

第 96 回産総研人工知能セミナー「数学と AI」

一つの重みは不明でも 百万個なら法則が見える

ランダム行列で見る深層学習

早瀬 友裕

産総研 AIRC／理研 AIP

同じ「手」でも，周囲との関係は変わる

けがした

手

を

すぐに

冷やす

次の

手

を

慎重に

考える

「手」は，文中のどの語と強く結びつくだろう？

「手」を Query にした重みは，文脈で変わりうる

両方とも Query = 「手」



「手」から文中の5トークンへの重みの並びを比べている

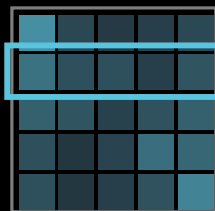
棒は直上のトークンへの重み。日本語 BERT (cl-tohoku/bert-base-japanese-v3) の実測（全層・全ヘッド平均）。各行の和は1

各トークンを Query にすると，Attention 行列 A になる

けがした 手 を すぐに 冷やす



この 5 個を「手」の行へ ↓



→ 2 行目：Query = 「手」

$$Q = XW^Q, \quad K = XW^K$$

$$S = d_{qk}^{-\frac{1}{2}} QK^\top$$

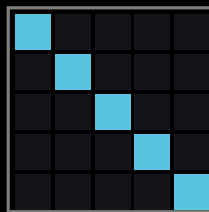
$$A = \text{softmax}(S)$$

A_{ij} トークン i から j への参照の重み

Vaswani et al., “Attention Is All You Need” (NeurIPS 2017)

