DEPTH WITH MEMORY: TOKEN DYNAMICS OF ATTENTION RESIDUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention Residuals aggregate stored layer outputs with normalized depth weights. This operation generally produces a history-dependent continuous-depth equation, rather than a closed ordinary differential equation. Starting from a normalized Volterra formulation, we identify a common-reweighting condition under which the dynamics close and derive the exact coupled equations for token directions and radii, with query, key, and value matrices fixed across depth. For spherical causal attention with a symmetric positive-definite value matrix having a simple largest eigenvalue, we impose a block structure on the query–key product and an explicit condition ensuring positive radial forcing. These conditions yield positive radial bounds and global well-posedness. When each history rate has an infinite integral, we prove by causal induction that, outside a Lebesgue-null set of admissible raw initial embeddings, every token direction converges to one of the two leading eigenvector poles. The prescribed history kernels are held fixed as the initial embeddings vary. The radial equation asymptotically averages the leading eigenvalue multiplied by the signed attention mass. We give conditions for radial convergence and an explicit softmax formula for the limiting lengths, which may differ within one directional cluster. The conclusions concern a specified normalized model and a sufficient clustering regime, not an arbitrarily trained Attention Residual networks.

1 INTRODUCTION

Residual connections are a basic component of the Transformer architecture: each layer adds its output to the representation received from the preceding layer (Vaswani et al., 2017). In a pre-normalized Transformer (Xiong et al., 2020), the hidden state is therefore the running sum of the embedding and of all earlier layer outputs, each with unit weight. This accumulation makes the norm of the hidden state grow with depth and dilutes the contribution of any single layer (Sun et al., 2025). Attention Residuals (Kimi Team, 2026) replace that fixed accumulation by learned attention over the initial embedding and preceding layer outputs, allowing each layer to select which earlier representations to reuse. This replacement changes the associated dynamics. A standard residual stream is a running sum, whereas an Attention Residual stream is a normalized, depth-dependent average of stored outputs whose weights may be revised at later layers. Consequently, a dynamical analysis of ordinary residual connections cannot be transferred to Attention Residuals without first accounting for this memory and reweighting mechanism.

Continuous-time analyses have modeled self-attention as an interacting token system. For self-attention with residual connections, tokens form clusters (Geshkovski et al., 2023; 2025; Karagodin et al., 2024), and the geometry of the limiting clusters is governed by the spectrum of the value matrix. For causal attention, Karagodin et al. (2024) prove single-cluster convergence for arbitrary query–key matrices when the value matrix is the identity, and they propose a broader conjectural classification for nonidentity value matrices. Abella et al. (2025) prove consensus for a symmetric value matrix and tokens that start in the open hemisphere of its simple leading eigenvector. Pham et al. (2025) study continuous token dynamics in pretrained Transformers and identify parameter-dependent convergence and divergence regimes. These works establish that token behavior depends strongly on the attention parameters and residual dynamics, but they do not model the receiving-layer-dependent reweighting of stored outputs that defines Attention Residuals. The resulting gap is the motivation for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

this paper: we ask whether spectral clustering persists after additive residual connections are replaced by Attention Residuals, and which assumptions are needed for such a conclusion.

Motivated by Attention Residuals, we study a normalized continuous-depth model in which prescribed history kernels reweight earlier outputs. The resulting Volterra equation retains memory of the trajectory. Common reweighting closes this equation as a coupled system for token directions and radii. Lemma 3.3 also gives a finite-horizon quadrature limit when the same prescribed kernel is used on every refined grid. It does not establish that the content-derived depth scores of the original architecture satisfy that scaling or that a fixed-grid network clusters at infinite depth.

Our main theoretical results follow the three stages of the analysis:

- **Coupled ODE for Attention Residuals (Section 3).** We derive an exact coupled ODE for token directions and radii under common reweighting of stored contributions (Theorem 3.4). The system captures how depth aggregation controls each token's directional update rate through its radius and prescribed history rate.
- **Explicit first-token dynamics (Section 4, Lemma 4.3).** We solve the first token's direction explicitly as a normalized matrix exponential in effective time. Its initial leading projection determines which leading-eigenvector pole it approaches, while its radius converges to the largest eigenvalue. This provides the starting point for causal induction.
- **Almost-everywhere two-cluster convergence (Section 5, Theorem 5.1).** We prove that every token direction converges to one of the two leading-eigenvector poles for almost every admissible raw initial configuration. We also determine each positive limiting radius explicitly from the limiting cluster memberships in its causal prefix. Thus, tokens with the same limiting direction can retain distinct raw embeddings; both poles need not be occupied.

These results give sufficient conditions for clustering in the normalized, one-head causal model with query, key, and value matrices shared across depth. The history kernels are prescribed and held fixed as the raw initial embeddings vary. The almost-everywhere conclusion concerns this raw initial state space, including the radii; it does not by itself give the same assertion for every fixed tuple of initial radii. The results do not claim a complete classification for arbitrary trained Transformers.

2 Background

Additive residual connections We consider a sequence of N tokens with representations in $\mathbb{R}^d$; i indexes token position and n indexes network depth. Let $\boldsymbol{u}_i \in \mathbb{R}^d$ be the initial embedding of token i, and let $\boldsymbol{Z}^n = (\boldsymbol{z}_1^n, \dots, \boldsymbol{z}_N^n)$ collect the token representations at depth n. Write $G_i^{(n)}(\boldsymbol{Z}^n)$ for the output of layer n at token i, including its input normalization. An additive residual update has the form $\boldsymbol{z}_i^{n+1} = \boldsymbol{z}_i^n + G_i^{(n)}(\boldsymbol{Z}^n)$, with $\boldsymbol{z}_i^0 = \boldsymbol{u}_i$. Unrolling the recurrence gives $\boldsymbol{z}_i^n = \boldsymbol{u}_i + \sum_{k=0}^{n-1} G_i^{(k)}(\boldsymbol{Z}^k)$. Thus each stored output enters the accumulated representation with unit coefficient. The update does not explicitly assign new weights to individual earlier outputs when forming the input to a later layer.

Attention over depth. Full Attention Residuals replace additive accumulation with a softmax-weighted combination of the initial embedding and preceding layer outputs (Kimi Team, 2026):

$$\begin{aligned} \boldsymbol{z}_i^n &= \alpha_{i,0}^n \boldsymbol{u}_i + \sum_{k=0}^{n-1} \alpha_{i,k+1}^n \boldsymbol{G}_i^{(k)}(\boldsymbol{Z}^k), \\ \alpha_{i,\ell}^n &= \frac{\exp\{\boldsymbol{w}_n^\top \boldsymbol{k}_{i,\ell}\}}{\sum_{j=0}^{n} \exp\{\boldsymbol{w}_n^\top \boldsymbol{k}_{i,j}\}}, \qquad 0 \le \ell \le n. \end{aligned} \tag{1}$$

Here $\boldsymbol{w}_n$ is the learned query at depth n, while $\boldsymbol{k}_{i,0}$ and $\boldsymbol{k}_{i,k+1}$ are the RMS-normalized keys of $\boldsymbol{u}_i$ and $\boldsymbol{G}_i^{(k)}(\boldsymbol{Z}^k)$, respectively. The stored outputs serve as values. A Block Attention Residual variant also exists; our analysis focuses on Full Attention Residuals (Kimi Team, 2026).

With a shared query $\boldsymbol{w}_n = \boldsymbol{w}$ and fixed keys for stored sources, adding one source rescales all existing weights by the same factor. Section 3 uses this common-reweighting property to motivate ODE closure. Sharing the depth query is an additional restriction and is distinct from sharing the

token-attention matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$. The discrete keys depend on the token representations; our continuous model instead prescribes history kernels and holds them fixed as the initial embeddings vary.

Spherical causal attention. Continuous-depth models describe token interactions as dynamics on the unit sphere (Geshkovski et al., 2023; Karagodin et al., 2024). For a single head with fixed $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, with $\boldsymbol{x}_i \in \mathbb{S}^{d-1}$. Causal attention restricts each token to sources $j \le i$. With $p_{ij}(\boldsymbol{X})$ obtained by applying softmax to the scores $d^{-1/2}\boldsymbol{x}_i^\top \boldsymbol{Q}^\top \boldsymbol{K}\boldsymbol{x}_j$ over these sources, the standard spherical dynamics are

$$\dot{\boldsymbol{x}}_i = \boldsymbol{P}_{\boldsymbol{x}_i}\left(\sum_{j\le i} p_{ij}(\boldsymbol{X})\boldsymbol{V}\boldsymbol{x}_j\right), \qquad \boldsymbol{P}_{\boldsymbol{x}}\mathbf{y} := \mathbf{y} - \langle \boldsymbol{x}, \mathbf{y}\rangle \boldsymbol{x}. \tag{2}$$

The projection $\boldsymbol{P}_{\boldsymbol{x}}$ keeps the velocity tangent to the sphere. Since $p_{11} = 1$, the first token evolves independently; when $\boldsymbol{V} = \boldsymbol{I}$, it is stationary and almost every initial configuration converges to its direction (Karagodin et al., 2024).

For an Attention Residual aggregate, both its radius $r_i = \|\boldsymbol{z}_i\|$ and direction $\boldsymbol{x}_i = \boldsymbol{z}_i/r_i$ evolve. Spherical normalization is a modeling idealization, distinct from RMSNorm. Under common reweighting, Section 3 derives a radial equation and multiplies the velocity in (2) by $\gamma_i(t)/r_i(t)$, where γ_i is the prescribed history rate. Positive radial bounds keep normalization defined; bounded radii and $\int_0^\infty \gamma_i(t)\,dt = \infty$ give each token unbounded effective time. We discuss related work in Appendix B.

3 Continuous-time ODE of Attention Residuals

3.1 Continuous-time integral equation

We introduce a continuous-depth memory model motivated by the Full Attention Residual recurrence. Fix a depth interval $[0, T]$ and a partition $0 = t_0 < t_1 < \cdots < t_m = T$, identifying layer k with t_k. This identification alone does not assert that a sequence of networks converges as the partition is refined; such a claim also requires control of the depth weights across partitions.

Starting from the Full Attention Residual recurrence (1), introduce a depth-dependent field $\boldsymbol{G}_i(t, \boldsymbol{Z})$ that agrees with the layer maps at the partition points: $\boldsymbol{G}_i(t_k, \boldsymbol{Z}) = \boldsymbol{G}_i^{(k)}(\boldsymbol{Z})$ for $k = 0, \dots, m-1$. The same recurrence then takes the time-indexed form

$$\boldsymbol{z}_i^n = \alpha_{i,0}^n \boldsymbol{u}_i + \sum_{k=0}^{n-1} \alpha_{i,k+1}^n \boldsymbol{G}_i(t_k, \boldsymbol{Z}^k). \tag{3}$$

The stored value $\boldsymbol{G}_i(t_k, \boldsymbol{Z}^k)$ is fixed after layer k; only its depth weight changes. Suppose, for a given partition, that functions a_i and K_i satisfy $\alpha_{i,0}^n = a_i(t_n)$ and

$$\alpha_{i,k+1}^n = \int_{t_k}^{t_{k+1}} K_i(t_n, s)\,ds. \tag{4}$$

Assume that these functions are nonnegative and that $a_i(t) + \int_0^t K_i(t, s)\,ds = 1$. The discrete recurrence is then a cellwise quadrature of the following memory equation:

$$\boldsymbol{z}_i(t) = a_i(t)\boldsymbol{u}_i + \int_0^t K_i(t, s)\boldsymbol{G}_i(s, \boldsymbol{Z}(s))\,ds. \tag{5}$$

The following theorem establishes an existence and uniqueness result for the solution of this equation, under some regularity conditions of the history kernel $K_i(t, s)$ and function $\boldsymbol{G}_i(t, \boldsymbol{Z})$.

Theorem 3.1 (Well-posedness of the continuous-depth memory equation)**.** *Fix $T > 0$ and define $D_T := \{(t, s) \in \mathbb{R}^2 : 0 \le s \le t \le T\}$. For every $i \in \{1, \dots, N\}$, suppose that $a_i \in C^1([0, T])$, that $K_i \in C(D_T)$, and that $\partial_t K_i$ exists in the interior of D_T and extends continuously to D_T.*

Suppose also that $\boldsymbol{G} : [0,\infty) \times \mathbb{R}^{Nd} \longrightarrow \mathbb{R}^{Nd}$ *is continuous and globally Lipschitz in its state variable, uniformly in time: there exists* $L \geq 0$ *such that*

$$\|\boldsymbol{G}(t,\boldsymbol{Y}) - \boldsymbol{G}(t,\boldsymbol{Z})\|_\infty \leq L\|\boldsymbol{Y} - \boldsymbol{Z}\|_\infty$$

for every $t \geq 0$ *and* $\boldsymbol{Y}, \boldsymbol{Z} \in \mathbb{R}^{Nd}$*. Then the Volterra equation*

$$\boldsymbol{z}_i(t) = a_i(t)\boldsymbol{u}_i + \int_0^t K_i(t,s)\boldsymbol{G}_i(s,\boldsymbol{Z}(s))\,ds, \qquad 0 \leq t \leq T, \tag{6}$$

has a unique solution $\boldsymbol{Z} \in C^1([0,T];\mathbb{R}^{Nd})$*. Its derivative is*

$$\dot{\boldsymbol{z}}_i(t) = \dot{a}_i(t)\boldsymbol{u}_i + K_i(t,t)\boldsymbol{G}_i(t,\boldsymbol{Z}(t)) + \int_0^t \partial_t K_i(t,s)\boldsymbol{G}_i(s,\boldsymbol{Z}(s))\,ds. \tag{7}$$

If, moreover, $a_i \in C^1_{\mathrm{loc}}([0,\infty))$ *and the stated regularity assumptions on* K_i *and* $\partial_t K_i$ *hold on* D_S *for every finite* $S > 0$*, then (5) admits a unique global solution* $\boldsymbol{Z} \in C^1_{\mathrm{loc}}([0,\infty);\mathbb{R}^{Nd})$*.*

The proof of Theorem 3.1 is given in Appendix C.1. We next identify a concrete condition on the Attention Residuals weights that give the common reweighting structure needed to reduce this memory equation to an ODE.

3.2 Coupled dynamics of attention layers

For convenience, define $\boldsymbol{W} = \boldsymbol{Q}^\top \boldsymbol{K}$ and

$$p_{ij}(\boldsymbol{X}) = \frac{\exp\{d^{-1/2}\boldsymbol{x}_i^\top \boldsymbol{W}\boldsymbol{x}_j\}}{\sum_{\ell \leq i}\exp\{d^{-1/2}\boldsymbol{x}_i^\top \boldsymbol{W}\boldsymbol{x}_\ell\}}; \boldsymbol{F}_i(\boldsymbol{X}) = \sum_{j\leq i} p_{ij}(\boldsymbol{X})\boldsymbol{V}\boldsymbol{x}_j. \tag{8}$$

Throughout this work, we impose the following assumptions.

Assumption 1. *(a) The matrices* $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{m\times d}$ *and the symmetric matrix* $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ *are fixed across depth. Write* $\boldsymbol{W} = \boldsymbol{Q}^\top \boldsymbol{K}$*. For a nonzero raw token* $\boldsymbol{z}_i$*, set* $\boldsymbol{x}_i = \boldsymbol{z}_i/\|\boldsymbol{z}_i\|$*. The layer output is the causal attention field* $\boldsymbol{G}_i(\boldsymbol{Z}) = \boldsymbol{F}_i(\boldsymbol{X}) := \sum_{j\leq i} p_{ij}(\boldsymbol{X})\boldsymbol{V}\boldsymbol{x}_j$*, with the softmax weights in (8) and score normalization* $1/\sqrt{d}$*.*

(b) By change of orthogonal basis, suppose that the matrix $\boldsymbol{V}$ *is diagonal, i.e.*

$$\boldsymbol{V} = \mathrm{diag}(\lambda_1,\ldots,\lambda_d), \qquad \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d, \qquad \lambda := \lambda_1 > 0. \tag{9}$$

(c) The score matrix has the block form

$$\boldsymbol{W} = \begin{pmatrix} m_0 & \boldsymbol{0}^\top \\ \boldsymbol{0} & m_1\boldsymbol{I}_{d-1} + \boldsymbol{B} \end{pmatrix}, \tag{10}$$

where the matrix $\boldsymbol{B}$ *and scalar* m_0, m_1 *satisfy* $\boldsymbol{B}^\top = -\boldsymbol{B}, m_0 > 0, m_1 < m_0 + \frac{\sqrt{d}}{2}$*. Moreover, it holds*

$$\eta := \lambda_d - \lambda_1\sqrt{\frac{2\sqrt{d}\log N + 4\rho}{m_0}} > 0. \tag{11}$$

where $\rho := \sqrt{(m_1 - m_0)^2 + \|\boldsymbol{B}\|_{\mathrm{op}}^2}$*.*

The condition $\eta > 0$ has several quantitative consequences. Since it implies $\lambda_d > 0$, write $\kappa = \lambda_1/\lambda_d \geq 1$. Rearranging (11) gives $m_0 > \kappa^2\big(2\sqrt{d}\log N + 4\rho\big)$. In particular, $\rho < m_0/4$ and $m_1 \geq m_0 - \rho > 3m_0/4 > 0$. Because $\boldsymbol{B}$ is skew-symmetric, $m_1\boldsymbol{I}_{d-1} + \boldsymbol{B}$ is invertible: if it annihilates a real vector $\boldsymbol{y}$, its inner product with $\boldsymbol{y}$ gives $m_1\|\boldsymbol{y}\|^2 = 0$. Hence $\boldsymbol{W}$ has rank d. Since $\boldsymbol{W} = \boldsymbol{Q}^\top\boldsymbol{K}$ with $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{m\times d}$, this regime requires $m \geq d$.

The certificate controls the attention score scale as well as the value-matrix condition number. For example, when $\boldsymbol{W} = m_0\boldsymbol{I}_d$, it requires $m_0/\sqrt{d} > 2\kappa^2\log N$. The regime is nonempty: $d = N = m = 2$, $\boldsymbol{V} = \mathrm{diag}(2,1)$, and $\boldsymbol{Q} = \boldsymbol{K} = \sqrt{10}\,\boldsymbol{I}_2$ give $\boldsymbol{W} = 10\boldsymbol{I}_2$ and $\eta \approx 0.1144 > 0$. The positivity of η supplies the uniform radial bound used to keep normalization defined. The block alignment makes the limiting field tractable along the leading axis, while $m_1 < m_0 + \sqrt{d}/2$ makes the scalar function G_μ in (24) strictly increasing, as shown in (25).

Assumption 2 (History kernel and coupled initial-value problem). *(a) The functions a_i and K_i are nonnegative. For every finite $T > 0$, $a_i \in C^1([0,T])$, and K_i and its receiving-time derivative $\partial_t K_i$ are continuous on $D_T = \{(t,s) : 0 \le s \le t \le T\}$, with one-sided boundary derivatives. For every $t \ge 0$,*

$$a_i(t) + \int_0^t K_i(t,s)\, ds = 1. \tag{12}$$

There is a continuous rate $\gamma_i : [0,\infty) \to (0,\infty)$ such that $\dot{a}_i = -\gamma_i a_i$ and $\partial_t K_i(t,s) = -\gamma_i(t) K_i(t,s)$ for $s < t$. Moreover, for all $i \in [N]$, it holds $\int_0^\infty \gamma_i(t)\, dt = \infty$.

