



Spectral Degeneration of Self-Attention under Isotropic Score Geometry

Mengda Li and Jianfeng Yao

The Chinese University of Hong Kong, Shenzhen

NeurIPS 2026

Score geometry and spectral concentration

Self-attention mixes token information through a row-stochastic matrix:

$$A = \mathrm{softmax}_{\mathrm{row}}(S + M).$$

Which score structures preserve dominant directions?

How does causal masking change the remaining spectrum?

- We isolate the effect of softmax in a canonical score model.
- We compare bidirectional attention with attention under a *causal* mask.
- We examine score geometry and attention spectra in GPT-2 and ViT.

Self-attention and score geometry

For a sequence of length ℓ, define the *input* matrix

$$X \in \mathbb{R}^{\ell \times d}.$$

The *query* matrix Q and *key* matrix K are

$$Q = XW_Q, \qquad K = XW_K, \qquad Q, K \in \mathbb{R}^{\ell \times d_{qk}}.$$

The *score* matrix has singular value decomposition

$$S = \frac{QK^\top}{\sqrt{d_{qk}}} = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(s_1, \ldots, s_\ell).$$

We separate score strength Σ from token-space orientation U, V.

The canonical score model

Retain the score singular values and align their directions with the token axes:

$$\Sigma = \mathrm{diag}(s_1, \ldots, s_\ell), \qquad s_1 \geq \cdots \geq s_\ell \geq 0.$$

Bidirectional attention

$$A_{\mathrm{b}} = \mathrm{softmax}_{\mathrm{row}}(\Sigma).$$

Every token can attend to every position.

Causal attention

$$A_{\mathrm{c}} = \mathrm{softmax}_{\mathrm{row}}(\Sigma + M),$$

$M_{ij} = -\infty$ for $j > i$, and zero otherwise.

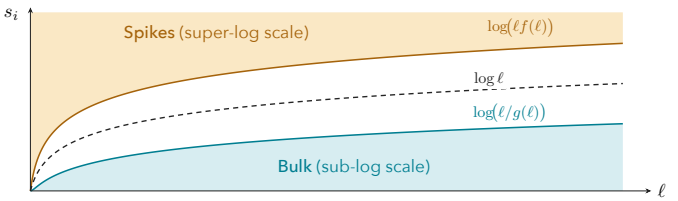
The diagonal surrogate is motivated by isotropy. The causal model places scores in descending token order.

Assumption 1. Appendix A discusses orientation dependence and the scope of the reduction.

The logarithmic softmax threshold

The diagonal weight competes with $\ell - 1$ off-diagonal unit weights:

$$(A_{\mathrm{b}})_{ii} = \frac{e^{s_i}}{e^{s_i} + \ell - 1}.$$



Fixed head dimension gives $k \le \mathrm{rank}(S) \le d_{qk}$, hence $k = o(\ell)$.

Assumption 2. The transition scale is $s_i \approx \log \ell$.

Spectral separation

$$s_i = \Omega(\log(\ell f(\ell)))\,, \quad i \le k,$$
$$s_i = O(\log(\ell/g(\ell)))\,, \quad i > k.$$

$f, g \to \infty$, $g = o(\ell)$.

No scores lie between the thresholds.

Bidirectional attention: singular values

Under canonical modeling and spectral separation, let

$$c_\ell = \frac{k}{\ell}, \qquad \varepsilon_\ell = \max\{f(\ell)^{-1}, g(\ell)^{-1}\}.$$

For $k = o(\ell)$, the singular values satisfy

$$\text{Spike:} \quad \max_{j \le k+1} |\sigma_j(A_{\mathrm{b}}) - 1| = O(\varepsilon_\ell + \sqrt{c_\ell}) \longrightarrow 0,$$

$$\text{Bulk:} \quad \max_{j \ge k+2} \sigma_j(A_{\mathrm{b}}) = O(\varepsilon_\ell) \longrightarrow 0.$$

Eigenvalues lie in $[0, 1]$: the leading $k + 1$ tend to 1, the rest to 0.

Softmax preserves the spike modes and one common average.

Theorems 1 and 2. The exact profile for a nonzero spike fraction is in the appendix.