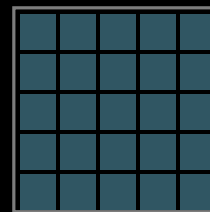
行と列は同じトークン列に対応し，各行の和は 1 になる

Attention 行列の作用を両極端で比べる



$$A = I_n \text{ 各位置を保つ}$$

どのトークン位置方向も同じ強さで残る

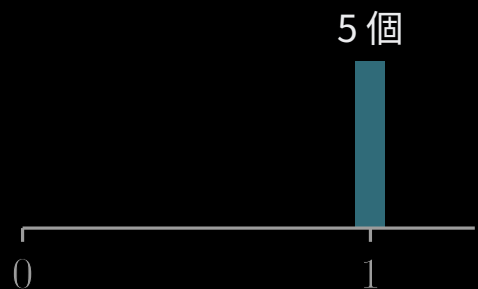


$$A = n^{-1} \mathbf{1}\mathbf{1}^\top \text{ 共通方向へ集約}$$

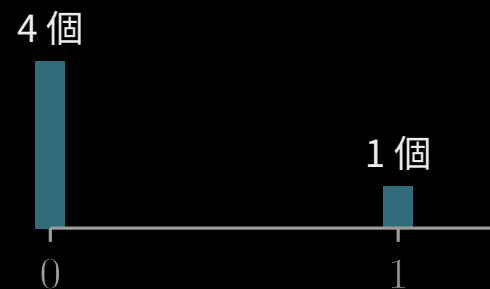
全トークンが同じ集約結果を受け取る

一様行列では，共通方向だけが残る，それに直交する位置方向は消える

トークン位置方向の作用を特異値で見る



$A = I$ 1 が 5 個



$A = n^{-1}\mathbf{1}\mathbf{1}^\top$ 1 が 1 個, 残りは 0

1 つの A が作用したとき, 特異ベクトルがトークン位置方向を定め, 特異値が
その方向を何倍にするか を測る

では, 複雑な A の特異値分布をどう計算するか

Attention を構成部品までほどく

$$A = \text{softmax}(S)$$

行方向 softmax

$$S = d_{qk}^{-\frac{1}{2}} XW^Q (W^K)^{\top} X^{\top}$$

入力と 2 つの射影を含む積

$$W^Q \quad W^K$$

Attention を構成する射影行列

まず 1 つの射影行列の特異値を計算できるようにする

正方基準モデルでは Query 行列が Gaussian 行列になる

一般形

$$Q = XW^Q$$

正方基準モデル

$$X = I_n, \quad d = d_{qk} = n$$

Gaussian 初期化の結果

$$Q = W^Q, \quad Q_{i,j} \sim \mathcal{N}(0, 1) \quad \text{i.i.d.}$$

$$C_Q := n^{-1} Q^\top Q \quad \text{Query の正規化 Gram 行列}$$

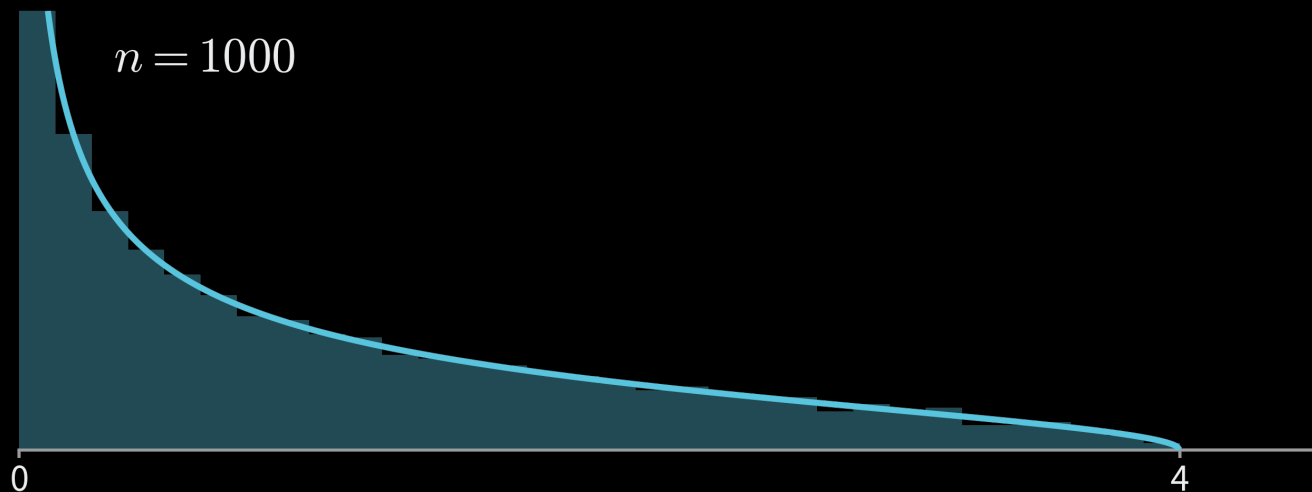
C_Q の第 i 固有値

$$\lambda_i(C_Q) = n^{-1} \sigma_i(Q)^2, \quad i = 1, \dots, n$$

個々の成分から離れ **二乗特異値の集団** を見る

次元を上げると二乗特異値の分布が安定する

$Q_{i,j} \sim \mathcal{N}(0, 1)$ とした C_Q の 1 標本の二乗特異値分布



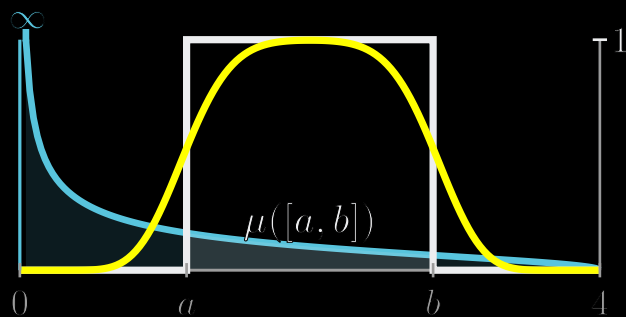
成分は標本ごとに変わる．それでも n を大きくすると，二乗特異値の経験分布には **安定した形** が現れる

実線は $n \rightarrow \infty$ の Marchenko-Pastur 極限分布 $\rho_{\text{MP}}(x) = \frac{1}{2\pi} \sqrt{\frac{4-x}{x}}$ ($0 < x < 4$) で， $x \rightarrow 0$ では発散する

多項式をプローブにして分布を探る

知りたい量 $\mu([a, b])$

MP 密度 ρ_{MP} 単関数 $1_{[a,b]}$ 多項式 p_K



区間の単関数を，積分の意味で多項式近似する

$$p_K(x) = \sum_{k=0}^K c_k x^k \approx 1_{[a,b]}(x)$$

$$\mu([a, b]) = \int 1_{[a,b]}(x) d\mu(x) \approx \int p_K(x) d\mu(x)$$

$$= \sum_{k=0}^K c_k \underbrace{\int x^k d\mu(x)}_{m_k}$$

$\text{tr}_n(M) := n^{-1} \text{Tr}(M)$ は行列内の平均 $\mathbb{E}$ は標本間の平均

$$m_k^{(n)} = \mathbb{E} \text{tr}_n [C_Q^k]$$

PART 1 / 3

ペアリングが現れる

n が大きいとき，どの項が残るのか

- 1. ペアリングが現れる
- 2. 組合せ構造から代数へ
- 3. Attention へ戻る

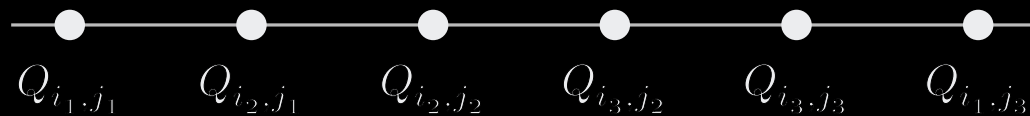
3 次モーメントを成分で書く

$$\begin{aligned} C_Q &= n^{-1} Q^\top Q \quad m_3^{(n)} := \mathbb{E} \operatorname{tr}_n [C_Q^3] \\ &= n^{-4} \sum_{i_1, i_2, i_3} \sum_{j_1, j_2, j_3} \mathbb{E} [Q_{i_1, j_1} Q_{i_2, j_1} Q_{i_2, j_2} Q_{i_3, j_2} Q_{i_3, j_3} Q_{i_1, j_3}] \end{aligned}$$

$Q^\top Q$ には Q が 2 個ある．3 乗すると 6 個の Gaussian 因子 が並ぶ

$\operatorname{tr}_n = n^{-1} \operatorname{Tr}$ ．各 i_r, j_r は $1, \dots, n$ を動く

6 個の因子をトレースの順に並べる

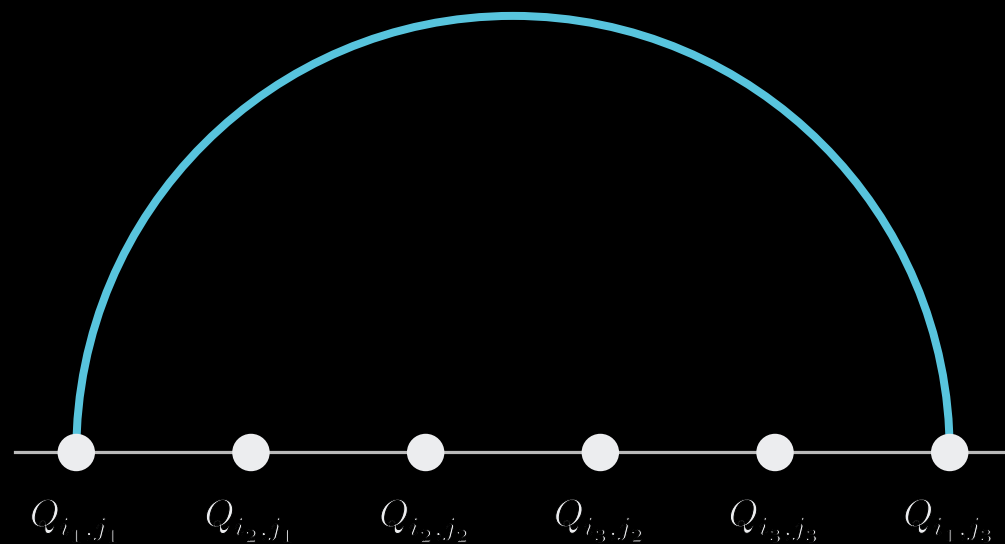


隣り合う因子は添字を共有する．左端と右端も i_1 で閉じる

点は因子，横位置はトレース内の順序

Wick の公式では 2 個の因子を組にする

$$\mathbb{E}[Q_{i_s, j_s} Q_{i_t, j_t}] = \delta_{i_s i_t} \delta_{j_s j_t}$$



弧 1 本が因子の組を表し，行添字と列添字がそれぞれ同一視される． n^{-3} は 3 個の組からではなく， $C_Q^3 = (n^{-1} Q^\top Q)^3$ の規格化から現れる