(b) The initial tokens are nonzero: $\boldsymbol{z}_i(0) = \boldsymbol{u}_i \neq \boldsymbol{0}$. Thus $r_i(0) = \|\boldsymbol{u}_i\|$ and $\boldsymbol{x}_i(0) = \boldsymbol{u}_i/\|\boldsymbol{u}_i\|$.

(c) The projection of $\boldsymbol{u}_1$ to the invariant subspace corresponding to eigenvalue λ_1, denote by $\mathcal{L}$, satisfies $\boldsymbol{P}_{\mathcal{L}}\boldsymbol{u}_1 \neq \boldsymbol{0}$. Set $\boldsymbol{\xi} := \boldsymbol{P}_{\mathcal{L}}\boldsymbol{u}_1/\|\boldsymbol{P}_{\mathcal{L}}\boldsymbol{u}_1\|$ and suppose $\dfrac{|\langle \boldsymbol{u}_i, \boldsymbol{\xi}\rangle|}{\|u_i\|} > 0$ for every i.

The next lemma concerns one fixed finite execution. Its shared-query condition makes the relative scores of already stored sources independent of the receiving layer.

Lemma 3.2. *Fix a finite partition and one realized trajectory of (1), including its stored source keys. Suppose that its depth query is shared across receiving layers, $\boldsymbol{w}_n = \boldsymbol{w}$. For each token i, there are functions a_i and K_i satisfying (4) and Assumption 2(a).*

The proof is given in Appendix C.2. A refinement result needs the same history kernel on every grid. We therefore first discretize the prescribed memory equation itself.

Lemma 3.3 (Finite-horizon quadrature limit). *Assume Assumption 1(a)–(c) and Assumption 2(a),(b). Fix the matrices, the initial embeddings, the history kernels, and $T > 0$. For each partition $0 = t_0 < \cdots < t_m = T$, let $h = \max_k (t_{k+1} - t_k)$ and prescribe the discrete depth weights by*

$$\alpha^n_{i,0} = a_i(t_n), \qquad \alpha^n_{i,k+1} = \int_{t_k}^{t_{k+1}} K_i(t_n, s)\, ds \quad (0 \le k < n \le m). \tag{13}$$

In (3), use the layer output $\boldsymbol{G}_i(t_k, \boldsymbol{Z}^k) = \boldsymbol{F}_i(\boldsymbol{X}^k)$ from (8). Let $\boldsymbol{Z}$ be the solution of (5) with the same fixed kernels and initial embeddings. Every discrete aggregate is nonzero, and there is a constant C_T, independent of the partition, such that

$$\max_{0 \le n \le m} \|\boldsymbol{Z}^n - \boldsymbol{Z}(t_n)\|_\infty \le C_T h.$$

The proof is given in Appendix C.3, along with some discussion on the relationship between the discrete recurrence and the derived continuous counterpart.

Theorem 3.4 (Directional and radial equations for attention layers). *Under Assumption 1(a) and Assumption 2(a),(b), the memory equation (5) is equivalent, while all aggregates are nonzero, to the Cauchy problem*

$$\begin{cases} \dot{\boldsymbol{z}}_i = \gamma_i(t) \left(\sum_{j \le i} p_{ij}(\boldsymbol{X}) \boldsymbol{V}\boldsymbol{x}_j - \boldsymbol{z}_i \right), \\ \boldsymbol{z}_i(0) = \boldsymbol{u}_i. \end{cases} \tag{14}$$

Writing $r_i = \|\boldsymbol{z}_i\|$ and $\boldsymbol{x}_i = \boldsymbol{z}_i/r_i$, the dynamics (14) is equivalent to

$$\begin{cases} \dot{\boldsymbol{x}}_i = \dfrac{\gamma_i(t)}{r_i} \sum_{j \le i} p_{ij}(\boldsymbol{X}) \left[\boldsymbol{V}\boldsymbol{x}_j - \langle \boldsymbol{x}_i, \boldsymbol{V}\boldsymbol{x}_j \rangle \boldsymbol{x}_i \right], \\ \dot{r}_i = \gamma_i(t) \left(\sum_{j \le i} p_{ij}(\boldsymbol{X}) \langle \boldsymbol{x}_i, \boldsymbol{V}\boldsymbol{x}_j \rangle - r_i \right), \\ \boldsymbol{x}_i(0) = \boldsymbol{u}_i/\|\boldsymbol{u}_i\|, \; r_i(0) = \|\boldsymbol{u}_i\|. \end{cases} \tag{15}$$

There is a unique maximal solution with positive radii. If its maximal forward endpoint $T_* < \infty$, then $\liminf_{t \uparrow T_*} \min_i r_i(t) = 0$.

Theorem 3.4 converts the history-dependent Attention Residuals model into coupled ODEs for token directions and radii. The proof of Theorem 3.4 is given in Appendix C.4.

4 LONG-TIME BEHAVIOUR OF FIRST TOKEN

This section establishes global existence for the coupled system and determines the long-time behavior of the first token. We give an explicit formula for its direction and prove that $\boldsymbol{x}_1(t) \to s_1\boldsymbol{e}$ and $r_1(t) \to \lambda$, where $s_1 = \operatorname{sign}\langle \boldsymbol{u}_1, \boldsymbol{e}\rangle$. These results provide the radial control and the base case for the causal induction in Section 5.

The following lemma gives a uniform positive lower bound on the radial forcing for every configuration of token directions. The proof of Lemma 4.1 is given in Appendix C.5.

Lemma 4.1. *For every configuration $\boldsymbol{X} \in (\mathbb{S}^{d-1})^N$ and every token i, it holds $\eta \le \langle \boldsymbol{x}_i, \boldsymbol{F}_i(\boldsymbol{X})\rangle \le \lambda$, where $\eta > 0$ is the constant in (11).*

Applying this estimate to the radial equation yields global positive radius bounds and an unbounded effective time, as established in the next lemma.

Lemma 4.2 (Global solution and effective time)**.** *The coupled system (15), initialized by $\boldsymbol{x}_i(0) = \boldsymbol{u}_i/\|\boldsymbol{u}_i\|$ and $r_i(0) = \|\boldsymbol{u}_i\|$, has a unique global solution. With $\Gamma_i(t) = \int_0^t \gamma_i(s)\,ds$, it holds*

$$e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + (1 - e^{-\Gamma_i(t)})\eta \le r_i(t) \le e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + (1 - e^{-\Gamma_i(t)})\lambda. \tag{16}$$

In particular, $r_i(t) \ge \min\{\|\boldsymbol{u}_i\|, \eta\} > 0$ and $r_i(t) \le M := \max\{\lambda, \max_j \|\boldsymbol{u}_j\|\}$. The function

$$\tau_i(t) := \int_0^t \frac{\gamma_i(s)}{r_i(s)}\,ds \ge \frac{\Gamma_i(t)}{M} \longrightarrow \infty, \tag{17}$$

and it has a continuously differentiable inverse t_i. The direction $\boldsymbol{y}(\tau) = \boldsymbol{x}_i(t_i(\tau))$ satisfies $\boldsymbol{y}' = (\boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^\top)\boldsymbol{F}_i(\boldsymbol{X}(t_i(\tau)))$.

The function τ_i is useful for studying a single direction. For comparisons between different initial raw tokens we will instead use $\Gamma_i(t)$, which depends only on the prescribed history rate. The first token attends only to itself. Its directional equation can be solved explicitly, so its limiting pole follows directly from the initial data.

Lemma 4.3. *Let $\boldsymbol{e} := \boldsymbol{e}_1$ denote the unit eigenvector corresponding to the simple leading eigenvalue λ. Then it holds*

$$\boldsymbol{x}_1(t) = \frac{e^{\tau_1(t)\boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau_1(t)\boldsymbol{V}}\boldsymbol{u}_1\|} \longrightarrow s_1\boldsymbol{e}, \qquad r_1(t) \longrightarrow \lambda, \tag{18}$$

where $s_1 = \operatorname{sign}\langle \boldsymbol{u}_1, \boldsymbol{e}\rangle$.

Lemma 4.3 completely determines the first-token dynamics: its initial leading projection selects its limiting pole, and its radius converges to the largest eigenvalue. This explicit limit supplies the base case for the causal induction developed in Section 5. The proof of Lemma 4.3 is given in Appendix C.7.

Remark 4.4. The directional limit in Lemma 4.3 agrees with the spectral description of single-token dynamics in Karagodin et al. (2024, Lemma 3.1) and the hemisphere attraction result for symmetric value matrices in Álvaro Rodríguez Abella et al. (2025, Lemma 5.2). In our model, the first direction follows the same spherical flow after the time change $\tau_1(t)$. The radial estimates ensure that $\tau_1(t) \to \infty$, even when $\gamma_1(t)$ is not bounded away from zero. They also yield $r_1(t) \to \lambda$, identifying the raw-token limit $\boldsymbol{z}_1(t) \to \lambda s_1\boldsymbol{e}$.

5 TWO-CLUSTERING REGIME FOR MULTIPLE-TOKEN SYSTEM

We now state our main result, regarding two-cluster behaviour of Attention Residuals recurrence. Under the stated matrix and history-kernel assumptions, almost every initial configuration with nonzero leading projections produces at most two directional clusters. Moreover, each radius converges, with a positive limiting radius determined explicitly by its limiting sign and the signs of its causal predecessors. The complete proof is given in Appendix C.13.

We now state the long-time result for the prescribed continuous system. It gives almost-everywhere directional clustering and explicit positive limiting radii. Its proof is in Appendix C.13.

Theorem 5.1 (Two-pole clustering by causal induction)**.** *Fix matrices satisfying Assumption 1(a)–(c) and one family of history kernels and rates satisfying Assumption 2(a). These matrices, kernels, and rates are held fixed as the raw initial embeddings vary. Among the initial configurations*

$$\mathcal{D} = \{(\boldsymbol{u}_1, \dots, \boldsymbol{u}_N) \in (\mathbb{R}^d)^N : \langle \boldsymbol{u}_i, \boldsymbol{e} \rangle \neq 0 \text{ for every } i\}, \tag{19}$$

there exists a Borel set $\mathcal{E}$, depending on the fixed model parameters, of zero Nd-dimensional Lebesgue measure such that, for every $U \in \mathcal{D} \setminus \mathcal{E}$, the unique global solution of (5) satisfies $\boldsymbol{x}_i(t) \to s_i \boldsymbol{e}$ for some $s_i \in \{-1, 1\}$ and every $1 \leq i \leq N$.

If $n_{i,s}$ counts the predecessors $j < i$ whose limiting sign is s, then

$$r_i(t) \longrightarrow R_i := \lambda \frac{(1 + n_{i,s_i}) e^{2m_0/\sqrt{d}} - n_{i,-s_i}}{(1 + n_{i,s_i}) e^{2m_0/\sqrt{d}} + n_{i,-s_i}} \geq \eta. \tag{20}$$

The proof follows the strategy as follow.

- Initially, the ith token evolves under a nonautonomous field $\boldsymbol{F}_i$ that can be approximated by a fixed limiting field $\overline{\boldsymbol{F}}_i$ determined only by the signs of those predecessors (Lemma 5.2).
- The limiting dynamics have a gradient-like structure, which forces the token either to approach one of the two leading unit eigenvectors, or to approach a equilibrium associated with a lower eigenspace (Lemma 5.3).
- We then show that the later case occur within a set of initial data that have zero Lebesgue measure (Lemma 5.5). Hence, outside a null set of initial states, token i must converge to one of two unit leading eigenvectors, whenever all its predecessors do (Lemma 5.6 and Theorem 5.1).

The limiting field. Fix an index $i \geq 2$ and suppose that $\boldsymbol{x}_j(t) \longrightarrow s_j \boldsymbol{e}, s_j \in \{-1, 1\}, j < i$. Thus, every predecessor of the ith token converges to one of the two poles $\boldsymbol{e}$ and $-\boldsymbol{e}$. Let $n_+ := \#\{j < i : s_j = 1\}, n_- := \#\{j < i : s_j = -1\}$. In particular, $n_+ + n_- = i - 1$. The *limiting field* $\overline{\boldsymbol{F}}_i$ is obtained by replacing each predecessor $\boldsymbol{x}_j(t)$, with $j < i$, by its limiting value $s_j \boldsymbol{e}$, while keeping the direction of the ith token as a free variable. More precisely, for every $\boldsymbol{x} \in \mathbb{S}^{d-1}$, define

$$\overline{\boldsymbol{F}}_i(\boldsymbol{x}) := \boldsymbol{F}_i(s_1 \boldsymbol{e}, \dots, s_{i-1} \boldsymbol{e}, \boldsymbol{x}). \tag{21}$$

Notice also that $\overline{\boldsymbol{F}}_i$ is still a function of the current direction $\boldsymbol{x}$. We now compute this field explicitly. Set $q := \langle \boldsymbol{x}, \boldsymbol{e} \rangle$. By the block structure of $\boldsymbol{W}$, it holds $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} = m_1 + (m_0 - m_1) q^2$, $\boldsymbol{x}^\top \boldsymbol{W} (\pm \boldsymbol{e}) = \pm m_0 q$. Consequently, the exponential score of the current token attending to itself is

$$E(q) := \exp\left(\frac{m_1 + (m_0 - m_1) q^2}{\sqrt{d}} \right).$$

Each predecessor whose limiting direction is $\boldsymbol{e}$ contributes the score $e^{m_0 q/\sqrt{d}}$, whereas each predecessor whose limiting direction is $-\boldsymbol{e}$ contributes the score $e^{-m_0 q/\sqrt{d}}$. Hence the limiting softmax denominator is

$$\mathcal{Z}(q) := E(q) + n_+ e^{m_0 q/\sqrt{d}} + n_- e^{-m_0 q/\sqrt{d}}.$$

It is convenient to write $\alpha(q) := \frac{E(q)}{\mathcal{Z}(q)}, \chi(q) := \frac{n_+ e^{m_0 q/\sqrt{d}} - n_- e^{-m_0 q/\sqrt{d}}}{\mathcal{Z}(q)}$. Here $\alpha(q)$ is the limiting attention weight assigned to the ith token itself. The quantity $\chi(q)$ is the total attention weight assigned to predecessors converging to $\boldsymbol{e}$, minus the total attention weight assigned to predecessors converging to $-\boldsymbol{e}$. Because $\boldsymbol{V}\boldsymbol{e} = \lambda \boldsymbol{e}$, we also have $\boldsymbol{V}(-\boldsymbol{e}) = -\lambda \boldsymbol{e}$. The self-attention contribution is therefore $\alpha(q) \boldsymbol{V} \boldsymbol{x}$, while the total contribution of all predecessors is

$$\frac{n_+ e^{m_0 q/\sqrt{d}} \lambda \boldsymbol{e} - n_- e^{-m_0 q/\sqrt{d}} \lambda \boldsymbol{e}}{\mathcal{Z}(q)} = \lambda \chi(q) \boldsymbol{e}.$$

Combining these two contributions gives the explicit formula

$$\overline{\boldsymbol{F}}_i(\boldsymbol{x}) = \alpha(q) \boldsymbol{V} \boldsymbol{x} + \lambda \chi(q) \boldsymbol{e}, \qquad q = \langle \boldsymbol{x}, \boldsymbol{e} \rangle. \tag{22}$$

The following lemma shows that, once the preceding tokens converge, the next raw token follows an asymptotically autonomous equation whose limiting field is $\overline{\boldsymbol{F}}_i$. We therefore analyze the equilibria and long-time behavior of this limiting field before returning to the original nonautonomous dynamics. The proof of Lemma 5.2 is given in Appendix C.8.

Lemma 5.2. *Fix the predecessor initial data and suppose $\boldsymbol{x}_j(t) \to s_j \boldsymbol{e}$ for $j < i$. Let $s = \Gamma_i(t) = \int_0^t \gamma_i(v)\, dv$ and let $\vartheta_i = \Gamma_i^{-1}$. Every choice of $\boldsymbol{u}_i \neq \boldsymbol{0}$ then solves the same nonautonomous raw equation*

$$\frac{d\boldsymbol{z}}{ds} = \boldsymbol{F}_i\left(\boldsymbol{x}_1(\vartheta_i(s)), \dots, \boldsymbol{x}_{i-1}(\vartheta_i(s)), \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\right) - \boldsymbol{z} = \overline{\boldsymbol{F}}_i\left(\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\right) - \boldsymbol{z} + \boldsymbol{R}_i(s, \boldsymbol{z}). \tag{23}$$

The remainder tends to zero in C^1 on every compact annulus $0 < a \le \|\boldsymbol{z}\| \le b < \infty$.

Equilibria of the limiting field. We next characterize its nonpolar directional equilibria of the limiting directional field. Let $\boldsymbol{x} \in \mathbb{S}^{d-1}$ be different from $\pm\boldsymbol{e}$, define $q := \langle \boldsymbol{x}, \boldsymbol{e}\rangle$. For $\boldsymbol{w} \perp \boldsymbol{e}, \quad \|\boldsymbol{w}\| = 1$, write

$$\boldsymbol{x} = q\boldsymbol{e} + \sqrt{1-q^2}\boldsymbol{w}.$$

At an equilibrium, the projected field $(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\top)\overline{\boldsymbol{F}}_i(\boldsymbol{x})$ vanishes, so $\overline{\boldsymbol{F}}_i(\boldsymbol{x})$ must be parallel to $\boldsymbol{x}$. Projecting this condition onto the eigenspaces of $\boldsymbol{V}$ shows that $\boldsymbol{w}$ must belong to a single lower eigenspace $H_\mu := \ker(\boldsymbol{V} - \mu\boldsymbol{I})$, where $\mu < \lambda$. Comparing the components along $\boldsymbol{e}$ and H_μ then gives $\alpha(q)(\lambda - \mu)q + \lambda\chi(q) = 0$. After multiplication by the positive denominator $\mathcal{Z}(q)$, this condition becomes

$$G_\mu(q) := (\lambda - \mu)qE(q) + \lambda\left(n_+ e^{m_0 q/\sqrt{d}} - n_- e^{-m_0 q/\sqrt{d}}\right) = 0. \tag{24}$$

Thus, the roots of G_μ determine the possible leading coordinates of nonpolar equilibria associated with H_μ. The block condition implies that G_μ is strictly increasing. Indeed,

$$G'_\mu(q) = (\lambda - \mu)E(q)\left[1 + \frac{2(m_0 - m_1)q^2}{\sqrt{d}}\right] + \frac{\lambda m_0}{\sqrt{d}}\left(n_+ e^{m_0 q/\sqrt{d}} + n_- e^{-m_0 q/\sqrt{d}}\right) > 0. \tag{25}$$

The strict inequality follows since $1 + \frac{2(m_0-m_1)q^2}{\sqrt{d}} \ge \min\left\{1,\, 1 + \frac{2(m_0-m_1)}{\sqrt{d}}\right\} > 0$. by the assumption $m_1 < m_0 + \sqrt{d}/2$. Consequently, G_μ has at most one root in $(-1, 1)$. If such a root exists, denote it by q_μ and define

$$\begin{aligned} \mathcal{M}_\mu &:= \left\{ q_\mu \boldsymbol{e} + \sqrt{1-q_\mu^2}\,\boldsymbol{w} : \boldsymbol{w} \in H_\mu,\ \|\boldsymbol{w}\| = 1 \right\}, \\ \mathcal{K}_\mu &:= \{\alpha(q_\mu)\mu\, \boldsymbol{x} : \boldsymbol{x} \in \mathcal{M}_\mu\}. \end{aligned} \tag{26}$$

The set $\mathcal{M}_\mu$ consists of the corresponding equilibrium directions. For every $\boldsymbol{x} \in \mathcal{M}_\mu$, the root identity gives $\overline{\boldsymbol{F}}_i(\boldsymbol{x}) = \alpha(q_\mu)\mu\boldsymbol{x}$; hence $\mathcal{K}_\mu$ is the associated equilibrium set for the raw limiting equation. If G_μ has no root in $(-1, 1)$, then no nonpolar equilibrium is associated with the eigenspace H_μ.