Exact bidirectional eigenvalue structure

Let $D = \mathrm{diag}(d_1, \ldots, d_\ell)$, with $d_i = \frac{e^{s_i}-1}{e^{s_i+\ell-1}}$ and $d_1 \geq \cdots \geq d_\ell$. The eigenvalues are real and satisfy

$$\lambda_1(A_\mathrm{b}) = 1, \qquad d_i \leq \lambda_i(A_\mathrm{b}) \leq d_{i-1}, \quad i \geq 2.$$

For adjacent diagonal values,

$$\begin{cases} \lambda_i(A_\mathrm{b}) = d_i, & d_i = d_{i-1}, \\ d_i < \lambda_i(A_\mathrm{b}) < d_{i-1}, & d_i < d_{i-1}. \end{cases}$$

Under separation, for $k < \ell$,

$$\begin{aligned} \text{Spike:} \quad & \lambda_i(A_\mathrm{b}) = 1 + O\big(f(\ell)^{-1}\big) \longrightarrow 1, \quad 2 \leq i \leq k, \\ & \lambda_{k+1}(A_\mathrm{b}) = 1 - c_\ell + O(\varepsilon_\ell), \\ \text{Bulk:} \quad & 0 \leq \lambda_i(A_\mathrm{b}) \leq d_{i-1} = O\big(g(\ell)^{-1}\big) \longrightarrow 0, \quad k+2 \leq i \leq \ell. \end{aligned}$$

Theorem 2. Equal adjacent values produce equality rather than strict interlacing.

Causal eigenvalues and singular values

Causal attention is lower triangular. Its eigenvalues are explicit:

$$\lambda_i(A_c) = \frac{e^{s_i}}{e^{s_i} + i - 1}, \qquad \lambda_1(A_c) = 1.$$

The canonical model also satisfies the universal singular-value bounds

$$\frac{1}{3i} \le \sigma_i(A_c) \le \|A_c\|_{\mathrm{op}} \le \sqrt{5}, \qquad 1 \le i \le \ell.$$

Under separation, the k spike eigenvalues approach one, and $\sigma_k(A_c) \ge 1 - o(1)$ for $k \ge 1$.

With uniformly bounded bulk scores $s_i \le B$ for $i > k$:

$$\lambda_i(A_c) = \Theta_B(1/i), \quad i > k, \qquad \#\{i > k : \sigma_i(A_c) > t\} \le \frac{C_B \log(\ell + 1)}{t^2}.$$

The counting bound holds for each fixed $t > 0$, for sufficiently large ℓ.

Theorems 3 and 4. The counting result also uses spectral separation.

Degeneration of the limiting spectral distributions

Under spectral separation, suppose $k/\ell \to c \in [0, 1]$.
For **both bidirectional and causal attention**,

$$\mu_A^{\text{eig}} = \frac{1}{\ell} \sum_{j=1}^{\ell} \delta_{\lambda_j(A)}, \qquad \mu_A^{\text{sv}} = \frac{1}{\ell} \sum_{j=1}^{\ell} \delta_{\sigma_j(A)},$$

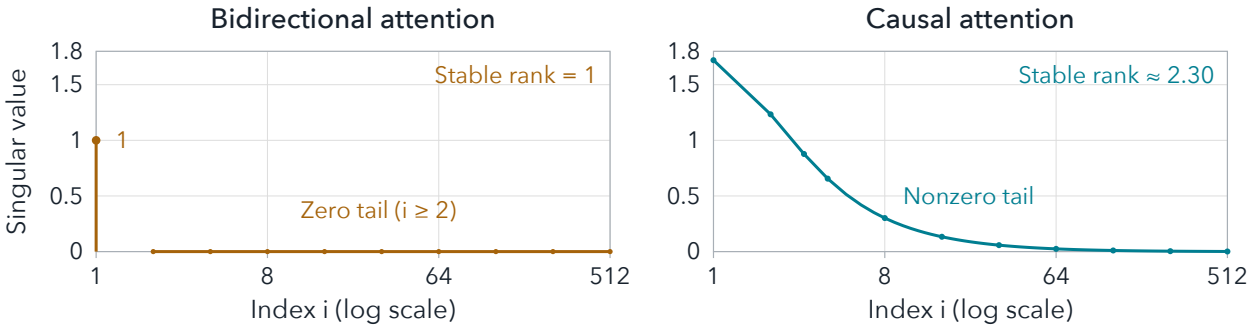
$$\mu_A^{\text{eig}},\ \mu_A^{\text{sv}} \implies c\delta_1 + (1-c)\delta_0.$$

For sublinear spikes $k = o(\ell)$, both limiting spectral distributions degenerate to δ_0.
Each mode has weight $1/\ell$, so $o(\ell)$ modes carry vanishing mass.