6 因子の期待値は 15 個のペアリングの和になる

$$\mathbb{E} \prod_{r=1}^6 Q_{i_r, j_r} = \sum_{\pi \in \mathcal{P}_2(6)} \prod_{\{s, t\} \in \pi} \delta_{i_s i_t} \delta_{j_s j_t}, \quad |\mathcal{P}_2(6)| = 15$$



入れ子

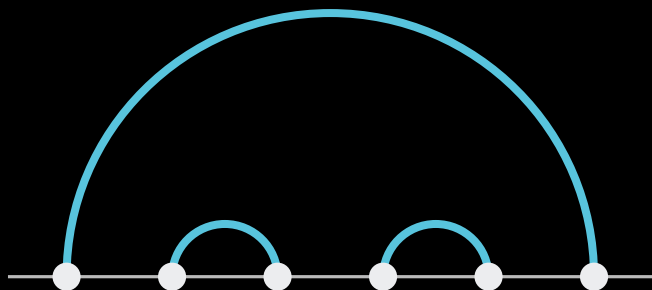


一部が交差



全部が交差

自由添字 4 個が主要項 n^0 を作る

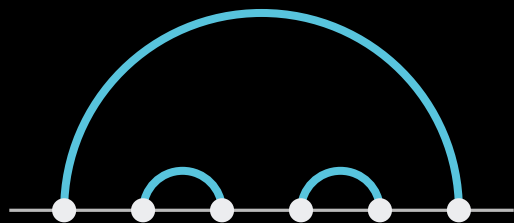


$$\pi = \{16, 23, 45\}$$

自由な添字は i_1, i_2, i_3, j の 4 個

$$\begin{aligned} & \text{添字和 } n^4 \\ & \text{Gram 規格化 } n^{-3} \\ & \text{規格化トレース } n^{-1} \\ & n^4 n^{-3} n^{-1} = 1 \end{aligned}$$

交差すると自由添字が 1 個減る



非交差: 自由添字 4 個

$$n^0$$



交差: 自由添字 3 個

$$n^{-1}$$

大次元では，自由添字を多く保つペアリングだけが残る

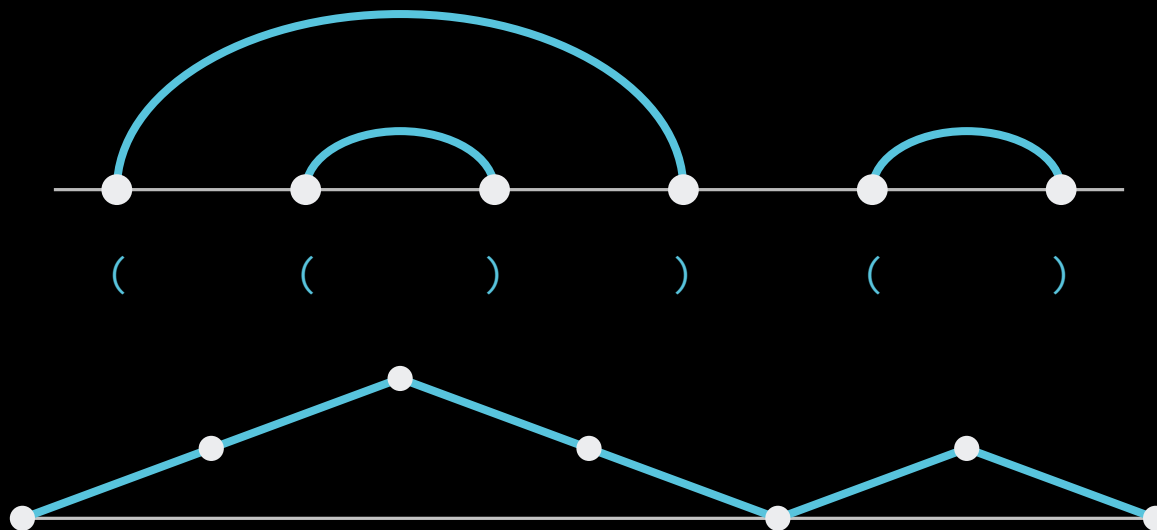
$n \rightarrow \infty$ では交差しないペアリングだけが残る

$$m_3^{(n)} = \sum_{\pi \in \mathcal{P}_2(6)} n^{c_i(\pi) + c_j(\pi) - 4} = 5 + 6n^{-1} + 4n^{-2}$$



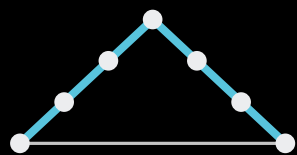
残るのは 弧が交差しない 5 個のペアリング である

非交差ペアリングは Dyck path と 1 対 1 に対応する

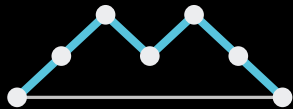


弧の始点を (, 終点を) と読む

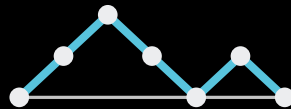
極限モーメントは Catalan 数になる



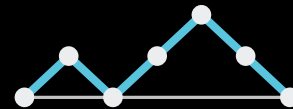
((()))



((())())



((())())



()(())



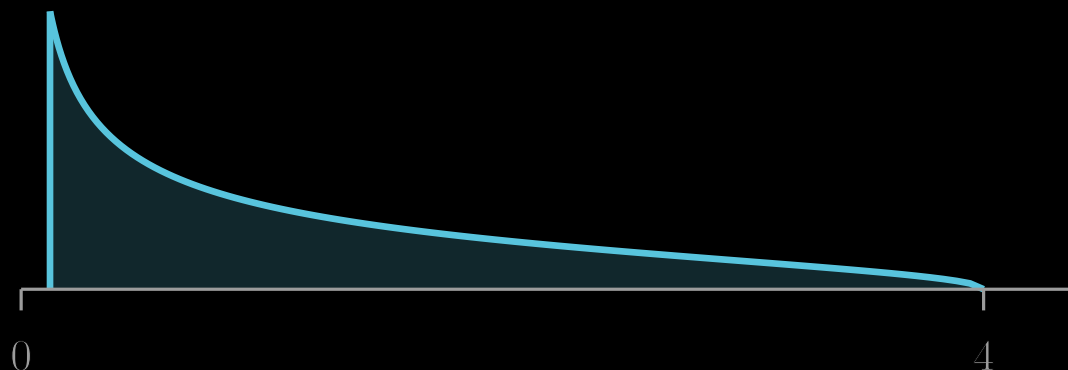
()()()

$$m_3^{(n)} \rightarrow |\text{NC}_2(6)| = C_3 = \frac{1}{3+1} \binom{6}{3} = 5$$