Lemma 5.3. *Fix $i \ge 2$ and suppose that the predecessor directions satisfy $\boldsymbol{x}_j(t) \longrightarrow s_j\boldsymbol{e}$, $s_j \in \{-1, 1\}$, $j < i$. Then one of the following alternatives holds*

1. *the ith direction converges to one of the two unit eigenvector, i.e. $\boldsymbol{x}_i(t) \longrightarrow s\boldsymbol{e}$ for some $s \in \{-1, 1\}$.*
2. *there is an eigenvalue $\mu < \lambda$ for which G_μ has a root $q_\mu \in (-1, 1)$, and $\operatorname{dist}\big(\boldsymbol{z}_i(t), \mathcal{K}_\mu\big) \longrightarrow 0$.*

Lemma 5.3 reduces the possible limits of the ith token to two alternatives: convergence to a leading-eigenvector or convergence toward a nonpolar equilibrium set $\mathcal{K}_\mu$. To prove almost-everywhere two-pole convergence, it remains to show that the second alternative occurs only for a null set of initial states. The expanding-coordinate comparison established next provides the required measure estimate. The proof of Lemma 5.3 is given in Appendix C.9.

The following cone comparison shows that, when one coordinate expands, the initial states whose trajectories remain in a bounded region form at most a graph over the nonexpanding coordinates; related graph characterizations appear in Zgliczyński (2009); Muşat & Boumal (2026). The proof of Lemma 5.4 is given in Appendix C.10.

Lemma 5.4. *Let U be a bounded region in $\mathbb{R}^a \times \mathbb{R}^b$, where $a \geq 1$, and consider the common nonautonomous equation*

$$\boldsymbol{p}' = \boldsymbol{A}\boldsymbol{p} + \boldsymbol{f}(s, \boldsymbol{p}, \boldsymbol{h}), \qquad \boldsymbol{h}' = \boldsymbol{C}\boldsymbol{h} + \boldsymbol{g}(s, \boldsymbol{p}, \boldsymbol{h}). \tag{27}$$

Assume that $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric, $\boldsymbol{A} \succeq \kappa \boldsymbol{I}$ for some $\kappa > 0$, and $\boldsymbol{C} \preceq 0$. Suppose there are $S \geq 0$ and $\epsilon < \kappa/4$ such that, for every $s \geq S$ and every $(\boldsymbol{p}, \boldsymbol{h}), (\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{h}}) \in U$,

$$\begin{aligned} \|\boldsymbol{f}(s, \boldsymbol{p}, \boldsymbol{h}) - \boldsymbol{f}(s, \widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{h}})\| &\leq \epsilon\big(\|\boldsymbol{p} - \widetilde{\boldsymbol{p}}\| + \|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|\big), \\ \|\boldsymbol{g}(s, \boldsymbol{p}, \boldsymbol{h}) - \boldsymbol{g}(s, \widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{h}})\| &\leq \epsilon\big(\|\boldsymbol{p} - \widetilde{\boldsymbol{p}}\| + \|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|\big). \end{aligned} \tag{28}$$

Fix $S_0 \geq S$, and let $\mathcal{T}_{S_0}$ be the set of initial states in U whose solutions exist and remain in U for every $s \geq S_0$. For each fixed $\boldsymbol{h}_0$, the section $\{\boldsymbol{p}_0 : (\boldsymbol{p}_0, \boldsymbol{h}_0) \in \mathcal{T}_{S_0}\}$ contains at most one point. Thus, if $\mathcal{T}_{S_0}$ is Borel, it has zero Lebesgue measure in $\mathbb{R}^a \times \mathbb{R}^b$.

Lemma 5.5 applies the preceding comparison argument to the Attention Residuals dynamics and shows that the basin of each nonpolar limiting set is null. Combining this fact with the alternatives in Lemma 5.3 yields the one-token induction step below. The proof of Lemma 5.5 is given in Appendix C.11.

Lemma 5.5. *Fix $i \geq 2$ and the predecessor initial data, and suppose that $\boldsymbol{x}_j(t) \to s_j \boldsymbol{e}$ for every $j < i$. For each $\mu < \lambda$ for which $\mathcal{K}_\mu$ is defined, the set*

$$\mathcal{E}_{i,\mu} := \big\{\boldsymbol{u}_i \in \mathbb{R}^d : \langle \boldsymbol{u}_i, \boldsymbol{e}\rangle \neq 0, \quad \operatorname{dist}\big(\boldsymbol{z}_i(t; \boldsymbol{u}_i), \mathcal{K}_\mu\big) \longrightarrow 0\big\}$$

is Borel and has zero d-dimensional Lebesgue measure.

Lemma 5.6 proves the conditional induction step: once all preceding tokens converge to the two unit leading eigenvectors, the next token does so for almost every choice of its initial state.

Lemma 5.6. *Fix the initial data $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}$ and suppose that their solutions satisfy $\boldsymbol{x}_j(t) \to s_j \boldsymbol{e}$ for every $j < i$. There is a Borel set $\mathcal{E}_i \subset \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ of zero Lebesgue measure such that every admissible $\boldsymbol{u}_i \notin \mathcal{E}_i$ satisfies $\boldsymbol{x}_i(t) \to \boldsymbol{e}$ or $-\boldsymbol{e}$.*

The proof of Lemma 5.6 is given in Appendix C.12. Applying this result successively along the causal order produces the global clustering result (Theorem 5.1).

Remark 5.7. The two-cluster behavior was conjectured for spherical causal attention by Karagodin et al. (2024, Conjecture 1), while Álvaro Rodríguez Abella et al. (2025, Theorem 5.2) established consensus for initial directions lying in a common hemisphere determined by the leading eigenvector. Under our matrix and history-kernel assumptions, we establish almost-everywhere two-pole convergence for the coupled Attention Residuals dynamics, allowing initial directions in both hemispheres. We also determine each token's positive limiting radius in terms of its limiting sign and those of its predecessors.

6 Conclusion

We developed a continuous-depth framework for Attention Residuals through a Volterra memory equation. Under common reweighting, this formulation admits an exact coupled equations for token directions and radii, revealing how history aggregation and token magnitudes jointly determine the dynamics. We established positive global well-posedness, solved the first-token dynamics explicitly, and used causal induction to prove almost-everywhere convergence to the two leading-eigenvector clusters. The result permits initial directions in both hemispheres and yields at most two occupied directional clusters.

Limitation. Our results only describe a sufficient almost-everywhere clustering regime for a single-head causal model. The analysis also leaves open a general rule for predicting the limiting cluster of later tokens from initialization.

Future work. Future work should extend the analysis to depth-dependent parameters and multiple attention heads, and characterize exceptional configurations more precisely. .

AI USE STATEMENT

We used generative AI tools to assist with LaTeX editing, manuscript organization and readability, and the identification of potentially relevant literature. We did not use these tools to develop the theoretical model or conceptual framework, formulate mathematical claims, construct or write proofs, propose or refine research hypotheses, or draw mathematical conclusions. All AI-assisted edits were critically reviewed by the authors, and all suggested references were read and verified before citation. The authors take full responsibility for the final content of this work, including its exposition, mathematical statements, proofs, figures, and accompanying artifacts.

ETHICS STATEMENT

Given the nature of the work, we do not foresee any negative societal and ethical impacts of our work.

REPRODUCIBILITY STATEMENT

All assumptions used in the theoretical results are stated explicitly in the main text. The paper contains no empirical experiments, datasets, or randomized computational results requiring separate reproduction.

REFERENCES

Álvaro Rodríguez Abella, João Pedro Silvestre, and Paulo Tabuada. Consensus is all you get: The role of attention in transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *SIAM Journal on Mathematics of Data Science*, 7(3):1367–1393, 2025. doi: 10.1137/24M167086X. URL https://doi.org/10.1137/24M167086X.

Martin Burger, Samira Kabri, Yury Korolev, Tim Roith, and Lukas Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization, 2025. URL https://arxiv.org/abs/2501.03096.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers, 2026. URL https://arxiv.org/abs/2501.18322.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, 2 edition, 1999. ISBN 978-0-471-31716-6.

Gerald B. Folland. *Advanced Calculus*. Author's online edition, 2 edition, 2023. Corrected online version.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model, 2024.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Mike Heddes, Adel Javanmard, Kyriakos Axiotis, Gang Fu, Mohammadhossein Bateni, and Vahab Mirrokni. DeepCrossAttention: Supercharging transformer residual connections. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri

Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22881–22903. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/heddes25a.html.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2411.04990.

Kimi Team. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026. doi: 10.48550/arXiv.2603.15031.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view, 2019. URL https://arxiv.org/abs/1906.02762.

Cheng Luo, Zefan Cai, and Junjie Hu. Delta attention residuals. *arXiv preprint arXiv:2605.18855*, 2026a.

Cheng Luo, Zefan Cai, and Junjie Hu. Multi-head attention residuals. *arXiv preprint arXiv:2607.27230*, 2026b.

Andreea-Alexandra Muşat and Nicolas Boumal. A non-autonomous center-stable set theorem for saddle avoidance in optimization, 2026.

Matteo Pagliardini, Amirkeivan Mohtashami, François Fleuret, and Martin Jaggi. Denseformer: Enhancing information flow in transformers via depth weighted averaging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2402.02622.

Duy-Tung Pham, An The Nguyen, Viet-Hoang Tran, Nhan-Phu Chung, Xin T. Tong, Tan M. Nguyen, and Thieu N. Vo. Dynamical properties of tokens in self-attention and effects of positional encoding. *arXiv preprint arXiv:2512.03058v1*, 2025. doi: 10.48550/arXiv.2512.03058.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3 edition, 1976. ISBN 978-0-07-054235-8.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention, 2022a.

Michael Eli Sander, Pierre Ablin, and Gabriel Peyré. Do residual neural networks discretize neural ordinary differential equations? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=1tCuRbPts3J.

Jonathan Su. Low-rank attention residuals. *arXiv preprint arXiv:2607.09694*, 2026.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *Advances in Neural Information Processing Systems*, 2025.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 9780821883280.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. doi: 10.48550/arXiv.1706.03762.

Thieu Vo, Duy-Tung Pham, Xin T. Tong, and Tan Minh Nguyen. Demystifying the token dynamics of deep selective state space models. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.03292.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 10524–10533. PMLR, 2020.

Xingda Yu, Yining Li, Xinzhang Liu, Zhihao Yang, Haowei He, Chao Wang, Yongxiang Li, and Shuangyong Song. Dual attention residuals. *arXiv preprint arXiv:2607.18730*, 2026.

Eberhard Zeidler. *Nonlinear Functional Analysis and Its Applications I: Fixed-Point Theorems*. Springer-Verlag, New York, 1986. ISBN 978-0-387-90914-1.

Piotr Zgliczyński. Covering relations, cone conditions and the stable manifold theorem. *Journal of Differential Equations*, 246(5):1774–1819, 2009. doi: 10.1016/j.jde.2008.12.019.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2409.19606.

Álvaro Rodríguez Abella, João Pedro Silvestre, and Paulo Tabuada. The asymptotic behavior of attention in transformers, 2025.

APPENDIX OF "DEPTH WITH MEMORY: TOKEN DYNAMICS OF ATTENTION RESIDUALS"

A Additional Background **14**

A.1 Existence, uniqueness, and continuation of ODE solutions 14

A.2 Grönwall inequality and scalar comparison 14

A.3 Banach fixed point theorem 15

A.4 Positive changes of time 15

A.5 Finite-time solution maps and null sets 16

B Related Work **16**

C Proofs of Theorems and Lemmas **17**

C.1 Proof of Theorem 3.1 17

C.2 Proof of Lemma 3.2 18

C.3 Proof of Lemma 3.3 19

C.4 Proof of Theorem 3.4 20

C.5 Proof of Lemma 4.1 20

C.6 Proof of Lemma 4.2 22

C.7 Proof of Lemma 4.3 24

C.8 Proof of Lemma 5.2 26

C.9 Proof of Lemma 5.3 27

C.10 Proof of Lemma 5.4 31

C.11 Proof of Lemma 5.5 32

C.12 Proof of Lemma 5.6 34

C.13 Proof of Theorem 5.1 34

A Additional Background

Throughout, $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^n$ and $\overline{B}_\delta(x_0) := \{x \in \mathbb{R}^n : |x - x_0| \le \delta\}$.

This appendix collects the background results used in the existence, radial-bound, and exceptional-set arguments. Textbook results are cited at their statements; the specialized consequences are derived below.

A.1 Existence, uniqueness, and continuation of ODE solutions

Let $U \subseteq \mathbb{R}^{n+1}$ be open, $f \in C(U, \mathbb{R}^n)$ and $(t_0, x_0) \in U$. Consider the initial value problem

$$\dot{x} = f(t, x), \qquad x(t_0) = x_0. \tag{29}$$

A solution of (29) on an interval $I \ni t_0$ is a function $x \in C^1(I, \mathbb{R}^n)$ with $(t, x(t)) \in U$ and $\dot{x}(t) = f(t, x(t))$ for all $t \in I$, and $x(t_0) = x_0$.

Definition A.1 (Local Lipschitz condition; Teschl, 2012, Section 2.2, eq. (2.18))**.** The map f is *locally Lipschitz continuous in the second argument, uniformly with respect to the first*, if for every compact set $V_0 \subset U$

$$L(V_0) := \sup_{(t,x)\neq(t,y)\in V_0} \frac{|f(t,x) - f(t,y)|}{|x - y|} < \infty.$$

Theorem A.2 (Picard–Lindelöf; Teschl, 2012, Theorem 2.2)**.** *Suppose $f \in C(U, \mathbb{R}^n)$, where $U \subseteq \mathbb{R}^{n+1}$ is open, and $(t_0, x_0) \in U$. If f is locally Lipschitz continuous in the second argument, uniformly with respect to the first, then there exists a unique local solution $x \in C^1(I)$ of (29), where I is some interval around t_0.*

More specifically, let $V = [t_0, t_0+T] \times \overline{B}_\delta(x_0) \subset U$ and let $M := \max_V |f|$. Then the solution exists at least for $t \in [t_0, t_0 + T_0]$ and remains in $\overline{B}_\delta(x_0)$, where $T_0 = \min\{T, \delta/M\}$ (with $\delta/M := \infty$ if $M = 0$). The analogous result holds on $[t_0 - T_0, t_0]$.

Theorem A.3 (Peano; Teschl, 2012, Theorem 2.19)**.** *Suppose f is continuous on $V = [t_0, t_0 + T] \times \overline{B}_\delta(x_0) \subset U$ and let $M := \max_V |f|$. Then there exists at least one solution of (29) on $[t_0, t_0 + T_0]$ which remains in $\overline{B}_\delta(x_0)$, where $T_0 = \min\{T, \delta/M\}$. The analogous result holds on $[t_0 - T_0, t_0]$.*

Theorem A.4 (Maximal extension and continuation; Teschl, 2012, Theorem 2.13 and Corollary 2.15)**.** *Under the hypotheses of Theorem A.2, the local solution has a unique maximal extension to an open interval (T_-, T_+) containing t_0. If a solution is defined up to a finite time b and, for some $a < b$ and compact $C \subset \mathbb{R}^n$, its trajectory satisfies*

$$[a, b] \times C \subset U, \qquad x(t) \in C \quad (a \le t < b),$$

then it extends beyond b. In particular, a finite maximal endpoint is incompatible with these compact-containment conditions.

For the raw attention equation, the state domain is $\Omega = \{(z_1, \dots, z_N) : z_i \neq 0 \text{ for every } i\}$. Bounds $0 < c \le |z_i(t)| \le C < \infty$ place the trajectory in a compact subset of Ω. On a finite time interval, these bounds permit continuation because the prescribed rates are continuous. An upper bound alone would not exclude approach to the boundary $z_i = 0$.

A.2 Grönwall inequality and scalar comparison

Lemma A.5 (Generalized Grönwall inequality; Teschl, 2012, Lemma 2.7)**.** *Let $\psi, \alpha, \beta \in C([0, T])$ with $\beta \ge 0$, and suppose*

$$\psi(t) \le \alpha(t) + \int_0^t \beta(s)\psi(s)\, ds, \qquad t \in [0, T].$$

Then

$$\psi(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp\Big(\int_s^t \beta(r)\, dr\Big)\, ds, \qquad t \in [0, T].$$

If, in addition, $\alpha(s) \le \alpha(t)$ for $s \le t$, then

$$\psi(t) \le \alpha(t) \exp\Big(\int_0^t \beta(s)\, ds\Big), \qquad t \in [0, T].$$

The scalar comparison used for the radii follows from the integrating-factor calculation underlying variation of constants; see Teschl (2012, Section 3.4, eq. (3.97)). We give the calculation to specify the inequality version precisely.

Lemma A.6 (Linear differential inequality)**.** *Let $a, b \in C([0,T];\mathbb{R})$ and $u \in C^1([0,T];\mathbb{R})$ satisfy $u'(t) + a(t)u(t) \le b(t)$. With $A(t) = \int_0^t a(s)\,ds$, one has*

$$u(t) \le e^{-A(t)}u(0) + \int_0^t e^{-[A(t)-A(s)]}b(s)\,ds.$$

Reversing the differential inequality reverses the conclusion; equality in the differential equation gives equality in this formula.

Proof. Multiply the differential inequality by the positive function $e^{A(t)}$. Its left side becomes $(e^{A(t)}u(t))'$. Integration from 0 to t and division by $e^{A(t)}$ give the result. □

In particular, if $r' = \gamma(t)(h(t) - r)$, where γ, h are continuous, $\gamma \ge 0$, and $\eta \le h(t) \le \Lambda$, then, writing $\Gamma(t) = \int_0^t \gamma(s)\,ds$,

$$e^{-\Gamma(t)}r(0) + (1 - e^{-\Gamma(t)})\eta \le r(t) \le e^{-\Gamma(t)}r(0) + (1 - e^{-\Gamma(t)})\Lambda.$$

This follows by applying the lemma twice and using $\int_0^t e^{-[\Gamma(t)-\Gamma(s)]}\gamma(s)\,ds = 1 - e^{-\Gamma(t)}$.

A.3 Banach fixed point theorem

Let $(X, \|\cdot\|)$ be a Banach space and $C \subseteq X$. A map $K : C \to C$ is a *contraction* if there is $\theta \in [0,1)$ such that $\|K(x) - K(y)\| \le \theta\|x - y\|$ for all $x, y \in C$. We write $K^0(x) = x$ and $K^n(x) = K(K^{n-1}(x))$.