Theorems 1–4 and Corollary 1. This weak-limit result does not require bounded bulk scores.

Example: attention at zero scores

$$\Sigma = 0: \qquad A_{\mathrm{b}} = \frac{\mathbf{1}\mathbf{1}^{\top}}{\ell}, \qquad (A_{\mathrm{c}})_{ij} = \frac{\mathbf{1}_{\{j \le i\}}}{i}.$$



$$\|A_{\mathrm{b}}\|_{\mathrm{F}}^2 = 1, \qquad \|A_{\mathrm{c}}\|_{\mathrm{F}}^2 = \textstyle\sum_{i=1}^{\ell} 1/i \sim \log \ell.$$

Section 3.4. Canonical zero-score attention, numerical singular values at $\ell = 512$.

Spectral ranks distinguish the two masks

Define the stable rank and effective rank:

$$\mathrm{srank}(A) = \frac{\|A\|_{\mathrm{F}}^2}{\|A\|_{\mathrm{op}}^2}, \qquad \mathrm{erank}_p(A) = \exp\left(-\sum_j q_j^{(p)} \log q_j^{(p)}\right), \quad q_j^{(p)} = \frac{\sigma_j(A)^p}{\sum_r \sigma_r(A)^p}.$$

With **fixed** k, spectral separation and **uniformly bounded bulk**:

	Bidirectional	Causal
srank	$k + 1 + o(1)$	$\Theta_B(\log \ell)$
erank_2	$k + 1 + o(1)$	$\Theta_B(\log \ell)$

Causal masking slows spectral compression as context grows.

Section 3.4 and Appendix D. The $p = 1$ and $p = 2$ entropy ranks have different tail sensitivity.

Headwise diagnostics in pretrained networks

We measure each head of the intact pretrained network, then aggregate.

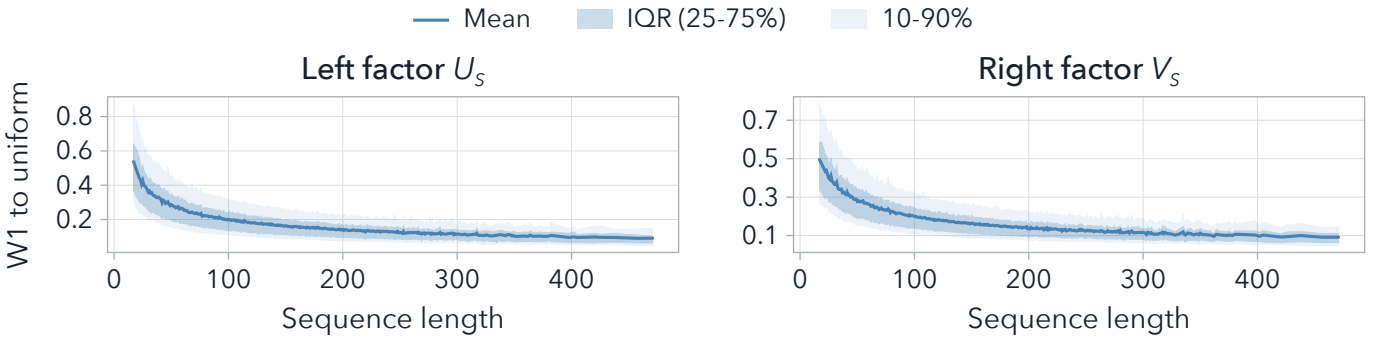
	GPT-2	ViT-B/16 (SWAG)
Data	WikiText-103	Imagenette
Attention pattern	Causal	Bidirectional
Layers / heads per layer	12 / 12	12 / 12
Score-factor diagnostics	1,646 sequences	3,925 images
Real/canonical comparison	12 documents	20 images
Comparison length	64–1,024 tokens	577 tokens

We examine score-factor angles, score rank, and real/canonical spectra.

Section 4 and Appendix F. Sample counts refer to distinct diagnostic groups.

Score-factor eigen-angle diagnostics

For actual scores $S_h = U_{S_h} \Sigma_h V_{S_h}^\top$, compare eigen-angles of U_{S_h} and V_{S_h} with a uniform reference.



