

自己注意機構における 逆温度スケーリングの統一的解析

早瀬友裕^{A, B} 唐木田亮^{B, A}
^A理研 AIP ^B産総研

Self-attention と softmax の定義

長さ n の入力ベクトル列 $x_1, \dots, x_n$ を, 学習する行列で変換する.

$$q_j = W^Q x_j, \quad k_j = W^K x_j, \quad v_j = W^V x_j$$

q : query k : key v : value d_{qk} : query / key の次元

$$z_{n,j} = \frac{q_i^\top k_j}{\sqrt{d_{qk}}}, \quad y_i = \sum_{j=1}^n p_{n,j}(\beta) v_j$$

$$p_{n,j}(\beta) = \text{softmax}(\beta z_n)_j = \frac{e^{\beta z_{n,j}}}{\sum_{\ell=1}^n e^{\beta z_{n,\ell}}}$$

位置 i を固定し, 文脈中の n 個の token から重み付き平均 y_i を求める.

逆温度 β は, 重みの集中度を調整するパラメータ.

文脈長 n が増えるとき, 逆温度 β_n をどう変えるか?

n : 文脈長 (context length) . Vaswani et al. (2017), § 3.2; Hayase & Karakida, § 2.1.

既存の逆温度スケーリングと指数の不一致

文脈長 n に応じて逆温度を $\beta_n \asymp (\log n)^\xi$ と伸ばす提案がある.
指数 ξ は仮定で異なる.

指数 ξ	スコアの仮定・補正	文献
$\frac{1}{2}$	Gaussian logits の解析	Giorlandino & Goldt (2026) Anson et al. (2025)
1	等相関配置 / LogN・SSMax	Chen et al. (2025) Bai et al. (2023), Nakanishi (2025)
2	YaRN の温度補正	Peng et al. (2023)

仮定が違えば指数も違う. スコア列から指数を読む枠組みが要る.

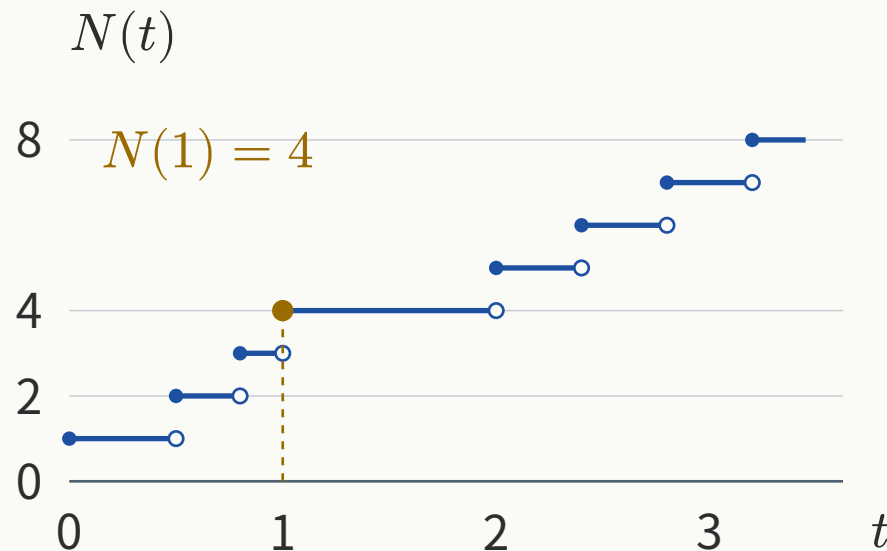
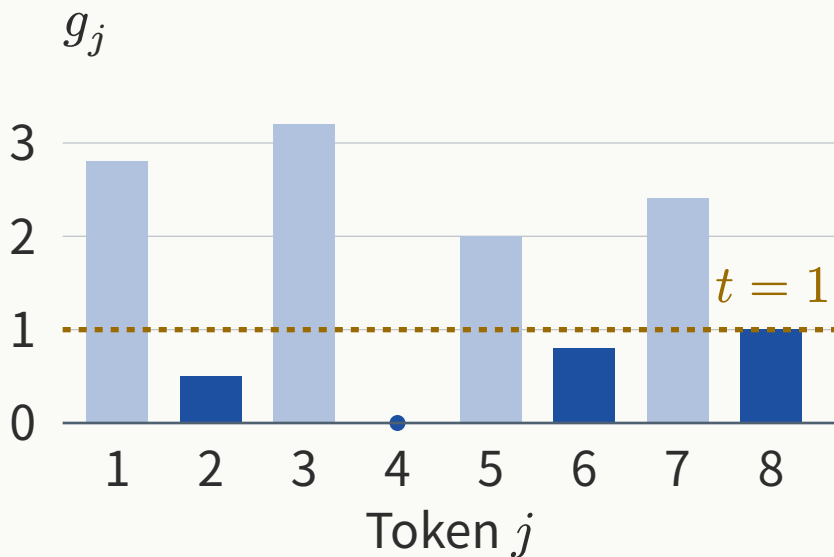
本研究：スコア列から臨界スケール Λ_n を定義し, 指数を統一的に導く.