Theorem A.7 (Contraction principle; Teschl, 2012, Theorem 2.1; see also Zeidler, 1986, Chapter 1)**.** *Let C be a nonempty closed subset of a Banach space X and let $K : C \to C$ be a contraction with constant $\theta \in [0,1)$. Then K has a unique fixed point $x^* \in C$, and*

$$\|K^n(x) - x^*\| \le \frac{\theta^n}{1-\theta}\,\|K(x) - x\|, \qquad x \in C,\ n \ge 0.$$

In particular, the Picard iterates $K^n(x)$ converge to x^* for every starting point $x \in C$.

Remark A.8 (Weighted supremum norm). For $X_T = C([0,T];\mathbb{R}^m)$ and $\beta > 0$, set

$$\|u\|_\beta = \sup_{0 \le t \le T} e^{-\beta t}|u(t)|.$$

The inequalities $e^{-\beta T}\|u\|_\infty \le \|u\|_\beta \le \|u\|_\infty$ show that this norm is equivalent to the supremum norm. A Cauchy sequence for $\|\cdot\|_\beta$ is therefore uniformly Cauchy; its pointwise limit exists in $\mathbb{R}^m$, and uniform convergence makes the limit continuous. Thus X_T is complete for this norm. If a Volterra operator satisfies

$$|(Ku)(t) - (Kv)(t)| \le H\int_0^t |u(s) - v(s)|\,ds,$$

then direct integration gives $\|Ku - Kv\|_\beta \le (H/\beta)\|u - v\|_\beta$. Choosing $\beta > H$ makes it a contraction, provided it maps the chosen nonempty closed set into itself. This is the application of Theorem A.7 used for the memory equation.

A.4 Positive changes of time

The following consequence of the inverse mapping theorem and chain rule explains the effective clocks; see Folland (2023, Theorems 3.18 and 2.26).

Lemma A.9 (Unbounded effective time)**.** *Let $a : [0,\infty) \to (0,\infty)$ be continuous, and set $\tau(t) = \int_0^t a(s)\,ds$. If $\int_0^\infty a(s)\,ds = \infty$, then τ is a strictly increasing bijection of $[0,\infty)$ onto itself. Its inverse $t = t(\tau)$ is continuously differentiable, with one-sided derivatives at zero, and*

$$\frac{dt}{d\tau} = \frac{1}{a(t(\tau))}.$$

If $x' = a(t)F(t,x)$, then $y(\tau) = x(t(\tau))$ satisfies $y' = F(t(\tau), y)$.

Proof. The fundamental theorem of calculus gives $\tau' = a > 0$. Monotonicity, continuity, and the divergent integral establish bijectivity. The inverse derivative formula and the equation for y follow from the inverse mapping theorem and the chain rule. □

In the paper, $a = \gamma_i/r_i$. The bounds $0 < r_i \le M$ and $\int_0^\infty \gamma_i = \infty$ imply $\tau_i(t) \ge M^{-1} \int_0^t \gamma_i(s)\, ds \to \infty$. Each token has its own clock; this calculation does not identify all clocks with a single common time change.

A.5 Finite-time solution maps and null sets

Write $\Phi_{t,s}(x)$ for the value at time t of the solution starting at x at time s. Under the hypotheses of Theorem A.2, continuous dependence and uniqueness give local inverse solution maps along each finite solution segment; see Teschl (2012, Theorems 2.8 and 2.9). More explicitly, on a compact neighborhood containing the nearby solution segments, let L be a uniform state Lipschitz constant. Applying Grönwall both forward and backward gives

$$e^{-L|t-s|}|x-y| \le |\Phi_{t,s}(x) - \Phi_{t,s}(y)| \le e^{L|t-s|}|x-y|.$$

The backward map is $\Phi_{s,t}$, since uniqueness gives $\Phi_{s,t} \circ \Phi_{t,s} = \mathrm{id}$ wherever these maps are defined. Thus the solution map and its local inverse are Lipschitz. This conclusion requires no differentiability of the coefficients in time.

Lemma A.10 (Transport of null sets)**.** *A locally Lipschitz map between open subsets of $\mathbb{R}^m$ sends sets of m-dimensional Lebesgue measure zero to sets of measure zero. Consequently, a locally bi-Lipschitz map also has null preimages of null sets in its target.*

Proof. First restrict to a region where the map has Lipschitz constant L. Cover a null set by cubes of side lengths ℓ_j with $\sum_j \ell_j^m < \varepsilon$. The image of the intersection with each cube has diameter at most $L\sqrt{m}\ell_j$, so it is contained in a cube of volume at most $(2L\sqrt{m})^m \ell_j^m$. Its outer measure is therefore at most $(2L\sqrt{m})^m \varepsilon$, which tends to zero with ε. A countable cover by Lipschitz neighborhoods proves the local assertion. Applying it to the local inverses proves the preimage assertion. □

Theorem A.11 (Tonelli, specialized to Borel sets; Folland, 1999, Theorem 2.37(a))**.** *Let $E \subset \mathbb{R}^p \times \mathbb{R}^q$ be Borel, and let $E_x = \{y : (x,y) \in E\}$. Writing $\mathcal{L}^k$ for k-dimensional Lebesgue measure, one has*

$$\mathcal{L}^{p+q}(E) = \int_{\mathbb{R}^p} \mathcal{L}^q(E_x)\, d\mathcal{L}^p(x).$$

In particular, if $\mathcal{L}^q(E_x) = 0$ for almost every x, then $\mathcal{L}^{p+q}(E) = 0$.

This is Tonelli's theorem applied to the indicator of E. It justifies passing from null exceptional sets for individual tokens, with their predecessors fixed, to a null set of initial token configurations. Measurability of the full exceptional set is required; fiberwise nullity alone is not the stated hypothesis.

B Related Work

Aggregation across network depth. Several architectures modify how representations are combined across layers. DenseFormer learns weighted combinations across depth (Pagliardini et al., 2024), while Hyper-Connections modify connections between residual streams and layer outputs (Zhu et al., 2025). DeepCrossAttention uses input-dependent weighting of layer outputs and depth-wise cross-attention to enhance residual learning (Heddes et al., 2025). Attention Residuals (Kimi Team, 2026) make the aggregation itself a softmax attention over depth; later variants attend to layer-wise differences (Luo et al., 2026a), use several routing heads (Luo et al., 2026b), route with low-rank keys (Su, 2026), or combine depth retrieval with multiple residual streams (Yu et al., 2026). We study a continuous-depth model with prescribed history kernels, whose common-reweighting structure permits ODE closure.

Continuous-depth attention, causal masking and clustering Continuous-depth analyses connect attention to interacting token systems. Sander et al. (2022a) give a Wasserstein gradient-flow interpretation under doubly stochastic normalization and a symmetry assumption on the attention parameters. Lu et al. (2019) interpret Transformer layers through a multi-particle dynamical system and numerical splitting methods. For residual networks, Sander et al. (2022b) quantify discrete-to-continuous approximation errors and show that increasing depth alone need not yield convergence to a neural ODE without additional regularity across depth. Geshkovski et al. (2023) relate clustering geometry to the value matrix for fixed parameters. Castin et al. (2026) develop a mean-field PDE framework for several attention variants, including masked and multi-head attention, and establish well-posedness and a particle-to-mean-field limit. Burger et al. (2025) study a gradient-flow formulation for a special case of self-attention dynamics on the unit sphere and analyze the associated stationary states.

For spherical causal attention, Karagodin et al. (2024) prove almost-everywhere singleton convergence with an identity value matrix and arbitrary fixed query and key matrices, and conjecture broader spectral classifications. Álvaro Rodríguez Abella et al. (2025) establish hemisphere consensus under specified time-varying parameters and analyze causal attention with symmetric value matrices having a simple largest eigenvalue. In a discrete hardmax setting, Alcalde et al. (2025) prove convergence to clustered equilibria determined by distinguished tokens called leaders. In a spherical self-attention model, Geshkovski et al. (2024) show that configurations with several clusters can persist for exponentially long times before eventual single-cluster collapse. This result highlights the distinction between finite-depth behavior and asymptotic clustering.

Our model additionally couples token directions to radii and prescribed history rates. Under our matrix and history-kernel assumptions, we prove two-pole convergence outside a Lebesgue-null set of admissible raw initial embeddings. This result does not resolve the conjectures for general residual causal attention.

Representation collapse and radial behavior. Clustering must also be distinguished from other forms of collapse. Dong et al. (2023) prove rapid rank collapse for pure attention without skip connections or multilayer perceptrons, while Pham et al. (2025) study parameter-dependent convergence and divergence, including effects of positional encoding. For selective state-space models, Vo et al. (2025) derive continuous-time token dynamics and characterize convergence to zero or divergence in a one-dimensional Mamba model. In our model, positive radial bounds keep normalization defined, and tokens approaching the same direction may retain different limiting lengths. Directional clustering therefore does not necessarily mean convergence to one raw embedding, nor does our analysis establish a loss of performance in trained networks.

C Proofs of Theorems and Lemmas

C.1 Proof of Theorem 3.1

Let $\mathcal{X}_T := C([0,T];\mathbb{R}^{Nd})$ and equip this space with the Bielecki norm

$$\|\boldsymbol{Z}\|_\beta := \sup_{0\le t\le T} e^{-\beta t}\|\boldsymbol{Z}(t)\|_\infty, \qquad \beta > 0.$$

This norm is equivalent to the usual supremum norm. Consequently, $\mathcal{X}_T$ is complete under $\|\cdot\|_\beta$ (Rudin, 1976, Chapters 4 and 7). Define a mapping $\mathcal{T} : \mathcal{X}_T \to \mathcal{X}_T$ by

$$(\mathcal{T}\boldsymbol{Z})_i(t) := a_i(t)\boldsymbol{u}_i + \int_0^t K_i(t,s)\boldsymbol{G}_i(s,\boldsymbol{Z}(s))\,ds.$$

The integrand is continuous, so $\mathcal{T}\boldsymbol{Z}$ is continuous. Since D_T is compact and K_i is continuous, the quantity $K_T := \max_{1\le i\le N}\sup_{(t,s)\in D_T}|K_i(t,s)|$ is finite. For $\boldsymbol{Y},\boldsymbol{Z}\in\mathcal{X}_T$, the Lipschitz condition on $\boldsymbol{G}$ gives

$$\begin{aligned} e^{-\beta t}\|(\mathcal{T}\boldsymbol{Y})(t)-(\mathcal{T}\boldsymbol{Z})(t)\|_\infty &\le LK_T e^{-\beta t}\int_0^t \|\boldsymbol{Y}(s)-\boldsymbol{Z}(s)\|_\infty\,ds \\ &\le LK_T\|\boldsymbol{Y}-\boldsymbol{Z}\|_\beta e^{-\beta t}\int_0^t e^{\beta s}\,ds \end{aligned}$$

$$\leq \frac{LK_T}{\beta}\|\boldsymbol{Y}-\boldsymbol{Z}\|_\beta.$$

Taking the supremum over $t \in [0,T]$ yields

$$\|\mathcal{T}\boldsymbol{Y}-\mathcal{T}\boldsymbol{Z}\|_\beta \leq \frac{LK_T}{\beta}\|\boldsymbol{Y}-\boldsymbol{Z}\|_\beta.$$

Choose $\beta > LK_T$. Then $\mathcal{T}$ is a contraction on the complete metric space $\mathcal{X}_T$. The Banach fixed-point theorem therefore gives a unique $\boldsymbol{Z} \in \mathcal{X}_T$ satisfying $\mathcal{T}\boldsymbol{Z} = \boldsymbol{Z}$ (Zeidler, 1986, Chapter 1). This fixed-point identity is precisely (5). Finally, suppose the hypotheses hold on every finite time triangle D_S. For each $S > 0$, the preceding argument produces a unique solution $\boldsymbol{Z}^{(S)}$ on $[0,S]$. If $0 < S_1 < S_2$, then $\boldsymbol{Z}^{(S_2)}|_{[0,S_1]}$ satisfies the same integral equation as $\boldsymbol{Z}^{(S_1)}$; finite-horizon uniqueness therefore gives

$$\boldsymbol{Z}^{(S_2)}|_{[0,S_1]} = \boldsymbol{Z}^{(S_1)}.$$

Thus the solution $\boldsymbol{Z}$ globally exists and $\boldsymbol{Z} \in C^1_{\mathrm{loc}}([0,\infty);\mathbb{R}^{Nd})$.

C.2 Proof of Lemma 3.2

Fix a token i. Under the shared-query assumption, the score assigned to a stored key does not change when later keys are added. Write $b_\ell := \exp\{\boldsymbol{w}^\top \boldsymbol{k}_\ell\} > 0$ and $Z_n := \sum_{\ell=0}^n b_\ell$. With the denominator in (1) normalized over all sources $0,\dots,n$, their weights are $\alpha^n_{i,\ell} = b_\ell/Z_n$ for $0 \leq \ell \leq n$. In particular, $Z_0 = b_0$ and $Z_{n+1} > Z_n$.

We construct a continuous positive rate whose accumulated value at each partition point records the growth of Z_n. For $k = 0,\dots,m-1$, let $\Delta t_k := t_{k+1} - t_k$ and $\delta_k := \log(Z_{k+1}/Z_k)$. Both numbers are positive. Choose a constant $c > 0$ smaller than every ratio $\delta_k/\Delta t_k$. On $[t_k, t_{k+1}]$, set

$$\gamma_i(t) := c + \frac{6(\delta_k - c\Delta t_k)}{\Delta t_k}\frac{t-t_k}{\Delta t_k}\left(1 - \frac{t-t_k}{\Delta t_k}\right),$$

and set $\gamma_i(t) := c$ for $t \geq t_m$. This function is strictly positive. It is continuous at each partition point because the added term vanishes at both ends of every cell. Its integral over $[t_k, t_{k+1}]$ is δ_k: the added term has integral $\delta_k - c\Delta t_k$. Hence, if $\Gamma_i(t) := \int_0^t \gamma_i(v)\,dv$, then $\Gamma_i(t_n) = \log(Z_n/Z_0)$ for every partition point. Moreover, $\int_0^\infty \gamma_i(t)\,dt = \infty$, since $\gamma_i(t) = c$ after t_m. This continuation verifies the required infinite accumulated rate, but it is chosen after the final observed layer. The finite execution does not determine the continued rate or the long-time behavior of the resulting continuous system.

Define the initial-embedding weight and history kernel by

$$a_i(t) := e^{-\Gamma_i(t)}, \qquad K_i(t,s) := \gamma_i(s)e^{-[\Gamma_i(t)-\Gamma_i(s)]}, \quad 0 \leq s \leq t. \tag{30}$$

Both are nonnegative. The continuity of γ_i makes a_i continuously differentiable and makes K_i continuous on every finite time triangle. For $s < t$, differentiation in the receiving time gives $\partial_t K_i(t,s) = -\gamma_i(t)K_i(t,s)$. The right-hand side is continuous up to the boundary of each triangle, so it also gives the required one-sided boundary derivatives. Likewise, $\dot{a}_i(t) = -\gamma_i(t)a_i(t)$.

The weights have total mass one. Indeed, integrating the derivative of $e^{\Gamma_i(s)}$ gives

$$a_i(t) + \int_0^t K_i(t,s)\,ds = e^{-\Gamma_i(t)}\left(1 + \int_0^t \gamma_i(s)e^{\Gamma_i(s)}\,ds\right) = 1.$$

Together with the positivity and divergent integral of γ_i, this verifies Assumption 2(a).

It remains to check the discrete weights, rather than merely their total mass. At a partition point, $a_i(t_n) = e^{-\Gamma_i(t_n)} = Z_0/Z_n = b_0/Z_n = \alpha^n_{i,0}$. For $0 \leq k < n$,

$$\begin{aligned}\int_{t_k}^{t_{k+1}} K_i(t_n,s)\,ds &= e^{-\Gamma_i(t_n)}\left(e^{\Gamma_i(t_{k+1})} - e^{\Gamma_i(t_k)}\right)\\ &= \frac{Z_{k+1}-Z_k}{Z_n} = \frac{b_{k+1}}{Z_n} = \alpha^n_{i,k+1}.\end{aligned}$$

This is exactly (4) for every receiving layer in the partition.

C.3 Proof of Lemma 3.3

This is a finite-horizon approximation of the prescribed memory equation. A shared-query Full Attention Residual network has the weights (13) if its content-derived keys and query happen to satisfy

$$\exp\Big\{(\boldsymbol{w}^{(h)})^\top\big(\boldsymbol{k}_{i,k+1}^{(h)}-\boldsymbol{k}_{i,0}^{(h)}\big)\Big\}=e^{\Gamma_i(t_{k+1})}-e^{\Gamma_i(t_k)},\qquad \Gamma_i(t)=\int_0^t\gamma_i(s)\,ds. \tag{31}$$

Here the query is shared within each network but may depend on the partition. Equation (31) is an additional compatibility condition, not a consequence of Lemma 3.2; its realization by the original content-key architecture is not established.

Indeed, admissible executions need not satisfy it. Take one token, $d=2$, $\boldsymbol{V}=\mathrm{diag}(2,1)$, $\boldsymbol{W}=10\boldsymbol{I}_2$, and $\boldsymbol{u}_1=2\boldsymbol{e}_1$. Every layer output and aggregate equals $2\boldsymbol{e}_1$, so the initial and stored content-derived keys coincide. The left-hand side of (31) is then 1 for every shared query, whereas its right-hand side on the first shrinking cell tends to zero for any fixed continuous history rate. Thus the lemma does not give a continuous-depth limit for every Full Attention Residual execution. Nor does its constant C_T, which may depend on T, connect fixed-grid dynamics to the $t\to\infty$ conclusion of Theorem 5.1.

Assumption 2(a) determines the form of the fixed kernel. Its mass identity at $t=0$ gives $a_i(0)=1$; hence $\dot a_i=-\gamma_i a_i$ gives $a_i(t)=e^{-\Gamma_i(t)}$. Differentiating the mass identity and using $\partial_t K_i=-\gamma_i(t)K_i$ gives $K_i(t,t)=\gamma_i(t)$. Solving the receiving-time equation from $t=s$ therefore gives $K_i(t,s)=\gamma_i(s)e^{-[\Gamma_i(t)-\Gamma_i(s)]}$.

The prescribed cell weights in (13) consequently have the explicit form

$$\alpha_{i,0}^n=e^{-\Gamma_i(t_n)},\qquad \alpha_{i,k+1}^n=e^{-\Gamma_i(t_n)}\big(e^{\Gamma_i(t_{k+1})}-e^{\Gamma_i(t_k)}\big).$$

The second identity follows by integrating $\frac{d}{ds}e^{\Gamma_i(s)}=\gamma_i(s)e^{\Gamma_i(s)}$ over the cell. These weights are nonnegative and sum to one by telescoping.