同じ機構はどの次数 k でも働き，主要項は非交差ペアリングの個数になる

全モーメントから極限分布を回収する

$$m_k^{(n)} \rightarrow C_k \quad (n \rightarrow \infty)$$



すべての次数で Catalan 数が現れ，コンパクトな台を持つ Marchenko–Pastur 極限分布が定まる

行列からペアリングへ，ペアリングから分布へ

行列

モーメントを成分に展開する

ペアリング

自由添字の数で主要項を選ぶ

分布

非交差構造を数えて極限へ戻る

1つのランダム行列を，**離散構造へ翻訳して分布まで戻る** 計算言語ができた

PART 2 / 3

組合せ構造から代数へ

- 1. ペアリングが現れる
- 2. 組合せ構造から代数へ
- 3. Attention へ戻る

複素行列の計算を，再利用できる代数的な規則へ
まとめる

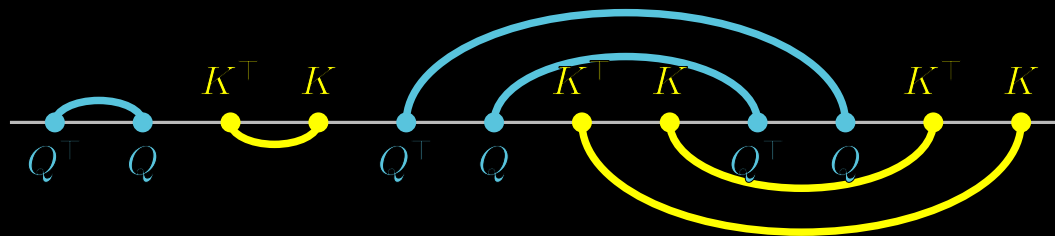
スコア行列の積は 2 つの Gram 行列へ移る

$K = (K_{i,j}), \quad K_{i,j} \sim \mathcal{N}(0, 1) \text{ i.i.d.}, \quad Q \text{ と } K \text{ は独立}$

$$C_Q = n^{-1} Q^\top Q \quad C_K := n^{-1} K^\top K$$

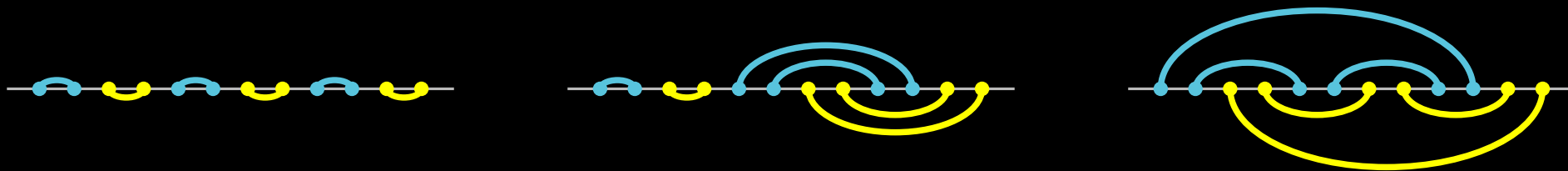
$$S = n^{-\frac{1}{2}} Q K^\top, \quad n^{-\frac{1}{2}} S = n^{-1} Q K^\top$$

$$\mathbb{E} \operatorname{tr}_n \left[\left((n^{-\frac{1}{2}} S) (n^{-\frac{1}{2}} S)^\top \right)^3 \right] = \mathbb{E} \operatorname{tr}_n \left[(C_Q C_K)^3 \right]$$