最大値からのギャップと累積個数 (gap counting)

$$g_{n,j} = z_n^* - z_{n,j}, \quad N_n(t) = \#\{j : g_{n,j} \leq t\}$$

各 token のギャップ

ギャップが t 以下の個数

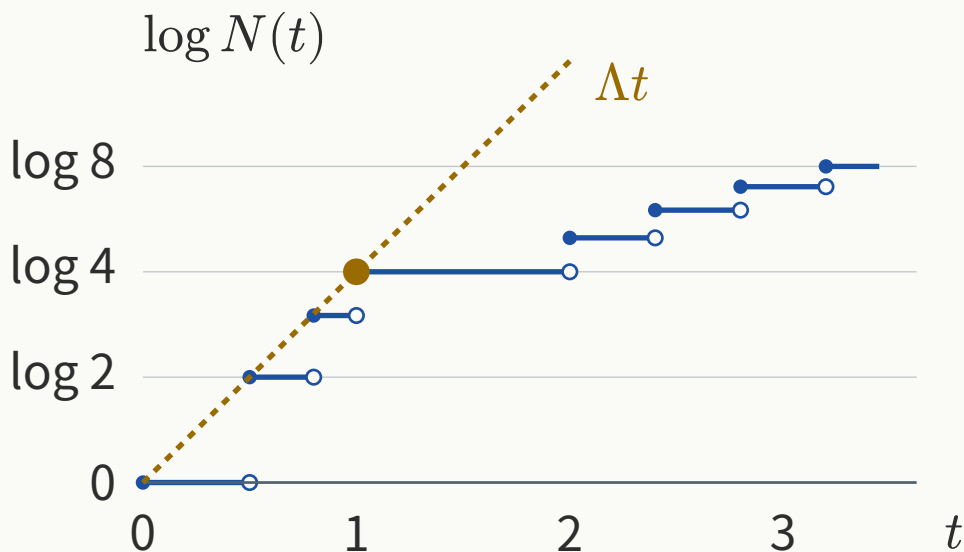


$N(1) = 4$: 最大スコアを持つ token も含めて数える.

Definition 4.3. $z_n^* = \max_j z_{n,j}$.

日本物理学会 第 81 回年次大会 · 14pL1222-12

臨界スケール Λ の定義



臨界スケールの定義

$$\Lambda_n := \sup_{t>0} \frac{\log N_n(t)}{t}$$

階段全体を上から覆う
最小の傾き.

$$N_n(t) \leq e^{\Lambda_n t}$$

この例では $\Lambda = \log 4$.

重みの減衰 $e^{-\beta t}$ と, 個数 $N(t)$ の増加が競合する.

Definition 4.4.

日本物理学会 第 81 回年次大会 · 14pL1222-12

逆温度と臨界スケール：二つの崩壊と予測

主定理

$\beta_n / \Lambda_n \rightarrow 0 \implies D_n := p_{n,(1)} - p_{n,(2)} \rightarrow 0$ 上位 2 つが分離しない.

$\beta_n / \Lambda_n \rightarrow \infty \implies H_n := -\sum_j p_{n,j} \log p_{n,j} \rightarrow 0$ 1 つに集中する.

スコアの Gaussian 性・独立性は不要. スコア列ごとに Λ_n が決まる.

予測（対数べきの仮定）

$\beta_n \asymp (\log n)^{\xi_\beta}, \Lambda_n \asymp (\log n)^{\xi_\Lambda}$ のとき, 二つの崩壊を避けるには $\xi_\beta = \xi_\Lambda$.