When layer n is added, every old weight is multiplied by $e^{-[\Gamma_i(t_{n+1})-\Gamma_i(t_n)]}$; the new output gets the remaining mass. Thus, with $\theta_{i,n}=1-e^{-[\Gamma_i(t_{n+1})-\Gamma_i(t_n)]}$, the quadrature recurrence becomes

$$\boldsymbol{z}_i^{n+1}=(1-\theta_{i,n})\boldsymbol{z}_i^n+\theta_{i,n}\boldsymbol{F}_i(\boldsymbol{X}^n),\qquad 0<\theta_{i,n}<1.$$

The subsequent radial and error estimates apply to this prescribed-weight recurrence. If a shared-query network satisfies (31), division of all its unnormalized scores by the initial-source score shows that its softmax weights equal (13). The proof does not require such a network to exist.

We next establish bounds that keep normalization defined. By Lemma 4.1, $\langle\boldsymbol{x}_i,\boldsymbol{F}_i(\boldsymbol{X})\rangle\ge\eta>0$ for every configuration of unit directions. Also $\|\boldsymbol{F}_i(\boldsymbol{X})\|\le\lambda$, since Assumption 1(c) makes $\boldsymbol{V}$ positive definite with operator norm λ. If $\boldsymbol{z}_i^n$ is nonzero, projecting the discrete update onto $\boldsymbol{x}_i^n$ gives its lower radial bound; the triangle inequality gives its upper bound:

$$(1-\theta_{i,n})\|\boldsymbol{z}_i^n\|+\theta_{i,n}\eta\le\|\boldsymbol{z}_i^{n+1}\|\le(1-\theta_{i,n})\|\boldsymbol{z}_i^n\|+\theta_{i,n}\lambda.$$

Induction from $\boldsymbol{z}_i^0=\boldsymbol{u}_i\neq\boldsymbol{0}$ shows that every discrete radius lies between $r_*:=\min\{\eta,\min_j\|\boldsymbol{u}_j\|\}>0$ and $R_*:=\max\{\lambda,\max_j\|\boldsymbol{u}_j\|\}$.

The continuous trajectory obeys the same bounds. Indeed, on the open set of nonzero aggregates, its raw equation $\dot{\boldsymbol{z}}_i=\gamma_i(t)(\boldsymbol{F}_i(\boldsymbol{X})-\boldsymbol{z}_i)$ has a unique local solution. Its radius satisfies $\dot r_i=\gamma_i(t)(\langle\boldsymbol{x}_i,\boldsymbol{F}_i(\boldsymbol{X})\rangle-r_i)$. Applying an integrating factor to the preceding bounds on the radial forcing yields

$$e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\|+(1-e^{-\Gamma_i(t)})\eta\le r_i(t)\le e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\|+(1-e^{-\Gamma_i(t)})\lambda.$$

The solution consequently stays in the bounded region where every radius is at least r_*, and extends throughout $[0,T]$. Variation of constants identifies it with the solution of (5).

On the region where all radii belong to $[r_*,R_*]$, normalization and the attention field together have a Lipschitz constant L in the maximum token norm. This constant does not depend on the partition. Set $D=\sup_{0\le t\le T}\|\dot{\boldsymbol{Z}}(t)\|_\infty$ and $K_T=\max_i\sup_{0\le s\le t\le T}K_i(t,s)$; both are finite. For

$e_n = \|\boldsymbol{Z}^n - \boldsymbol{Z}(t_n)\|_\infty$, subtract the continuous equation at t_n from the discrete recurrence and use the exact cell-weight identity established above. The Lipschitz bound gives

$$e_n \le LK_T \sum_{k<n} \int_{t_k}^{t_{k+1}} \|\boldsymbol{Z}^k - \boldsymbol{Z}(s)\|_\infty \, ds \le LK_T \sum_{k<n} (t_{k+1} - t_k) e_k + LDK_T T h.$$

In the last step, $\|\boldsymbol{Z}(s) - \boldsymbol{Z}(t_k)\|_\infty \le Dh$ on cell k. Induction in this inequality, using $(t_{k+1} - t_k)e^{LK_T t_k} \le \int_{t_k}^{t_{k+1}} e^{LK_T s}\, ds$, gives $e_n \le LDK_T T h\, e^{LK_T t_n}$. Since $t_n \le T$, the assertion follows with $C_T = LDK_T T e^{LK_T T}$.

C.4 Proof of Theorem 3.4

For a continuous trajectory in Ω, the integrand in (5) and its partial derivative with respect to the receiving time are continuous on every finite triangle. Differentiating (12) and using Assumption 2(a) first gives

$$K_i(t,t) = \gamma_i(t).$$

The Leibniz rule then gives

$$\dot{\boldsymbol{z}}_i = -\gamma_i a_i \boldsymbol{u}_i + \gamma_i \boldsymbol{F}_i(\boldsymbol{X}) - \gamma_i \int_0^t K_i(t,s) \boldsymbol{F}_i(\boldsymbol{X}(s))\, ds = \gamma_i(\boldsymbol{F}_i(\boldsymbol{X}) - \boldsymbol{z}_i),$$

which is (14). Conversely, multiplying this ODE by $e^{\Gamma_i(t)}$, where $\Gamma_i(t) := \int_0^t \gamma_i(s)\, ds$, and integrating recovers (5): indeed, (12) gives $a_i(0) = 1$, while Assumption 2(a) and $K_i(s,s) = \gamma_i(s)$ yield

$$a_i(t) = e^{-\Gamma_i(t)}, \qquad K_i(t,s) = \gamma_i(s) e^{-[\Gamma_i(t) - \Gamma_i(s)]}.$$

Historical outputs are evaluated at their production time s, and the shared matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are fixed across depth. The chain rule gives, for $r_i > 0$,

$$\dot{r}_i = \langle \boldsymbol{x}_i, \dot{\boldsymbol{z}}_i \rangle, \qquad \dot{\boldsymbol{x}}_i = \frac{1}{r_i}(\boldsymbol{I} - \boldsymbol{x}_i \boldsymbol{x}_i^\top)\dot{\boldsymbol{z}}_i.$$

Substituting the explicit sum in (14) yields (15). Conversely, these two equations give (14) after differentiating $\boldsymbol{z}_i = r_i \boldsymbol{x}_i$. Their directional velocity has zero inner product with $\boldsymbol{x}_i$ on the unit sphere, so the constructed directional solution remains on $\mathbb{S}^{d-1}$.

On compact subsets of $[0,\infty) \times \Omega$, normalization is smooth and the softmax denominator is positive. The raw vector field is therefore continuous in time and locally Lipschitz in the state, uniformly on such sets. The Picard–Lindelof theorem and maximal-solution theorem give local existence, uniqueness, and a unique maximal extension (Teschl, 2012, Theorems 2.2 and 2.13). At $t = 0$ these are understood one-sidedly; equivalently, the continuous coefficient paths can be extended to a small negative interval.

Since $p_{ij} \ge 0$ and $\sum_{j \le i} p_{ij} = 1$, $\|\boldsymbol{F}_i(\boldsymbol{X})\| \le \|\boldsymbol{V}\|_{\mathrm{op}}$. Thus (15) implies $\dot{r}_i \le \gamma_i(\|\boldsymbol{V}\|_{\mathrm{op}} - r_i)$, and integration after multiplication by e^{Γ_i} gives

$$r_i(t) \le e^{-\Gamma_i(t)} r_i(0) + (1 - e^{-\Gamma_i(t)})\|\boldsymbol{V}\|_{\mathrm{op}} \le \max\{r_i(0), \|\boldsymbol{V}\|_{\mathrm{op}}\}.$$

If $T_* < \infty$ and the radii were bounded below near T_*, these bounds would keep $\boldsymbol{Z}(t)$ in a compact subset of Ω up to T_*. The continuation criterion (Teschl, 2012, Corollary 2.15) would extend the solution past T_*. This proves the final assertion.

C.5 Proof of Lemma 4.1

Fix a configuration $\boldsymbol{X} \in (\mathbb{S}^{d-1})^N$ and a token $i \in \{1, \dots, N\}$. We prove the lower bound first. We begin with two consequences of the assumptions. Since

$$\eta = \lambda_d - \lambda \sqrt{\frac{2\sqrt{d}\log N + 4\rho}{m_0}} > 0,$$

we necessarily have $\lambda_d > 0$. Together with $\lambda_1 \ge \cdots \ge \lambda_d$, this shows that $\boldsymbol{V}$ is positive definite. In particular, because $\lambda = \lambda_1$ is the largest eigenvalue of $\boldsymbol{V}$, $\|\boldsymbol{V}\|_{\mathrm{op}} = \lambda$. The proof rests on the fact

that the softmax assigns limited total weight to tokens that are far from $\boldsymbol{x}_i$. To make this precise, write

$$\boldsymbol{E} := \boldsymbol{W} - m_0\boldsymbol{I} = \begin{pmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & (m_1 - m_0)\boldsymbol{I}_{d-1} + \boldsymbol{B} \end{pmatrix}.$$

We first bound the operator norm of $\boldsymbol{E}$. Let $\boldsymbol{y} = (y_1, \boldsymbol{y}_\perp) \in \mathbb{R} \times \mathbb{R}^{d-1}$ be arbitrary. The first coordinate of $\boldsymbol{E}\boldsymbol{y}$ is zero, while its remaining coordinates are $((m_1 - m_0)\boldsymbol{I}_{d-1} + \boldsymbol{B})\boldsymbol{y}_\perp$. Hence

$$\begin{aligned} \|\boldsymbol{E}\boldsymbol{y}\|^2 &= \|(m_1 - m_0)\boldsymbol{y}_\perp + \boldsymbol{B}\boldsymbol{y}_\perp\|^2 \\ &= (m_1 - m_0)^2\|\boldsymbol{y}_\perp\|^2 + 2(m_1 - m_0)\langle \boldsymbol{y}_\perp, \boldsymbol{B}\boldsymbol{y}_\perp\rangle + \|\boldsymbol{B}\boldsymbol{y}_\perp\|^2. \end{aligned}$$

The middle term vanishes because $\boldsymbol{B}$ is skew-symmetric. Indeed, $\boldsymbol{B}^\top = -\boldsymbol{B}$ implies

$$\langle \boldsymbol{y}_\perp, \boldsymbol{B}\boldsymbol{y}_\perp\rangle = \langle \boldsymbol{B}^\top\boldsymbol{y}_\perp, \boldsymbol{y}_\perp\rangle = -\langle \boldsymbol{B}\boldsymbol{y}_\perp, \boldsymbol{y}_\perp\rangle = -\langle \boldsymbol{y}_\perp, \boldsymbol{B}\boldsymbol{y}_\perp\rangle,$$

and therefore $\langle \boldsymbol{y}_\perp, \boldsymbol{B}\boldsymbol{y}_\perp\rangle = 0$. It follows that

$$\begin{aligned} \|\boldsymbol{E}\boldsymbol{y}\|^2 &= (m_1 - m_0)^2\|\boldsymbol{y}_\perp\|^2 + \|\boldsymbol{B}\boldsymbol{y}_\perp\|^2 \\ &\le \big((m_1 - m_0)^2 + \|\boldsymbol{B}\|_{\mathrm{op}}^2\big)\|\boldsymbol{y}_\perp\|^2 \\ &\le \rho^2\|\boldsymbol{y}\|^2. \end{aligned}$$

Thus $\|\boldsymbol{E}\|_{\mathrm{op}} \le \rho$. For the remainder of the proof, abbreviate $p_j := p_{ij}(\boldsymbol{X})$ and let $D_j := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$, $j \le i$. Thus D_j is the Euclidean distance between the direction of token i and the direction of source token j. Subtracting the self score of token i from the score of source j gives

$$\frac{1}{\sqrt{d}}\boldsymbol{x}_i^\top\boldsymbol{W}\boldsymbol{x}_j - \frac{1}{\sqrt{d}}\boldsymbol{x}_i^\top\boldsymbol{W}\boldsymbol{x}_i = \frac{1}{\sqrt{d}}\boldsymbol{x}_i^\top(m_0\boldsymbol{I} + \boldsymbol{E})(\boldsymbol{x}_j - \boldsymbol{x}_i) = -\frac{m_0}{2\sqrt{d}}D_j^2 + \epsilon_j,$$

where $\epsilon_j := \frac{1}{\sqrt{d}}\boldsymbol{x}_i^\top\boldsymbol{E}(\boldsymbol{x}_j - \boldsymbol{x}_i)$. This error term measures the contribution of the part of $\boldsymbol{W}$ that differs from $m_0\boldsymbol{I}$. By Cauchy–Schwarz and the operator-norm bound proved above,

$$|\epsilon_j| \le \frac{1}{\sqrt{d}}\|\boldsymbol{x}_i\|\,\|\boldsymbol{E}\|_{\mathrm{op}}\,\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \le \frac{\rho}{\sqrt{d}}D_j \le \frac{2\rho}{\sqrt{d}}.$$

The last inequality follows from $D_j \le \|\boldsymbol{x}_i\| + \|\boldsymbol{x}_j\| = 2$. Subtracting the same self score from every logit does not change the softmax probabilities. Therefore

$$p_j = \frac{\exp\!\Big(-\frac{m_0}{2\sqrt{d}}D_j^2 + \epsilon_j\Big)}{Z_i}, \qquad Z_i := \sum_{\ell \le i}\exp\!\left(-\frac{m_0}{2\sqrt{d}}D_\ell^2 + \epsilon_\ell\right).$$

For the self term $\ell = i$, we have $D_i = 0$ and $\epsilon_i = 0$. That term is therefore equal to one, which shows that $Z_i \ge 1$ and hence $\log Z_i \ge 0$. Taking logarithms in the preceding softmax formula gives

$$\log p_j = -\frac{m_0}{2\sqrt{d}}D_j^2 + \epsilon_j - \log Z_i.$$

Multiplying this identity by p_j, summing over $j \le i$, and using $\sum_{j\le i} p_j = 1$, we obtain the important identity

$$\frac{m_0}{2\sqrt{d}}\sum_{j\le i} p_j D_j^2 = -\sum_{j\le i} p_j \log p_j + \sum_{j\le i} p_j\epsilon_j - \log Z_i. \tag{32}$$

We now estimate the three terms on the right-hand side. The first term is the entropy of a probability vector with i entries, and it is at most $\log i$. The second term satisfies

$$\sum_{j\le i} p_j\epsilon_j \le \sum_{j\le i} p_j|\epsilon_j| \le \frac{2\rho}{\sqrt{d}},$$

because the weights are nonnegative and sum to one. Finally, $-\log Z_i \le 0$ because $Z_i \ge 1$. Substituting these estimates into (32) yields

$$\frac{m_0}{2\sqrt{d}}\sum_{j\le i} p_j D_j^2 \le \log i + \frac{2\rho}{\sqrt{d}}.$$

Since $m_0 > 0$ and $i \le N$, we conclude that

$$\sum_{j\le i} p_j D_j^2 \le \frac{2\sqrt{d}\log i + 4\rho}{m_0} \le \frac{2\sqrt{d}\log N + 4\rho}{m_0}. \tag{33}$$

Thus the block structure of $\boldsymbol{W}$ gives a uniform bound on the attention-weighted squared distance between $\boldsymbol{x}_i$ and the source directions to which it attends. For every $j \le i$,

$$\langle \boldsymbol{x}_i, \boldsymbol{V}\boldsymbol{x}_j\rangle = \langle \boldsymbol{x}_i, \boldsymbol{V}\boldsymbol{x}_i\rangle + \langle \boldsymbol{x}_i, \boldsymbol{V}(\boldsymbol{x}_j - \boldsymbol{x}_i)\rangle \ge \lambda_d - \|\boldsymbol{x}_i\|\,\|\boldsymbol{V}\|_{\mathrm{op}}\,\|\boldsymbol{x}_j - \boldsymbol{x}_i\| = \lambda_d - \lambda D_j.$$

Here the first term is at least λ_d by the Rayleigh quotient bound for the symmetric matrix $\boldsymbol{V}$, and the second term is bounded below using Cauchy–Schwarz, $\|\boldsymbol{x}_i\| = 1$, and $\|\boldsymbol{V}\|_{\mathrm{op}} = \lambda$. Using the definition $\boldsymbol{F}_i(\boldsymbol{X}) = \sum_{j\le i} p_j \boldsymbol{V}\boldsymbol{x}_j$, we therefore obtain

$$\langle \boldsymbol{x}_i, \boldsymbol{F}_i(\boldsymbol{X})\rangle \ge \lambda_d - \lambda \sum_{j\le i} p_j D_j \ge \lambda_d - \lambda \left(\sum_{j\le i} p_j D_j^2\right)^{1/2} \ge \lambda_d - \lambda\sqrt{\frac{2\sqrt{d}\log N + 4\rho}{m_0}} = \eta.$$

It remains to prove the upper bound. Since the attention weights are nonnegative and sum to one,

$$\|\boldsymbol{F}_i(\boldsymbol{X})\| = \left\|\sum_{j\le i} p_j \boldsymbol{V}\boldsymbol{x}_j\right\| \le \sum_{j\le i} p_j \|\boldsymbol{V}\boldsymbol{x}_j\| \le \sum_{j\le i} p_j \|\boldsymbol{V}\|_{\mathrm{op}}\|\boldsymbol{x}_j\| = \lambda.$$

Because $\|\boldsymbol{x}_i\| = 1$, another application of Cauchy–Schwarz gives

$$\langle \boldsymbol{x}_i, \boldsymbol{F}_i(\boldsymbol{X})\rangle \le |\langle \boldsymbol{x}_i, \boldsymbol{F}_i(\boldsymbol{X})\rangle| \le \|\boldsymbol{x}_i\|\,\|\boldsymbol{F}_i(\boldsymbol{X})\| \le \lambda.$$

Combining the lower and upper bounds proves that $\eta \le \langle \boldsymbol{x}_i, \boldsymbol{F}_i(\boldsymbol{X})\rangle \le \lambda$ for every token i and every configuration $\boldsymbol{X} \in (\mathbb{S}^{d-1})^N$.