GPT-2 band means: about 0.31 at 17–64 tokens and 0.10 at 385–471 tokens. ViT averages about 0.08 for each factor.

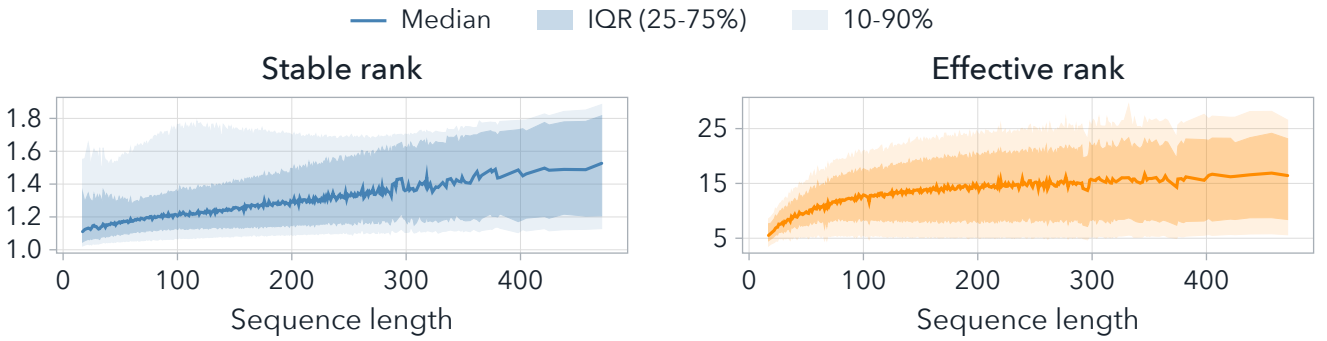
Near-uniform angles motivate an isotropic description.

This diagnostic alone does not establish a joint Haar law.

Section 4.4 and Appendix F.3.1. GPT-2 endpoint length bands contain 254 and 8 sequences.

A few effective score directions

GPT-2 score rank summaries remain modest over the tested lengths.



$$k \leq \operatorname{rank}(S_h) \leq d_{qk} = 64.$$

A few dominant score directions motivate spike/bulk modeling.

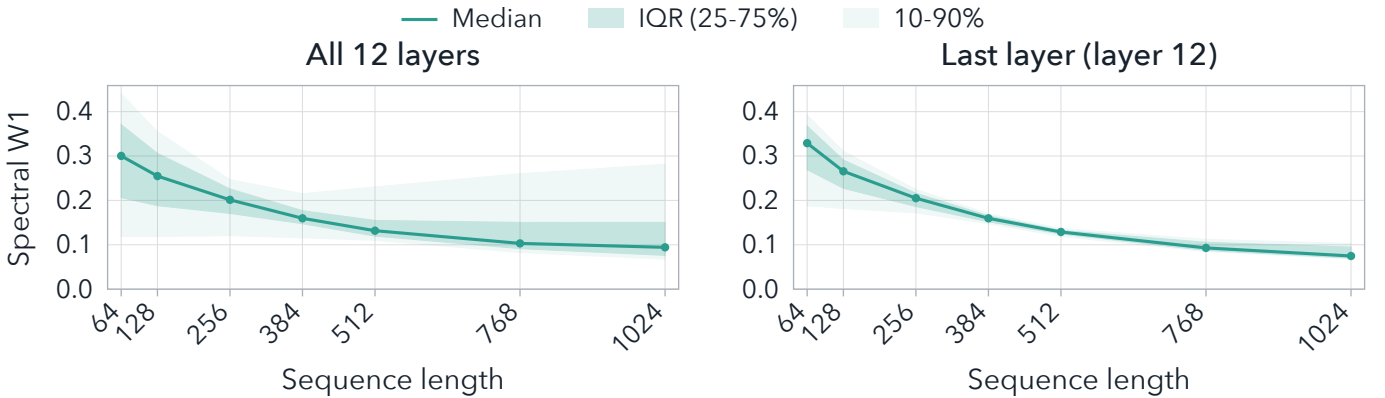
These finite-length summaries do not verify the asymptotic separation thresholds.

Section 4.2 and Appendix F. Curves: medians with interquartile and wider percentile bands.

Real versus canonical attention in GPT-2

Matched attention: $A_{\text{real}} = \text{softmax}_{\text{row}}(S + M)$, $A_{\Sigma} = \text{softmax}_{\text{row}}(\Sigma + M)$.