逆温度の指数 ξ_β は, スコアから読める Λ の指数 ξ_Λ と一致するはず.

Theorem 4.6（十分大きい n で最大スコアの候補が 1 つ）, Corollary 4.8. 指数の一致であり, $\beta_n = \Lambda_n$ ではない.

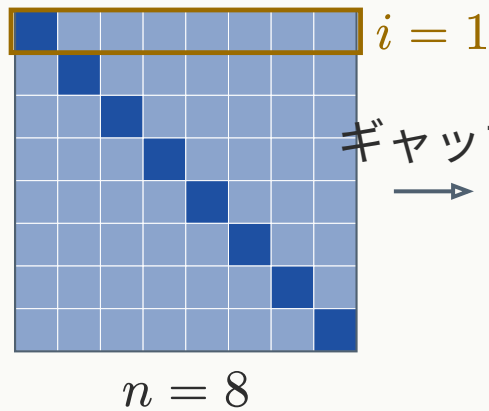
等相関 (simplex) 配置のスケーリング指数

全 token のノルム q と内積 ρ が共通. ギャップは $\Delta_n = q - \rho$.

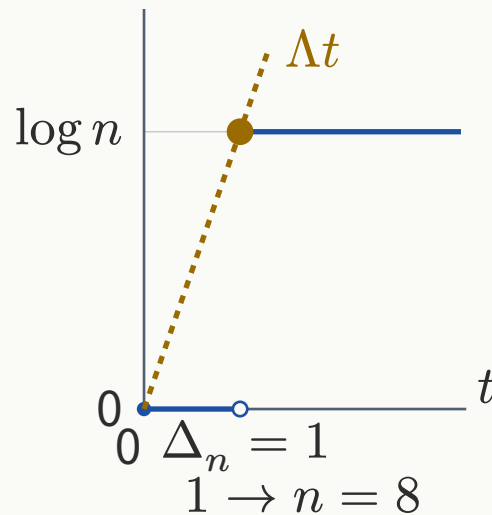
スコア行列 Σ_n^{simp}

$\log N_n(t)$

$$\Lambda_n = \frac{\log n}{\Delta_n}$$



ギャップ
→



ギャップ	Λ_n	指数 ξ
$\sqrt{\log n}$	$\sqrt{\log n}$	$\frac{1}{2}$
1	$\log n$	1
$(\log n)^{-1}$	$(\log n)^2$	2

1 行目 : 2, 1, ..., 1

$\Lambda_n \asymp (\log n)^\xi$ の指数 ξ は, ギャップの変化で決まる.

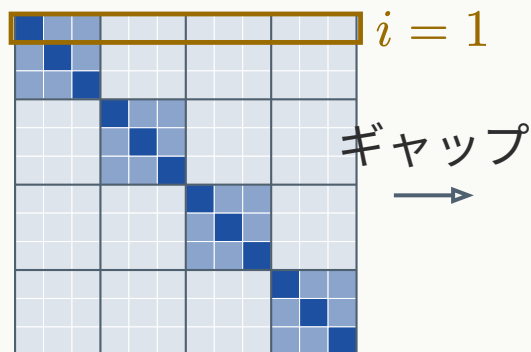
§ 4.4. $\Sigma_n^{\text{simp}} = (r_n^2 - \Delta_n)J_n + \Delta_n I_n$, Cowsik et al. (2025), Chen et al. (2025). $\asymp$: 定数倍を除いた比較.

二階層ブロック配置の臨界スケール

大きさ m_n のブロックに分割. ギャップはブロック内 δ_n , ブロック間 τ_n .