Q 由来どうし, K 由来どうしを別々に組にする

2色の計算には，再利用できる規則が必要になる



各色に 15 通り $15 \times 15 = 225$

個別のペアリングを越えて，**再利用できる計算規則**へまとめる

225 通りを追わず，3 次モーメントを求めたい



各色に 15 通り $15 \times 15 = 225$

計算済み

$$\mathbb{E} \operatorname{tr}_n [C_Q^k]$$

$$\mathbb{E} \operatorname{tr}_n [C_K^k]$$

$k = 1, 2, \dots$

知りたい量

$$\mathbb{E} \operatorname{tr}_n \left[(C_Q C_K)^3 \right]$$

各行列の情報を保ったまま **混合だけを計算する規則** がほしい

1 つでは消えた向きが, 2 つの積には残る

1 つの行列

$$C_Q = O_Q \Lambda_Q O_Q^\top$$

$$\mathrm{tr}_n [C_Q^k] = \mathrm{tr}_n (\Lambda_Q^k)$$

$$O_Q^\top O_Q = I_n \text{ で相殺}$$

2 つの行列

$$\mathrm{tr}_n \left[(C_Q C_K)^3 \right]$$

$$C_Q \ C_K \ C_Q \ C_K \ C_Q \ C_K$$

O_Q と O_K を含む積が残る

混合モーメントでは, 2 つの向きの積を調べる必要がある

ガウス分布の回転対称性から、一様な固有基底が現れる

$$UC_QU^\top \stackrel{d}{=} C_Q \quad UC_KU^\top \stackrel{d}{=} C_K$$

Query の Gram 行列

$$C_Q = O_Q \Lambda_Q O_Q^\top$$

O_Q

Key の Gram 行列

$$C_K = O_K \Lambda_K O_K^\top$$

O_K

O_Q, O_K は 独立な Haar 直交行列 : 直交群から一様に選んだ行列

どちらの Haar 行列の積が、出発点へ戻るだろう

語 A

$$O_Q O_K O_K^\top O_Q^\top$$

$$g_Q g_K g_K^{-1} g_Q^{-1} = e$$

$$\mathbb{E} \operatorname{tr}_n [O_Q O_K O_K^\top O_Q^\top] = 1$$

語 B

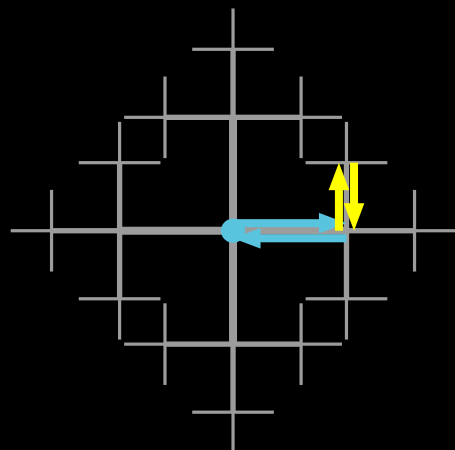
$$O_Q O_K O_Q^\top O_K^\top$$

$$g_Q g_K g_Q^{-1} g_K^{-1} \neq e$$

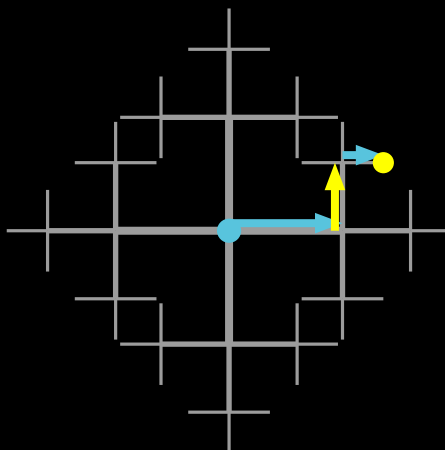
$$\lim_{n \rightarrow \infty} \mathbb{E} \operatorname{tr}_n [O_Q O_K O_Q^\top O_K^\top] = 0$$

転置を逆元とみなすと隣り合う元と逆元だけが消去される

根へ戻る経路だけが，トレースに残る



根へ戻る



根以外で止まる

$$\langle \delta_u, \delta_v \rangle = \mathbf{1}_{u=v}$$

$$L_g \delta_w = \delta_{gw}$$

$$\tau_{\mathbb{F}_2}(T) = \langle T\delta_e, \delta_e \rangle$$

根 e の係数だけを読む

$\mathbb{F}_2 = \langle g_Q, g_K \rangle$ の語は，自由群の Cayley グラフ（木）の経路として読める

固有値は分布を， Haar 方向は混合を決める

$$\mathrm{tr}_n \left[(C_Q C_K)^3 \right] = \mathrm{tr}_n \left[(O_Q \Lambda_Q O_Q^\top O_K \Lambda_K O_K^\top)^3 \right]$$

固有値

$$\Lambda_Q \quad \Lambda_K$$

各行列だけの MP モーメント

向き

$$O_Q \quad O_K$$

2 色を混ぜた積のトレース規則

固有値の情報を保ったまま，混合の規則を整理する

平均を分けると，未知の混合は交互積へ集約される

$$M^\circ := M - \mathbb{E} \operatorname{tr}_n(M) I_n \quad \mathbb{E} \operatorname{tr}_n(M^\circ) = 0$$

$$M = C_Q$$

$$M = C_K$$

$$C_Q^\circ \ C_K^\circ \ C_Q^\circ \ C_K^\circ$$

計算済み

スカラー項と，同じ色だけのモーメント

新しい混合

中心化した 2 色の因子が交互に残る

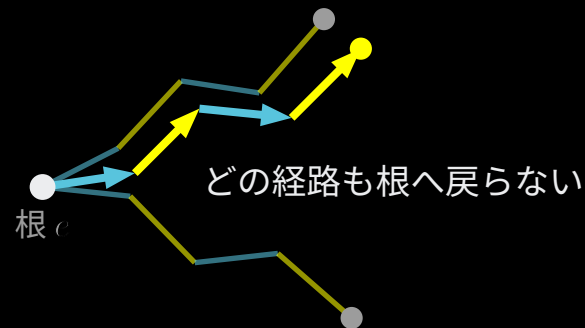
各色のブロックから期待正規化トレース倍の I_n を引く操作を **中心化** と呼ぶ

根へ戻らない交互積は トレースに残らない

中心化した 6 因子

$$C_Q^\circ \ C_K^\circ \ C_Q^\circ \ C_K^\circ \ C_Q^\circ \ C_K^\circ$$

各色から平均成分を除く



$$\lim_{n \rightarrow \infty} \mathbb{E} \operatorname{tr}_n \left[\left(C_Q^\circ C_K^\circ \right)^3 \right] = 0$$

自由群の木では根へ戻らず 大次元 Wishart 行列では トレースに残らない

この交互中心化積の消失則を自由性と呼ぶ

6 因子を 平均と残りへ分ける

$$C_Q \ C_K \ C_Q \ C_K \ C_Q \ C_K$$

Query の Gram 行列

$$C_Q = \mathbb{E} \operatorname{tr}_n(C_Q) I_n + C_Q^\circ$$

$$\text{平均} \quad \mathbb{E} \operatorname{tr}_n(C_Q) I_n$$

$$\text{残り} \quad C_Q^\circ$$

Key の Gram 行列

$$C_K = \mathbb{E} \operatorname{tr}_n(C_K) I_n + C_K^\circ$$

$$\text{平均} \quad \mathbb{E} \operatorname{tr}_n(C_K) I_n$$

$$\text{残り} \quad C_K^\circ$$

各因子で平均か残りを選んで展開する

展開項は 単色モーメントが消える交互積へ整理される

- | | | |
|---|------------|---------------------------------------|
| 1 | 平均を外へ出す | 期待正規化トレース倍の I_n をスカラーとして扱う |
| 2 | 同色をまとめる | 隣り合う Query 側または Key 側の因子を 1 つのブロックにする |
| 3 | 各ブロックを再中心化 | まとまった同色ブロックを平均と残りへもう一度分ける |

単色だけなら計算済みのモーメント

中心化した色が交互ならトレースは 0

どの展開項も同じ 3 段階で処理できる

次数が変わっても使う情報は同じ

各次数の展開で次の2つのモーメント列を繰り返し参照する

Query 側

$$\left\{ \mathbb{E} \operatorname{tr}_n \left[C_Q^k \right] \right\}_{k \geq 1}$$

Key 側

$$\left\{ \mathbb{E} \operatorname{tr}_n \left[C_K^k \right] \right\}_{k \geq 1}$$

$k = 3, 4, 5, \dots$ 次数ごとに中心化と自由性の再帰を組み直す

この2列を全次数について1つの関数にまとめたい

S 変換では混合分布の計算が掛け算になる

3つの極限スペクトル分布を次で表す

$$\mu_Q$$

Query 側の Gram 行列の極限分布

$$\mu_K$$

Key 側の Gram 行列の極限分布

μ_{prod} 2つの Gram 行列の積の極限分布

$$\mathcal{S}_{\mu_{\text{prod}}}(z) = \mathcal{S}_{\mu_Q}(z) \mathcal{S}_{\mu_K}(z)$$

