021).
- L. Benigni, S. Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. arXiv:2201.04753 (2022).
- J. Pennington, P. Worah. Nonlinear random matrix theory for deep learning. **NeurIPS 2017**.
- U. Haagerup, F. Larsen. Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras. **J. Funct. Anal.** 176 (2000).
- F. Benaych-Georges. Rectangular random matrices, related convolution. **Probab. Theory Relat. Fields** 144 (2009).
- R. Speicher, J. Wendel. Entrywise application of non-linear functions on orthogonally invariant matrices. arXiv:2412.06943 (2024).
- B. Collins, T. Hayase. Asymptotic freeness of layerwise Jacobians caused by invariance of MLP. **Comm. Math. Phys.** 397 (2023).
- Y. Dong, J.-B. Cordonnier, A. Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. **ICML 2021**.
- L. Noci et al. Signal propagation in Transformers: theoretical perspectives and the role of rank collapse. **NeurIPS 2022**.
- T. Nait Saada et al. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. **ICML 2025**.