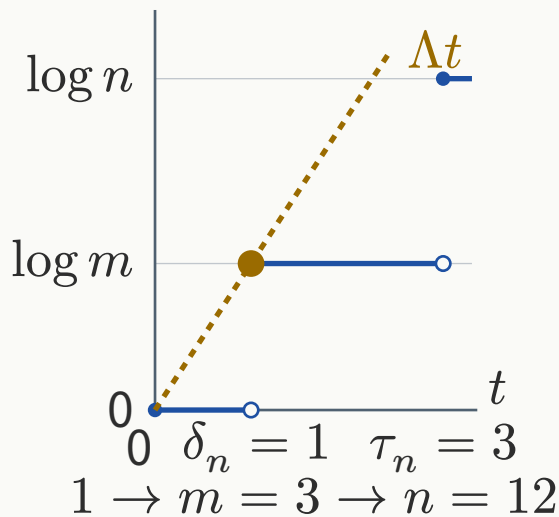
スコア行列 Σ_n^{block}

$\log N_n(t)$



$n = 12, m = 3$

1 行目 : 4, 3, 3, 1, ..., 1



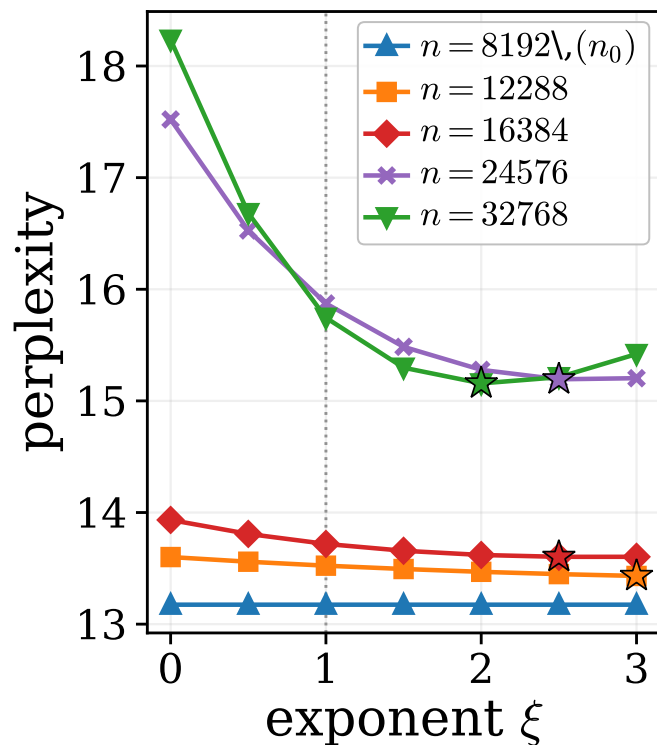
$$\Lambda_n = \max \left\{ \frac{\log m_n}{\delta_n}, \frac{\log n}{\tau_n} \right\}$$

近い集団の寄与 $\log m_n / \delta_n$ と
全体の寄与 $\log n / \tau_n$ の
大きい方.

$\log 3 > \frac{\log 12}{3}$: ブロック内の近い 3 候補が Λ を決める.

Qwen : PPL 最小の指数と Λ の指数

Qwen-7B-Chat / PG-19



書籍 PG-19 で, 次の token の予測性能 PPL を測る.

PPL は小さいほど良い.

ξ_{β}^* : 逆温度の倍率 $[\log n / \log 8192]^{\xi}$ の ξ を 0.5 刻みで掃引し, PPL が最小になる ξ .

$\hat{\xi}_{\Lambda}$: attention スコアの gap counting から計算した理論値.

文脈長	ξ_{β}^* (PPL 掃引)	$\hat{\xi}_{\Lambda}$ (理論)
16,384	2.5	2.31 ± 0.10
32,768	2.0	1.76 ± 0.10

差は 0.3 以下で, 探索の刻みより小さい.
ともに文脈長とともに減少.

Fig. 1(a), Table 2. 学習時の文脈長は 8192. Proof-Pile-2 でも同様 (補足 2) .

