SCALING LIMITS OF LONG-CONTEXT TRANSFORMERS

GIUSEPPE BRUNO, SHI CHEN, ZHENGJIANG LIN, YURY POLYANSKIY, AND PHILIPPE RIGOLLET

Abstract. We study the long-context limit of softmax self-attention with a fixed query and a random context of n i.i.d. keys on the sphere, viewing the inverse temperature β_n as the scaling parameter that decides whether attention degenerates into uniform averaging or collapses onto the single closest key. We show that the critical scale at which selectivity emerges is determined by the local exponent of the distance-to-query distribution near zero rather than by global features of the context, and scales like $\beta_n^* \asymp n^{2/(d-1)}$ for uniform keys on $\mathbb{S}^{d-1}$. Furthermore, we characterize the limiting laws of the ordered attention weights and of the attention output across all regimes of β_n: a subcritical regime in which the output reduces to a local average around q with explicit deterministic bias and Gaussian fluctuations; a critical regime in which a finite collection of nearest keys retains macroscopic mass without single-key collapse; and a supercritical regime in which all mass concentrates on the closest key. Of notable interest is the subcritical case with identity value matrix where the attention map approximately implements a backward heat equation.

Contents

1. Introduction

The push toward ever longer contexts has placed transformers on a trajectory in which the context length n is no longer a fixed architectural detail but a scaling parameter in its own right [DYY+19, BPC20, ZGD+20, Goo24]. From this vantage point, a finite-context attention head is naturally understood as a finite-sample approximation of an idealized *infinite-context* limit. Context length then plays the role of a resource controlling how faithfully the finite system resolves this idealization, and its $n \to \infty$ scaling limit becomes the natural object to study.

Self-attention [VSP+17] with Post-LayerNorm[1] routes a query $q \in \mathbb{S}^{d-1}$ against a context of keys $x_1, \dots, x_n \in \mathbb{S}^{d-1}$ through the softmax attention weights

$$A_j \;=\; \frac{\exp\big(\beta_n \langle q, x_j\rangle\big)}{\sum_{l=1}^{n} \exp\big(\beta_n \langle q, x_l\rangle\big)}, \qquad j = 1, \dots, n, \tag{1.1}$$

which form a probability distribution over the context and are then used to aggregate the x_j. The inverse temperature β_n is a classical device employed in long-context transformers [BBC+23, PQFS24, ZCL+24, AWA25, BPA25, Nak25, PLAS+25, CLPR26, GG26, QHVM26] that tunes how sharply $A = (A_1, \dots, A_n)$ separates keys with highest scores from the bulk. In all settings, β_n is set to grow with the context length n. With no rescaling ($\beta_n = 1$), even a key whose alignment $\langle q, x_j\rangle$ exceeds the bulk by a constant margin still produces a weight $A_j = \Theta(1/n)$: the exponential gap such a key introduces is dwarfed by the n competing terms in the softmax denominator. The head loses the ability to distinguish relevant keys from the bulk—a phenomenon now commonly called *attention dilation* or *attention fading* [VPBP25, Nak25]—and β_n must grow with n at the right rate to preserve selectivity.

A first wave of theoretical work has examined the scaling limits as $n \to \infty, \beta_n \to \infty$ through simple, fully tractable models in which the fluctuations across keys are suppressed by design. These analyses connect attention sharpness to the contraction properties of self-attention layers [GLPR25, CLPR25] and identify a logarithmic critical scale [Nak25, CLPR26]: when $\beta_n \asymp \log n$, attention concentrates on $\Theta(1)$ tokens while remaining content-adaptive. Such results pin down the right scale, but the price of tractability is that the random objects which actually live at this scale—the size of the largest weight, the gap to the runner-up, the effective number of attended tokens, and the corresponding fluctuations of the attention map—are by construction invisible. A finer description requires that the keys themselves be treated as random and that the entire random profile of attention be analyzed at large n.

Probability theory has a long tradition of answering such questions: take a random system indexed by a size n, find the rescaling under which it admits a non-degenerate limit, and identify the law that emerges. Familiar instances include Gaussian fluctuations at scale $\sqrt{n}$ in the central limit theorem, the Fisher–Tippett–Gnedenko classification of extremal laws, and the $\log n$-scale free-energy transition of the Random Energy Model [Der81], recently connected to long-context attention by [GG26].

For a fixed query vector $q \in \mathbb{S}^{d-1}$ and context vectors $x_1, \dots, x_n \in \mathbb{S}^{d-1}$, define the *attention weights* as in (1.1) and the corresponding *attention output*

$$\mathsf{ATT}(n) = V \sum_{j=1}^{n} A_j x_j,$$

where $V \in \mathbb{R}^{d \times d}$ is a fixed *value matrix*. The natural scalar coordinate for these objects is the *distance-to-query*, or D2Q, $T_i := 1 - \langle q, x_i\rangle$, which is half the squared

[1]We show how PostLN attention with general Key and Query matrices reduces to this model in Section 2.

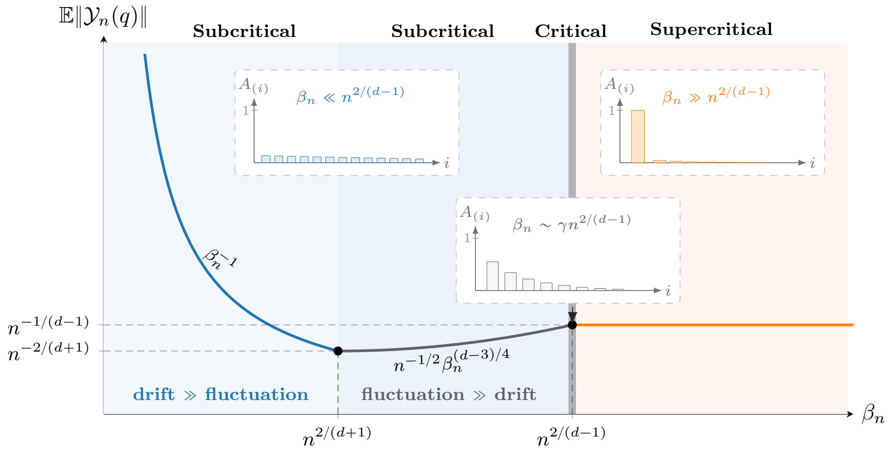

Figure 1. Unified schematic of the ordered-weight and attention-output regimes as the inverse temperature β_n varies, with $\mathcal{Y}_n(q) = \mathsf{ATT}(n) - Vq$

Euclidean distance from key x_i to the query. Two qualitatively distinct failure modes flank the regime of interest: if β_n is too small, the smallest D2Qs are blurred into the bulk, the weights A_j are essentially uniform, and the attention map fails to distinguish relevant tokens; if β_n is too large, the entire output collapses onto the single closest key, and the map ceases to integrate the context. The question addressed in this paper is:

> *As $n \to \infty$, which scalings of β_n produce non-degenerate limiting laws for the ordered attention weights $\{A_j\}_j$ and the attention output $\mathsf{ATT}(n)$, and what are those laws?*

To make the limit precise, we model the keys $x_1, \ldots, x_n$ as i.i.d. samples from a common distribution on the sphere, which induces a distribution of the D2Qs $T_1, \ldots, T_n$ on $[0, 2]$. The behavior at large n is then governed by the local behavior of this distribution near 0—the distribution of close neighbors—through the same mechanism that controls right-tail extremes in classical extreme-value theory. Within this model, our results below paint a complete picture of the scaling limits of attention across regimes of β_n, with explicit limit theorems at and around the critical scale.

Our contributions. We answer the question above through a statistical scaling theory for the attention distribution of a fixed query q over a random context. The unifying principle is that long-context behavior is controlled not by the global shape of the context distribution but by the small-T tail of the D2Q—i.e., the local geometry of the context around q—and by the point process of the closest keys that this tail generates. Our main results are the following.

(C1) *The critical scale is set by the small-D2Q tail.* We identify the inverse temperature β_n^* at which softmax first resolves the closest keys to q from the bulk of the context, and show that this scale is determined by the local exponent of the D2Q distribution near zero, rather than by global features of the context. For keys uniform on $\mathbb{S}^{d-1}$, this principle yields the polynomial law $\beta_n^* \asymp n^{2/(d-1)}$, and for more structured key distributions it singles out local dimension and density as the relevant quantities.

(C2) *Limit laws for the ordered attention weights.* We characterize the asymptotic law of the ordered profile $A_{(1)} \geqslant A_{(2)} \geqslant \cdots$ across the full range of β_n. Three regimes emerge within a single framework: a subcritical regime in which the weights flatten and the head approaches uniform averaging; a critical regime around β_n^* at which the weights converge to a non-degenerate random profile, with a finite collection of influential neighbors retaining macroscopic mass; and a supercritical regime in which the weight on the single closest key tends to one. The critical limit is the principal object of the paper: it is the unique scale at which selectivity emerges without single-key collapse.

(C3) *Limit laws for the attention output.* We extend the analysis to the attention output $\mathsf{ATT}(n) = V \sum_j A_j x_j$. The subcritical phase is not merely a regime of vanishing individual weights: as long as β_n still diverges, $\mathsf{ATT}(n)$ behaves as a local average of the keys around q, and we obtain its limiting law as the sum of a deterministic bias, a Gaussian fluctuation term, and a density-sensitive tangential drift. At and above the critical scale, the corresponding limits of $\mathsf{ATT}(n)$ follow from those of the weights in (C2).

(C4) *A heat-type interpretation of subcritical residual updates.* The deterministic vector field that drives the subcritical output in (C3) admits a transparent interpretation: in the isotropic case, its tangential component is the gradient of the log token density. Attention layers operating below the critical scale therefore push representations from low-density toward high-density regions of the context, which corresponds approximately to a backward heat evolution.

Related work. Empirical work on long-context attention provides much of the motivation for the present study. A number of recent methods modify the attention logits in order to preserve useful sparsity as the context grows, including Qwen-style scaling, YaRN, SWAN-GPT, scalable-softmax, selective attention, and scale-invariant attention [BBC⁺23, PQFS24, PLAS⁺25, Nak25, ZCL⁺24, AWA25]. These works show that attention sharpness is an important practical degree of freedom in long contexts. Our goal is complementary: we do not propose a new implementation, but instead ask what limiting law a rescaled attention row should obey once the context itself is modeled statistically.

Theoretical explanations of such scaling laws are still limited. The deterministic critical-scaling theory of [CC22, Nak25, CLPR26] identifies a logarithmic scale $\beta_n \asymp \log n$, while the signal-propagation theory of [GG26] and the scale-invariant-attention framework of [AWA25] point, in different Gaussian settings, to $\sqrt{\log n}$-type score scales. Our polynomial scale arises from a different score geometry: the relevant competitors are rare nearest neighbors to a fixed query vector, and the scale is set by the local lower tail of the D2Q. Thus the contrast between logarithmic, square-root-logarithmic, and polynomial laws should be viewed as evidence that attention scaling depends on the geometry of the competing scores, rather than as a disagreement between models.

Notation. We introduce the notation used throughout the paper. We denote the ordered attention scores as $A_{(1)} \geqslant A_{(2)} \geqslant \cdots \geqslant A_{(n)}$. Define the *D2Q* $T_i :=$

$1 - \langle q, x_i \rangle \in [0,2]$, $i = 1, \dots, n$. The attention weights can be written as

$$A_i = \frac{e^{-\beta_n T_i}}{Z_n}, \qquad Z_n := \sum_{j=1}^{n} e^{-\beta_n T_j}. \tag{1.2}$$

Let $T_{(1)} \leqslant T_{(2)} \leqslant \cdots \leqslant T_{(n)}$ be the order statistics of the D2Qs. Finally, let the D2Q CDF be $F(t) := \mathbb{P}(T_1 \leqslant t)$, $t \in [0,2]$. We set $\alpha := \frac{2}{d-1}$ throughout the paper. We define the constant $C(q)$ to be the local intensity of near matches to the query $C(q) := \lim_{t \to 0^+} F(t) t^{-1/\alpha}$.
Organization. The remainder of the paper is organized as follows. Section 2 introduces the fixed-query attention model and the spherical i.i.d. baseline. Section 3 establishes the subcritical, critical, and supercritical limits for the ordered attention weights. Section 4 derives the corresponding attention-output limits, including the marked Poisson functional at criticality and the subcritical drift/fluctuation trichotomy. Section 5 presents numerical experiments illustrating the predicted phase transitions. The appendices collect the proofs and extend the framework to correlated tokens and RoPE.

2. Setup

We recall that a PostLN attention head [VSP+17, GLPR25] with context $x_1, \dots, x_n$ is parametrized by three $d \times d$ matrices Q, K, and V. It implements the map

$$x \in \mathbb{S}^{d-1} \mapsto V \sum_{j=1}^{n} A_j x_j,$$

where the attention weights are defined by

$$A_j \;=\; \frac{\exp\big(\beta \langle Qx, Kx_j \rangle\big)}{\sum_{l=1}^{n} \exp\big(\beta \langle Qx, Kx_l \rangle\big)}, \qquad j = 1, \dots, n\,.$$

Setting $q = K^\top Qx / \|K^\top Qx\| \in \mathbb{S}^{d-1}$ and $\beta' = \beta\, \|K^\top Qx\|$, we can rewrite

$$A_j \;=\; \frac{\exp\big(\beta' \langle q, x_j \rangle\big)}{\sum_{l=1}^{n} \exp\big(\beta' \langle q, x_l \rangle\big)}, \qquad j = 1, \dots, n,$$

so that, up to rescaling β, the attention weights take the form (1.1).

Throughout the main theorems, we assume a random context $x_1, \dots, x_n$ sampled i.i.d from some distribution on the sphere.

Assumption 1 (i.i.d. spherical context)**.** *The query $q \in \mathbb{S}^{d-1}$ is fixed. The context directions $x_1, \dots, x_n$ are i.i.d. with common law $\rho(x)\, \mathrm{d}\sigma(x)$ on $\mathbb{S}^{d-1}$, where σ is surface measure. The density ρ is positive at q and is C^2.*

A few remarks regarding Assumption 1 should be noted.
Regularity of ρ, intrinsic dimension, and off-support geometry. The C^2 regularity is used only for the attention-output expansions in Section 4. The scaling limits of the ordered weights in Section 3 require only continuity of ρ at q and $\rho(q) > 0$. If ρ is supported on a smooth $(d'-1)$-dimensional submanifold, with $d'-1 < d-1$, through q with positive local density and non-degenerate contact, then our results hold with d replaced by d'. If q is not contained in the support of ρ, the same analysis applies after subtracting the smallest attainable D2Q from all D2Qs.
Correlations and positional encoding. Independence is also not essential to our results. The proofs of the scaling limits use only two inputs: a one-point lower-tail

law for the D2Q $T_i = 1 - \langle q, x_i \rangle$ and a Poisson, or more generally extremal, limit for close neighbors to the query q. Under stationary weak dependence with anti-clustering, the same exponent and Poisson softmax limit persist. We also discuss how positional encodings such as RoPE [SAL+24] can also be handled with minor technical adjustments in Section F.

3. Scaling limits of attention weights

In this section, we consider the scaling limits for the ordered attention weights. We start with an informal derivation of the critical scaling $\beta_n^* \asymp n^{2/(d-1)}$ (equivalently $\beta_n^* \asymp n^\alpha$ with $\alpha = 2/(d-1)$).

3.1. **A heuristic explanation for the critical scaling.** The critical scale $\beta_n^* \asymp n^{2/(d-1)}$ arises precisely at the balance point of two opposing effects. As the context length n grows, more random tokens from the input naturally fall close to the fixed query direction q. At the same time, increasing the rescaling factor β_n narrows the effective softmax window: the factor $e^{-\beta_n T_i}$ assigns comparable weight only to tokens whose distance-to-query (D2Q) values T_i differ on the scale $1/\beta_n$.

For simplicity, first assume that the distribution ρ in Assumption 1 is the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. Geometrically, a D2Q window of width $1/\beta_n$ corresponds to a small ball around q. Indeed, for tokens x_i sufficiently close to q, we have the local approximation $T_i \approx \frac{1}{2}\mathrm{dist}(q, x_i)^2$. Consequently, the active softmax window has radius of order $\beta_n^{-1/2}$. Since this region is intrinsically $(d-1)$-dimensional, the expected number of context tokens falling inside the window scales as

$$n \cdot \mathrm{Vol}\big(B_{d-1}(q, \beta_n^{-1/2})\big) \asymp n\beta_n^{-(d-1)/2} = n\beta_n^{-1/\alpha}.$$

Here, $\mathrm{Vol}\big(B_{d-1}(q, \beta_n^{-1/2})\big)$ is the volume of the geodesic ball around q of radius $\beta_n^{-1/2}$. If β_n grows too rapidly with n, the window shrinks faster than new tokens enter it and the expected count tends to zero: attention collapses onto the single nearest token. Conversely, if β_n grows too slowly, this number diverges and attention is effectively averaged over a diverging number of tokens. The critical regime occurs precisely when these two effects balance, i.e.,

$$n\beta_n^{-1/\alpha} \asymp 1, \qquad \text{or equivalently} \qquad \beta_n \asymp n^\alpha.$$

The same scaling holds for general distributions ρ satisfying Assumption 1. Indeed, $\mathbb{P}(T_i < t) \asymp t^{1/\alpha}$ (see Theorem A.1), so the expected number of tokens lying in a D2Q window of width β_n^{-1} is

$$n\mathbb{P}(T_i < \beta_n^{-1}) \asymp n\beta_n^{-1/\alpha}.$$

The next three results make this intuitive picture precise. In each case, the main technical step is to pass from the tail law $\mathbb{P}(T_i < t) \asymp t^{1/\alpha}$ to the corresponding limit law for the softmax denominator.