C.6 Proof of Lemma 4.2

By Theorem 3.4, the coupled system has a unique maximal solution on an interval $[0, T_*)$, where every radius $r_i(t)$ is positive. Here $T_* \in (0, \infty]$ denotes the maximal forward existence time. We first derive bounds that hold throughout this maximal interval and then use them to prove that $T_* = \infty$. For each token i, define its radial forcing by

$$f_i(t) := \langle \boldsymbol{x}_i(t), \boldsymbol{F}_i(\boldsymbol{X}(t))\rangle.$$

This is the scalar quantity toward which the radial equation drives r_i. Indeed, the second equation in (15) can be written as

$$\dot r_i(t) = \gamma_i(t)\big(f_i(t) - r_i(t)\big).$$

Lemma 4.1 applies to every configuration of unit directions, so along the entire maximal solution we have $\eta \le f_i(t) \le \lambda$ for $0 \le t < T_*$. Recall that $\Gamma_i(t) = \int_0^t \gamma_i(s)\,ds$. Since $\dot\Gamma_i(t) = \gamma_i(t)$, multiplying the radial equation by $e^{\Gamma_i(t)}$ gives

$$\begin{aligned}\frac{d}{dt}\left(e^{\Gamma_i(t)} r_i(t)\right) &= e^{\Gamma_i(t)}\left(\dot r_i(t) + \gamma_i(t) r_i(t)\right)\\ &= e^{\Gamma_i(t)}\gamma_i(t) f_i(t).\end{aligned}$$

Integrating from 0 to t and using $\Gamma_i(0) = 0$ and $r_i(0) = \|\boldsymbol{u}_i\|$, we obtain the exact representation

$$r_i(t) = e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + \int_0^t e^{-[\Gamma_i(t) - \Gamma_i(s)]}\gamma_i(s) f_i(s)\,ds. \tag{34}$$

We apply the bounds $\eta \le f_i(s) \le \lambda$ to (34). The lower bound gives

$$\begin{aligned} r_i(t) &\ge e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + \eta\int_0^t e^{-[\Gamma_i(t) - \Gamma_i(s)]}\gamma_i(s)\,ds\\ &= e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + \big(1 - e^{-\Gamma_i(t)}\big)\eta.\end{aligned}$$

Similarly, the upper bound gives

$$r_i(t) \le e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + \lambda \int_0^t e^{-[\Gamma_i(t)-\Gamma_i(s)]}\gamma_i(s)\,ds$$
$$= e^{-\Gamma_i(t)}\|\boldsymbol{u}_i\| + \big(1 - e^{-\Gamma_i(t)}\big)\lambda.$$

This proves (16) on the maximal interval $[0, T_*)$. Because $\gamma_i(t) > 0$, we have $\Gamma_i(t) \ge 0$, and hence $e^{-\Gamma_i(t)} \in (0, 1]$. Each expression in (16) is therefore a convex combination: the lower comparison function is a convex combination of $\|\boldsymbol{u}_i\|$ and η, while the upper comparison function is a convex combination of $\|\boldsymbol{u}_i\|$ and λ. Consequently, $r_i(t) \ge \min\{\|\boldsymbol{u}_i\|, \eta\} > 0$ and

$$r_i(t) \le \max\{\|\boldsymbol{u}_i\|, \lambda\} \le M, \qquad M := \max\left\{\lambda, \max_{1\le j\le N}\|\boldsymbol{u}_j\|\right\}.$$

The strict positivity of the lower bound follows from the assumptions $\boldsymbol{u}_i \neq \boldsymbol{0}$ and $\eta > 0$. We next show that the maximal solution is global. Suppose, for contradiction, that $T_* < \infty$. Set $r_{\min} := \min_{1\le j\le N} \min\{\|\boldsymbol{u}_j\|, \eta\}$. The number $r_{\min}$ is strictly positive because there are only finitely many tokens, every initial token is nonzero, and $\eta > 0$. The bounds proved above imply that, for every $t < T_*$ and every i,

$$r_{\min} \le \|\boldsymbol{z}_i(t)\| = r_i(t) \le M.$$

Thus the raw state $\boldsymbol{Z}(t)$ remains in the set $\mathcal{K} := \prod_{i=1}^N \{\boldsymbol{z} \in \mathbb{R}^d : r_{\min} \le \|\boldsymbol{z}\| \le M\}$. This set is compact, and it is contained in the open domain $\Omega = \{\boldsymbol{Z} \in (\mathbb{R}^d)^N : \boldsymbol{z}_i \neq \boldsymbol{0} \text{ for every } i\}$, on which all normalized directions $\boldsymbol{x}_i = \boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ are well defined.

On Ω, the normalization map $\boldsymbol{z}_i \mapsto \boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ is smooth. The softmax weights and the attention field are smooth functions of the normalized directions, so the vector field in (14) is locally Lipschitz in the state. Since $\mathcal{K}$ is a compact subset of Ω, the vector field is uniformly locally Lipschitz and bounded on the portion of the state space visited by the solution up to time T_*. The standard continuation criterion for ordinary differential equations therefore extends the solution beyond T_*. This contradicts the definition of T_* as the maximal forward existence time. Hence $T_* = \infty$, and the solution is global. It remains to study the effective directional time. For each token, define

$$\tau_i(t) := \int_0^t \frac{\gamma_i(s)}{r_i(s)}\,ds.$$

The global upper bound $r_i(s) \le M$ and the positivity of γ_i imply

$$\frac{\gamma_i(s)}{r_i(s)} \ge \frac{\gamma_i(s)}{M}.$$

Integrating this inequality gives

$$\tau_i(t) \ge \frac{1}{M}\int_0^t \gamma_i(s)\,ds = \frac{\Gamma_i(t)}{M}.$$

Assumption 2(a) states that $\int_0^\infty \gamma_i(s)\,ds = \infty$, so $\Gamma_i(t) \to \infty$ as $t \to \infty$. It follows that $\tau_i(t) \to \infty$ as well. Moreover,

$$\tau_i'(t) = \frac{\gamma_i(t)}{r_i(t)} > 0.$$

Both γ_i and r_i are continuous, and r_i is bounded away from zero. Thus τ_i is continuously differentiable and strictly increasing. Since $\tau_i(0) = 0$ and $\tau_i(t) \to \infty$, it maps $[0, \infty)$ bijectively onto $[0, \infty)$. Its inverse, denoted by t_i, is continuously differentiable by the inverse function theorem, with the derivative understood one-sidedly at the endpoint $\tau = 0$. Finally, define the direction of token i in its effective time by

$$\boldsymbol{y}(\tau) := \boldsymbol{x}_i(t_i(\tau)).$$

Because $\tau_i(t_i(\tau)) = \tau$, differentiation of this identity gives

$$t_i'(\tau) = \frac{1}{\tau_i'(t_i(\tau))} = \frac{r_i(t_i(\tau))}{\gamma_i(t_i(\tau))}.$$

Applying the chain rule to $\boldsymbol{y}(\tau)$ and using the directional equation in (15), we obtain

$$\begin{aligned}\boldsymbol{y}'(\tau) &= \dot{\boldsymbol{x}}_i(t_i(\tau))\, t_i'(\tau) \\ &= \frac{\gamma_i(t_i(\tau))}{r_i(t_i(\tau))} \left(\boldsymbol{I} - \boldsymbol{y}(\tau)\boldsymbol{y}(\tau)^\top\right) \boldsymbol{F}_i(\boldsymbol{X}(t_i(\tau))) \frac{r_i(t_i(\tau))}{\gamma_i(t_i(\tau))} \\ &= \left(\boldsymbol{I} - \boldsymbol{y}(\tau)\boldsymbol{y}(\tau)^\top\right) \boldsymbol{F}_i(\boldsymbol{X}(t_i(\tau))).\end{aligned}$$

This proves the claimed equation in effective time and completes the proof.

C.7 Proof of Lemma 4.3

Because the attention is causal, the first token can attend only to itself. Thus $p_{11} = 1$, and consequently $\boldsymbol{F}_1(\boldsymbol{X}) = \boldsymbol{V}\boldsymbol{x}_1$. By Lemma 4.2, the effective time $\tau_1(t)$ is strictly increasing and satisfies $\tau_1(t) \to \infty$ as $t \to \infty$. If t_1 denotes its inverse and $\boldsymbol{y}(\tau) := \boldsymbol{x}_1(t_1(\tau))$, then the directional equation from Lemma 4.2 reduces to

$$\boldsymbol{y}'(\tau) = \left(\boldsymbol{I} - \boldsymbol{y}(\tau)\boldsymbol{y}(\tau)^\top\right) \boldsymbol{V}\boldsymbol{y}(\tau), \qquad \boldsymbol{y}(0) = \frac{\boldsymbol{u}_1}{\|\boldsymbol{u}_1\|}. \tag{35}$$

We first solve this equation explicitly. Consider the curve

$$\widehat{\boldsymbol{y}}(\tau) := \frac{e^{\tau \boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau \boldsymbol{V}}\boldsymbol{u}_1\|}.$$

This curve is well defined for every $\tau \geq 0$. Indeed, the matrix $e^{\tau \boldsymbol{V}}$ is invertible, and $\boldsymbol{u}_1 \neq \boldsymbol{0}$ by Assumption 2(b), so $e^{\tau \boldsymbol{V}}\boldsymbol{u}_1$ cannot vanish. We now verify that $\widehat{\boldsymbol{y}}$ solves (35). Let $\boldsymbol{g}(\tau) := e^{\tau \boldsymbol{V}}\boldsymbol{u}_1$. Since $\boldsymbol{V}$ is independent of τ, the derivative of the matrix exponential gives $\boldsymbol{g}'(\tau) = \boldsymbol{V}\boldsymbol{g}(\tau)$. Moreover,

$$\frac{d}{d\tau}\|\boldsymbol{g}(\tau)\| = \frac{\langle \boldsymbol{g}(\tau), \boldsymbol{g}'(\tau)\rangle}{\|\boldsymbol{g}(\tau)\|} = \frac{\langle \boldsymbol{g}(\tau), \boldsymbol{V}\boldsymbol{g}(\tau)\rangle}{\|\boldsymbol{g}(\tau)\|}.$$

Differentiating $\widehat{\boldsymbol{y}}(\tau) = \boldsymbol{g}(\tau)/\|\boldsymbol{g}(\tau)\|$ therefore yields

$$\begin{aligned}\widehat{\boldsymbol{y}}'(\tau) &= \frac{\boldsymbol{V}\boldsymbol{g}(\tau)}{\|\boldsymbol{g}(\tau)\|} - \frac{\boldsymbol{g}(\tau)}{\|\boldsymbol{g}(\tau)\|^2} \frac{\langle \boldsymbol{g}(\tau), \boldsymbol{V}\boldsymbol{g}(\tau)\rangle}{\|\boldsymbol{g}(\tau)\|} \\ &= \boldsymbol{V}\widehat{\boldsymbol{y}}(\tau) - \langle \widehat{\boldsymbol{y}}(\tau), \boldsymbol{V}\widehat{\boldsymbol{y}}(\tau)\rangle \widehat{\boldsymbol{y}}(\tau) \\ &= \left(\boldsymbol{I} - \widehat{\boldsymbol{y}}(\tau)\widehat{\boldsymbol{y}}(\tau)^\top\right) \boldsymbol{V}\widehat{\boldsymbol{y}}(\tau).\end{aligned}$$

At $\tau = 0$, this curve satisfies $\widehat{\boldsymbol{y}}(0) = \frac{\boldsymbol{u}_1}{\|\boldsymbol{u}_1\|}$. Thus $\widehat{\boldsymbol{y}}$ satisfies the same initial-value problem as $\boldsymbol{y}$. The vector field in (35) is smooth, so this initial-value problem has a unique solution. Hence

$$\boldsymbol{y}(\tau) = \frac{e^{\tau \boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau \boldsymbol{V}}\boldsymbol{u}_1\|}.$$

Returning to the original time variable gives the claimed formula

$$\boldsymbol{x}_1(t) = \frac{e^{\tau_1(t) \boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau_1(t) \boldsymbol{V}}\boldsymbol{u}_1\|}. \tag{36}$$

We next determine the limit of this expression. Let $\boldsymbol{e}_1 = \boldsymbol{e}, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ be the orthonormal eigenbasis in which $\boldsymbol{V}$ is diagonal, and write

$$\boldsymbol{u}_1 = \sum_{k=1}^{d} a_k \boldsymbol{e}_k, \qquad a_k := \langle \boldsymbol{u}_1, \boldsymbol{e}_k\rangle.$$

Because the leading eigenspace is $\operatorname{span}\{\boldsymbol{e}\}$, Assumption 2(c) implies that the projection of $\boldsymbol{u}_1$ onto this eigenspace is nonzero. Equivalently, $a_1 = \langle \boldsymbol{u}_1, \boldsymbol{e}\rangle \neq 0$. Therefore the sign $s_1 := \operatorname{sign}\langle \boldsymbol{u}_1, \boldsymbol{e}\rangle$ is well defined. Since $\boldsymbol{V}\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k$, the action of the matrix exponential on $\boldsymbol{u}_1$ is

$$e^{\tau \boldsymbol{V}}\boldsymbol{u}_1 = \sum_{k=1}^{d} a_k e^{\lambda_k \tau} \boldsymbol{e}_k = e^{\lambda \tau} \left(a_1 \boldsymbol{e} + \sum_{k=2}^{d} a_k e^{-(\lambda - \lambda_k)\tau} \boldsymbol{e}_k \right).$$

The positive scalar factor $e^{\lambda\tau}$ cancels when this vector is normalized. Thus

$$\frac{e^{\tau \boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau \boldsymbol{V}}\boldsymbol{u}_1\|} = \frac{a_1\boldsymbol{e} + \sum_{k=2}^{d} a_k e^{-(\lambda-\lambda_k)\tau}\boldsymbol{e}_k}{\left\|a_1\boldsymbol{e} + \sum_{k=2}^{d} a_k e^{-(\lambda-\lambda_k)\tau}\boldsymbol{e}_k\right\|}.$$

The leading eigenvalue is simple, so $\lambda - \lambda_k > 0$ for every $k \geq 2$. It follows that $e^{-(\lambda-\lambda_k)\tau} \to 0$ as $\tau \to \infty$. Consequently, the numerator in the preceding expression converges to $a_1\boldsymbol{e}$, while its denominator converges to $|a_1|$. Hence

$$\frac{e^{\tau \boldsymbol{V}}\boldsymbol{u}_1}{\|e^{\tau \boldsymbol{V}}\boldsymbol{u}_1\|} \longrightarrow \frac{a_1}{|a_1|}\boldsymbol{e} = s_1\boldsymbol{e}.$$

Since $\tau_1(t) \to \infty$, substituting $\tau = \tau_1(t)$ proves $\boldsymbol{x}_1(t) \longrightarrow s_1\boldsymbol{e}$ as $t \to \infty$.

It remains to prove that $r_1(t) \to \lambda$. We first identify the limit of the forcing term in the radial equation. Because the first token attends only to itself, $p_{11} = 1$, and therefore

$$\boldsymbol{F}_1(\boldsymbol{X}(t)) = \boldsymbol{V}\boldsymbol{x}_1(t).$$

Consequently, the radial forcing for the first token is $f_1(t) := \langle \boldsymbol{x}_1(t), \boldsymbol{F}_1(\boldsymbol{X}(t))\rangle = \langle \boldsymbol{x}_1(t), \boldsymbol{V}\boldsymbol{x}_1(t)\rangle$. We have already proved that $\boldsymbol{x}_1(t) \to s_1\boldsymbol{e}$, where $s_1 \in \{-1, 1\}$. Since the map $\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{V}\boldsymbol{x}\rangle$ is continuous, we may pass to the limit in the preceding expression and get $f_1(t) \to \langle \boldsymbol{e}, \boldsymbol{V}\boldsymbol{e}\rangle = \lambda$.

Subtract λ from the radius and rewrite the radial equation as

$$\frac{d}{dt}\big(r_1(t) - \lambda\big) = \gamma_1(t)\left(f_1(t) - \lambda - \big(r_1(t) - \lambda\big)\right).$$

Fix $T \geq 0$. Applying the integrating-factor formula on the interval $[T, t]$ gives

$$r_1(t) - \lambda = e^{-\int_T^t \gamma_1(v)\,dv}\big(r_1(T) - \lambda\big) + \int_T^t e^{-\int_s^t \gamma_1(v)\,dv}\gamma_1(s)\big(f_1(s) - \lambda\big)\,ds.$$

The first term records the remaining influence of the radius at time T. The integral term records the influence of the difference between the radial forcing and its limit λ. The kernel in the integral is nonnegative. Its total mass is

$$\int_T^t e^{-\int_s^t \gamma_1(v)\,dv}\gamma_1(s)\,ds = \Big[e^{-\int_s^t \gamma_1(v)\,dv}\Big]_{s=T}^{s=t} = 1 - e^{-\int_T^t \gamma_1(v)\,dv} \leq 1.$$

Indeed, when the expression inside the brackets is differentiated with respect to s, its derivative is $\gamma_1(s)e^{-\int_s^t \gamma_1(v)\,dv}$. Taking absolute values in the integrating-factor identity and using the kernel bound gives

$$\begin{aligned} |r_1(t) - \lambda| &\leq e^{-\int_T^t \gamma_1(s)\,ds}|r_1(T) - \lambda| \\ &\quad + \sup_{s\geq T}|f_1(s) - \lambda| \int_T^t e^{-\int_s^t \gamma_1(v)\,dv}\gamma_1(s)\,ds \\ &\leq e^{-\int_T^t \gamma_1(s)\,ds}|r_1(T) - \lambda| + \sup_{s\geq T}|f_1(s) - \lambda|. \end{aligned} \tag{37}$$

We now show directly that the right-hand side tends to zero. Let $\varepsilon > 0$. Since $f_1(t) \to \lambda$, there exists $T \geq 0$ such that

$$\sup_{s\geq T}|f_1(s) - \lambda| < \frac{\varepsilon}{2}.$$

For this fixed T, Assumption 2(a) implies

$$\int_T^\infty \gamma_1(s)\,ds = \infty.$$

To see this, note that $\int_0^\infty \gamma_1(s)\,ds = \infty$ by assumption, while $\int_0^T \gamma_1(s)\,ds$ is finite because γ_1 is continuous on the compact interval $[0, T]$. Therefore

$$e^{-\int_T^t \gamma_1(s)\,ds} \longrightarrow 0 \qquad \text{as } t \to \infty.$$

We may consequently choose $t_\varepsilon \geq T$ such that, for every $t \geq t_\varepsilon$,

$$e^{-\int_T^t \gamma_1(s)\,ds}|r_1(T) - \lambda| < \frac{\varepsilon}{2}.$$

Substituting the last two estimates into (37) gives

$$|r_1(t) - \lambda| < \varepsilon \qquad \text{for every } t \geq t_\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this proves

$$r_1(t) \longrightarrow \lambda.$$

Together with the explicit formula for $\boldsymbol{x}_1(t)$ and the limit $\boldsymbol{x}_1(t) \to s_1\boldsymbol{e}$, this completes the proof.

C.8 Proof of Lemma 5.2

By Assumption 2(a), the function $\Gamma_i(t) = \int_0^t \gamma_i(v)\,dv$ is continuously differentiable, satisfies $\Gamma_i'(t) = \gamma_i(t) > 0$, and tends to infinity. Since $\Gamma_i(0) = 0$, it is a bijection from $[0, \infty)$ onto itself. Its inverse ϑ_i is continuously differentiable, with $\vartheta_i'(s) = 1/\gamma_i(\vartheta_i(s))$; at zero the derivative is understood from the right. Also, $\vartheta_i(s) \to \infty$: for every $T \geq 0$, the inequality $s > \Gamma_i(T)$ implies $\vartheta_i(s) > T$. Both Γ_i and ϑ_i depend only on the prescribed rate, so they are independent of $\boldsymbol{u}_i$.

For a solution with initial value $\boldsymbol{u}_i$, write $\boldsymbol{z}(s) := \boldsymbol{z}_i(\vartheta_i(s))$. The raw equation (14) and the chain rule give

$$\begin{aligned}\frac{d\boldsymbol{z}}{ds} &= \dot{\boldsymbol{z}}_i(\vartheta_i(s))\vartheta_i'(s) \\ &= \boldsymbol{F}_i\left(\boldsymbol{x}_1(\vartheta_i(s)), \dots, \boldsymbol{x}_{i-1}(\vartheta_i(s)), \frac{\boldsymbol{z}(s)}{\|\boldsymbol{z}(s)\|}\right) - \boldsymbol{z}(s).\end{aligned}$$

Here $\boldsymbol{F}_i$ is written as a function of its first i directions, since later directions do not enter the causal row. Changing $\boldsymbol{u}_i$ changes only the initial condition $\boldsymbol{z}(0) = \boldsymbol{u}_i$ in this equation. This proves the first equality in (23) and explains why it is one common nonautonomous equation.