正方 Ginibre 基準模型では

$$\mathcal{S}_{\mu_Q}(z) = \mathcal{S}_{\mu_K}(z) = \frac{1}{1+z}, \quad \mathcal{S}_{\mu_{\text{prod}}}(z) = \frac{1}{(1+z)^2}$$

PART 3 / 3

Attention へ戻る

- 1. ペアリングが現れる
- 2. 組合せ構造から代数へ
- 3. Attention へ戻る

前半の組合せ論と自由性を非線形模型へ広げ、
Attention へ戻る

組合せ論と自由性から Gaussian equivalence が導かれる

非線形な Attention

線形ランダム行列模型

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

$$Y_{\text{lin}}^f$$

背後にある計算

非線形なランダム行列でも、モーメントを組合せ論的に展開する

高次元で残る項を自由性によって整理する

その結果、中心化後の極限二乗特異値分布が一致する

Attention のスペクトルを、計算可能な線形模型で調べられる

無マスク・単一ヘッド・ランダム初期化・固定 β の比例極限 | Theorem 3.2

中心化後の Attention は，スコア項と独立ノイズの和に対応する

共通方向を分ける

$$u_n = n^{-\frac{1}{2}} \mathbf{1}_n, \quad A^\perp = A - u_n u_n^\top$$

softmax から決まる係数

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

$$\boxed{Y_{\text{lin}}^f} = n^{-\frac{1}{2}} \sqrt{\theta_2} \begin{matrix} W^Q & W^K \\ \swarrow & \searrow \\ \begin{matrix} \text{スコア項} \end{matrix} \end{matrix} S + n^{-\frac{1}{2}} \sqrt{\theta_1 - \theta_2} \begin{matrix} W^Q, W^K \text{ とは独立} \\ \downarrow \\ \begin{matrix} \text{独立ノイズ} \end{matrix} \end{matrix} Z$$

線形模型

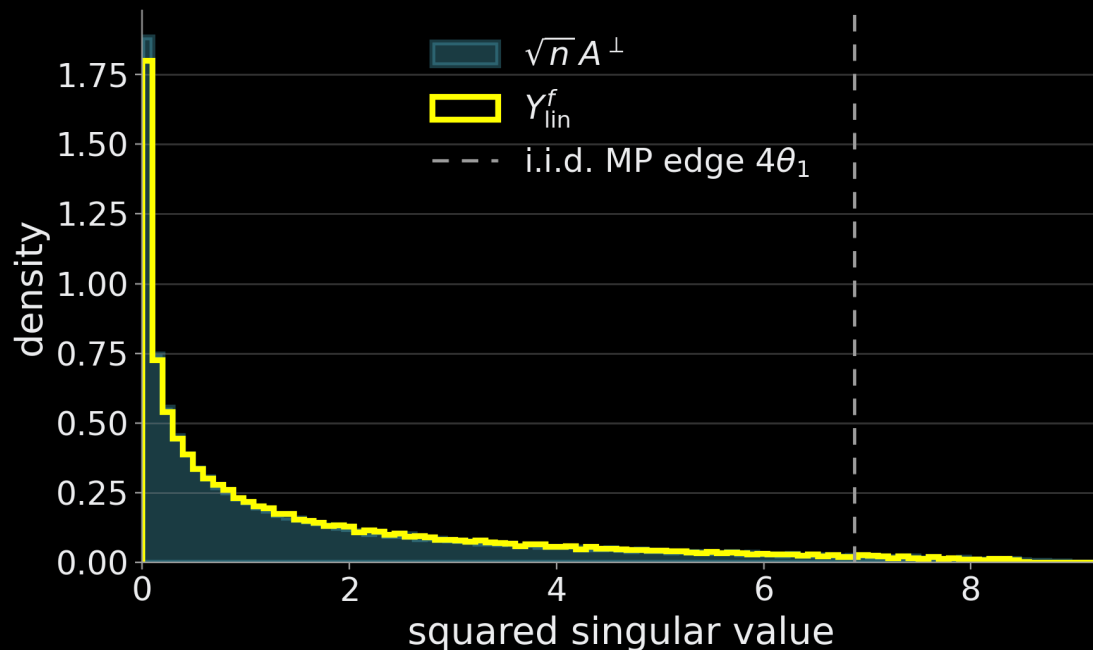
スコア項

独立ノイズ

↑
↓ 同じ極限二乗特異値分布
 $\sqrt{n} A^\perp$

Attention と近似模型は，ほぼ同じバルクを描く

固定 β の比例極限では， $\sqrt{n}A^\perp$ と Y_{lin}^f の経験二乗特異値分布が同じ極限へモーメント収束する



Hayase et al., AISTATS 2026, Thm. 3.2 / Fig. 1 | $n = d = d_{\text{qk}} = 1000$, $\beta = 1$, 10 試行 | 各試行の上位 3 つを除外

Attention のスペクトルを計算できる問題へ移す

$$\sqrt{n}A^\perp$$

$$Y_{\text{lin}}^f$$

極限二乗特異値分布が一致

橋渡し

非線形な softmax を含む Attention と線形ランダム行列模型の二乗特異値分布を対応させる

計算

線形模型に現れる Gaussian 行列の和と積をモーメントと自由性の計算言語で調べる

複雑な非線形模型をスペクトルを計算できる模型へ翻訳する

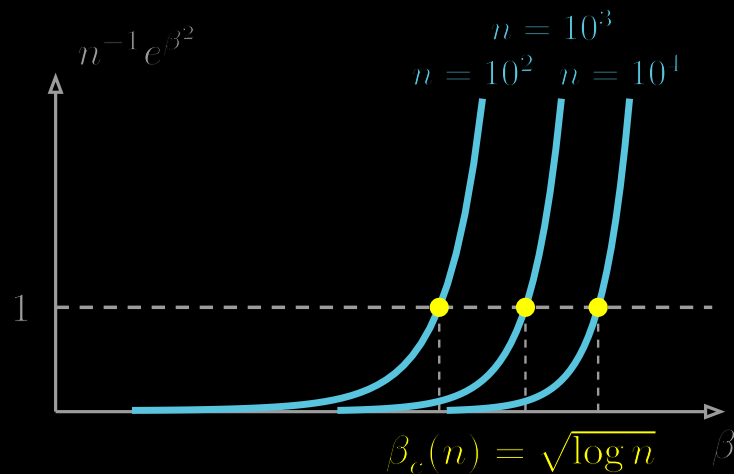
バルク端のスケールは $\beta_c = \Theta(\sqrt{\log n})$ を示唆する