3.2. **Supercritical regime.** When $\beta_n \gg n^\alpha$, the temperature is high enough to distinguish the top D2Q from all its competitors. The attention function therefore becomes a hard nearest-neighbor selector.

Theorem 3.1 (Supercritical regime)**.** *Under Assumption 1, choose β_n in* (1.1) *such that $\beta_n n^{-\alpha} \to \infty$ as $n \to \infty$. We have that for every fixed $i \geqslant 2$,*

$$\lim_{n\to\infty} A_{(1)} = 1, \qquad \lim_{n\to\infty} A_{(i)} = 0, \qquad \textit{in probability.} \tag{3.1}$$

The proof of Theorem 3.1 is deferred to Section B.

3.3. **Critical regime.** At the critical scale $\beta_n \asymp n^\alpha$, the expected number of tokens in the active softmax window remains of order one. The correct limiting object is the point process of rescaled D2Qs.

Theorem 3.2 (Critical regime)**.** *Under Assumption 1, choose* β_n *in* (1.1) *such that* $\beta_n n^{-\alpha} \to \gamma \in (0, \infty)$*. Let* $\mathcal{N}$ *be a Poisson point process on* $\mathbb{R}_+$ *with intensity measure* $\Lambda([0, y]) = C(q)(y/\gamma)^{1/\alpha}$ *, for* $y \geqslant 0$*. Write* $0 < Y_1 < Y_2 < \cdots$ *for its atoms in increasing order. Then, for every fixed* $k \geqslant 1$*,*

$$(A_{(1)}, \ldots, A_{(k)}) \to (W_1, \ldots, W_k) \qquad \text{in distribution,} \tag{3.2}$$

where

$$W_i = \frac{e^{-Y_i}}{\sum_{j=1}^{\infty} e^{-Y_j}},$$

and $\{Y_i\}_{i \geqslant 1}$ *are the increasing atoms of the Poisson point process* $\mathcal{N}$*, i.e.,* Y_i *is the* i*-th arrival of* $\mathcal{N}$*.*

This is the selective regime: finitely many close neighbors retain nonzero mass, but no deterministic winner is selected. The proof is given in Section C.

3.4. **Subcritical regime.** When $\beta_n \ll n^\alpha$, the active softmax window contains a diverging number of near-query tokens. Define the active softmax window size by $m_n(q) := C(q) n \beta_n^{-1/\alpha}$. In the subcritical regime with $\beta_n \to \infty$, we have $m_n(q) \to \infty$. Hence no individual token can retain macroscopic mass. Nevertheless, after ordering by score, the top $O(m_n(q))$ weights secure a deterministic amount of total weight.

Theorem 3.3 (Subcritical regime)**.** *Under Assumption 1, choose* β_n *in* (1.1) *such that* $\beta_n n^{-\alpha} \to 0$ *as* $n \to \infty$*. We have that for every* $i \geqslant 1$*,*

$$\lim_{n\to\infty} A_{(i)} = 0 \qquad \text{in probability}. \tag{3.3}$$

If $\beta_n \to \infty$*, and let* $\{k_n\}_{n \geqslant 1}$ *be a sequence of integers such that* $\lim_{n\to\infty} \frac{k_n}{m_n(q)} = x \geqslant 0$*, for some nonnegative constant* x *(for instance* $k_n = \lfloor x m_n(q) \rfloor$*). We have that*

$$\lim_{n\to\infty} \frac{A_{(k_n)}}{A_{(1)}} = \exp(-x^\alpha) \qquad \text{in probability,} \tag{3.4}$$

Indeed, we have that

$$\lim_{n\to\infty} m_n(q) A_{(k_n)} = \frac{\exp(-x^\alpha)}{\Gamma\left(\frac{1}{\alpha} + 1\right)} \qquad \text{in probability,} \tag{3.5}$$

where $\Gamma(\cdot)$ *is the gamma function. Moreover,*

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} A_{(i)} = \frac{\gamma_{\mathrm{inc}}\left(\frac{1}{\alpha}, x^\alpha\right)}{\Gamma\left(\frac{1}{\alpha}\right)}\ , \qquad \text{in probability}\ , \tag{3.6}$$

where $\gamma_{\mathrm{inc}}(s, z) := \int_0^z u^{s-1} e^{-u} \, \mathrm{d}u$ *is the lower incomplete gamma function.*

The proofs for Theorem 3.3 is included in Section D.

4. Scaling limits of attention outputs

Section 3 describes how the ordered attention weights $A_{(1)} \geqslant A_{(2)} \geqslant \cdots \geqslant A_{(n)}$ are distributed. We now track the vector returned by the attention layer $\mathsf{ATT}(n)$. Under Assumption 1, we show in Section 4.1 and Section 4.2 that all three regimes with $\beta_n \to \infty$ concentrate on those tokens approaching q, so that the first-order limit of $\mathsf{ATT}(n)$ is always Vq. We therefore study the displacement

$$\mathcal{Y}_n(q) := \mathsf{ATT}(n) - Vq = V \sum_{i=1}^{n} A_i(x_i - q). \tag{4.1}$$

The two relevant scales in Section 4.1 and Section 4.2 are $\beta_n \asymp n^{\alpha}$, the ordered-weight transition, and $\beta_n \asymp n^{\alpha/(1+\alpha)} = n^{2/(d+1)}$, the drift/fluctuation transition inside the subcritical regime.

4.1. **Extreme-value output limits.**

In this subsection, we record attention outputs in the supercritical and the critical regimes in which only finitely many near-query tokens are visible at the output scale.

First, for the supercritical regime, we prove in the following Theorem 4.1 that the displacement $\mathcal{Y}_n(q)$ scales as $n^{-\alpha/2}$.

Theorem 4.1 (Supercritical output). *Under Assumption 1, assume that $\beta_n n^{-\alpha} \to \infty$. Then for $a_n = (C(q)n)^{\alpha}$, we have*

$$\lim_{n\to\infty} \sqrt{a_n}\, \mathcal{Y}_n(q) = V\Phi \qquad \textit{in distribution,}$$

where the random vector $\Phi = \sqrt{2R_1}\,\Theta_1$, R_1 is a Weibull random variable with

$$\mathbb{P}(R_1 > r) = e^{-r^{1/\alpha}}, \qquad r \geqslant 0,$$

and Θ_1 is uniformly distributed over $\mathbb{S}_q^{d-2} := \{\theta \in T_q\mathbb{S}^{d-1} : \|\theta\| = 1\}$, and is independent of R_1. In particular, $\mathcal{Y}_n(q) \to 0$ in probability, and hence $\mathsf{ATT}(n) \to Vq$ in probability.

Theorem 4.1 says that, once the weights have collapsed to a single winner, the only remaining randomness is the nearest-neighbor geometry: its radial D2Q has a Weibull-type extreme-value law, while its angular mark is asymptotically uniform in the tangent space.

We next show the attention outputs within the critical regime.

Theorem 4.2 (Critical output). *Under Assumption 1, assume that $\beta_n n^{-\alpha} \to \gamma \in (0,\infty)$. There exists a Poisson point process $\mathcal{M}_\gamma$ on $[0,\infty) \times \mathbb{S}_q^{d-2}$ with intensity measure depending on γ and $\rho(q)$ in Assumption 1, such that as $n \to \infty$*

$$\lim_{n\to\infty} \sqrt{\beta_n}\, \mathcal{Y}_n(q) = V\Xi \qquad \textit{in distribution.}$$

where

$$\Xi := \frac{\displaystyle\int_0^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}\sqrt{2y}\,\theta\, \mathcal{M}_\gamma(dy, d\theta)}{\displaystyle\int_0^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}\, \mathcal{M}_\gamma(dy, d\theta)}. \tag{4.2}$$

In particular, $\mathcal{Y}_n(q) \to 0$ in probability, and hence $\mathsf{ATT}(n) \to Vq$ in probability.

The marked Poisson process in Theorem 4.2 is the output analogue of the Poisson process governing the critical ordered weights in Theorem 3.2. Its scalar coordinate records how far a token is from the optimal score, while the mark records the local direction in which that token perturbs the output. The denominator in (4.2) is the limiting softmax partition function; the numerator is the corresponding tangent displacement.

4.2. **Subcritical local averaging.** When $\beta_n \ll n^\alpha$, the number of relevant near-query tokens diverges. The output is then governed by local averaging rather than by a finite extreme process. Define

(4.3)
$$\mathcal{Y}(q) := \nabla_{\mathbb{S}^{d-1}} \log \rho(q) - \frac{d-1}{2} q, \qquad \Sigma(q) := \frac{c_d}{\rho(q)} \mathbf{P}_q, \qquad c_d := 2^{-d} \pi^{-(d-1)/2},$$

where $\mathbf{P}_q = \mathrm{Id} - qq^\top$ is the orthogonal projection onto the tangent space $T_q \mathbb{S}^{d-1}$.

Theorem 4.3 (Subcritical output)**.** *Under Assumption 1, choose β_n in* (1.1) *such that $\beta_n \to \infty$ and $\beta_n n^{-\alpha} \to 0$ as $n \to \infty$.*

(1) ***Drift dominated.*** *If $\beta_n^{1+\alpha} n^{-\alpha} \to 0$, then*
$$\lim_{n\to\infty} \beta_n \mathcal{Y}_n(q) = V\mathcal{Y}(q) \qquad \text{in probability}.$$
(2) ***Mixed drift–fluctuation regime.*** *If $\beta_n^{1+\alpha} n^{-\alpha} \to \tau \in (0, \infty)$, then*
$$\lim_{n\to\infty} \beta_n \mathcal{Y}_n(q) = V\mathcal{Y}(q) + G_\tau \qquad \text{in distribution}.$$
where $G_\tau \sim \mathcal{N}(0, \tau^{1/\alpha} V \Sigma(q) V^\top)$.

(3) ***Fluctuation dominated.*** *If $\beta_n^{1+\alpha} n^{-\alpha} \to \infty$, then*
$$\lim_{n\to\infty} \sqrt{n \beta_n^{-1/\alpha+1}}\, \mathcal{Y}_n(q) = G_1 \qquad \text{in distribution}.$$
where $G_1 \sim \mathcal{N}(0, V\Sigma(q)V^\top)$

In all three regimes, we have that $\mathcal{Y}_n(q) \to 0$ in probability, and hence $\mathsf{ATT}(n) \to Vq$ in probability.

The condition $\beta_n^{1+\alpha} n^{-\alpha} \asymp 1$ is equivalent to $\beta_n \asymp n^{\alpha/(1+\alpha)} = n^{2/(d+1)}$. Thus the subcritical regime contains its own output-level phase transition: for β_n at or below this critical scale, the deterministic local-density drift is present, while it is swamped by Gaussian fluctuations above it.

4.3. **Residual dynamics.** We now study how one attention layer changes the relative geometry of the tokens in the drift-dominated subcritical regime, tracking the cosine similarity between normalized tokens before and after one layer. For $V = \mathrm{Id}$, the attention output can be written as
$$\mathsf{ATT}(n) = q + \mathcal{Y}_n(q),$$
where, by Theorem 4.3(1), $\mathcal{Y}_n(q)$ is of order β_n^{-1}. Therefore the layer-normalized attention output can be approximated as
$$q' = \frac{\mathsf{ATT}(n)}{\|\mathsf{ATT}(n)\|} = \frac{q + \mathcal{Y}_n(q)}{\|q + \mathcal{Y}_n(q)\|} \simeq q + \frac{1}{\beta_n} \nabla_{\mathbb{S}^{d-1}} \log \rho(q).$$
The proposition below formalizes this observation and makes this claim rigorous.

Proposition 4.4. *Assume $V = \mathrm{Id}$. Let $q \in \mathbb{S}^{d-1}$ satisfy $\rho(q) > 0$, and define q' as*

$$q' := \frac{q + \gamma \mathcal{Y}_n(q)}{\|q + \gamma \mathcal{Y}_n(q)\|}, \tag{4.4}$$

for some fixed $\gamma > 0$. In the drift-dominated regime where $\beta_n^{1+\alpha} n^{-\alpha} \to 0$ and $\beta_n \to \infty$, it holds

$$\beta_n (q' - q) \to \gamma \nabla_{\mathbb{S}^{d-1}} \log \rho(q), \tag{4.5}$$

in probability.

The above proposition indicates that attention, together with PostLN, approximately implements the update

$$q \mapsto q + \frac{\gamma}{\beta_n} \nabla_{\mathbb{S}^{d-1}} \log \rho(q),$$

which is precisely a one-step Euler discretization of $\dot{q} = \nabla_{\mathbb{S}^{d-1}} \log \rho(q)$. In turn, it is a celebrated result of Jordan, Kinderlehrer, and Otto [JKO98] that this Wasserstein gradient flow of the entropy functional [CNWR25] implements the reverse heat equation.

The proof of Theorem 4.4 is postponed to Appendix E.

5. Numerical Experiments

The following experiments illustrate three consequences of the scaling theory from the main results on the model introduced in (1.1) and (4.1).

Output field on the sphere. We first visualize the output vector fields from Theorems 4.1 to 4.3. We take $d = 3$, so $\alpha = 1$, set $V = \mathrm{Id}$, and draw $n = 10^4$ context tokens from

$$\rho(x) \propto \exp(x_1 x_2), \qquad x \in \mathbb{S}^2.$$

For each query direction q we compute $\mathcal{Y}_n(q) = \sum_i A_i(x_i - q)$ and plot the first coordinate of $\mathbf{P}_q \mathcal{Y}_n(q)$ after the corresponding scaling. With this choice of parameters the critical scaling is $\beta_n \asymp n$, while the subcritical drift/fluctuation scaling from Theorem 4.3 is $\beta_n \asymp n^{1/2}$. Thus the columns of Figure 2 show the supercritical regime ($\beta_n = n^{5/4}$), the critical regime ($\beta_n = n$), the fluctuation-dominated subcritical regime ($\beta_n = n^{3/4}$), the mixed drift–fluctuation regime ($\beta_n = n^{1/2}$), and the drift-dominated subcritical regime ($\beta_n = n^{1/4}$).

The first two empirical columns are visibly rough: in the supercritical regime the output is essentially determined by one nearest token, while at criticality only a finite number of close neighbors contributes. Moving into the subcritical regime, more tokens enter the local average and the field becomes smoother. The fluctuation-dominated panel still shows sampling noise; the mixed panel begins to show the large-scale drift; and the drift-dominated panel has the same structure as the deterministic density-gradient field $\mathbf{P}_q \mathcal{Y}(q)$ from (4.3), shown in the last column.

Ordered weights. We next check the ordered attention weight statements in Theorems 3.1 to 3.3. Contexts are uniform on $\mathbb{S}^4$, so $d = 5$ and $\alpha = 1/2$. The left panel of Figure 3 estimates $A_{(1)}$ over a logarithmic grid in (n, β_n), averaging over 100 trials. In agreement with Theorems 3.1 and 3.3, the largest weight is small below the critical scaling and approaches one above it; the transition occurs at $\beta_n \asymp n^{\alpha}$. The right panel fixes $n = 1000$, $q = e_1$, and $\beta_n = n^{\alpha/4}$. It plots the ratios $A_{(k)}/A_{(1)}$ against the rescaled rank $x = k/m_n(q)$, and compares them with the prediction $e^{-x^{\alpha}}$ in Theorem 3.3.

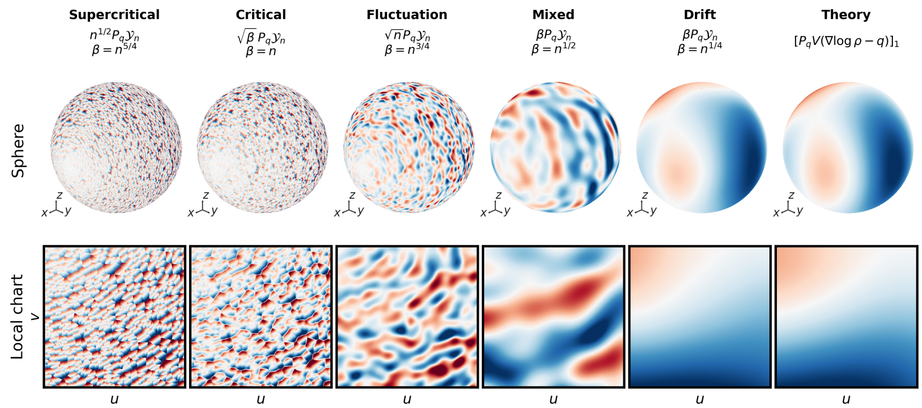

Figure 2. Rescaled output displacement on $\mathbb{S}^2$ for $n = 10^4$, $V = \mathrm{Id}$, and $\rho(x) \propto \exp(x_1 x_2)$. The columns use $\beta_n = n^{5/4}, n, n^{3/4}, n^{1/2}, n^{1/4}$, with the regime-dependent scalings shown above the panels, and the last column is the deterministic drift field. The bottom row is a geodesic chart centered at $(1, 1, 1)/\sqrt{3}$.