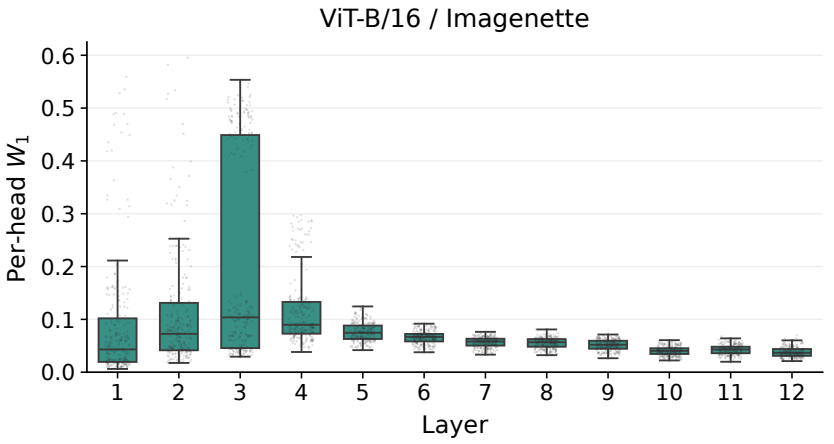
$$W_1 = \frac{1}{\ell} \sum_{j=1}^{\ell} \left| \sigma_j(A_{\text{real}}) - \sigma_j(A_{\Sigma}) \right|.$$



At 64 and 1,024 tokens, median W_1 is **0.3002 and 0.0942** across all layers, and **0.3290 and 0.0748** in the last layer.

Section 4.3. 12 documents at matched context lengths. Bands: IQR and 10–90% variation across pairs.

Real versus canonical attention in ViT



Matched bidirectional comparison

20 images, $\ell = 577$.
12 layers and 12 heads.
Layer median W_1:
0.0367–0.1037.

The empirical spectral laws are close across the tested layers.

Section 4.3. Boxes show median and IQR. Whiskers extend to 1.5 times the IQR.

Summary, and Question?

Softmax creates spectral compression.
In the canonical model, separated *spikes* survive while the *bulk* concentrates near zero.

The *causal* mask changes retained spectral mass.
With fixed *spikes* and bounded *bulk*, stable rank grows as $\log \ell$ instead of approaching $k + 1$.

Experiments connect the model to pretrained attention.
Score geometry and real/canonical spectral comparisons in GPT-2 and ViT support the spectral analysis.

Linktree

Mengda Li and Jianfeng Yao. Spectral Degeneration of Self-Attention under Isotropic Score Geometry. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026.

github.com/Mengda-Li/attention

li.mengda@outlook.com

The full bidirectional singular-value profile

For $1 \le k < \ell$, let $c_\ell = k/\ell$ and define

$$\alpha_\pm(x) = \sqrt{1 - \frac{x}{2} \pm \sqrt{x - \frac{3}{4}x^2}}.$$

Under spectral separation, the ordered singular values are uniformly $O(\varepsilon_\ell)$-close to

$$\left(\alpha_+(c_\ell),\ \underbrace{1, \dots, 1}_{k-1},\ \alpha_-(c_\ell),\ \underbrace{0, \dots, 0}_{\ell-k-1} \right).$$

At $k = 0$, the profile is $(1, 0, \dots, 0)$.

At $k = \ell$, the profile is $(1, \dots, 1)$.

The two exceptional values encode the spike/bulk coupling.

Theorem 1. As $c_\ell \to 0$, both $\alpha_+(c_\ell)$ and $\alpha_-(c_\ell)$ approach one.

A shared averaging direction

Bidirectional attention admits an exact decomposition:

$$A_{\mathrm{b}} = D + z\mathbf{1}^{\top}, \qquad D = \mathrm{diag}(d_i).$$

$$d_i = \frac{e^{s_i} - 1}{e^{s_i} + \ell - 1}, \quad z_i = \frac{1}{e^{s_i} + \ell - 1}.$$

Spike rows

$d_i \to 1$, $\ell z_i \to 0$.

Each row approaches a token selector.

Bulk rows

$d_i \to 0$, $\ell z_i \to 1$.

Bulk rows approach the same global average.

k spike directions plus one shared averaging direction

Section 3.2, exact decomposition underlying Theorems 1 and 2. A bulk is present when $k < \ell$.