線形モデルのスケール評価

バルク端 $\asymp e^{\beta^2}$

$A^\perp$ の二乗特異値へ戻す

バルク端 $\asymp n^{-1}e^{\beta^2}$



釣り合い

$$n^{-1}e^{\beta_c^2} \asymp 1$$

対数を取る

$$\beta_c^2 \asymp \log n$$

境界候補

$$\beta_c = \Theta(\sqrt{\log n}) = \Theta(\sqrt{\log d})$$

固定 β の Gaussian equivalence を手がかりに、増大する β のスケールを推定 | Hayase, Collins & Karakida, Sec. 6.1

長い文脈に応じて温度を調整する経験則

Qwen LogN-Scaling

YaRN

SSMax

極限分布の計算から，Attention のスケール候補を得た

01

経験二乗特異値分布に，安定した極限が現れる

ランダム性と高次元極限の中から，スペクトルの形が立ち上がる

02

モーメントと自由性から，その極限分布を回収できる

全モーメントを計算し，分布の形へ戻す

03

Gaussian 模型から，Attention のスケール候補が見える

$\beta_c = \Theta(\sqrt{\log n})$ 今回の模型における逆温度の境界候補

代数へ移すと，複雑な組合せを計算できる

1 | ほどく

ランダム行列を
組合せへ

Wick 展開と次数計算

大次元で残る接続を取り出す

2 | まとめる

接続と消去の規則を
代数へ

自由性

混合モーメントを整理する

3 | 計算する

$\mathcal{S}$ 変換へ移すと

積の計算が掛け算になる

$$\mathcal{S}_{\mu_{\text{prod}}}(z) = \mathcal{S}_{\mu_Q}(z) \mathcal{S}_{\mu_K}(z)$$

全次数の再帰を 1 本にまとめる

Gaussian equivalence を介して スペクトルの計算を Attention へ返す

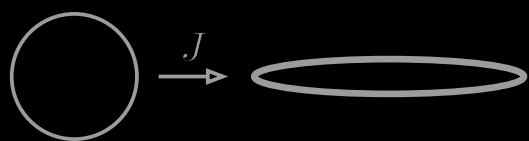
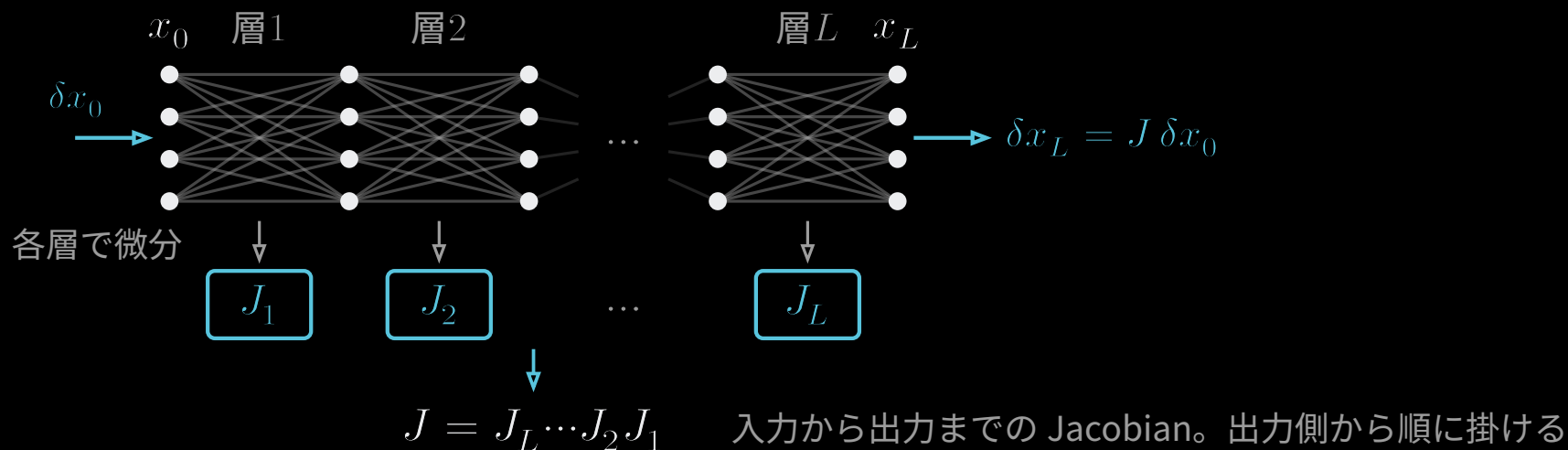
OUTLOOK