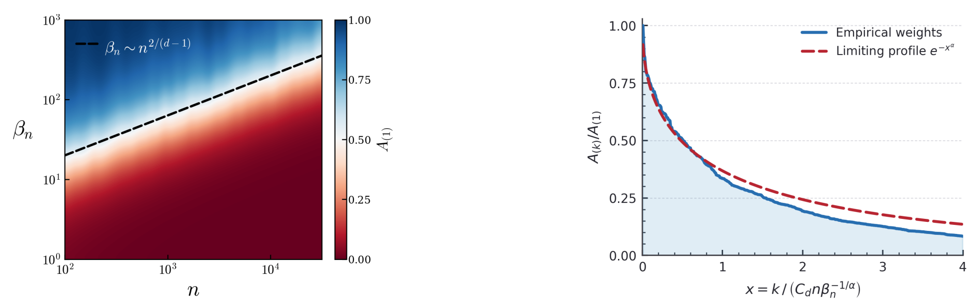

Figure 3. Ordered attention weights for uniform contexts on $\mathbb{S}^4$. Left: the heatmap shows the empirical mean of $A_{(1)}$ over 100 trials; the dashed curve $\beta_n = 2n^{\alpha}$ marks the critical scaling. Right: with $\beta_n = n^{\alpha/4}$, the rescaled ordered weights follow the subcritical prediction $e^{-x^{\alpha}}$, where $x = k/m_n(q)$.

6. Conclusion

We introduced a tractable probabilistic model to better understand attention scaling in the long-context regime. Within this framework we obtained a complete classification of the scaling limits for both the ordered attention weights and the resulting attention outputs. The central result is that the optimal rescaling takes the form $\beta_n \asymp n^{2/(d-1)}$, governed by the local extreme-value tail of the distance-to-query distribution $\mathbb{P}(T_i < t)$. To the best of our knowledge, this is the first scaling prescription for attention that exhibits an explicit dependence on the embedding dimension d, reflecting the intrinsic geometry of the key space. Below this threshold every attention weight vanishes; above it the mechanism collapses to a single-winner selector; only at the critical scale do the ordered weights converge to a non-degenerate Poisson point process.

Appendix A. Introduction to the Appendix, Notation, and Shared Lemmas

The appendices are organized as follows. Appendix B proves the supercritical single-winner regime. Appendix C proves the critical Poisson point process limit and the induced limiting ordered weights. Appendix D proves the subcritical results in the same order as the main text: scalar softmax estimates and ordered weights first, then the local moment estimates and central limit argument for the attention output, and finally the backward heat-type behavior of long-context attention. Appendix F treats the RoPE extension with correlated tokens.

We use the following notation throughout the appendices. The query $q \in \mathbb{S}^{d-1}$ is fixed, $x_1, x_2, \ldots$ are drawn according to the density ρ in Assumption 1. The D2Qs and their CDF are

$$T_i := 1 - \langle q, x_i \rangle \in [0,2], \qquad F(t) := \mathbb{P}(T_i \leqslant t).$$

We write

$$T_{(1)} \leqslant T_{(2)} \leqslant \cdots \leqslant T_{(n)}$$

for the order statistics of the D2Qs and

$$A_i = \frac{e^{-\beta_n T_i}}{Z_n}, \qquad Z_n := \sum_{j=1}^{n} e^{-\beta_n T_j}, \qquad A_{(1)} \geqslant A_{(2)} \geqslant \cdots \geqslant A_{(n)}$$

for the corresponding attention weights and their decreasing rearrangement. We also set

$$\alpha := \frac{2}{d-1}, \qquad \eta := \frac{1}{\alpha} - 1 = \frac{d-3}{2}, \qquad \mathbf{P}_q := \mathrm{Id} - qq^\top.$$

We use $\Theta(h)$ to denote a quantity comparable to h, i.e., there are constants C_1, C_2 depending on ρ as introduced in Assumption 1 such that $C_1 h \leqslant \Theta(h) \leqslant C_2 h$. Similarly, we use $x \approx y$ if there are two positive constants C_1, C_2, which depend at most on d and ρ in Assumption 1, such that $C_1 x \leqslant y \leqslant C_2 x$. The symbol $o_{\mathbb{P}}(1)$ denotes a random sequence converging to 0 in probability. More generally, $R_n = o_{\mathbb{P}}(a_n)$ means $R_n/a_n \to 0$ in probability, and $R_n = O_{\mathbb{P}}(a_n)$ means that R_n/a_n is tight.

The CDF $F(t)$ is of order $\Theta(t^{1/\alpha})$, because

$$F(t) := \mathbb{P}\left(T_i < t\right) = \mathbb{P}\left(\mathrm{dist}(q,x) < 2 \arcsin\left(t^{1/2} 2^{-1/2}\right)\right) = \Theta\left(t^{\frac{d-1}{2}}\right) = \Theta\left(t^{1/\alpha}\right).$$

Here, $\mathrm{dist}(q,x)$ is the geodesic distance between q and x on the unit sphere $\mathbb{S}^{d-1}$, and hence exactly equals the angle between q and x. In particular, we denote

$$C(q) := \lim_{t \to 0^+} F(t) t^{-1/\alpha}.$$

We summarize the asymptotics of $F(t)$ in the following two lemmas.

Lemma A.1 (CDF asymptotics)**.** *Under Assumption 1, the law of T_i is absolutely continuous on $[0,2]$, and has a density $f_T(t) := F'(t)$. Moreover, there are constants $0 < c_- < c_+ < \infty$ such that*

$$c_- t^{1/\alpha} \leqslant F(t) \leqslant c_+ t^{1/\alpha}, \quad c_- t^{\eta} \leqslant f_T(t) \leqslant c_+ t^{\eta} \qquad \forall t \in [0,2].$$

In particular, as $t \to 0^+$,

$$F(t) = C(q) t^{1/\alpha} + o\big(t^{1/\alpha}\big), \ \textit{with} \ C(q) = \frac{2^{1/\alpha} \sigma_{d-2}}{d-1} \rho(q),$$

where σ_{d-2} denotes the surface area of $\mathbb{S}^{d-2}$.

Lemma A.2 (Quantile asymptotics)**.** *As $u \to 0^+$,*

$$F^{-1}(u) = \left(\frac{u}{C(q)}\right)^{\alpha} (1 + o(1)).$$

The same relation holds with u replaced by any random sequence $U_n \to 0$ in probability:

$$F^{-1}(U_n) = \left(\frac{U_n}{C(q)}\right)^{\alpha} (1 + o_{\mathbb{P}}(1)).$$

We use the following form of Rényi's representation for uniform order statistics; see [Dev06, Chapter 5, Theorem 2.2].

Lemma A.3 (Rényi's Representation)**.** *Let $U_1, U_2, \dots, U_n$ be i.i.d.* Uniform$(0,1)$ *random variables, and let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ denote the associated order statistics.*

Let $E_1, E_2, \dots, E_{n+1}$ be i.i.d. exponential random variables with rate 1, i.e., $E_i \sim \mathrm{Exp}(1)$. Define the partial sums

$$\Gamma_k = \sum_{i=1}^{k} E_i, \quad k = 1, 2, \dots, n+1.$$

Then

$$\big(U_{(1)}, U_{(2)}, \dots, U_{(n)}\big) \overset{d}{=} \left(\frac{\Gamma_1}{\Gamma_{n+1}}, \frac{\Gamma_2}{\Gamma_{n+1}}, \dots, \frac{\Gamma_n}{\Gamma_{n+1}}\right).$$

We will repeatedly use the standard point-process formulation. Let Ω be a locally compact Polish space. We denote by $M_p(\Omega)$ the space of boundedly finite counting measures on Ω, equipped with the vague topology; see [DVJ08]. A point process on Ω can be viewed as an $M_p(\Omega)$-valued random element. Thus, when

$$\mathcal{N}_n \Rightarrow \mathcal{N} \qquad \text{in } M_p(\Omega),$$

we mean weak convergence of the laws of the random counting measures $\mathcal{N}_n$ to the law of $\mathcal{N}$ in this topology.

We will use the Laplace-functional characterization of convergence of point processes. Namely, to prove $\mathcal{N}_n \Rightarrow \mathcal{N}$ in $M_p(\Omega)$, it is enough to show that, for every $\varphi \in C_c^+(\Omega)$,

$$\mathbb{E}\left[\exp\left\{-\int_\Omega \varphi(y)\, \mathcal{N}_n(dy)\right\}\right] \longrightarrow \mathbb{E}\left[\exp\left\{-\int_\Omega \varphi(y)\, \mathcal{N}(dy)\right\}\right],$$

where $C_c^+(\Omega)$ denotes the set of nonnegative continuous functions on Ω with compact support; see, e.g., [DVJ08, Chapter 11].

Appendix B. Supercritical Regime: Proof of Theorem 3.1

We use the notation from Appendix A. In particular,

$$A_i = \frac{e^{-\beta_n T_i}}{Z_n}, \qquad Z_n = \sum_{j=1}^{n} e^{-\beta_n T_j}, \qquad T_{(1)} \leqslant \cdots \leqslant T_{(n)}.$$

By Theorem A.1,

$$F(t) = C(q) t^{1/\alpha} + o\big(t^{1/\alpha}\big), \qquad C(q) = \frac{2^{1/\alpha} \sigma_{d-2}}{d-1} \rho(q),$$

and the density $f(t) = F'(t)$ satisfies $f(t) = \Theta(t^\eta)$ near 0, where $\eta = 1/\alpha - 1$. The largest attention weight can be written as

$$A_{(1)} = \frac{e^{-\beta_n T_{(1)}}}{\sum_{k=1}^n e^{-\beta_n T_k}} = \frac{1}{1+S_n}, \qquad S_n := \sum_{k=2}^n e^{-\beta_n (T_{(k)} - T_{(1)})}.$$

We first prove the following lemma for S_n.

Lemma B.1. *For any $n \to \infty$ and any $\beta_n \to \infty$, we have that*

$$\mathbb{E}[S_n] \approx n^2 \left(O\left(e^{-\beta_n}\right) + O\left(e^{-C'n}\right) + \beta_n^{-1/\alpha} \mathbf{B}\left(1, n-1\right) + \beta_n^{-1} \mathbf{B}\left(2-\alpha, n-1\right) \right). \tag{B.1}$$

Here, C' is a positive constant depending on ρ in Assumption 1, and $\mathbf{B}(\cdot,\cdot)$ is the beta function.

We defer the proof of Theorem B.1 after proving Theorem 3.1.

Proof of Theorem 3.1. We show that if $\beta_n n^{-\alpha} \to \infty$, we have that

$$\lim_{n\to\infty} \mathbb{E}[S_n] = 0, \tag{B.2}$$

and thus $S_n \to 0$ in probability.

By Theorem B.1 and the classical asymptotics of the beta function [Olv97], we have that as $n \to \infty$,

$$\mathbf{B}\left(1, n-1\right) \approx n^{-1}, \text{ and } \mathbf{B}\left(2-\alpha, n-1\right) \approx n^{-2+\alpha}. \tag{B.3}$$

Hence, we have that

$$n^2 \beta_n^{-1/\alpha} \mathbf{B}\left(1, n-1\right) \approx \beta_n^{-1/\alpha} n, \text{ and } n^2 \beta_n^{-1} \mathbf{B}\left(2-\alpha, n-1\right) \approx \beta_n^{-1} n^\alpha, \tag{B.4}$$

both of which go to 0 if $\beta_n n^{-\alpha} \to \infty$. Because $O\left(e^{-\beta_n}\right) + O\left(e^{-C'n}\right)$ also goes to 0, we thus get that $\lim_{n\to\infty} \mathbb{E}[S_n] = 0$. □

Proof of Theorem B.1. Recall that $\eta = \frac{d-3}{2} = \frac{1-\alpha}{\alpha}$. Hence, by the arguments at the beginning of Section B, we see that $f(t) = \Theta(t^\eta)$.

Let $K_n := \{i \in [\![1, n]\!] \mid T_i = T_{(1)}\}$ and $R_n := \# K_n$. First, we see that

$$\sum_{i=1}^n \mathbf{1}_{i \in K_n} \sum_{j \neq i} e^{-\beta_n (T_j - T_i)} = R_n S_n \,. \tag{B.5}$$

Because T_i's are i.i.d. and have continuous density, we have that for any $i \neq j$, $\mathbb{P}\left(T_i = T_j\right) = 0$. Hence, $\mathbb{P}(R_n > 1) = 0$ and thus $R_n = 1$ almost surely. We have that

$$\mathbb{E}[S_n] = \mathbb{E}[R_n S_n] = \mathbb{E}\left[\sum_{i=1}^n \mathbf{1}_{i \in K_n} \sum_{j \neq i} e^{-\beta_n (T_j - T_i)} \right] = \sum_{i=1}^n \mathbb{E}\left[\mathbf{1}_{i \in K_n} \sum_{j \neq i} e^{-\beta_n (T_j - T_i)} \right]. \tag{B.6}$$

Because T_i's are i.i.d., we have that

$$\begin{aligned}\mathbb{E}[S_n] &= n\mathbb{E}\left[\mathbf{1}_{1\in K_n}\sum_{j\geqslant 2} e^{-\beta_n(T_j-T_1)}\right] = n\sum_{j\geqslant 2}\mathbb{E}\left[\mathbf{1}_{1\in K_n} e^{-\beta_n(T_j-T_1)}\right]\\ &= n(n-1)\mathbb{E}\left[\mathbf{1}_{1\in K_n} e^{-\beta_n(T_2-T_1)}\right]\\ &= n(n-1)\int_0^2 f(t)\left(\int_t^2 f(s)\mathrm{d}s\right)^{n-2}\left(\int_t^2 e^{-\beta_n(s-t)}f(s)\mathrm{d}s\right)\mathrm{d}t,\end{aligned} \tag{B.7}$$

where in the last equality, we used the fact that

$$1\in K_n \iff T_1\leqslant T_j \text{ for all } j\geqslant 2.$$

We first estimate the term $f(t)\left(\int_t^2 f(s)\mathrm{d}s\right)^{n-2}$ for $t\in[0,2]$. We know that $C_1 t^{\eta+1}\leqslant F(t)\leqslant C_2 t^{\eta+1}$ for some C_1, C_2 depending on ρ in Assumption 1. We pick a $\delta<1$ depending on ρ, d, such that when $t<\delta$, $C_2 t^{\eta+1}<1$. Because $\int_\delta^2 f(s)\mathrm{d}s = F(2)-F(\delta) = 1-F(\delta)<1$, we have that $\left(\int_\delta^2 f(s)\mathrm{d}s\right)^{n-2}\approx e^{-Cn}$ for some positive C depending on ρ in Assumption 1. Hence, we see that

$$\begin{aligned}f(t)\left(\int_t^2 f(s)\mathrm{d}s\right)^{n-2} &\approx e^{-Cn}+\mathbf{1}_{t<\delta}f(t)\left(\int_t^2 f(s)\mathrm{d}s\right)^{n-2}\\ = e^{-Cn}+\mathbf{1}_{t<\delta}f(t)\left(1-F(t)\right)^{n-2} &\approx e^{-Cn}+\mathbf{1}_{t<\delta}t^\eta\left(1-C_1t^{\eta+1}\right)^{n-2}.\end{aligned} \tag{B.8}$$

We next estimate the term $\left(\int_t^2 e^{-\beta_n(s-t)}f(s)\mathrm{d}s\right)$. By (B.8), we only need to discuss the case when $t\in[0,\delta]$ as we can use the trivial bound 1 when $t>\delta$. Because $f(s)=\Theta(s^\eta)$, we have that

$$\begin{aligned}&\int_t^2 e^{-\beta_n(s-t)}f(s)\mathrm{d}s \approx \int_t^2 e^{-\beta_n(s-t)}s^\eta\mathrm{d}s = \int_0^{2-t} e^{-\beta_n s}(s+t)^\eta\mathrm{d}s\\ &\approx \int_0^{2-t} e^{-\beta_n s}(s^\eta+t^\eta)\mathrm{d}s \approx O\left(e^{-\beta_n}\right)+\int_0^\infty e^{-\beta_n s}(s^\eta+t^\eta)\mathrm{d}s\\ &\approx O\left(e^{-\beta_n}\right)+\beta_n^{-\eta-1}+\beta_n^{-1}t^\eta.\end{aligned} \tag{B.9}$$

Hence, combining (B.8) and (B.9), we see that $\mathbb{E}[S_n]$ in (B.7) can be estimated as

$$\mathbb{E}[S_n]\approx n^2\left(e^{-Cn}+O\left(e^{-\beta_n}\right)+\int_0^\delta\left(\beta_n^{-\eta-1}+\beta_n^{-1}t^\eta\right)t^\eta\left(1-C_1t^{\eta+1}\right)^{n-2}\mathrm{d}t\right). \tag{B.10}$$

In the above integral, we do a change of variable for $t=sC_1^{-\frac{1}{\eta+1}}$ to s and by the choice of δ, the integral domain for s is $[0,\delta']$ for some $\delta'<1$. Hence,

$$\begin{aligned}\mathbb{E}[S_n] &\approx n^2\left(e^{-Cn}+O\left(e^{-\beta_n}\right)+\int_0^{\delta'}\left(\beta_n^{-\eta-1}+\beta_n^{-1}s^\eta\right)s^\eta\left(1-s^{\eta+1}\right)^{n-2}\mathrm{d}s\right)\\ &\approx n^2\left(e^{-Cn}+O\left(e^{-\beta_n}\right)+O\left(e^{-C'n}\right)+\int_0^1\left(\beta_n^{-\eta-1}+\beta_n^{-1}s^\eta\right)s^\eta\left(1-s^{\eta+1}\right)^{n-2}\mathrm{d}s\right).\end{aligned} \tag{B.11}$$

The above estimate follows because when $s \in [\delta', 1]$, the term $\left(1 - s^{\eta+1}\right)^{n-2}$ is an $O(e^{-C'n})$ term for some C' depending on ρ. Finally, we make a change of variable $s = r^{\frac{1}{\eta+1}}$, and we combine the two terms e^{-Cn} and $O(e^{-C'n})$, we then have that

(B.12)

$$\begin{aligned}\mathbb{E}[S_n] &\approx n^2 \left(O\left(e^{-\beta_n}\right) + O\left(e^{-C'n}\right) + \int_0^1 \left(\beta_n^{-\eta-1} + \beta_n^{-1} r^{\frac{\eta}{\eta+1}}\right) (1-r)^{n-2}\, \mathrm{d}r \right) \\ &= n^2 \left(O\left(e^{-\beta_n}\right) + O\left(e^{-C'n}\right) + \beta_n^{-\eta-1} \mathbf{B}\left(1, n-1\right) + \beta_n^{-1} \mathbf{B}\left(\frac{2\eta+1}{\eta+1}, n-1\right)\right).\end{aligned}$$

Because $\eta = \frac{1}{\alpha} - 1$, we finish the proof of Theorem B.1.