To estimate the effect of moving predecessors, we use the same softmax formula to extend $\boldsymbol{F}_i$ to arbitrary vector arguments in $(\mathbb{R}^d)^i$, without normalizing those arguments. This is a smooth extension: its numerator and denominator are smooth, and its denominator is a sum of strictly positive exponentials. On the compact product of closed unit balls, every score lies between $-\|\boldsymbol{W}\|_{\mathrm{op}}/\sqrt{d}$ and $\|\boldsymbol{W}\|_{\mathrm{op}}/\sqrt{d}$. In particular, the denominator is at least $ie^{-\|\boldsymbol{W}\|_{\mathrm{op}}/\sqrt{d}} > 0$ there. The first derivatives and the mixed second derivatives of this extension are consequently bounded on that product.

Define the field difference, before normalizing the last argument, by

$$\boldsymbol{H}_i(t, \boldsymbol{x}) := \boldsymbol{F}_i(\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_{i-1}(t), \boldsymbol{x}) - \boldsymbol{F}_i(s_1\boldsymbol{e}, \dots, s_{i-1}\boldsymbol{e}, \boldsymbol{x}).$$

Denote $D_{\boldsymbol{x}}\boldsymbol{F}$ and $D_j\boldsymbol{F}_i$ for the derivative in its $\boldsymbol{x}$ and jth vector argument. respectively. For fixed t, interpolate each predecessor along the segment $\boldsymbol{y}_j(\theta) := s_j\boldsymbol{e} + \theta(\boldsymbol{x}_j(t) - s_j\boldsymbol{e})$, where $0 \leq \theta \leq 1$ and $j < i$. Every point of this segment lies in the closed unit ball, by the triangle inequality. The fundamental theorem of calculus gives

$$\begin{aligned}\boldsymbol{H}_i(t, \boldsymbol{x}) &= \int_0^1 \sum_{j<i} D_j\boldsymbol{F}_i(\boldsymbol{y}_1(\theta), \dots, \boldsymbol{y}_{i-1}(\theta), \boldsymbol{x})[\boldsymbol{x}_j(t) - s_j\boldsymbol{e}]\,d\theta, \\ D_{\boldsymbol{x}}\boldsymbol{H}_i(t, \boldsymbol{x}) &= \int_0^1 \sum_{j<i} D_jD_{\boldsymbol{x}}\boldsymbol{F}_i(\boldsymbol{y}_1(\theta), \dots, \boldsymbol{y}_{i-1}(\theta), \boldsymbol{x})[\boldsymbol{x}_j(t) - s_j\boldsymbol{e}]\,d\theta.\end{aligned}$$

The second identity applies the same argument to the matrix-valued function $D_{\boldsymbol{x}}\boldsymbol{F}_i$; the bracket contracts the predecessor argument of its mixed derivative. The bounded derivatives above show that there is a constant C, depending only on the fixed matrices, d, and i, such that both $\|\boldsymbol{H}_i(t, \boldsymbol{x})\|$ and $\|D_{\boldsymbol{x}}\boldsymbol{H}_i(t, \boldsymbol{x})\|_{\mathrm{op}}$ are at most $C\sum_{j<i}\|\boldsymbol{x}_j(t) - s_j\boldsymbol{e}\|$ for all $t \geq 0$ and all unit $\boldsymbol{x}$. Operator norms are induced by the Euclidean norms. The same constant works for every t and $\boldsymbol{x}$ because all the interpolated arguments lie in the same compact product of balls.

We now transfer these estimates to the raw state. Let $\nu(\boldsymbol{z}) := \boldsymbol{z}/\|\boldsymbol{z}\|$ for $\boldsymbol{z} \neq \boldsymbol{0}$, and define $\boldsymbol{R}_i(s, \boldsymbol{z}) := \boldsymbol{H}_i(\vartheta_i(s), \nu(\boldsymbol{z}))$. The definition of $\boldsymbol{H}_i$ proves the second equality in (23). To estimate its derivative, for an arbitrary increment $\boldsymbol{h} \in \mathbb{R}^d$ differentiate normalization:

$$D\nu(\boldsymbol{z})[\boldsymbol{h}] = \frac{\boldsymbol{h}}{\|\boldsymbol{z}\|} - \frac{\boldsymbol{z}\,\langle \boldsymbol{z}, \boldsymbol{h}\rangle}{\|\boldsymbol{z}\|^3} = \frac{1}{\|\boldsymbol{z}\|}\big(\boldsymbol{I} - \nu(\boldsymbol{z})\nu(\boldsymbol{z})^\top\big)\boldsymbol{h}.$$

The matrix $\big(\boldsymbol{I} - \nu(\boldsymbol{z})\nu(\boldsymbol{z})^\top\big)$ is an orthogonal projection, so $\|D\nu(\boldsymbol{z})\|_{\mathrm{op}} \le 1/\|\boldsymbol{z}\|$. Fix $0 < a \le b < \infty$ and denote the corresponding compact annulus by $A_{a,b} := \{\boldsymbol{z} \in \mathbb{R}^d : a \le \|\boldsymbol{z}\| \le b\}$. The chain rule, with s held fixed, now gives

$$D_{\boldsymbol{z}}\boldsymbol{R}_i(s, \boldsymbol{z}) = D_{\boldsymbol{x}}\boldsymbol{H}_i(\vartheta_i(s), \nu(\boldsymbol{z}))\, D\nu(\boldsymbol{z}),$$

$$\|\boldsymbol{R}_i(s, \boldsymbol{z})\|_{C^1_{\boldsymbol{z}}} := \sup_{\boldsymbol{z}\in A_{a,b}} \|\boldsymbol{R}_i(s, \boldsymbol{z})\| + \sup_{\boldsymbol{z}\in A_{a,b}} \|D_{\boldsymbol{z}}\boldsymbol{R}_i(s, \boldsymbol{z})\|_{\mathrm{op}} \quad \le C(1 + a^{-1}) \sum_{j<i} \|\boldsymbol{x}_j(\vartheta_i(s)) - s_j \boldsymbol{e}\|.$$

Each term in this finite sum tends to zero by the predecessor convergence hypothesis and $\vartheta_i(s) \to \infty$. This proves the asserted convergence on every such annulus. In particular, the conclusion concerns state derivatives; no convergence of time derivatives is required.

Finally, the directional comparison used below follows from the same estimates. On the unit sphere, the difference of the projected attention fields is $(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{H}_i(t, \boldsymbol{x})$. For a tangent increment $\boldsymbol{h}$ at $\boldsymbol{x}$, its derivative is $-(\boldsymbol{h}\boldsymbol{x}^\top + \boldsymbol{x}\boldsymbol{h}^\top)\boldsymbol{H}_i(t, \boldsymbol{x}) + (\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\top)D_{\boldsymbol{x}}\boldsymbol{H}_i(t, \boldsymbol{x})[\boldsymbol{h}]$. Since $\|\boldsymbol{x}\| = 1$, its norm is at most $(2\|\boldsymbol{H}_i(t, \boldsymbol{x})\| + \|D_{\boldsymbol{x}}\boldsymbol{H}_i(t, \boldsymbol{x})\|_{\mathrm{op}})\|\boldsymbol{h}\|$. Thus this projected difference also tends to zero in C^1 uniformly on the sphere. Substituting any inverse effective time $t_i(\tau) \to \infty$ preserves that convergence, as needed in the subsequent directional argument.

C.9 Proof of Lemma 5.3

We prove the result in four stages. We first express the directional equation as a vanishing perturbation of the limiting field. We then show that this limiting field is a gradient field on the unit sphere and identify all of its critical points. Next, we prove that the directional trajectory approaches one connected component of the critical set. Finally, in the nonpolar case, we use the radial equation to recover the corresponding limit of the raw token.

Step 1: Reduction to a perturbed limiting directional flow. By Lemma 4.2, the individual effective time of the ith token is

$$\tau_i(t) = \int_0^t \frac{\gamma_i(s)}{r_i(s)}\, ds,$$

and satisfies $\tau_i(t) \to \infty$. Since its derivative is strictly positive, τ_i has an inverse, which we denote by t_i. Define the direction in this effective time by $\boldsymbol{y}(\tau) := \boldsymbol{x}_i(t_i(\tau))$. The directional equation in Lemma 4.2, together with the field estimate established in Lemma 5.2, gives

$$\begin{aligned} \boldsymbol{y}' &= (\boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^\top)\,[\alpha(q)\boldsymbol{V}\boldsymbol{y} + \lambda\chi(q)\boldsymbol{e}] + \mathcal{E}_i(\tau, \boldsymbol{y}), \\ q &= \langle \boldsymbol{y}, \boldsymbol{e}\rangle, \qquad \|\mathcal{E}_i(\tau, \cdot)\|_{C^1(\mathbb{S}^{d-1})} \longrightarrow 0. \end{aligned} \tag{38}$$

Here $\mathcal{E}_i$ is the difference between the actual projected attention field and the projected field obtained by fixing every predecessor at its limiting pole. In particular, the error tends to zero uniformly on the sphere. Only this uniform convergence, rather than the full C^1 convergence, is needed in the present proof.

Step 2: A potential for the limiting field. Define $\Phi : \mathbb{S}^{d-1} \to \mathbb{R}$ by

$$\Phi(\boldsymbol{y}) := \frac{1}{2}\langle \boldsymbol{y}, \boldsymbol{V}\boldsymbol{y}\rangle + \lambda \int_0^q \frac{\chi(v)}{\alpha(v)}\, dv.$$

This function is smooth. Indeed, α and χ are smooth on $[-1, 1]$, and $\alpha(q) > 0$ for every $q \in [-1, 1]$ because $\alpha(q) = E(q)/\mathcal{Z}(q)$ and both $E(q)$ and $\mathcal{Z}(q)$ are strictly positive. Since $\boldsymbol{V}$ is symmetric, the Euclidean gradient of Φ is

$$\nabla\Phi(\boldsymbol{y}) = \boldsymbol{V}\boldsymbol{y} + \lambda\frac{\chi(q)}{\alpha(q)}\boldsymbol{e}.$$

The tangent space of the unit sphere at $\boldsymbol{y}$ is $\boldsymbol{y}^{\perp}$, and the orthogonal projection onto this tangent space is $\boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^{\top}$. Therefore, the spherical gradient of Φ is

$$\nabla_{\mathbb{S}}\Phi(\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^{\top})\left[\boldsymbol{V}\boldsymbol{y} + \lambda\frac{\chi(q)}{\alpha(q)}\boldsymbol{e}\right].$$

Multiplying this equality by $\alpha(q)$ shows that the autonomous part of (38) is precisely $\alpha(q)\nabla_{\mathbb{S}}\Phi(\boldsymbol{y})$. Hence,

$$\boldsymbol{y}' = \alpha(q)\nabla_{\mathbb{S}}\Phi(\boldsymbol{y}) + \mathcal{E}_i(\tau, \boldsymbol{y}).$$

Along the trajectory, the chain rule consequently gives the important identity

$$\frac{d}{d\tau}\Phi(\boldsymbol{y}(\tau)) = \alpha(q)\|\nabla_{\mathbb{S}}\Phi(\boldsymbol{y})\|^2 + \langle\nabla_{\mathbb{S}}\Phi(\boldsymbol{y}), \mathcal{E}_i(\tau, \boldsymbol{y})\rangle. \tag{39}$$

Without the error term, this identity says that Φ is nondecreasing and is strictly increasing away from its critical set. Because the error tends uniformly to zero, the same strict increase holds at sufficiently large times on every closed set that stays a positive distance from the critical set.

We now identify the critical points of Φ. A unit vector $\boldsymbol{y}$ is critical if and only if

$$(\boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^{\top})\left[\boldsymbol{V}\boldsymbol{y} + \lambda\frac{\chi(q)}{\alpha(q)}\boldsymbol{e}\right] = \boldsymbol{0}.$$

Equivalently, there is a scalar c such that

$$\boldsymbol{V}\boldsymbol{y} + \lambda\frac{\chi(q)}{\alpha(q)}\boldsymbol{e} = c\boldsymbol{y}. \tag{40}$$

The two unit eigenvectors $\boldsymbol{e}$ and $-\boldsymbol{e}$ satisfy this condition because $\boldsymbol{V}\boldsymbol{e} = \lambda\boldsymbol{e}$. Suppose that $\boldsymbol{y}$ is not $\pm\boldsymbol{e}$. Decompose it according to the eigenspaces of $\boldsymbol{V}$:

$$\boldsymbol{y} = q\boldsymbol{e} + \sum_{\nu<\lambda}\boldsymbol{y}_\nu, \qquad \boldsymbol{y}_\nu \in H_\nu.$$

Because $\boldsymbol{y}$ is not $\pm\boldsymbol{e}$, at least one transverse component $\boldsymbol{y}_\mu$ is nonzero. Projecting (40) onto eigenspace H_μ gives

$$(\mu - c)\boldsymbol{y}_\mu = \boldsymbol{0}.$$

Hence $\mu = c$. Then we can write $\boldsymbol{y} = q\boldsymbol{e} + \boldsymbol{y}_\mu + \sum_{\nu<\lambda,\nu\neq\mu}\boldsymbol{y}_\nu$, $\boldsymbol{y}_\nu \in H_\nu$. For any eigenvalue $\nu \neq \mu$, project (40) onto H_ν yields $(\nu - \mu)\boldsymbol{y}_\nu = 0$, which implies $y_\nu = 0$. It follows that there is a lower eigenvalue $\mu < \lambda$ such that

$$\boldsymbol{y} = q\boldsymbol{e} + \sqrt{1 - q^2}\,\boldsymbol{w},$$

for some $\boldsymbol{w} \in H_\mu$, $\|\boldsymbol{w}\| = 1$. Projecting (40) onto the leading direction $\boldsymbol{e}$ then gives

$$\lambda q + \lambda\frac{\chi(q)}{\alpha(q)} = \mu q.$$

After multiplication by $\alpha(q)$, this becomes

$$\alpha(q)(\lambda - \mu)q + \lambda\chi(q) = 0.$$

Using $\alpha(q) = E(q)/\mathcal{Z}(q)$ and the definition of $\chi(q)$, multiplication by $\mathcal{Z}(q)$ yields

$$(\lambda - \mu)qE(q) + \lambda\left(n_+e^{m_0q/\sqrt{d}} - n_-e^{-m_0q/\sqrt{d}}\right) = G_\mu(q) = 0.$$

The strict monotonicity established in (25) implies that G_μ has at most one root in $(-1, 1)$. Therefore, if such a root exists, it must be $q = q_\mu$, and the corresponding critical directions are exactly the vectors in $\mathcal{M}_\mu$. We have thus proved that the critical set of Φ is

$$\{\boldsymbol{e}, -\boldsymbol{e}\} \cup \bigcup_{\substack{\mu<\lambda:\\ G_\mu(q_\mu)=0,\ q_\mu\in(-1,1)}} \mathcal{M}_\mu. \tag{41}$$

There are only finitely many distinct eigenvalues μ, and there is at most one q_μ for each of them. Hence this critical set has only finitely many connected components. The function Φ is constant on

each set $\mathcal{M}_\mu$. To see this, take $\boldsymbol{y} = q_\mu \boldsymbol{e} + \sqrt{1-q_\mu^2}\,\boldsymbol{w} \in \mathcal{M}_\mu$. Since $\boldsymbol{w} \in H_\mu$ and $\|\boldsymbol{w}\| = 1$, we have

$$\langle \boldsymbol{y}, \boldsymbol{V}\boldsymbol{y} \rangle = \lambda q_\mu^2 + \mu(1-q_\mu^2).$$

This value depends only on q_μ and μ, and the integral term in Φ also depends only on q_μ. It is therefore independent of the choice of $\boldsymbol{w}$. In particular, Φ has only finitely many critical values.

Step 3: Convergence toward one critical component. We first prove that the scalar function $\tau \mapsto \Phi(\boldsymbol{y}(\tau))$ converges. Since the sphere is compact and Φ is continuous, this function is bounded.

Let $I \subset \mathbb{R}$ be a closed interval containing no critical value of Φ, and suppose that $\Phi^{-1}(I)$ is nonempty. The set $\Phi^{-1}(I)$ is compact and contains no critical point. Therefore, the continuous function $\|\nabla_{\mathbb{S}}\Phi\|$ has a strictly positive minimum on that set. Denote this minimum by $c_I > 0$. Also set

$$\alpha_* := \min_{q\in[-1,1]} \alpha(q) > 0, \qquad L_\Phi := \max_{\boldsymbol{y}\in\mathbb{S}^{d-1}} \|\nabla_{\mathbb{S}}\Phi(\boldsymbol{y})\| < \infty.$$

Because $\mathcal{E}_i(\tau,\cdot) \to 0$ uniformly, there is a time τ_I such that, for every $\tau \geq \tau_I$,

$$L_\Phi \|\mathcal{E}_i(\tau,\cdot)\|_{C^0(\mathbb{S}^{d-1})} \leq \frac{1}{2}\alpha_* c_I^2.$$

Whenever $\tau \geq \tau_I$ and $\Phi(\boldsymbol{y}(\tau)) \in I$, (39) and the Cauchy–Schwarz inequality give

$$\begin{aligned}
\frac{d}{d\tau}\Phi(\boldsymbol{y}(\tau)) &\geq \alpha_* \|\nabla_{\mathbb{S}}\Phi(\boldsymbol{y})\|^2 - \|\nabla_{\mathbb{S}}\Phi(\boldsymbol{y})\|\|\mathcal{E}_i(\tau,\boldsymbol{y})\| \\
&\geq \alpha_* c_I^2 - L_\Phi \|\mathcal{E}_i(\tau,\cdot)\|_{C^0(\mathbb{S}^{d-1})} \\
&\geq \frac{1}{2}\alpha_* c_I^2 > 0.
\end{aligned}$$

Thus, at sufficiently large times, the potential crosses every closed interval containing no critical value only in the increasing direction. It cannot remain in such an interval indefinitely because its derivative there has a fixed positive lower bound. After it exits through the upper endpoint, it cannot return through that endpoint in the decreasing direction. Consequently, no noncritical value of Φ can be a limit value of $\Phi(\boldsymbol{y}(\tau))$.

The set of limit values of a bounded continuous real-valued function on $[0,\infty)$ is the interval between its lower and upper limits. Indeed, if the lower and upper limits were different, continuity and the intermediate value theorem would produce arbitrarily late visits to every value strictly between them. In the present case, every limit value must belong to the finite set of critical values of Φ. A finite set cannot contain a nondegenerate interval. Therefore,

$$\lim_{\tau\to\infty} \Phi(\boldsymbol{y}(\tau))$$

exists.