数学的構造は 深層回路の別の場所にも現れる

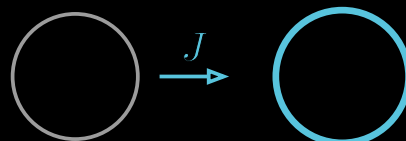
解析・アルゴリズム的能力・表現設計に現れた 3 つの例

以下の 3 例は互いに独立

超深層回路の勾配解析: Dynamical isometry



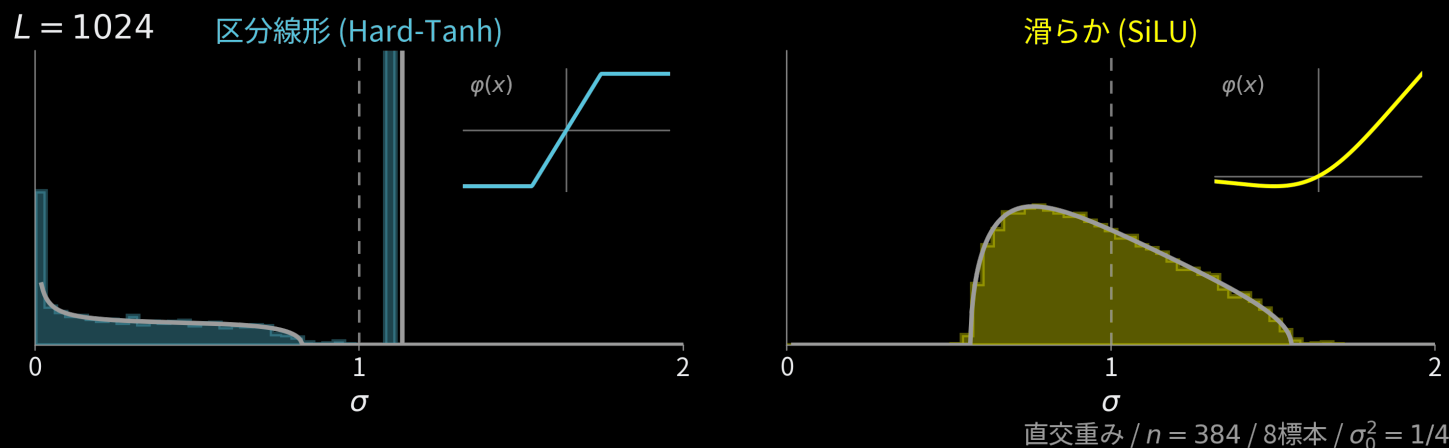
平均が 1 でも、深さとともに広がる



全特異値が 1 付近 = どの方向も長さを保つ

初期化時の解析. このスペクトルは前半の $\mathcal{S}$ 変換で計算できる | Pennington et al., NeurIPS 2017. Collins & Hayase, CMP 397, 2023. | 独立した例 1/3

層を重ねるほどスペクトルは普遍的な形へ収束する



$$\mathcal{S}_{\mu_{JJ^T}}(z) = \left(\mathcal{S}_{\mu_{D^2}}(z) \mathcal{S}_{\mu_{W^T W}}(z) \right)^L \xrightarrow{L \rightarrow \infty} \begin{cases} e^{-z\sigma_0^2/(1+z)} & \text{区分線形} \\ e^{-z\sigma_0^2} & \text{滑らか} \end{cases}$$

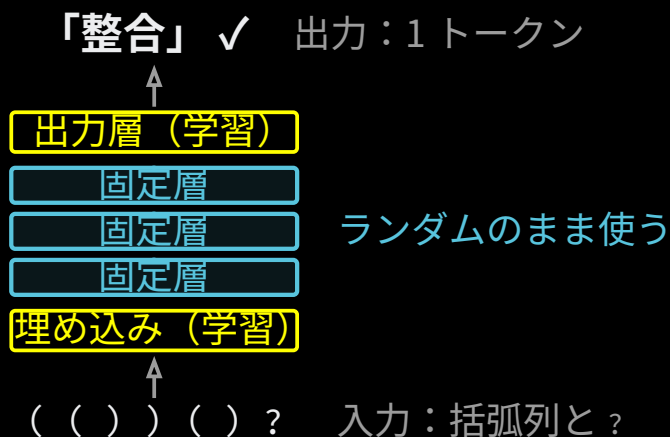
前半の掛け算規則の L 乗が、 $(1 + x/L)^L \rightarrow e^x$ の形で指数関数の極限に化ける

安定な極限には直交性が必要． Gaussian 重みではどの活性化でも，ReLU でも不可

Pennington, Schoenholz & Ganguli, AISTATS 2018 (arXiv:1802.09979), Sec. 5. | 二重スケーリング $\sigma_0^2 = 1/4$ 固定

ランダムな Transformer にもアルゴリズム的能力がある

Random Transformer



Transformer 内部をランダムに固定し，入力・位置埋め込みと出力層だけを学習する

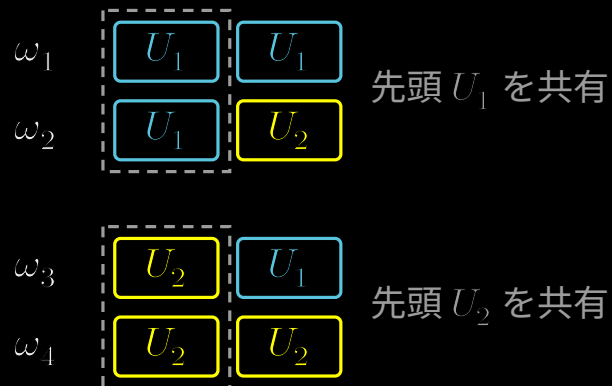
例：Parenthesis Balancing——括弧列に ? を付けて入力し，

整合か不整合かを 1 トークンで答える（長さ 60 まで）

この課題で精度 100%。モジュラー加算・10 桁の加算・連想再生でも 100% に達する

自由群の構造を入力射影へ入れる

長さ 2 の語は 4 パターン



Free Random Projection

自由群の語をランダム直交行列の積へ写すと，共通接頭辞が共有行列因子になる

この階層的依存を入力射影へ導入し，論文で検証した ICRL 課題で未見環境への汎化性能が向上した

Hayase, Collins & Inoue, AISTATS 2026, arXiv:2504.06983. | 独立した例 3/3

今回の方法

深層回路から構造を抽出・分解し

代数計算から性質を再構成する

共通モチーフ

再帰と消去 木上の経路・Dyck path・自由群の語

兆しが現れた3つの独立例

スペクトルの安定性

アルゴリズム的能力

表現設計

Dynamical Isometry

Random Transformer

Free Random Projection

次の問い これらの関係をたどると、深層回路の何が見えてくるだろうか

参考文献 | 付録

- Vaswani et al., “Attention Is All You Need,” NeurIPS 2017.
- Bishop, Del Moral, Niclas, “An introduction to Wishart matrix moments,” 2018. [arXiv:1710.10864](#)
- Nica, Speicher, Lectures on the Combinatorics of Free Probability, 2006.
- Hayase, Collins, Karakida, “Gaussian Equivalence for Self-Attention,” AISTATS 2026. [arXiv:2510.06685](#)
- Pennington, Schoenholz, Ganguli, “Resurrecting the sigmoid in deep learning through dynamical isometry,” NeurIPS 2017.
- Collins, Hayase, “Asymptotic Freeness of Layerwise Jacobians Caused by Invariance of Multilayer Perceptron,” CMP 397, 2023.
- Zhong, Andreas, “Algorithmic Capabilities of Random Transformers,” NeurIPS 2024.
- Hayase, Collins, Inoue, “Free Random Projection for In-Context Reinforcement Learning,” AISTATS 2026. [arXiv:2504.06983](#)
- Bai et al., “Qwen Technical Report,” 2023.
- Peng et al., “YaRN: Efficient Context Window Extension of Large Language Models,” 2023.
- Nakanishi, “Scalable-Softmax Is Superior for Attention,” 2025.