□

Appendix C. Critical Regime: Proof of Theorem 3.2

We prove Theorem 3.2 in this section. Let $U_i = F(T_i)$. Then U_i are i.i.d. uniform on $(0, 1)$, and $T_{(i)} = F^{-1}(U_{(i)})$. By Theorem A.3, we may write

$$(U_{(1)}, \ldots, U_{(n)}) \overset{d}{=} \left(\frac{\Gamma_1}{\Gamma_{n+1}}, \ldots, \frac{\Gamma_n}{\Gamma_{n+1}}\right),$$

where $\Gamma_i = E_1 + \cdots + E_i$ and E_i are i.i.d. exponential random variables of mean one. Hence

$$\beta_n T_{(i)} \overset{d}{=} \beta_n F^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right).$$

Recall from Theorem A.2, we have

$$F^{-1}(u) = \left(\frac{u}{C(q)}\right)^{\alpha} (1 + o(1)), \qquad u \downarrow 0.$$

Because $\Gamma_{n+1}/n \to 1$ almost surely and $\beta_n n^{-\alpha} \to \gamma$, it follows that for every fixed i,

$$\beta_n T_{(i)} \to Y_i := \gamma \left(\frac{\Gamma_i}{C(q)}\right)^{\alpha} \qquad \text{a.s.}$$

It remains to pass this convergence through the softmax denominator. By Theorem A.2, there is a $c > 0$ such that $F^{-1}(u) \geqslant cu^{\alpha}, \forall u \in [0, 1]$. On the event $\Gamma_{n+1}/n \to 1$, for all large n,

$$\beta_n F^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right) \geqslant c'\Gamma_i^{\alpha},$$

for some c' independent of n. We obtain the domination

$$\exp\left\{-\beta_n F^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\right\} \leqslant e^{-c'\Gamma_i^{\alpha}}, \qquad 1 \leqslant i \leqslant n,$$

for all sufficiently large n.

We now check that this dominating sequence is summable almost surely. By the strong law of large numbers,

$$\frac{\Gamma_i}{i} = \frac{E_1 + \cdots + E_i}{i} \longrightarrow 1 \qquad \text{a.s.}$$

Therefore, almost surely, there exists i_0 such that $\Gamma_i \geqslant \frac{i}{2}$, for all $i \geqslant i_0$. Consequently, for all $i \geqslant i_0$, $e^{-c'\Gamma_i^\alpha} \leqslant e^{-c'2^{-\alpha}i^\alpha}$. Hence

$$\sum_{i\geqslant 1} e^{-c'\Gamma_i^\alpha} < \infty \qquad \text{a.s.}$$

We may therefore apply dominated convergence for series, and get

$$\sum_{i=1}^{n} \exp\left\{-\beta_n F^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right)\right\} \longrightarrow \sum_{i=1}^{\infty} e^{-Y_i} \qquad \text{a.s.}$$

The limiting denominator is finite by the previous summability argument and is strictly positive because $e^{-Y_1} > 0$. Therefore, for every fixed k,

$$(A_{(1)}, \dots, A_{(k)}) \Rightarrow \left(\frac{e^{-Y_1}}{\sum_{j\geqslant 1} e^{-Y_j}}, \dots, \frac{e^{-Y_k}}{\sum_{j\geqslant 1} e^{-Y_j}}\right).$$

Finally, we identify the limit with the Poisson point process: the variables $Y_i = \gamma(\Gamma_i/C(q))^\alpha$ are the increasing atoms of the Poisson point process on $\mathbb{R}_+$ with intensity measure $\Lambda([0,y]) = C(q)\,(y/\gamma)^{1/\alpha}$. This proves Theorem 3.2.

Appendix D. Subcritical Regime: Proof of Theorem 3.3

This appendix proves the attention weight results in the subcritical regime.

D.1. Subcritical partition function.

Recall from Appendix A. We use

$$Z_n = \sum_{i=1}^{n} e^{-\beta_n T_i}, \qquad A_i = \frac{e^{-\beta_n T_i}}{Z_n}.$$

The following lemma characterizes the limit of Z_n in the subcritical regime.

Lemma D.1. *As $\beta \to \infty$,*

$$\mathbb{E}[e^{-\beta T_1}] = C(q)\Gamma\left(\frac{1}{\alpha}+1\right)\beta^{-1/\alpha}(1+o(1)).$$

Proof. Let $F(t) = \mathbb{P}(T_1 \leqslant t)$. Integration by parts gives

$$\mathbb{E}[e^{-\beta T_1}] = e^{-2\beta} + \beta\int_0^2 e^{-\beta t}F(t)\,dt.$$

Theorem A.1 gives

$$F(t) = C(q)t^{1/\alpha}(1+o(1)), \qquad t \to 0^+.$$

After the change of variables $u = \beta t$, dominated convergence yields

$$\beta\int_0^2 e^{-\beta t}F(t)\,dt = C(q)\beta^{-1/\alpha}\int_0^\infty e^{-u}u^{1/\alpha}\,du\,(1+o(1)).$$

The integral is $\Gamma(1/\alpha+1)$, which proves the claim. $\square$

Corollary D.2 (Subcritical partition function)**.** *If $\beta_n \to \infty$ and $\beta_n n^{-\alpha} \to 0$, then*

$$\frac{Z_n}{\mathbb{E}[Z_n]} \xrightarrow{\mathbb{P}} 1, \qquad \frac{Z_n}{C(q)n\beta_n^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)} \xrightarrow{\mathbb{P}} 1.$$

Proof. By Lemma D.1,

$$\mathbb{E}[Z_n] = C(q)n\beta_n^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)(1+o(1)), \qquad \mathrm{Var}(Z_n) \leqslant n\mathbb{E}[e^{-2\beta_n T_1}] = O(n\beta_n^{-1/\alpha}).$$

Therefore

$$\frac{\mathrm{Var}(Z_n)}{\mathbb{E}[Z_n]^2} = O\Big((n\beta_n^{-1/\alpha})^{-1}\Big) \to 0,$$

because $\beta_n n^{-\alpha} \to 0$. Chebyshev's inequality gives $Z_n/\mathbb{E}[Z_n] \to 1$ in probability, and the displayed deterministic asymptotic for $\mathbb{E}[Z_n]$ gives the second claim. □

D.2. **Proof of Theorem 3.3.** The ordered-weight proof uses the quantile asymptotics from Theorem A.2 to handle the denominator Z_n in $A_{(i)}$. To handle the numerator $e^{-\beta_n T_{(i)}}$ in $A_{(i)}$, we prove in the following that the best D2Q is so small that $\beta_n T_{(1)}$ vanishes, and the k_n-th D2Q is governed by the quantile level $k_n/(n\beta_n^{-1/\alpha})$.

Lemma D.3. *If $\beta_n \to \infty$ and $\beta_n n^{-\alpha} \to 0$, then*

$$\beta_n T_{(1)} \xrightarrow{\mathbb{P}} 0.$$

Proof. For every $\varepsilon > 0$,

$$\mathbb{P}(\beta_n T_{(1)} > \varepsilon) = \left(1 - F\left(\frac{\varepsilon}{\beta_n}\right)\right)^n \leqslant \exp\left(-nF\left(\frac{\varepsilon}{\beta_n}\right)\right).$$

By Theorem A.1,

$$nF\left(\frac{\varepsilon}{\beta_n}\right) = C(q)\varepsilon^{1/\alpha}n\beta_n^{-1/\alpha}(1+o(1)) \to \infty,$$

so the probability tends to zero. □

Lemma D.4. *Assume $\beta_n \to \infty$, $\beta_n n^{-\alpha} \to 0$, and*

$$\frac{k_n}{C(q)n\beta_n^{-1/\alpha}} \to x \in [0,\infty).$$

Then

$$\beta_n T_{(k_n)} \xrightarrow{\mathbb{P}} x^\alpha.$$

Proof. First suppose $x > 0$. Let $U_{(k_n)}$ be the k_n-th order statistic of an i.i.d. uniform sample on $(0,1)$. Then

$$T_{(k_n)} \stackrel{d}{=} F^{-1}(U_{(k_n)}).$$

Since $k_n \to \infty$ and $k_n/n \to 0$, the beta distribution formulas for uniform order statistics give

$$\mathbb{E}[U_{(k_n)}] = \frac{k_n}{n+1}, \qquad \mathrm{Var}(U_{(k_n)}) = \frac{k_n(n-k_n+1)}{(n+1)^2(n+2)}.$$

Therefore,

$$\lim_{n\to\infty} \frac{\mathrm{Var}(U_{(k_n)})}{\mathbb{E}[U_{(k_n)}]^2} = 0.$$

Hence $(n/k_n)U_{(k_n)} \to 1$ in probability by Chebyshev's inequality. Applying Lemma A.2,

$$\begin{aligned}\beta_n T_{(k_n)} &\overset{d}{=} \beta_n F^{-1}\left(\frac{k_n}{n}\frac{nU_{(k_n)}}{k_n}\right)\\ &= \beta_n\left(\frac{k_n}{C(q)n}\frac{nU_{(k_n)}}{k_n}\right)^{\alpha}(1+o_{\mathbb{P}}(1)) \xrightarrow{\mathbb{P}} x^{\alpha}.\end{aligned}$$

A priori the convergence is in distribution, but since the limit is deterministic then $\beta_n T_{(k_n)}$ converges also in probability. It remains to treat $x = 0$. Fix $\varepsilon > 0$ and set

$$M_n(\varepsilon) := \#\left\{i : T_i \leqslant \frac{\varepsilon}{\beta_n}\right\}.$$

Then $M_n(\varepsilon)$ is binomial with mean

$$\mu_n = nF\left(\frac{\varepsilon}{\beta_n}\right) = C(q)\varepsilon^{1/\alpha} n\beta_n^{-1/\alpha}(1+o(1)) \to \infty.$$

The assumption $k_n/(C(q)n\beta_n^{-1/\alpha}) \to 0$ gives $k_n/\mu_n \to 0$. Hence, for all large n, $k_n \leqslant \mu_n/2$, and Chebyshev's inequality yields

$$\mathbb{P}(\beta_n T_{(k_n)} > \varepsilon) = \mathbb{P}(M_n(\varepsilon) < k_n) \leqslant \mathbb{P}(|M_n(\varepsilon)-\mu_n| > \mu_n/2) \leqslant \frac{4\mathrm{Var}(M_n(\varepsilon))}{\mu_n^2} \leqslant \frac{4}{\mu_n} \to 0,$$

where we used the fact that $\mathrm{Var}(M_n(\varepsilon)) \leqslant \mu_n$ because $M_n(\varepsilon)$ is binomial. Since $\varepsilon > 0$ was arbitrary, $\beta_n T_{(k_n)} \to 0$ in probability. □

Proof of Theorem 3.3. First, $A_{(i)} \leqslant A_{(1)}$. To prove $A_{(i)} \to 0$ for fixed i, it suffices to consider subsequences with either $\beta_n \leqslant M < \infty$ or $\beta_n \to \infty$. If $\beta_n \leqslant M$, then

$$A_{(1)} \leqslant \frac{1}{\sum_{j=1}^{n} e^{-\beta_n T_j}} \leqslant \frac{e^{2M}}{n} \to 0,$$

since $T_j \in [0, 2]$. If $\beta_n \to \infty$, then Theorem D.2 implies $Z_n \to \infty$ in probability, and therefore $A_{(1)} \leqslant 1/Z_n \to 0$ in probability.

Now assume $\beta_n \to \infty$ and

$$\frac{\beta_n^{1/\alpha} k_n}{C(q)n} \to x, \qquad x \geqslant 0.$$

By Lemma D.3, and Lemma D.4,

$$\beta_n T_{(1)} \xrightarrow{\mathbb{P}} 0, \qquad \beta_n T_{(k_n)} \xrightarrow{\mathbb{P}} x^{\alpha}.$$

Thus

$$\frac{A_{(k_n)}}{A_{(1)}} = \exp\bigl(-\beta_n T_{(k_n)} + \beta_n T_{(1)}\bigr) \xrightarrow{\mathbb{P}} \exp(-x^{\alpha}).$$

For the absolute normalization,

$$C(q)n\beta_n^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right) A_{(k_n)} = \frac{C(q)n\beta_n^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right)}{Z_n} e^{-\beta_n T_{(k_n)}}.$$

The first factor converges to 1 in probability by Corollary D.2; the second factor converges to $\exp\{-x^{\alpha}\}$ by Lemma D.4. Slutsky's theorem gives

$$C(q)n\beta_n^{-1/\alpha}\Gamma\left(\frac{1}{\alpha}+1\right) A_{(k_n)} \xrightarrow{\mathbb{P}} \exp(-x^{\alpha}),$$

which is equivalent to the stated formula.

Finally, for (3.6), by assumption, we have that $m_n \to \infty$ and $\lim_{n\to\infty} \frac{k_n}{m_n(q)} = x \geqslant 0$. We can rewrite the sum in (3.6) as:

$$\sum_{i=1}^{k_n} A_{(i)} = \frac{m_n}{Z_n} \int_0^{k_n/m_n} e^{-\beta_n T_{(\lceil s m_n \rceil)}} \, \mathrm{d}s.$$

We now identify the limit of the integral. For each fixed $s \in (0, x]$, $\frac{\lceil s m_n \rceil}{m_n} \to s$. Therefore, by Theorem D.4, $\beta_n T_{(\lceil s m_n \rceil)} \xrightarrow{\mathbb{P}} s^\alpha$, and hence $e^{-\beta_n T_{(\lceil s m_n \rceil)}} \xrightarrow{\mathbb{P}} e^{-s^\alpha}$. Since the random variables are bounded by 1, this convergence also holds in L^1. Thus, by Fubini and dominated convergence,

$$\mathbb{E} \int_0^x \left| e^{-\beta_n T_{(\lceil s m_n \rceil)}} - e^{-s^\alpha} \right| \mathrm{d}s = \int_0^x \mathbb{E} \left| e^{-\beta_n T_{(\lceil s m_n \rceil)}} - e^{-s^\alpha} \right| \mathrm{d}s \longrightarrow 0.$$

Consequently,

$$\int_0^x e^{-\beta_n T_{(\lceil s m_n \rceil)}} \, \mathrm{d}s \xrightarrow{L^1} \int_0^x e^{-s^\alpha} \, \mathrm{d}s,$$

and hence also in probability. Moreover,

$$\left| \int_0^{k_n/m_n} e^{-\beta_n T_{(\lceil s m_n \rceil)}} \, \mathrm{d}s - \int_0^x e^{-\beta_n T_{(\lceil s m_n \rceil)}} \, \mathrm{d}s \right| \leqslant \left| \frac{k_n}{m_n} - x \right| \to 0.$$

Therefore

$$\int_0^{k_n/m_n} e^{-\beta_n T_{(\lceil s m_n \rceil)}} \, \mathrm{d}s \xrightarrow{\mathbb{P}} \int_0^x e^{-s^\alpha} \, \mathrm{d}s.$$

Furthermore, by Theorem D.2, $\frac{m_n}{Z_n} \xrightarrow{\mathbb{P}} \frac{1}{\Gamma\left(\frac{1}{\alpha}+1\right)}$, and using the Continuous Mapping Theorem we get:

$$\sum_{i=1}^{k_n} A_{(i)} \xrightarrow{\mathbb{P}} \frac{1}{\Gamma\left(\frac{1}{\alpha}+1\right)} \int_0^x e^{-s^\alpha} \, \mathrm{d}s = \frac{1}{\Gamma\left(\frac{1}{\alpha}+1\right)} \frac{1}{\alpha} \int_0^{x^\alpha} e^{-u} u^{1/\alpha - 1} \, \mathrm{d}u = \frac{\gamma_{\mathrm{inc}}\left(\frac{1}{\alpha}, x^\alpha\right)}{\Gamma\left(\frac{1}{\alpha}\right)}.$$

□

Appendix E. Attention Outputs in the Supercritical, Critical, and Subcritical Regimes

This appendix proves the attention-output limits from Section 4. The proofs of the supercritical and critical regimes first identify the relevant local point process near the query, then apply the softmax functional to that process. The subcritical regime is different and is treated later by local moment estimates and a triangular-array central limit theorem.