nanoGPT：学習した逆温度の指数と Λ の指数

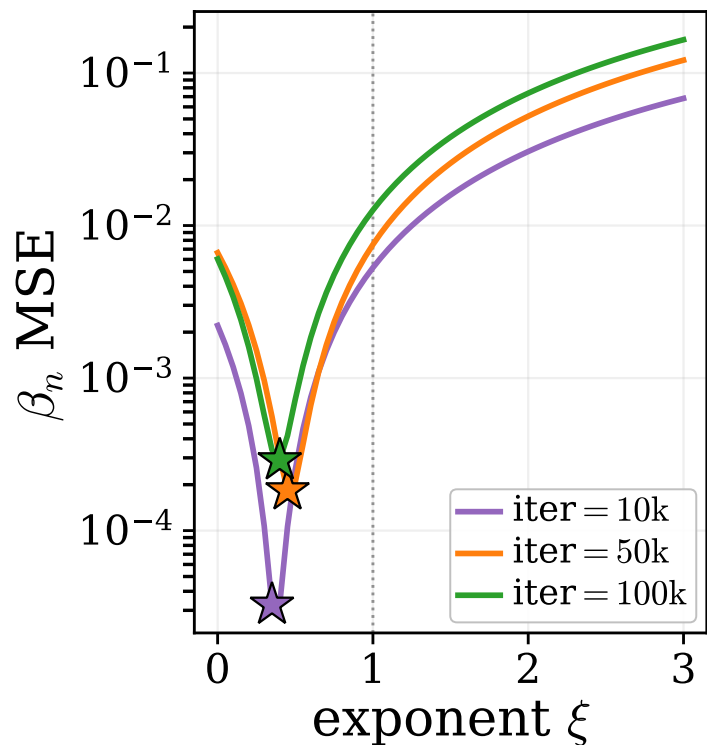
GPT-124M / SlimPajama

文脈長 1024 で 100k step 学習.

位置 n ごとの逆温度 β_n も学習する.

ξ_β^* ：学習した β_n を $(\log n)^\xi$ で当てはめた最適値.

ξ_Λ ：スコアの gap counting から計算した理論値.



Dataset	ξ_β^* (学習した β_n)	$\hat{\xi}_\Lambda$ (理論)
SlimPajama	0.35 ± 0.04	0.52 ± 0.03
OpenWebText	0.42 ± 0.05	0.69 ± 0.05

差は 0.17 と 0.27. 指数はどれも 1 未満で,
1 を超える Qwen と対照的.

Fig. 1(c), Table 2. 図は全位置の当てはめ, 表は 6 点の回帰. 25k step では差 0.12~0.14 (補足 3) .

まとめ

1. 臨界スケール Λ_n は, スコアの gap counting だけから決まる.

$\Lambda_n = \sup_{t>0} \log N_n(t)/t$. Gaussian 性や独立性は不要.

2. 二つの崩壊を避けるには, 指数 $\xi_\beta = \xi_\Lambda$.

等相関配置では, ギャップの取り方で $\xi_\Lambda = 1/2, 1, 2$ が実現する.

3. Qwen ($\xi \approx 2$) と nanoGPT ($\xi \approx 0.5$) で,

理論値 $\hat{\xi}_\Lambda$ が掃引の最適値 ξ_β^* と 0.3 以内で一致.

逆温度のスケールは文脈長だけでは決まらず, スコアの構造で決まる.

今後の課題：モデルの構造や学習過程からスコアの分布を導く.

参考文献

本研究

Hayase & Karakida (2026). A unified framework for critical scaling of inverse temperature in self-attention. arXiv:2605.12697.

定義

Vaswani et al. (2017). Attention is all you need. NeurIPS 30.

指数 1/2

Giorlandino & Goldt (2026). Two failure modes of deep transformers and how to avoid them. ICLR 2026, arXiv:2505.24333.

Anson, Wang & Aitchison (2025). Scale-invariant attention. NeurIPS 2025, arXiv:2505.17083.

指数 1

Nakanishi (2025). Scalable-Softmax is superior for attention. arXiv:2501.19399.

Bai et al. (2023). Qwen technical report. arXiv:2309.16609.

Chen, Lin, Polyanskiy & Rigollet (2025). Critical attention scaling in long-context transformers. arXiv:2510.05554.

等相關配置

Cowsik, Nebabu, Qi & Ganguli (2025). Geometric dynamics of signal propagation predict trainability of transformers. Phys. Rev. E 112, 055301.

指数 2

Peng, Quesnelle, Fan & Shippole (2023). YaRN: Efficient context window extension of large language models. ICLR 2024, arXiv:2309.00071.

データ・コード : PG-19 (Rae et al., 2020), Proof-Pile-2 (Azerbayev et al., 2024), nanoGPT (Karpathy, 2022), SlimPajama (Soboleva et al., 2023), OpenWebText (Gokaslan et al., 2019).