E.1. **Convergence of a marked point process.** We recall from Section 4,

$$a_n = (C(q)n)^\alpha, \qquad \mathbf{P}_q = \mathrm{Id} - qq^\top,$$

and, whenever $\mathbf{P}_q x_i \neq 0$, define the tangent direction

$$\Theta_i := \frac{\mathbf{P}_q x_i}{\|\mathbf{P}_q x_i\|} \in \mathbb{S}_q^{d-2}, \qquad \mathbb{S}_q^{d-2} := \{\theta \in T_q \mathbb{S}^{d-1} : \|\theta\| = 1\}.$$

On the null event $\mathbf{P}_q x_i = 0$, choose Θ_i arbitrarily. Let ν_q denote the uniform probability measure on $\mathbb{S}_q^{d-2}$.

We use the following marked version of the point-process limit. It is the only local extreme-value input needed in Theorem 4.1 and Theorem 4.2.

Lemma E.1 (Local marked point process)**.** *Let $b_n > 0$ satisfy $b_n n^{-\alpha} \to \lambda \in (0, \infty)$, and define*

$$\mathcal{M}_n^{(b)} := \sum_{i=1}^{n} \delta_{(b_n T_i, \Theta_i)} \qquad \text{on } [0,\infty) \times \mathbb{S}_q^{d-2}.$$

Then

$$\mathcal{M}_n^{(b)} \Rightarrow \mathcal{M}_\lambda,$$

where $\mathcal{M}_\lambda$ is a Poisson point process with intensity measure

$$\Lambda_\lambda([0,y] \times B) = C(q) \left(\frac{y}{\lambda}\right)^{1/\alpha} \nu_q(B)$$

for every $y \geqslant 0$ and every ν_q-continuity set $B \subset \mathbb{S}_q^{d-2}$.

Proof. Let

$$\mu_n(A) := n\,\mathbb{P}\big((b_n T_1, \Theta_1) \in A\big)$$

be the one-particle mean measure. We first show $\mu_n \to \Lambda_\lambda$ vaguely. Let $g \in C_c([0,\infty) \times \mathbb{S}_q^{d-2})$, with support contained in $[0,R] \times \mathbb{S}_q^{d-2}$. In geodesic polar coordinates around q,

$$x = (\cos r)q + (\sin r)\theta, \qquad T(x) = 1 - \cos r, \qquad \mathrm{d}\sigma(x) = (\sin r)^{d-2}\mathrm{d}r\,\mathrm{d}\sigma_q(\theta),$$

where $\mathrm{d}\sigma_q$ is surface measure on $\mathbb{S}_q^{d-2}$. With the change of variables $y = b_n(1-\cos r)$,

$$(\sin r)^{d-2}\mathrm{d}r = b_n^{-1/\alpha}\left(2y - \frac{y^2}{b_n}\right)^{1/\alpha-1}\mathrm{d}y.$$

Therefore

$$\int g\,\mathrm{d}\mu_n = n b_n^{-1/\alpha} \int_{\mathbb{S}_q^{d-2}} \int_0^R g(y,\theta)\rho\left(\left(1 - \frac{y}{b_n}\right)q + b_n^{-1/2}\left(2y - \frac{y^2}{b_n}\right)^{1/2}\theta\right) \times \left(2y - \frac{y^2}{b_n}\right)^{1/\alpha-1}\mathrm{d}y\,\mathrm{d}\sigma_q(\theta).$$

Since $n b_n^{-1/\alpha} \to \lambda^{-1/\alpha}$, ρ is continuous at q, and $y^{1/\alpha-1}$ is integrable at zero, dominated convergence gives

$$\int g\,\mathrm{d}\mu_n \longrightarrow \frac{C(q)}{\alpha\lambda^{1/\alpha}} \int_{\mathbb{S}_q^{d-2}} \int_0^R g(y,\theta) y^{1/\alpha-1}\mathrm{d}y\,\nu_q(\mathrm{d}\theta).$$

Here we used

$$C(q) = \frac{2^{1/\alpha}\sigma_{d-2}}{d-1}\rho(q), \qquad \alpha = \frac{2}{d-1}, \qquad \mathrm{d}\sigma_q = \sigma_{d-2}\,\mathrm{d}\nu_q.$$

Thus $\mu_n \to \Lambda_\lambda$ vaguely.

Now take $\varphi \in C_c([0,\infty) \times \mathbb{S}_q^{d-2})$, $\varphi \geqslant 0$. Since the marked pairs $(b_n T_i, \Theta_i)$ are i.i.d.,

$$\begin{aligned}\mathbb{E}\exp\left\{-\int \varphi\,\mathrm{d}\mathcal{M}_n^{(b)}\right\} &= \left(\mathbb{E}e^{-\varphi(b_n T_1, \Theta_1)}\right)^n \\ &= \left(1 - \frac{1}{n}\int (1 - e^{-\varphi})\,\mathrm{d}\mu_n\right)^n.\end{aligned}$$

By vague convergence of μ_n, the integral converges to $\int(1-e^{-\varphi})\,\mathrm{d}\Lambda_\lambda$. Hence the Laplace functionals converge to those of a Poisson point process with intensity Λ_λ. $\square$

E.2. **Supercritical output: proof of Theorem 4.1.** In the supercritical regime, the softmax output is asymptotically the nearest token itself. The only extra point beyond Theorem 3.1 is that this replacement is valid at the finer scale $a_n^{-1/2}$, where we recall the definition $a_n = (C(q)n)^\alpha$.

Let

$$I_n := \arg\min_{1\leqslant i\leqslant n} T_i, \qquad T_n^\star := T_{I_n} = T_{(1)}, \qquad x_n^\star := x_{I_n}, \qquad \Theta_n^\star := \Theta_{I_n}.$$

The minimizer is unique almost surely because the law of T_i is continuous.

Lemma E.2 (Nearest marked token)**.** *Under Assumption 1,*

$$(a_n T_n^\star, \Theta_n^\star) \Rightarrow (R_1, \Theta_1),$$

where

$$\mathbb{P}(R_1 > r) = e^{-r^{1/\alpha}}, \qquad r \geqslant 0,$$

and $\Theta_1 \sim \nu_q$ is independent of R_1.

Proof. Apply Theorem E.1 with $b_n = a_n$. Since $a_n n^{-\alpha} = C(q)^\alpha$, the limiting intensity is

$$\Lambda_\star(\mathrm{d}y, \mathrm{d}\theta) = \frac{1}{\alpha} y^{1/\alpha - 1} \mathrm{d}y\, \nu_q(\mathrm{d}\theta).$$

The first atom map is continuous at the limiting process almost surely, because the process is simple and its y-marginal is diffuse. Thus the first marked atom converges. Its radial tail is

$$\mathbb{P}(R_1 > r) = \exp\{-\Lambda_\star([0,r]\times \mathbb{S}_q^{d-2})\} = e^{-r^{1/\alpha}},$$

and the product form of $\Lambda_\star$ gives independence of R_1 and Θ_1. $\square$

Lemma E.3 (Nearest-token reduction)**.** *If $\beta_n n^{-\alpha} \to \infty$, then*

$$\sqrt{a_n}\left\| \sum_{i=1}^n A_i x_i - x_n^\star \right\| \xrightarrow{\mathbb{P}} 0.$$

Proof. Set $\lambda_n := \beta_n / a_n$, so $\lambda_n \to \infty$. Since the denominator of the softmax is at least $e^{-\beta_n T_n^\star}$,

$$\sqrt{a_n}\left\| \sum_{i=1}^n A_i x_i - x_n^\star \right\| \leqslant D_n := \sum_{i\neq I_n} e^{-\beta_n (T_i - T_n^\star)} \sqrt{a_n} \|x_i - x_n^\star\|.$$

Fix $M > 0$ and write $E_{n,M} = \{a_n T_n^\star \leqslant M\}$. By Theorem E.2,

$$\lim_{n\to\infty} \mathbb{P}(E_{n,M}^c) = e^{-M^{1/\alpha}}.$$

We show that $\mathbb{E}[D_n \mathbf{1}_{E_{n,M}}] \to 0$. Conditionally on $T_n^\star = t$, the remaining D2Qs have law $\mathrm{d}F(s)/(1-F(t))$ on $[t,2]$. Also

$$\|x_i - x_n^\star\| \leqslant \|x_i - q\| + \|x_n^\star - q\| = \sqrt{2T_i} + \sqrt{2t}.$$

For $t \leqslant M/a_n$, and for all large n, $1 - F(t) \geqslant 1/2$. Using the local density bound $F'(s) \leqslant Cs^{1/\alpha-1}$ in Theorem A.1, $\mathbb{E}[D_n \mathbf{1}_{E_{n,M}}]$ is bounded by

$$Cn\sqrt{a_n} \int_t^2 e^{-\beta_n(s-t)}(\sqrt{s} + \sqrt{t})s^{1/\alpha-1}\,\mathrm{d}s.$$

Put $t = y/a_n$, $s = t + v/a_n$. Since $0 \leqslant y \leqslant M$ and $na_n^{-1/\alpha} = C(q)^{-1}$, the last display is at most a constant times

$$\sup_{0\leqslant y\leqslant M} \int_0^\infty e^{-\lambda_n v}\big(\sqrt{y+v} + \sqrt{y}\big)(y+v)^{1/\alpha-1}\,\mathrm{d}v.$$

This supremum tends to zero as $\lambda_n \to \infty$: for $d = 2$ the integrand is bounded by a constant, and for $d \geqslant 3$ it is bounded on $0 \leqslant y \leqslant M$ by $C_M(1 + v^{1/\alpha-1/2})$. Hence $\mathbb{E}[D_n \mathbf{1}_{E_{n,M}}] \to 0$.

By Markov's inequality,

$$\limsup_{n\to\infty} \mathbb{P}(D_n > \varepsilon) \leqslant e^{-M^{1/\alpha}},$$

for every $\varepsilon > 0$. Letting $M \to \infty$ proves $D_n \to 0$ in probability. $\square$

Proof of Theorem 4.1. Write

$$x_n^\star = (\cos r_n^\star)q + (\sin r_n^\star)\Theta_n^\star, \qquad T_n^\star = 1 - \cos r_n^\star.$$

Then

$$x_n^\star - q = -T_n^\star q + \sqrt{2T_n^\star - (T_n^\star)^2}\,\Theta_n^\star.$$

By Theorem E.2, $a_n T_n^\star$ is tight, and therefore

$$\sqrt{a_n}(x_n^\star - q) = \sqrt{2a_n T_n^\star}\,\Theta_n^\star + o_{\mathbb{P}}(1) \Rightarrow \sqrt{2R_1}\,\Theta_1.$$

By Theorem E.3,

$$\sqrt{a_n}\sum_{i=1}^n A_i(x_i - q) = \sqrt{a_n}(x_n^\star - q) + o_{\mathbb{P}}(1).$$

Applying the fixed linear map V gives

$$\sqrt{a_n}\,\mathcal{Y}_n(q) \Rightarrow V(\sqrt{2R_1}\Theta_1) = V\Phi.$$

Since $a_n \to \infty$, this also implies $\mathcal{Y}_n(q) \to 0$ in probability and hence $\mathsf{ATT}(n) \to Vq$ in probability. $\square$

E.3. **Critical output: proof of Theorem 4.2.** In the critical regime, the output is no longer determined by a single nearest token. Instead, the finitely many near-query competitors surviving in the Poisson limit all contribute through the same softmax functional as in Appendix C.

Assume $\beta_n n^{-\alpha} \to \gamma \in (0, \infty)$. Let

$$\mathcal{M}_n := \sum_{i=1}^n \delta_{(\beta_n T_i, \Theta_i)}.$$

By Theorem E.1 with $b_n = \beta_n$,

$$\mathcal{M}_n \Rightarrow \mathcal{M}_\gamma, \qquad \Lambda_\gamma(\mathrm{d}y, \mathrm{d}\theta) = \frac{C(q)}{\alpha\gamma^{1/\alpha}} y^{1/\alpha-1}\,\mathrm{d}y\,\nu_q(\mathrm{d}\theta).$$

Define

$$H_n := \int_0^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}\sqrt{2y}\,\theta\,\mathcal{M}_n(\mathrm{d}y,\mathrm{d}\theta), \qquad Z_n := \int_0^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}\,\mathcal{M}_n(\mathrm{d}y,\mathrm{d}\theta),$$

and define H and Z by the same formulas with $\mathcal{M}_\gamma$ in place of $\mathcal{M}_n$.

Lemma E.4 (Critical softmax functional)**.** *We have*

$$(H_n, Z_n) \Rightarrow (H, Z), \qquad 0 < Z < \infty \quad \textit{almost surely.}$$

Consequently,

$$\frac{H_n}{Z_n} \Rightarrow \frac{H}{Z}.$$

Proof. For fixed $R < \infty$, the truncated maps

$$\mu \mapsto \int_0^R \int_{\mathbb{S}_q^{d-2}} e^{-y}\sqrt{2y}\,\theta\,\mu(\mathrm{d}y,\mathrm{d}\theta), \qquad \mu \mapsto \int_0^R \int_{\mathbb{S}_q^{d-2}} e^{-y}\,\mu(\mathrm{d}y,\mathrm{d}\theta)$$

are continuous at every point measure with no atom on $\{R\} \times \mathbb{S}_q^{d-2}$. Since the limiting intensity has diffuse y-marginal, the continuous mapping theorem gives convergence of the truncated pair.

It remains to remove the truncation. The required tail bound is

$$\lim_{R\to\infty} \limsup_{n\to\infty} \mathbb{E} \int_R^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}(1+\sqrt{y})\,\mathcal{M}_n(\mathrm{d}y,\mathrm{d}\theta) = 0, \tag{E.1}$$

and the same bound holds with $\mathcal{M}_\gamma$ in place of $\mathcal{M}_n$. Indeed,

$$\begin{aligned}\mathbb{E} \int_R^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}(1+\sqrt{y})\,\mathcal{M}_n(\mathrm{d}y,\mathrm{d}\theta) &= n\mathbb{E}\big[e^{-\beta_n T_1}(1+\sqrt{\beta_n T_1})\mathbf{1}_{\{\beta_n T_1 > R\}}\big] \\ &\leqslant \sum_{\ell=\lfloor R\rfloor}^\infty e^{-\ell}(1+\sqrt{\ell+1})n\mathbb{P}(\beta_n T_1 \leqslant \ell+1).\end{aligned}$$

For $\ell + 1 \leqslant t_0\beta_n$, the small-cap bound gives

$$n\mathbb{P}(\beta_n T_1 \leqslant \ell+1) \leqslant C\gamma^{-1/\alpha}(\ell+1)^{1/\alpha},$$

while the remaining range $\ell + 1 > t_0\beta_n$ is exponentially small. Hence the limsup is bounded by the tail of a convergent exponential series. The limiting version follows directly from Campbell's theorem:

$$\mathbb{E} \int_R^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}(1+\sqrt{y})\,\mathcal{M}_\gamma(\mathrm{d}y,\mathrm{d}\theta) = \frac{C(q)}{\alpha\gamma^{1/\alpha}} \int_R^\infty e^{-y}(1+\sqrt{y})y^{1/\alpha-1}\,\mathrm{d}y \to 0.$$

The same $\varepsilon/3$ truncation argument used in Appendix C now yields $(H_n, Z_n) \Rightarrow (H, Z)$.

Finally, $Z < \infty$ almost surely by Campbell's theorem. Also, for every $R > 0$,

$$\mathbb{P}(Z = 0) \leqslant \mathbb{P}\big(\mathcal{M}_\gamma([0,R] \times \mathbb{S}_q^{d-2}) = 0\big) = \exp\left\{-C(q)\left(\frac{R}{\gamma}\right)^{1/\alpha}\right\},$$

which tends to zero as $R \to \infty$. Thus $Z > 0$ almost surely, and the ratio convergence follows from the continuous mapping theorem. $\square$

Proof of Theorem 4.2. The ideal tangent numerator is H_n, but the true centered numerator is

$$\widehat{H}_n := \sum_{i=1}^n e^{-\beta_n T_i}\sqrt{\beta_n}(x_i - q), \qquad \sqrt{\beta_n}\sum_{i=1}^n A_i(x_i - q) = \frac{\widehat{H}_n}{Z_n}.$$

We first show that

$$\widehat{H}_n - H_n \xrightarrow{\mathbb{P}} 0. \tag{E.2}$$

For an atom with $y_i = \beta_n T_i$, write

$$x_i = \left(1 - \frac{y_i}{\beta_n}\right) q + \left(\frac{2y_i}{\beta_n} - \frac{y_i^2}{\beta_n^2}\right)^{1/2} \Theta_i.$$

Therefore, uniformly over $0 \leqslant y_i \leqslant R$,

$$\left\| \sqrt{\beta_n}(x_i - q) - \sqrt{2y_i}\,\Theta_i \right\| = o_{\mathbb{P}}(1).$$

The contribution from $\{y_i \leqslant R\}$ is thus $o_{\mathbb{P}}(1)$, because $\mathcal{M}_n([0,R]\times \mathbb{S}_q^{d-2})$ is tight. On the tail $\{y_i > R\}$,

$$\left\| \sum_{y_i > R} e^{-y_i}\big(\sqrt{\beta_n}(x_i - q) - \sqrt{2y_i}\Theta_i\big) \right\| \leqslant 2\sqrt{2}\int_R^\infty \int_{\mathbb{S}_q^{d-2}} e^{-y}\sqrt{y}\,\mathcal{M}_n(\mathrm{d}y,\mathrm{d}\theta),$$

and the expectation of the right-hand side vanishes after $n \to \infty$ and then $R \to \infty$ by (E.1). This proves (E.2).

By Theorem E.4,

$$\sqrt{\beta_n}\sum_{i=1}^n A_i(x_i - q) = \frac{\widehat{H}_n}{Z_n} \Rightarrow \frac{H}{Z}.$$

Applying V gives

$$\sqrt{\beta_n}\,\mathcal{Y}_n(q) \Rightarrow V\frac{\displaystyle\int e^{-y}\sqrt{2y}\,\theta\,\mathcal{M}_\gamma(\mathrm{d}y,\mathrm{d}\theta)}{\displaystyle\int e^{-y}\,\mathcal{M}_\gamma(\mathrm{d}y,\mathrm{d}\theta)} = V\Xi.$$

Since $\beta_n \to \infty$, this also implies $\mathcal{Y}_n(q) \to 0$ in probability and hence $\mathsf{ATT}(n) \to Vq$ in probability. $\square$

E.4. **Local moment estimates for the attention output.** The attention output theorem is different from the ordered-weight theorem, since it requires the vector numerator, not just the scalar denominator. Write

$$B_n(q) := \sum_{i=1}^n (x_i - q)e^{-\beta_n T_i}, \qquad Z_n = \sum_{i=1}^n e^{-\beta_n T_i}.$$

Then, when $V = \mathrm{Id}$

$$\mathcal{Y}_n(q) = \frac{B_n(q)}{Z_n}$$

The following local estimates identify the deterministic bias and covariance of $B_n(q)$. In this subsection, $\mathbf{P}_q := \mathrm{Id} - qq^\top$ denotes the orthogonal projection onto $T_q\mathbb{S}^{d-1}$, and $\nabla_{\mathbb{S}^{d-1}}$ denotes the spherical gradient. We also recall the two deterministic quantities used in Theorem 4.3:

$$\mathcal{Y}(q) = \nabla_{\mathbb{S}^{d-1}} \log\rho(q) - \frac{d-1}{2}q, \qquad \Sigma(q) = \frac{c_d}{\rho(q)}\mathbf{P}_q.$$

The proof below uses the additional C^2-regularity of ρ near q assumed in Theorem 4.3.

Lemma E.5 (Local output moments)**.** *As $\beta \to \infty$,*

$$\mathbb{E}[(x_1 - q)e^{-\beta T_1}] = C(q)\Gamma\left(\frac{1}{\alpha} + 1\right)\beta^{-1/\alpha-1}\mathcal{Y}(q)(1 + o(1)), \tag{E.3}$$

$$\mathbb{E}[(x_1 - q)(x_1 - q)^\top e^{-2\beta T_1}] = C(q)\Gamma\left(\frac{1}{\alpha} + 1\right)(2\beta)^{-1/\alpha-1}\,\mathbf{P}_q(1 + o(1)), \tag{E.4}$$

$$\mathbb{E}[\|x_1 - q\|^3 e^{-3\beta T_1}] = O\Big(\beta^{-1/\alpha-3/2}\Big). \tag{E.5}$$

Proof. Rotate coordinates so that $q = e_d$, and put $m = d-1 = 2/\alpha$. We parametrize a neighborhood of q by tangent coordinates $v \in \mathbb{R}^m$:

$$\Psi(v) = \big(v, \sqrt{1 - \|v\|^2}\big), \qquad \|v\| < \delta.$$

In this chart,

$$T(\Psi(v)) = 1 - \langle q, \Psi(v)\rangle = 1 - \sqrt{1 - \|v\|^2} = \frac{\|v\|^2}{2} + O(\|v\|^4),$$

$$\mathrm{d}\sigma(\Psi(v)) = (1 - \|v\|^2)^{-1/2}\,dv = (1 + O(\|v\|^2))\,dv,$$

and

$$\Psi(v) - q = \big(v, -\|v\|^2/2\big) + O(\|v\|^4),$$

where $\mathrm{d}\sigma$ denotes surface measure on $\mathbb{S}^{d-1}$, Moreover, since ρ is C^2 near q,

$$\rho(\Psi(v)) = \rho(q) + \langle \nabla_{\mathbb{S}^{d-1}}\rho(q), v\rangle + O(\|v\|^2),$$

where $T_q\mathbb{S}^{d-1}$ is identified with $\mathbb{R}^m \times \{0\}$. The contribution from the complement of this coordinate patch is exponentially small because T is bounded below there by a positive constant.

For the first moment (E.3), the tangent component comes from multiplying the linear term v in $\Psi(v) - q$ by the linear term $\langle \nabla_{\mathbb{S}^{d-1}}\rho(q), v\rangle$ in the density expansion:

$$\int_{\mathbb{R}^m} v\,\langle \nabla_{\mathbb{S}^{d-1}}\rho(q), v\rangle\, e^{-\beta\|v\|^2/2}\,\mathrm{d}v = (2\pi)^{m/2}\beta^{-m/2-1}\nabla_{\mathbb{S}^{d-1}}\rho(q).$$

The normal component comes from $-\|v\|^2 q/2$, giving

$$-\frac{\rho(q)}{2}q\int_{\mathbb{R}^m}\|v\|^2 e^{-\beta\|v\|^2/2}\,\mathrm{d}v = -\frac{m}{2}\rho(q)(2\pi)^{m/2}\beta^{-m/2-1}q.$$

All other terms are lower order: odd Gaussian integrals vanish, and the remaining nonzero terms contain at least three tangent powers or the factor $\beta\|v\|^4$ from the exponential expansion. Thus

$$\mathbb{E}[(x_1 - q)e^{-\beta T_1}] = \rho(q)(2\pi)^{m/2}\left(\nabla_{\mathbb{S}^{d-1}}\log\rho(q) - \frac{m}{2}q\right)\beta^{-m/2-1}(1 + o(1)).$$

Since

$$C(q)\Gamma\left(\frac{1}{\alpha} + 1\right) = \rho(q)(2\pi)^{1/\alpha},$$

this proves (E.3).

For the second moment (E.4) with $e^{-2\beta T_1}$, the leading contribution is tangent–tangent:

$$\rho(q)\int_{\mathbb{R}^m} vv^\top e^{-\beta\|v\|^2}\,\mathrm{d}v = \rho(q)(2\pi)^{m/2}(2\beta)^{-m/2-1}\mathrm{Id}_m.$$

Embedding the tangent covariance into $\mathbb{R}^d$ gives $\mathbf{P}_q$. The normal–normal part is order $\beta^{-m/2-2}$, and the tangent–normal mixed part is lower order by parity or by degree. This proves (E.4).

Finally, for the third moment (E.5), in the same chart, $\|x-q\| \asymp \|v\|$ and $T(\Psi(v)) \asymp \|v\|^2$. Hence

$$\|x-q\|^3 e^{-3\beta T(x)} \leqslant C\|v\|^3 e^{-c\beta\|v\|^2}$$

locally, and so the local contribution is bounded by

$$C\int_{\mathbb{R}^m} \|v\|^3 e^{-c\beta\|v\|^2}\,\mathrm{d}v = O(\beta^{-(m+3)/2}) = O(\beta^{-1/\alpha-3/2}).$$

The contribution away from the chart is exponentially small, proving (E.5). □

E.5. **Subcritical output: proof of Theorem 4.3.** We now prove the output theorem. The proof has two reductions. The first replaces the random denominator Z_n by its expectation using Theorem D.2. The second applies a triangular-array central limit theorem to the numerator $B_n(q)$.

Lemma E.6 (Denominator reduction)**.** *Assume $V = \mathrm{Id}$. If $r_n > 0$ and*

$$r_n \frac{B_n(q)}{\mathbb{E}[Z_n]} \Rightarrow Z,$$

then

$$r_n \mathcal{Y}_n(q) \Rightarrow Z.$$

The same implication holds with convergence in probability in place of convergence in distribution.

Proof. By Theorem D.2,

$$\frac{Z_n}{\mathbb{E}[Z_n]} \xrightarrow{\mathbb{P}} 1, \qquad \frac{\mathbb{E}[Z_n]}{Z_n} \xrightarrow{\mathbb{P}} 1.$$

Since

$$r_n \mathcal{Y}_n(q) = r_n \frac{B_n(q)}{\mathbb{E}[Z_n]} \cdot \frac{\mathbb{E}[Z_n]}{Z_n},$$

Slutsky's theorem finishes the proof. □

Lemma E.7 (Numerator fluctuation lemma)**.** *Set*

$$X_{n,i} := (x_i - q)e^{-\beta_n T_i}, \qquad m_n := \mathbb{E}[B_n(q)] = n\mathbb{E}[X_{n,1}], \qquad \Gamma_n := n\,\mathrm{Cov}(X_{n,1}).$$

Suppose $r_n > 0$, $\mu \in \mathbb{R}^d$, and Γ is a symmetric nonnegative matrix such that

$$r_n \frac{m_n}{\mathbb{E}[Z_n]} \to \mu, \tag{E.6}$$

$$\frac{r_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n \to \Gamma, \tag{E.7}$$

$$\frac{r_n^3 n\mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} \to 0. \tag{E.8}$$

Then

$$r_n \frac{B_n(q)}{\mathbb{E}[Z_n]} \Rightarrow \mu + \mathcal{N}(0,\Gamma).$$

If $\Gamma = 0$, this means convergence in probability to μ.

Proof. Define
$$Y_{n,i} := \frac{r_n}{\mathbb{E}[Z_n]}\big(X_{n,i} - \mathbb{E}[X_{n,1}]\big).$$
Then the $Y_{n,i}$ are independent and centered, and
$$r_n \frac{B_n(q)}{\mathbb{E}[Z_n]} = r_n \frac{m_n}{\mathbb{E}[Z_n]} + \sum_{i=1}^n Y_{n,i}.$$
Moreover,
$$\sum_{i=1}^n \mathrm{Cov}(Y_{n,i}) = \frac{r_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n \to \Gamma$$
by (E.7), and
$$\sum_{i=1}^n \mathbb{E}[\|Y_{n,i}\|^3] \leqslant C \frac{r_n^3 n \mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} \to 0$$
by (E.8). The Lyapunov central limit theorem for triangular arrays (for instance, Proposition 2.27 in [VdV00]) gives
$$\sum_{i=1}^n Y_{n,i} \Rightarrow \mathcal{N}(0,\Gamma).$$
Combining this with (E.6) proves the claim. □

Proof of Theorem 4.3. It is enough to prove the theorem when $V = \mathrm{Id}$, because the general case follows by applying the linear map V to the identity-value limits. We therefore work with
$$\mathcal{Y}_n(q) = \frac{B_n(q)}{Z_n}.$$
By Theorem D.1 and (E.3),
$$\mathbb{E}[Z_n] = C(q)\Gamma\left(\frac{1}{\alpha}+1\right) n\beta_n^{-1/\alpha}(1+o(1)), \tag{E.9}$$
$$m_n = C(q)\Gamma\left(\frac{1}{\alpha}+1\right) \mathcal{Y}(q) n\beta_n^{-1/\alpha-1}(1+o(1)). \tag{E.10}$$
Moreover, (E.4) implies
$$n\mathbb{E}[X_{n,1}X_{n,1}^\top] = C(q)\Gamma\left(\frac{1}{\alpha}+1\right) 2^{-1/\alpha-1}\mathbf{P}_q\, n\beta_n^{-1/\alpha-1}(1+o(1)).$$
Since (E.3) gives $\mathbb{E}[X_{n,1}] = O(\beta_n^{-1/\alpha-1})$,
$$n\mathbb{E}[X_{n,1}]\mathbb{E}[X_{n,1}]^\top = O(n\beta_n^{-2/\alpha-2}) = o(n\beta_n^{-1/\alpha-1}).$$
Therefore
$$\Gamma_n = C(q)\Gamma\left(\frac{1}{\alpha}+1\right) 2^{-1/\alpha-1}\mathbf{P}_q\, n\beta_n^{-1/\alpha-1}(1+o(1)).$$
Recalling that $c_d = 2^{-d}\pi^{-(d-1)/2}$ and $\Sigma(q) = c_d\rho(q)^{-1}\mathbf{P}_q$, we get
$$\frac{\beta_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n = \Sigma(q)\frac{\beta_n^{1+1/\alpha}}{n}(1+o(1)), \tag{E.11}$$
$$\frac{\beta_n^3 n\mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} = O\left(\beta_n^{-1/2}\left(\frac{\beta_n^{1+1/\alpha}}{n}\right)^2\right). \tag{E.12}$$

We now apply the numerator fluctuation lemma in the three possible subcritical output regimes.

Drift-dominated regime. Assume $\beta_n^{1+\alpha}n^{-\alpha} \to 0$, equivalently $\beta_n^{1+1/\alpha}/n \to 0$, and set $r_n = \beta_n$. From (E.9) and (E.10),

$$r_n \frac{m_n}{\mathbb{E}[Z_n]} \to \mathcal{Y}(q).$$

By (E.11),

$$\frac{r_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n \to 0,$$

and by (E.12),

$$\frac{r_n^3 n \mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} \to 0.$$

Thus Theorems E.6 and E.7 yield

$$\beta_n \mathcal{Y}_n(q) \xrightarrow{\mathbb{P}} \mathcal{Y}(q).$$

Mixed drift–fluctuation regime. Assume $\beta_n^{1+\alpha}n^{-\alpha} \to \tau \in (0,\infty)$, equivalently $\beta_n^{1+1/\alpha}/n \to \tau^{1/\alpha}$, and again set $r_n = \beta_n$. Then

$$r_n \frac{m_n}{\mathbb{E}[Z_n]} \to \mathcal{Y}(q), \qquad \frac{r_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n \to \tau^{1/\alpha}\Sigma(q).$$

Moreover, (E.12) gives

$$\frac{r_n^3 n \mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} = O(\beta_n^{-1/2}) \to 0,$$

because $n \asymp \beta_n^{1+1/\alpha}$ in this regime. Hence

$$\beta_n \mathcal{Y}_n(q) \Rightarrow \mathcal{Y}(q) + \mathcal{N}(0, \tau^{1/\alpha}\Sigma(q)).$$

Fluctuation-dominated regime. Assume $\beta_n^{1+\alpha}n^{-\alpha} \to \infty$ while still $\beta_n n^{-\alpha} \to 0$, and set

$$r_n = \sqrt{n\beta_n^{-1/\alpha+1}}.$$

Then

$$r_n \frac{m_n}{\mathbb{E}[Z_n]} = \sqrt{n\beta_n^{-1/\alpha-1}}\,\mathcal{Y}(q)(1+o(1)) \to 0,$$

because $\beta_n^{1+1/\alpha}/n \to \infty$. Also,

$$\frac{r_n^2}{\mathbb{E}[Z_n]^2}\Gamma_n \to \Sigma(q).$$

Finally, by (E.5),

$$\frac{r_n^3 n \mathbb{E}[\|X_{n,1}\|^3]}{\mathbb{E}[Z_n]^3} = O\Big((n\beta_n^{-1/\alpha})^{-1/2}\Big) \to 0.$$

Thus

$$\sqrt{n\beta_n^{-1/\alpha+1}}\,\mathcal{Y}_n(q) \Rightarrow \mathcal{N}(0, \Sigma(q)).$$

In all three cases the normalizing factor diverges, so $\mathcal{Y}_n(q) \to 0$ in probability in the identity-value case. For general V, apply the linear map $z \mapsto Vz$ to the identity-value limits. This gives the three conclusions of Theorem 4.3, with covariance $V\Sigma(q)V^\top$, and also $\mathcal{Y}_n(q) \to 0$ in probability. □

E.6. **Residual dynamics: proof of Theorem 4.4.** The residual dynamics use only the drift-dominated output regime and the first-order expansion of the normalization map $u \mapsto u/\|u\|$.