補足 1 | 分配関数による主定理の証明

$$Z_n(\beta) = \sum_j e^{-\beta g_{n,j}} = \beta \int_0^\infty e^{-\beta t} N_n(t) dt$$

接触点から下界を得る

$$Z_n \geq N_n(\Delta_n) e^{-\beta_n \Delta_n}$$

$$D_n \leq \left(\frac{2}{e}\right) \frac{\beta_n}{\Lambda_n} \rightarrow 0$$

包絡線から上界を得る

$$1 \leq Z_n \leq \frac{\beta_n}{\beta_n - \Lambda_n} \rightarrow 1$$

上界には $\beta_n > \Lambda_n$ が必要.

$$H_n = \log Z_n - \beta_n \frac{Z'_n}{Z_n}$$

$H_n \rightarrow 0$ を示すため, 微分項も評価する.

Theorem 4.6 の証明. 十分大きい n での評価. 左は逆温度と Λ の比が 0 に, 右は無限大に向かう極限.

補足 2 | Qwen の評価条件と推定方法

Dataset	文脈長	ξ_{β}^*	$\hat{\xi}_{\Lambda}$	除外率
PG-19	16k	2.5	2.31 ± 0.10	7.4%
PG-19	32k	2.0	1.76 ± 0.10	8.8%
Proof-Pile-2	16k	2.5	2.33 ± 0.11	6.8%
Proof-Pile-2	32k	2.0	1.74 ± 0.11	7.2%

補正倍率： $m(n; \xi) = \left[\max\left(1, \frac{\log n}{\log 8192}\right) \right]^{\xi}$.
各評価文脈長で 4 位置を測定し, $\log X$ の平均を $\log \log n$ に回帰.
現在の層の倍率を掛ける前のスコアを用いる. 前の層の温度は表現に影響する.
± : bootstrap IQR の半幅, B=200, 再標本化の単位は(layer, head).
最大値が同点となる行を除外： $\varepsilon = 10^{-6}, \log_{10} \Lambda > 5$.

Table 2, § 4.5, 実験付録. 16k=16,384, 32k=32,768. ξ_{β}^* (PPL 最小の指数) に区間推定はない.
日本物理学会 第 81 回年次大会・14pL1222-12

補足 3 | nanoGPT の学習逆温度と臨界スケール

Dataset	学習 step	ξ_{β}^*	$\hat{\xi}_{\Lambda}$
SlimPajama	25k	0.40 ± 0.03	0.54 ± 0.03
SlimPajama	100k	0.35 ± 0.04	0.52 ± 0.03
OpenWebText	25k	0.46 ± 0.05	0.57 ± 0.04
OpenWebText	100k	0.42 ± 0.05	0.69 ± 0.05

学習した逆温度の指数を推定する. PPL 最小化とは異なる.

6 位置 $n = 64, 128, 256, 512, 768, 1024$ で OLS 回帰.

β の土は OLS 標準誤差, Λ の土は bootstrap IQR の半幅.

学習時の seed は各設定 1 つ. 3 つの評価 seed は独立な学習試行ではない.

Fig. 1 の全位置 MSE フィットによる指数 0.4–0.5 とは, 推定方法が異なる.

Table 2, 実験付録. GPT-124M, 文脈長 1024, learned absolute PE. 最大値の同点を除外.

補足 4 | 対数べきスケーリングと指数の一致

対数のべきで表されるスケーリングを仮定する.

$$\beta_n \asymp (\log n)^{\xi_\beta}, \quad \Lambda_n \asymp (\log n)^{\xi_\Lambda}$$

$$\xi_\beta < \xi_\Lambda \implies \beta_n / \Lambda_n \rightarrow 0, \quad \xi_\beta > \xi_\Lambda \implies \beta_n / \Lambda_n \rightarrow \infty$$

この仮定の下で, 二つの崩壊をともに避けるには $\xi_\beta = \xi_\Lambda$ が必要.

これは指数についての必要条件であり, $\beta_n = \Lambda_n$ とは異なる.
係数の値や, 崩壊を避けられることまでは保証しない.

接点からの分解: $\Lambda_n = \frac{\alpha_n \log n}{\Delta_n}, \xi_\Lambda = 1 + \xi_\alpha - \xi_\Delta.$

Δ_n : 最大の接触ギャップ, $N_n(\Delta_n) = n^{\alpha_n}$. 各指数が存在する場合.

Definitions 4.4–4.5; Corollaries 4.8–4.9. 漸近的な指数と有限範囲の推定値も区別する.