Proof of Theorem 4.4. Because $\beta_n^{1+\alpha} n^{-\alpha} \to 0$, Theorem 4.3(1) specialized to $V = \mathrm{Id}$ gives

$$\beta_n \mathcal{Y}_n(q) \xrightarrow{\mathbb{P}} \mathcal{Y}(q) = \nabla_{\mathbb{S}^{d-1}} \log \rho(q) - \frac{d-1}{2} q.$$

Consequently, $\mathcal{Y}_n(q) = O_{\mathbb{P}}(\beta_n^{-1})$.

For $q \in \mathbb{S}^{d-1}$ and a small perturbation h,

$$\|q+h\|^{-1} = \big(1 + 2\langle q, h\rangle + \|h\|^2\big)^{-1/2} = 1 - \langle q, h\rangle + O(\|h\|^2).$$

Thus

$$\frac{q+h}{\|q+h\|} = q + h - \langle q, h\rangle q + O(\|h\|^2) = q + \mathbf{P}_q h + O(\|h\|^2).$$

Setting $h = \gamma \mathcal{Y}_n(q)$, we obtain

$$q' = q + \gamma \mathbf{P}_q \mathcal{Y}_n(q) + R_n, \qquad R_n = O_{\mathbb{P}}(\beta_n^{-2}).$$

Since $\mathbf{P}_q \mathcal{Y}(q) = \nabla_{\mathbb{S}^{d-1}} \log \rho(q)$,

$$\beta_n \mathbf{P}_q \mathcal{Y}_n(q) \xrightarrow{\mathbb{P}} \nabla_{\mathbb{S}^{d-1}} \log \rho(q). \tag{E.13}$$

and this concludes the proof. $\square$

An immediate consequence is the following corollary:

Corollary E.8. *Assume $V = \mathrm{Id}$ and let q^* be a critical point of ρ on $\mathbb{S}^{d-1}$. For any $q \neq q^*$ in a sufficiently small neighborhood of q^*, the limiting update in Theorem 4.4 increases $\langle q', (q^*)'\rangle$ if q^* is a strict local maximum of ρ, and decreases it if q^* is a strict local minimum.*

Proof. Consider $q_1, q_2 \in \mathbb{S}^{d-1}$. Expanding the inner product of the updated queries, and using Theorem 4.4 yields:

$$\beta_n\big(\langle q_1', q_2'\rangle - \langle q_1, q_2\rangle\big) \xrightarrow{\mathbb{P}} \gamma\Big(\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q_1), q_2\rangle + \langle q_1, \nabla_{\mathbb{S}^{d-1}} \log \rho(q_2)\rangle\Big). \tag{E.14}$$

Apply Equation (E.14) with $q_1 = q$ and $q_2 = q^*$. Since $\nabla_{\mathbb{S}^{d-1}} \log \rho(q^*) = 0$,

$$\beta_n\big(\langle q', (q^*)'\rangle - \langle q, q^*\rangle\big) \xrightarrow{\mathbb{P}} \gamma \langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q), q^*\rangle.$$

It remains to identify the sign of the limiting inner product. Parameterize the geodesic issuing from q^* by

$$q(\theta) = (\cos\theta) q^* + (\sin\theta) u,$$

where $u \in T_{q^*}\mathbb{S}^{d-1}$ is a unit tangent vector and $\theta > 0$ is small. Since $\nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta))$ is tangent to the sphere at $q(\theta)$, it is orthogonal to $q(\theta)$, so

$$\cos\theta \,\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), q^*\rangle + \sin\theta \,\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), u\rangle = 0.$$

Therefore

$$\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), q^*\rangle = -\tan\theta \,\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), u\rangle.$$

A Taylor expansion along the geodesic around the critical point q^* gives

$$\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), u\rangle = \theta \,\langle \nabla^2_{\mathbb{S}^{d-1}} \log \rho(q^*) u, u\rangle + o(\theta).$$

Using $\tan\theta = \theta + o(\theta)$, we obtain

$$\langle \nabla_{\mathbb{S}^{d-1}} \log \rho(q(\theta)), q^* \rangle = -\theta^2 \langle \nabla^2_{\mathbb{S}^{d-1}} \log \rho(q^*) u, u \rangle + o(\theta^2).$$

If q^* is a strict local maximum of ρ, then the spherical Hessian of $\log \rho$ at q^* is negative definite, so the right-hand side is strictly positive for θ small. Thus the limiting update increases $\langle q', (q^*)' \rangle$. Conversely, if q^* is a strict local minimum of ρ, then the spherical Hessian is positive definite, so the limiting update decreases $\langle q', (q^*)' \rangle$. Finally, since $\nabla_{\mathbb{S}^{d-1}} \log \rho(q^*) = 0$, Equation (E.14) applied with both queries equal to q^* shows that q^* does not move to first order. □

Appendix F. RoPE with Correlated Tokens

In this appendix, we consider the critical scaling problem when the attention scores incorporate Rotary Positional Embeddings (RoPE) [SAL+24] and the context tokens are correlated.

RoPE encodes the positions by rotating query and key coordinates in frequency-dependent two-dimensional planes before taking their inner product. Since these rotations are orthogonal, a RoPE score can be rewritten as an ordinary dot product between the key and a query direction that moves deterministically with relative position. Thus, for the one-query asymptotic studied in this paper, RoPE changes the fixed target direction q into a deterministic orbit (u_i) on the sphere. The random part remains the token sequence (x_i), while the RoPE rotations are deterministic.

We use the standard block-diagonal RoPE form. For an even rotated dimension $d = 2L$, let

$$R_p = \operatorname{diag}\big(r(p\vartheta_1), \dots, r(p\vartheta_L)\big), \qquad r(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}.$$

The same notation also covers the common case where only a rotated subspace is used, by leaving the remaining coordinates unrotated. If a query at position p attends to a key at position i, then

$$\langle R_p q, R_i x_i \rangle = \langle R_i^\top R_p q, x_i \rangle = \langle R_{i-p}^\top q, x_i \rangle.$$

After relabeling relative positions, we write

$$u_i := R_i^\top q, \qquad \langle R_0 q, R_i x_i \rangle = \langle u_i, x_i \rangle.$$

Changing the reference position or the sign convention only shifts the initial phase or replaces the frequency vector by its negative, and does not affect the limiting phase law below.

It is useful to encode this deterministic orbit as a torus rotation. Let

$$\mathbb{T} := \mathbb{R}/\mathbb{Z}, \qquad \theta := \left(\frac{\vartheta_1}{2\pi}, \dots, \frac{\vartheta_L}{2\pi} \right) \in \mathbb{T}^L.$$

Define the continuous map

$$\Phi_q : \mathbb{T}^L \longrightarrow \mathbb{S}^{d-1}, \qquad \Phi_q(x) := \operatorname{diag}\big(r(-2\pi x_1), \dots, r(-2\pi x_L)\big) q.$$

Then the RoPE target sequence is simply

$$u_i = \Phi_q(i\theta \bmod 1), \qquad i \in \mathbb{Z}.$$

This is the only structural property of RoPE used in the proof: the target direction follows a deterministic translation orbit on a torus, mapped continuously into the sphere.

Let

$$\mathcal{H}_q := \overline{\{i\theta \bmod 1 : i \in \mathbb{Z}\}} \subset \mathbb{T}^L.$$

This is a closed subgroup of the torus. Let $m_{\mathcal{H}_q}$ denote Haar probability measure on $\mathcal{H}_q$ [Fol16], and define the deterministic phase law

$$\pi_q := (\Phi_q)_{\#} m_{\mathcal{H}_q}.$$

The orbit closure in the sphere is

$$\mathcal{O}_q := \overline{\{u_i : i \in \mathbb{Z}\}} = \Phi_q(\mathcal{H}_q),$$

and Proposition F.7 proves that

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{u_i} \Rightarrow \pi_q.$$

Thus π_q is the long-run deterministic distribution of RoPE target directions. If the frequencies are rational multiples of 2π, then the orbit is periodic and π_q is uniform on that finite orbit. If $1, \theta_1, \ldots, \theta_L$ are linearly independent over $\mathbb{Q}$, then $\mathcal{H}_q = \mathbb{T}^L$, and $m_{\mathcal{H}_q}$ is the uniform measure over $\mathbb{T}^L$. Intermediate arithmetic relations give the corresponding Haar measure on a proper closed subgroup.

With this notation, the score, D2Q, and weights are

$$a_i = \beta_n \langle u_i, x_i \rangle, \qquad U_i := 1 - \langle u_i, x_i \rangle \in [0, 2],$$

and

$$A_i := \frac{e^{-\beta_n U_i}}{\sum_{j=1}^{n} e^{-\beta_n U_j}}, \qquad A_{(1)} \geqslant A_{(2)} \geqslant \cdots \geqslant A_{(n)}.$$

Consequently, as in the i.i.d. fixed-query case, the asymptotic behavior of the ordered weights is controlled by the lower tail of the D2Q variables. We make the following assumptions in this appendix.

Assumption 2 (Correlated tokens with RoPE)**.** *The sequence $(x_i)_{i\in\mathbb{Z}}$ is stationary*[2] *with values in $\mathbb{S}^{d-1}$. Moreover:*

(i) there exists an integer $m \geqslant 0$ such that (x_i) is m-dependent, i.e., for any two index sets $I, J \subset \mathbb{Z}$ with $\mathrm{dist}(I, J) > m$, we have $\{x_i : i \in I\}$ and $\{x_j : j \in J\}$ are independent.

(ii) the common law of x_i has a density ρ with respect to surface measure on $\mathbb{S}^{d-1}$, and ρ is continuous and strictly positive on a neighborhood of $\mathcal{O}_q$.

(iii) for each $1 \leqslant \ell \leqslant m$, the pair (x_0, x_ℓ) has a density with respect to product surface measure, bounded on a neighborhood of $\mathcal{O}_q \times \mathcal{O}_q$.

Remark F.1 (On stationarity)**.** *Stationarity in Assumption 2 is a convenient sufficient condition, not an essential one. The same argument should extend whenever the rescaled D2Q has the same Poisson Point Process limit.*

Remark F.2 (When anti-clustering fails)**.** *When Assumption 2(iii) fails, the first-order scale often remains n^α, but the limiting point process is typically no longer Poisson. In that case the limiting ordered weights are obtained by applying the same softmax functional to the appropriate cluster-Poisson extremal process.*

---

[2] A sequence of random variables $(x_k)_{k\in\mathbb{Z}}$ is called *stationary* if for every $m \geqslant 1$, every choice of indices $k_1, \ldots, k_m \in \mathbb{Z}$, and every $h \in \mathbb{Z}$, one has

$$(x_{k_1}, \ldots, x_{k_m}) \overset{d}{=} (x_{k_1+h}, \ldots, x_{k_m+h}).$$

Under Assumption 2, the critical exponent remains $\alpha = \frac{2}{d-1}$, and RoPE changes only the constant in the limiting point process through the orbit average $\overline{C}(q)$.

$$\overline{C}(q) := \frac{2^{1/\alpha}\sigma_{d-2}}{d-1} \int_{\mathbb{S}^{d-1}} \rho(u)\, \pi_q(du),$$

Theorem F.3 (Critical regime with correlated tokens and RoPE)**.** *Under Assumption 2, assume that $\beta_n n^{-\alpha} \longrightarrow \gamma \in (0,\infty)$, and let $\mathcal{X}$ be a Poisson point process on $\mathbb{R}_+ := [0,\infty)$ with intensity measure*

$$\overline{\Lambda}([0,y]) = \overline{C}(q)\Big(\frac{y}{\gamma}\Big)^{1/\alpha}, \qquad y \geqslant 0.$$

Write $0 < Y_1 < Y_2 < \cdots$ for the atoms of $\mathcal{X}$ arranged in increasing order. Then, for every fixed $k \geqslant 1$,

$$\big(A_{(1)}, \dots, A_{(k)}\big) \Rightarrow \big(W_1, \dots, W_k\big) \qquad \textit{as } n \to \infty,$$

where

$$W_i := \frac{e^{-Y_i}}{\sum_{j\geqslant 1} e^{-Y_j}}, \qquad i \geqslant 1.$$

Remark F.4. *For simplicity, we state and prove only the critical RoPE limit. Under the same assumptions, the supercritical and subcritical conclusions from the i.i.d. case carry over to the RoPE setting. In particular, similar to Theorems 3.1 and 3.3, $\beta_n n^{-\alpha} \to \infty$ gives $A_{(1)} \to 1$, while $\beta_n n^{-\alpha} \to 0$ gives $A_{(i)} \to 0$ for every fixed i.*

Remark F.5. *We note that if we denote $F_i(t) = \mathbb{P}(U_i < t)$, then $\overline{C}(q)$ equals the average of local density along the RoPE orbit:*

$$\overline{C}(q) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \lim_{t\to 0^+} F_i(t) t^{-1/\alpha}$$

In particular, if ρ is invariant under the RoPE rotations along the orbit of q, then

$$\int_{\mathbb{S}^{d-1}} \rho(u)\, \pi_q(du) = \rho(q),$$

so that $\overline{C}(q) = C(q)$ and the first-order critical normalization is the same as in the i.i.d. theorem.

Define

$$\mathcal{X}_n := \sum_{i=1}^{n} \delta_{\beta_n U_i} \qquad \text{on } \mathbb{R}_+.$$

We first establish the Poisson limit for $\mathcal{X}_n$.

Lemma F.6 (Poisson point process limit)**.** *Assume that $\beta_n n^{-\alpha} \longrightarrow \gamma \in (0,\infty)$. Under Assumption 2, we have $\mathcal{X}_n \Rightarrow \mathcal{X}$ in $M_p(\mathbb{R}_+)$, where $\mathcal{X}$ is a Poisson point process on $\mathbb{R}_+$ with intensity measure $\overline{\Lambda}$ given by*

$$\overline{\Lambda}([0,y]) = \overline{C}(q)\Big(\frac{y}{\gamma}\Big)^{1/\alpha}, \qquad y \geqslant 0.$$

Proof of Theorem F.6. We prove the convergence by Laplace functionals.

Fix a nonnegative function $\varphi \in C_c(\mathbb{R}_+)$. Let $R < \infty$ be such that $\operatorname{supp}\varphi \subset [0, R]$, and set

$$G(y) := 1 - e^{-\varphi(y)}, \qquad X_{i,n} := G(\beta_n U_i) = 1 - e^{-\varphi(\beta_n U_i)}.$$

Then $0 \leqslant X_{i,n} \leqslant \mathbf{1}_{\{\beta_n U_i \leqslant R\}}$, and

$$\mathbb{E}\exp\Big(-\int_{\mathbb{R}_+} \varphi(y)\, \mathcal{X}_n(dy)\Big) = \mathbb{E}\prod_{i=1}^{n}(1 - X_{i,n}).$$

We will show that this converges to

$$\exp\Big(-\int_{\mathbb{R}_+} (1 - e^{-\varphi(y)})\, \overline{\Lambda}(dy)\Big).$$

Step 1. Uniform one-point tail asymptotics. Note that by Assumption 2, ρ is continuous on a neighborhood of the compact set $\mathcal{O}_q$, so it is uniformly continuous there. Hence

$$\sup_{u \in \mathcal{O}_q} \Big|\mathbb{P}(1 - \langle u, x_1\rangle \leqslant t) - C(u)t^{1/\alpha}\Big| = o(t^{1/\alpha}).$$

Since $u_i \in \mathcal{O}_q$, this gives, uniformly in i,

$$\mathbb{P}(U_i \leqslant t) = C(u_i)t^{1/\alpha} + o(t^{1/\alpha}) \qquad \text{as} \qquad t \downarrow 0. \tag{F.1}$$

In particular, for every fixed $R < \infty$ there is a constant $C_R < \infty$ such that, for all large n and all $1 \leqslant i \leqslant n$,

$$\mathbb{P}(\beta_n U_i \leqslant R) \leqslant \frac{C_R}{n}. \tag{F.2}$$

Step 2. Convergence of the mean measures. Let

$$\mu_n(B) := \sum_{i=1}^{n} \mathbb{P}(\beta_n U_i \in B), \qquad B \subset \mathbb{R}_+ \text{ Borel}.$$

For every fixed $y \geqslant 0$, (F.1) gives

$$\mu_n([0, y]) = \sum_{i=1}^{n} \mathbb{P}\Big(U_i \leqslant \frac{y}{\beta_n}\Big) = \frac{y^{1/\alpha}}{\beta_n^{1/\alpha}} \sum_{i=1}^{n} C(u_i) + n\, o(\beta_n^{-1/\alpha}). \tag{F.3}$$

Recall from Lemma A.1 that, for $u \in \mathbb{S}^{d-1}$

$$C(u) = \frac{2^{1/\alpha}\sigma_{d-2}}{d-1}\rho(u)\,. \tag{F.4}$$

We need the following lemma for the limit of the mean measure:

Proposition F.7. *With $\pi_q = (\Phi_q)_{\#} m_{H_q}$ defined at the beginning of Section F,*

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{u_i} \Rightarrow \pi_q \qquad \textit{weakly on } \mathcal{P}(\mathbb{S}^{d-1}).$$

We defer the proof of Proposition F.7 to the end of this appendix. By Proposition F.7, we obtain directly that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(u_i) \longrightarrow \int_{\mathbb{S}^{d-1}} \rho(u)\, \pi_q(du). \tag{F.5}$$

Therefore, combining (F.3), (F.4) and (F.5) and using $\beta_n n^{-\alpha} \to \gamma$, we have

$$\mu_n([0,y]) \longrightarrow \frac{2^{1/\alpha}\sigma_{d-2}}{d-1}\gamma^{-1/\alpha}y^{1/\alpha}\int_{\mathbb{S}^{d-1}} \rho(u)\,\pi_q(du) = \overline{C}(q)\Big(\frac{y}{\gamma}\Big)^{1/\alpha} = \overline{\Lambda}([0,y]).$$

The limit distribution function is continuous in y, so μ_n converges vaguely to Λ on $\mathbb{R}_+$. Hence, since $G = 1 - e^{-\varphi}$ is continuous and compactly supported,

$$L_n := \sum_{i=1}^{n} \mathbb{E}X_{i,n} = \int G(y)\,\mu_n(dy) \longrightarrow L := \int G(y)\,\overline{\Lambda}(dy). \tag{F.6}$$

Step 3. Short-range two-hit bound. Fix $1 \leqslant \ell \leqslant m$. By Assumption 2(iii), the pair $(x_i, x_{i+\ell})$ has a density bounded on a neighborhood of $\mathcal{O}_q \times \mathcal{O}_q$. Therefore, for all sufficiently small t,

$$\mathbb{P}(U_i \leqslant t,\, U_{i+\ell} \leqslant t) \leqslant C' t^{2/\alpha},$$

uniformly in i and $1 \leqslant \ell \leqslant m$. Taking $t = R/\beta_n$ yields

$$\mathbb{E}[X_{i,n}X_{i+\ell,n}] \leqslant \mathbb{P}(\beta_n U_i \leqslant R,\, \beta_n U_{i+\ell} \leqslant R) \leqslant \frac{C_R}{n^2} \tag{F.7}$$

for all large n, uniformly in i and $1 \leqslant \ell \leqslant m$.

On the other hand, if $|i-j| > m$, then m-dependence in Assumption 2(i) gives independence, and (F.2) gives

$$\mathbb{E}[X_{i,n}X_{j,n}] = \mathbb{E}X_{i,n}\,\mathbb{E}X_{j,n} \leqslant \frac{C_R}{n^2}. \tag{F.8}$$

Step 4. Blocking reduction to independent blocks. Let $b_n := \lfloor n^{1/2} \rfloor$. Partition $\{1,\ldots,n\}$ into consecutive large blocks of length b_n, separated by gaps of length m. More explicitly, set the block

$$B_{r,n} := \{(r-1)(b_n+m)+1, \ldots, (r-1)(b_n+m)+b_n\}$$

and we let $\mathcal{I}_n \subset \{1,\ldots,n\}$ collect those r such that $B_{r,n} \subset \{1,\ldots,n\}$. Let $\mathcal{B}_n = \bigcup_{r\in\mathcal{I}_n} B_{r,n}$ be the union of these large blocks; see Figure 4 for an illustration.

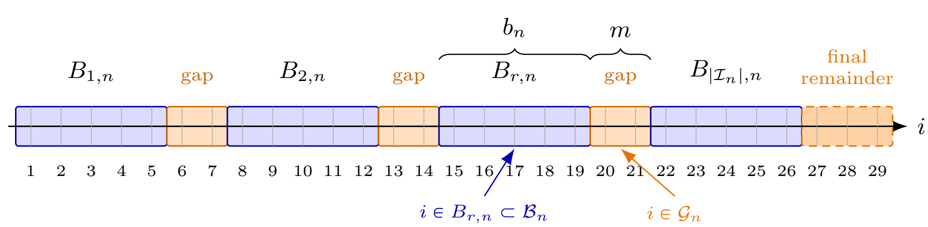

Figure 4. Large blocks are separated by gaps of length m, so the corresponding block variables are independent under m-dependence.

Let $\mathcal{G}_n := \{1,\ldots,n\}\backslash\mathcal{B}_n$ be the union of the gaps and the final remainder. Then $|\mathcal{G}_n| = O\Big(\frac{n}{b_n} + b_n\Big) = o(n)$. Since $0 \leqslant X_{i,n} \leqslant 1$,

$$\left|\prod_{i=1}^{n}(1-X_{i,n}) - \prod_{i\in\mathcal{B}_n}(1-X_{i,n})\right| \leqslant \sum_{i\in\mathcal{G}_n} X_{i,n}.$$

Using (F.2),

$$\mathbb{E}\sum_{i\in\mathcal{G}_n} X_{i,n} \leqslant \frac{C_R|\mathcal{G}_n|}{n} \longrightarrow 0. \tag{F.9}$$

Therefore

$$\mathbb{E}\prod_{i=1}^{n}(1-X_{i,n}) = \mathbb{E}\prod_{i\in\mathcal{B}_n}(1-X_{i,n}) + o(1). \tag{F.10}$$

The large blocks are separated by more than m indices, so the corresponding collections of variables are independent. Hence

$$\mathbb{E}\prod_{i\in\mathcal{B}_n}(1-X_{i,n}) = \prod_r Q_{r,n}, \qquad Q_{r,n} := \mathbb{E}\prod_{i\in B_{r,n}}(1-X_{i,n}).$$

For each block, define $a_{r,n} := \sum_{i\in B_{r,n}} \mathbb{E}X_{i,n}$. The standard Bonferroni bound

$$\left|\prod_{i\in B}(1-x_i) - \Big(1-\sum_{i\in B} x_i\Big)\right| \leqslant \sum_{\substack{i<j\\ i,j\in B}} x_i x_j, \qquad 0\leqslant x_i\leqslant 1,$$

gives

$$Q_{r,n} = 1 - a_{r,n} + \varepsilon_{r,n}, \qquad |\varepsilon_{r,n}| \leqslant \sum_{\substack{i<j\\ i,j\in B_{r,n}}} \mathbb{E}[X_{i,n}X_{j,n}].$$

By (F.7) and (F.8),

$$|\varepsilon_{r,n}| \leqslant C_R b_n^2/n^2. \tag{F.11}$$

Since the number of large blocks is $O(n/b_n)$,

$$\sum_{r\in\mathcal{I}_n} |\varepsilon_{r,n}| \leqslant C_R\frac{b_n}{n} \longrightarrow 0. \tag{F.12}$$

Moreover, by (F.2),

$$\max_{r\in\mathcal{I}_n} a_{r,n} \leqslant C_R\frac{b_n}{n} \longrightarrow 0, \tag{F.13}$$

and by (F.6) together with (F.9),

$$\sum_{r\in\mathcal{I}_n} a_{r,n} = \sum_{i\in\mathcal{B}_n} \mathbb{E}X_{i,n} = L_n + o(1) \longrightarrow L.$$

By (F.11) and (F.13), for all large n, $|-a_{r,n}+\varepsilon_{r,n}| \leqslant 1/2$ uniformly in r, and therefore

$$\sum_{r\in\mathcal{I}_n}\log Q_{r,n} = \sum_{r\in\mathcal{I}_n}\log(1-a_{r,n}+\varepsilon_{r,n}) = -\sum_{r\in\mathcal{I}_n} a_{r,n} + \sum_{r\in\mathcal{I}_n}\varepsilon_{r,n} + O\Big(\sum_{r\in\mathcal{I}_n}(a_{r,n}^2+\varepsilon_{r,n}^2)\Big).$$

The last error term tends to zero because

$$\sum_{r\in\mathcal{I}_n} a_{r,n}^2 \leqslant (\max_r a_{r,n})\sum_{r\in\mathcal{I}_n} a_{r,n} \to 0, \qquad \sum_{r\in\mathcal{I}_n}\varepsilon_{r,n}^2 \leqslant (\max_{r\in\mathcal{I}_n}|\varepsilon_{r,n}|)\sum_{r\in\mathcal{I}_n}|\varepsilon_{r,n}| \to 0.$$

Combining this with (F.12) gives

$$\sum_{r\in\mathcal{I}_n}\log Q_{r,n} \longrightarrow -L.$$

Consequently,

$$\mathbb{E}\prod_{i\in\mathcal{B}_n}(1-X_{i,n}) = \prod_{r\in\mathcal{I}_n} Q_{r,n} \longrightarrow e^{-L}. \tag{F.14}$$

Finally, combining (F.10) and (F.14), we get

$$\mathbb{E}\exp\Big(-\int_{\mathbb{R}_+}\varphi(y)\,\mathcal{N}_n(dy)\Big) \longrightarrow \exp\Big(-\int_{\mathbb{R}_+}(1-e^{-\varphi(y)})\,\overline{\Lambda}(dy)\Big),$$

which implies the point process convergence $\mathcal{X}_n \Rightarrow \mathcal{X}$ in $M_p(\mathbb{R}_+)$. □

Proof of Theorem F.7. Define the empirical measures on $\mathbb{T}^L$ by

$$\eta_n := \frac{1}{n}\sum_{j=1}^{n}\delta_{j\theta \bmod 1}.$$

For $k \in \mathbb{Z}^L$, their Fourier coefficients are

$$\widehat{\eta}_n(k) := \int_{\mathbb{T}^L} e^{2\pi \mathrm{i} k\cdot x}\,\eta_n(dx) = \frac{1}{n}\sum_{j=1}^{n} e^{2\pi \mathrm{i} jk\cdot\theta}.$$

If $k\cdot\theta \in \mathbb{Z}$, then every term equals 1, so $\widehat{\eta}_n(k) = 1$. If $k\cdot\theta \notin \mathbb{Z}$, then the right-hand side is a normalized geometric sum and therefore converges to 0. Hence

$$\widehat{\eta}_n(k) \longrightarrow \mathbf{1}_{\{k\cdot\theta\in\mathbb{Z}\}}.$$

These are exactly the Fourier coefficients of Haar probability measure $m_{\mathcal{H}_q}$ on $\mathcal{H}_q$ [Rud17]. Therefore $\eta_n \Rightarrow m_{\mathcal{H}_q}$ on $\mathbb{T}^L$. Pushing forward by the continuous map Φ_q yields

$$\frac{1}{n}\sum_{j=1}^{n}\delta_{u_j} = (\Phi_q)_{\#}\eta_n \Rightarrow (\Phi_q)_{\#}m_{\mathcal{H}_q} = \pi_q.$$

This proves the claim. □

Proof of Theorem F.3. To pass from the point-process limit to the softmax denominator, we use a strengthened Laplace-functional convergence, which essentially follows the proof of Theorem F.6. Let

$$Z_n := \int e^{-y}\mathcal{X}_n(\mathrm{d}y), \qquad Z := \int e^{-y}\mathcal{X}(\mathrm{d}y).$$

For every $\varphi \in C_c(\mathbb{R}_+)$, $\varphi \geqslant 0$ and every $s \geqslant 0$, set

$$G_s(y) := 1 - \exp\{-\varphi(y) - se^{-y}\}.$$

Then

$$0 \leqslant G_s(y) \leqslant 1 - e^{-\varphi(y)} + se^{-y},$$

and the right-hand side is integrable against the limiting intensity, since $\int_0^\infty e^{-y}y^{1/\alpha-1}\,dy < \infty$. The same blocking argument used for the compactly supported Laplace functional in the proof of Theorem F.6 therefore gives

$$\begin{aligned}\mathbb{E}\exp\left\{-\int\varphi(y)\mathcal{X}_n(\mathrm{d}y) - sZ_n\right\} &= \mathbb{E}\prod_i\big(1 - G_s(\beta_n U_i)\big)\\ &\longrightarrow \exp\left\{-\int G_s(y)\Lambda(\mathrm{d}y)\right\}\\ &= \mathbb{E}\exp\left\{-\int\varphi(y)\mathcal{X}(\mathrm{d}y) - sZ\right\}.\end{aligned}$$

This is exactly convergence of the joint Laplace transform of $\left(\mathcal{X}_n, \int e^{-y}\mathcal{X}_n(\mathrm{d}y)\right)$. Varying φ over a convergence-determining class in $C_c(\mathbb{R}_+)$ and $s \geqslant 0$, this gives

$$\left(\mathcal{X}_n, \int e^{-y}\mathcal{X}_n(\mathrm{d}y)\right) \Rightarrow \left(\mathcal{X}, \int e^{-y}\mathcal{X}(\mathrm{d}y)\right).$$

It remains to pass from the point process and its denominator to the ordered softmax weights. Let

$$0 < Y^n_{(1)} < Y^n_{(2)} < \cdots < Y^n_{(n)}$$

denote the atoms of $\mathcal{X}_n$ in increasing order, so that $Y^n_{(i)} = \beta_n U_{(i)}$ and $A_{(i)} = \frac{e^{-Y^n_{(i)}}}{Z_n}$. The limiting intensity measure $\overline{\Lambda}$ is diffuse, and $\overline{\Lambda}([0,R]) \to \infty$ as $R \to \infty$. Hence the limiting Poisson point process $\mathcal{X}$ is simple, has infinitely many atoms, and its first k atoms $0 < Y_{(1)} < \cdots < Y_{(k)}$ are finite almost surely. Moreover

$$Z = \int e^{-y}\mathcal{X}(\mathrm{d}y) = \sum_{j \geqslant 1} e^{-Y_{(j)}}$$

is almost surely finite and strictly positive: finiteness follows from $\int_0^\infty e^{-y}\overline{\Lambda}(\mathrm{d}y) < \infty$, while positivity follows from the existence of atoms. Therefore the map

$$(\mu, z) \longmapsto \left(\frac{e^{-y_{(1)}(\mu)}}{z}, \ldots, \frac{e^{-y_{(k)}(\mu)}}{z}\right),$$

where $y_{(i)}(\mu)$ is the i-th atom of μ, is continuous at $(\mathcal{X}, Z)$ almost surely. Applying the continuous mapping theorem to the joint convergence above gives

$$(A_{(1)}, \ldots, A_{(k)}) = \left(\frac{e^{-Y^n_{(1)}}}{Z_n}, \ldots, \frac{e^{-Y^n_{(k)}}}{Z_n}\right) \Rightarrow \left(\frac{e^{-Y_{(1)}}}{\sum_{j\geqslant 1} e^{-Y_{(j)}}}, \ldots, \frac{e^{-Y_{(k)}}}{\sum_{j\geqslant 1} e^{-Y_{(j)}}}\right).$$

This is exactly the claimed convergence to $(W_1, \ldots, W_k)$, and proves Theorem F.3. $\square$

References

[AWA25] Ben Anson, Xi Wang, and Laurence Aitchison. Scale-invariant attention. In *NeurIPS*, 2025.

[BBC+23] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report, 2023. arXiv:2309.16609.

[BPA25] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *NeurIPS*, 2025.

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. arXiv:2004.05150.

[CC22] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention, 2022. arXiv:2202.12172.

[CLPR25] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models, 2025. arXiv:2504.14697.

[CLPR26] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *ICLR*, 2026.

[CNWR25] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025.

[Der81] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, 1981.

[Dev06] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

[DVJ08] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer, 2008.

[DYY+19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

[Fol16] Gerald B Folland. *A course in abstract harmonic analysis*. CRC press, 2016.

[GG26] Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *ICLR*, 2026.

[GLPR25] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc.*, 2025.

[Goo24] Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. arXiv:2403.05530.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[Nak25] Ken M Nakanishi. Scalable-softmax is superior for attention, 2025. arXiv:2501.19399.

[Olv97] Frank W. J. Olver. *Asymptotics and Special Functions*. A K Peters, Wellesley, MA, 1997. Reprint, with corrections, of the 1974 Academic Press edition.

[PLAS+25] Krishna C. Puvvada, Faisal Ladhak, Santiago Akle Serano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, and Boris Ginsburg. SWAN: An efficient and scalable approach for long-context language modeling. In *EMNLP*, 2025.

[PQFS24] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *ICLR*, 2024.

[QHVM26] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model, 2026. arXiv:2602.11139.

[Rud17] Walter Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

[SAL+24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VPBP25] Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *ICML*, 2025.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[ZCL+24] Xuechen Zhang, Xiangyu Chang, Mingchen Li, Amit Roy-Chowdhury, Jiasi Chen, and Samet Oymak. Selective attention: Enhancing transformer through principled context control. In *NeurIPS*, 2024.

[ZGD+20] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020.

(GB) Departement Mathematik und Statistik, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
Email address: giuseppe.bruno@unibe.ch

(SC) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: schen636@mit.edu

(ZL) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: linzj@mit.edu

(YP) Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: yp@mit.edu

(PR) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: rigollet@math.mit.edu