We next prove that every limit point of $\boldsymbol{y}(\tau)$ is a critical point of Φ. Suppose, for contradiction, that a noncritical point $\boldsymbol{y}_*$ were a limit point. Choose neighborhoods $U_0 \Subset U$ of $\boldsymbol{y}_*$ such that $\|\nabla_{\mathbb{S}}\Phi\|$ is bounded away from zero on the closure of U. The same estimate as above shows that, for all sufficiently large τ, there is a constant $\delta > 0$ such that

$$\frac{d}{d\tau}\Phi(\boldsymbol{y}(\tau)) \geq \delta \qquad \text{whenever } \boldsymbol{y}(\tau) \in U.$$

The velocity in (38) is uniformly bounded at large times because the limiting field is continuous on the compact sphere and the error is uniformly small. Hence there is a fixed time length $\ell > 0$ such that every sufficiently late visit to U_0 keeps the trajectory inside U for at least time ℓ. During each such visit, the potential increases by at least $\delta\ell$. Since $\boldsymbol{y}_*$ is assumed to be a limit point, there would be arbitrarily late visits to U_0. This contradicts the already established convergence of $\Phi(\boldsymbol{y}(\tau))$. Therefore, every limit point of $\boldsymbol{y}$ is critical.

Let

$$\omega(\boldsymbol{y}) := \bigcap_{T\geq 0} \overline{\{\boldsymbol{y}(\tau) : \tau \geq T\}}$$

denote the limit set of the directional trajectory. This set is nonempty and compact because the trajectory remains on the compact sphere. It is also connected: for each T, the trajectory tail

$\{\boldsymbol{y}(\tau) : \tau \geq T\}$ is the continuous image of the connected interval $[T, \infty)$, so its closure is connected; the sets obtained as T increases form a nested family of nonempty compact connected sets, whose intersection is connected.

We have proved that $\omega(\boldsymbol{y})$ is contained in the critical set (41). That critical set has only finitely many disjoint compact connected components. Since $\omega(\boldsymbol{y})$ is connected, it must lie in one of those components. The trajectory must then approach that component. Indeed, otherwise there would be a sequence $\tau_n \to \infty$ staying a fixed positive distance from it; compactness would produce a convergent subsequence whose limit belongs to $\omega(\boldsymbol{y})$ but not to that component.

If the selected component is the isolated point $\{\boldsymbol{e}\}$ or $\{-\boldsymbol{e}\}$, then

$$\boldsymbol{y}(\tau) \longrightarrow \boldsymbol{e} \qquad \text{or} \qquad \boldsymbol{y}(\tau) \longrightarrow -\boldsymbol{e}.$$

Since $\tau_i(t) \to \infty$ and $\boldsymbol{x}_i(t) = \boldsymbol{y}(\tau_i(t))$, this gives $\boldsymbol{x}_i(t) \to \boldsymbol{e}$ or $\boldsymbol{x}_i(t) \to -\boldsymbol{e}$, which is the first alternative of the lemma. Suppose instead that the selected component is contained in $\mathcal{M}_\mu$ for some lower eigenvalue μ. Then

$$\operatorname{dist}\big(\boldsymbol{x}_i(t), \mathcal{M}_\mu\big) \longrightarrow 0.$$

Every vector in $\mathcal{M}_\mu$ has leading coordinate q_μ. It follows immediately that $\langle \boldsymbol{x}_i(t), \boldsymbol{e}\rangle \longrightarrow q_\mu$.

Step 4: The radial and raw-state limits in the nonpolar case. Set $L_\mu := \alpha(q_\mu)\mu$. We first show that the radial forcing converges to L_μ. Let $\boldsymbol{x} \in \mathcal{M}_\mu$. Then $\boldsymbol{x} = q_\mu \boldsymbol{e} + \sqrt{1 - q_\mu^2}\, \boldsymbol{w}$ for some $\boldsymbol{w} \in H_\mu,\ \|\boldsymbol{w}\| = 1$ and hence

$$\boldsymbol{V}\boldsymbol{x} = \lambda q_\mu \boldsymbol{e} + \mu\sqrt{1 - q_\mu^2}\, \boldsymbol{w}.$$

The root identity $G_\mu(q_\mu) = 0$ can be written as

$$\lambda\chi(q_\mu) = -\alpha(q_\mu)(\lambda - \mu)q_\mu.$$

Substituting these two identities into the formula for the limiting field gives

$$\begin{aligned}
\overline{\boldsymbol{F}}_i(\boldsymbol{x}) &= \alpha(q_\mu)\boldsymbol{V}\boldsymbol{x} + \lambda\chi(q_\mu)\boldsymbol{e} \\
&= \alpha(q_\mu)\left[\lambda q_\mu \boldsymbol{e} + \mu\sqrt{1 - q_\mu^2}\, \boldsymbol{w}\right] - \alpha(q_\mu)(\lambda - \mu)q_\mu \boldsymbol{e} \\
&= \alpha(q_\mu)\mu\left[q_\mu \boldsymbol{e} + \sqrt{1 - q_\mu^2}\, \boldsymbol{w}\right] = L_\mu \boldsymbol{x}.
\end{aligned}$$

Therefore $\langle \boldsymbol{x}, \overline{\boldsymbol{F}}_i(\boldsymbol{x})\rangle = L_\mu$ for every $\boldsymbol{x} \in \mathcal{M}_\mu$. By Lemma 5.2, the actual attention field converges uniformly to the limiting field as the predecessors converge to $\pm\boldsymbol{e}$. The map $\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \overline{\boldsymbol{F}}_i(\boldsymbol{x})\rangle$ is continuous on the compact unit sphere and is equal to L_μ on $\mathcal{M}_\mu$. Together with $\operatorname{dist}(\boldsymbol{x}_i(t), \mathcal{M}_\mu) \to 0$, these facts imply that the actual radial forcing $f_i(t) := \langle \boldsymbol{x}_i(t), \boldsymbol{F}_i(\boldsymbol{X}(t))\rangle$ satisfies $f_i(t) \longrightarrow L_\mu$. The exact radial equation is

$$\dot{r}_i(t) = \gamma_i(t)\big(f_i(t) - r_i(t)\big).$$

Subtracting L_μ and applying the integrating-factor formula on $[T, t]$ gives

$$r_i(t) - L_\mu = e^{-\int_T^t \gamma_i(v)\, dv}\big(r_i(T) - L_\mu\big) + \int_T^t e^{-\int_s^t \gamma_i(v)\, dv}\gamma_i(s)\big(f_i(s) - L_\mu\big)\, ds.$$

The integral kernel is nonnegative and has total mass

$$\int_T^t e^{-\int_s^t \gamma_i(v)\, dv}\gamma_i(s)\, ds = 1 - e^{-\int_T^t \gamma_i(v)\, dv} \leq 1.$$

Taking absolute values therefore yields

$$|r_i(t) - L_\mu| \leq e^{-\int_T^t \gamma_i(v)\, dv}|r_i(T) - L_\mu| + \sup_{s \geq T}|f_i(s) - L_\mu|. \tag{42}$$

Assumption 2(a) gives $\int_T^\infty \gamma_i(v)\, dv = \infty$. Hence, for fixed T, the first term on the right-hand side of (42) tends to zero as $t \to \infty$. We obtain

$$\limsup_{t \to \infty}|r_i(t) - L_\mu| \leq \sup_{s \geq T}|f_i(s) - L_\mu|.$$

Since $f_i(t) \to L_\mu$, the right-hand side tends to zero as $T \to \infty$. Thus $r_i(t) \longrightarrow L_\mu = \alpha(q_\mu)\mu$. This limit is positive. Indeed, the condition (11) implies $\lambda_d > 0$, and therefore every eigenvalue μ of $\boldsymbol{V}$ satisfies $\mu \geq \lambda_d > 0$. Since $\alpha(q_\mu) > 0$, we have $L_\mu > 0$.

It remains to translate the directional and radial convergence into the asserted convergence toward $\mathcal{K}_\mu$. Since $\mathcal{M}_\mu$ is compact, for each t choose $\widehat{\boldsymbol{x}}(t) \in \mathcal{M}_\mu$ such that

$$\|\boldsymbol{x}_i(t) - \widehat{\boldsymbol{x}}(t)\| = \operatorname{dist}\big(\boldsymbol{x}_i(t), \mathcal{M}_\mu\big).$$

Because $L_\mu \widehat{\boldsymbol{x}}(t) \in \mathcal{K}_\mu$, we have

$$\operatorname{dist}\big(\boldsymbol{z}_i(t), \mathcal{K}_\mu\big) \leq \|r_i(t)\boldsymbol{x}_i(t) - L_\mu \widehat{\boldsymbol{x}}(t)\| \leq r_i(t)\|\boldsymbol{x}_i(t) - \widehat{\boldsymbol{x}}(t)\| + |r_i(t) - L_\mu|.$$

The radii are uniformly bounded by Lemma 4.2, the first difference tends to zero because the direction approaches $\mathcal{M}_\mu$, and the second difference tends to zero by the radial limit just proved. Therefore,

$$\operatorname{dist}\big(\boldsymbol{z}_i(t), \mathcal{K}_\mu\big) \longrightarrow 0.$$

This proves the nonpolar alternative and completes the proof.

C.10 Proof of Lemma 5.4

Fix $S_0 \geq S$. Suppose that two solutions $(\boldsymbol{p}, \boldsymbol{h})$ and $(\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{h}})$ remain in the bounded region for all $s \geq S_0$, with $\boldsymbol{h}(S_0) = \widetilde{\boldsymbol{h}}(S_0)$ but $\boldsymbol{p}(S_0) \neq \widetilde{\boldsymbol{p}}(S_0)$. We will show that their $\boldsymbol{p}$-coordinates must separate exponentially, contradicting boundedness.

Write $\Delta\boldsymbol{p} := \boldsymbol{p} - \widetilde{\boldsymbol{p}}$ and $\Delta\boldsymbol{h} := \boldsymbol{h} - \widetilde{\boldsymbol{h}}$, and let $\Delta\boldsymbol{f}$ and $\Delta\boldsymbol{g}$ denote the differences of the corresponding remainder terms at these two states at the same time. Subtracting their equations gives

$$(\Delta\boldsymbol{p})' = \boldsymbol{A}\Delta\boldsymbol{p} + \Delta\boldsymbol{f}, \qquad (\Delta\boldsymbol{h})' = \boldsymbol{C}\Delta\boldsymbol{h} + \Delta\boldsymbol{g}.$$

Since both states remain in the region, each remainder difference has norm at most $\epsilon(\|\Delta\boldsymbol{p}\| + \|\Delta\boldsymbol{h}\|)$.

Set $Q(s) := \|\Delta\boldsymbol{p}(s)\|^2 - \|\Delta\boldsymbol{h}(s)\|^2$. Thus $Q \geq 0$ describes the cone $\|\Delta\boldsymbol{p}\| \geq \|\Delta\boldsymbol{h}\|$, and $Q(S_0) = \|\Delta\boldsymbol{p}(S_0)\|^2 > 0$. The matrix inequalities and Cauchy–Schwarz yield

$$\begin{aligned} Q' &= 2\langle\Delta\boldsymbol{p}, \boldsymbol{A}\Delta\boldsymbol{p}\rangle - 2\langle\Delta\boldsymbol{h}, \boldsymbol{C}\Delta\boldsymbol{h}\rangle + 2\langle\Delta\boldsymbol{p}, \Delta\boldsymbol{f}\rangle - 2\langle\Delta\boldsymbol{h}, \Delta\boldsymbol{g}\rangle \\ &\geq 2\kappa\|\Delta\boldsymbol{p}\|^2 - 2\|\Delta\boldsymbol{p}\|\,\|\Delta\boldsymbol{f}\| - 2\|\Delta\boldsymbol{h}\|\,\|\Delta\boldsymbol{g}\| \\ &\geq 2\kappa\|\Delta\boldsymbol{p}\|^2 - 2\epsilon(\|\Delta\boldsymbol{p}\| + \|\Delta\boldsymbol{h}\|)^2. \end{aligned}$$

At a nonzero boundary point, $Q = 0$ implies $\|\Delta\boldsymbol{p}\| = \|\Delta\boldsymbol{h}\| > 0$. Hence

$$Q' \geq 2\kappa\|\Delta\boldsymbol{p}\|^2 - 8\epsilon\|\Delta\boldsymbol{p}\|^2 = (2\kappa - 8\epsilon)\|\Delta\boldsymbol{p}\|^2 > 0, \tag{43}$$

where the strict inequality uses $\epsilon < \kappa/4$. While $Q > 0$, the vector $\Delta\boldsymbol{p}$ is nonzero and $\|\Delta\boldsymbol{h}\| < \|\Delta\boldsymbol{p}\|$. We may therefore differentiate its norm and obtain

$$\frac{d}{ds}\|\Delta\boldsymbol{p}\| = \frac{\langle\Delta\boldsymbol{p}, \boldsymbol{A}\Delta\boldsymbol{p} + \Delta\boldsymbol{f}\rangle}{\|\Delta\boldsymbol{p}\|} \geq \kappa\|\Delta\boldsymbol{p}\| - \|\Delta\boldsymbol{f}\| \geq (\kappa - 2\epsilon)\|\Delta\boldsymbol{p}\| \geq \frac{\kappa}{2}\|\Delta\boldsymbol{p}\|.$$

Integration gives $\|\Delta\boldsymbol{p}(s)\| \geq e^{\kappa(s-S_0)/2}\|\Delta\boldsymbol{p}(S_0)\|$ as long as Q remains positive.

This estimate also rules out contact with the cone boundary. Indeed, if $T > S_0$ were the first time with $Q(T) = 0$, continuity and the preceding estimate would give $\|\Delta\boldsymbol{p}(T)\| \geq e^{\kappa(T-S_0)/2}\|\Delta\boldsymbol{p}(S_0)\| > 0$. The contact would therefore occur at a nonzero boundary point. Since Q is positive just before T and vanishes at T, its derivative there must satisfy $Q'(T) \leq 0$, contradicting (43). Thus $Q(s) > 0$ for every $s \geq S_0$.

The exponential lower bound now holds for all future times and forces $\|\Delta\boldsymbol{p}(s)\| \to \infty$. But both solutions remain in a bounded region, so their $\boldsymbol{p}$-coordinates, and hence their difference, remain bounded. This contradiction proves that there is at most one initial $\boldsymbol{p}$ for each initial $\boldsymbol{h}$ among solutions that remain in the region for all later times.

Let $\mathcal{T}$ denote this set of states at time S_0, and suppose it is Borel. If $\boldsymbol{h} \in \mathbb{R}^b$, where b is the dimension of the remaining coordinates, each section $\mathcal{T}_{\boldsymbol{h}} := \{\boldsymbol{p} \in \mathbb{R}^a : (\boldsymbol{p}, \boldsymbol{h}) \in \mathcal{T}\}$ is Borel and contains at

most one point. Because $a \geq 1$, its a-dimensional Lebesgue measure is zero. Tonelli's theorem, applied to the indicator of $\mathcal{T}$, therefore gives

$$\mathrm{vol}_{a+b}(\mathcal{T}) = \int_{\mathbb{R}^b} \mathrm{vol}_a(\mathcal{T}_{\boldsymbol{h}})\, d\boldsymbol{h} = 0.$$

If $\boldsymbol{h}$ instead ranges over a smooth submanifold M, the comparison above still applies: all differences and norms are taken in its ambient Euclidean space. Each section $\mathcal{T}_{\boldsymbol{h}}$ consequently still has zero a-dimensional measure. The induced volume on M is σ-finite, as follows from a countable cover by coordinate patches of finite volume. Applying Tonelli with this induced volume in the outer integral proves the same zero-product-volume conclusion.

C.11 Proof of Lemma 5.5

Fix the predecessor trajectories as in the statement. We will show that, at each sufficiently late time, the states that remain near $\mathcal{K}_\mu$ have zero volume. We will then transfer this conclusion back to the initial value $\boldsymbol{u}_i$.

Put $q_* := q_\mu$ and $r_* := \alpha(q_*)\mu$. The radial certificate implies $\lambda_d > 0$, so $\mu > 0$ and $r_* > 0$. Let $H_\mu = \ker(\boldsymbol{V} - \mu \boldsymbol{I})$ and $m = \dim H_\mu$. Because $q_* \in (-1, 1)$, every direction in $\mathcal{M}_\mu$ has a nonzero projection onto H_μ. Thus, throughout a neighborhood of $\mathcal{K}_\mu$, we can write

$$\boldsymbol{z} = r\left(q\boldsymbol{e} + \boldsymbol{v} + \sqrt{1 - q^2 - \|\boldsymbol{v}\|^2}\,\boldsymbol{w}\right), \qquad \boldsymbol{v} \in (\mathrm{span}\{\boldsymbol{e}\} \oplus H_\mu)^\perp, \quad \boldsymbol{w} \in H_\mu, \quad \|\boldsymbol{w}\| = 1. \tag{44}$$

Here $r = \|\boldsymbol{z}\|$, $q = \langle \boldsymbol{z}/r, \boldsymbol{e}\rangle$, and $\boldsymbol{v}$ is the projection of $\boldsymbol{z}/r$ onto $(\mathrm{span}\{\boldsymbol{e}\} \oplus H_\mu)^\perp$. The vector $\boldsymbol{w}$ is the normalized projection of $\boldsymbol{z}/r$ onto H_μ. These formulas recover the coordinates smoothly whenever $r > 0$ and $1 - q^2 - \|\boldsymbol{v}\|^2 > 0$. In these coordinates, $\mathcal{K}_\mu$ is given by $r = r_*$, $q = q_*$, $\boldsymbol{v} = \boldsymbol{0}$, and arbitrary $\boldsymbol{w}$ on the unit sphere of H_μ.

We first examine the limiting equation in the common time $s = \Gamma_i(t)$. Its raw state satisfies $\boldsymbol{z}' = \overline{\boldsymbol{F}}_i(\boldsymbol{z}/\|\boldsymbol{z}\|) - \boldsymbol{z}$; a prime denotes an s-derivative. Set $\boldsymbol{x} = \boldsymbol{z}/r$. Taking the radial component and the component tangent to the unit sphere gives $r' = \langle \boldsymbol{x}, \overline{\boldsymbol{F}}_i(\boldsymbol{x})\rangle - r$ and $\boldsymbol{x}' = r^{-1}(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^\top)\overline{\boldsymbol{F}}_i(\boldsymbol{x})$.

For each $k \geq 2$ with $\lambda_k \neq \mu$, write $v_k = \langle \boldsymbol{v}, \boldsymbol{e}_k\rangle$. Orthogonality of the eigenspaces yields

$$\langle \boldsymbol{x}, \boldsymbol{V}\boldsymbol{x}\rangle = \lambda q^2 + \mu(1 - q^2) + \sum_{\substack{k \geq 2 \\ \lambda_k \neq \mu}} (\lambda_k - \mu) v_k^2.$$

For the radial equation, write $\psi_\mu(q) := \alpha(q)[\lambda q^2 + \mu(1 - q^2)] + \lambda\chi(q)q$. Substituting the preceding identity into the radial and directional equations gives

$$\begin{aligned} q' &= \frac{1}{r}\left[\frac{(1 - q^2)G_\mu(q)}{\mathcal{Z}(q)} - \alpha(q)q \sum_{\substack{k \geq 2 \\ \lambda_k \neq \mu}} (\lambda_k - \mu) v_k^2\right], \\ v_k' &= \frac{\alpha(q)(\lambda_k - \langle \boldsymbol{x}, \boldsymbol{V}\boldsymbol{x}\rangle) - \lambda\chi(q)q}{r}\, v_k, \\ r' &= \psi_\mu(q) + \alpha(q) \sum_{\substack{k \geq 2 \\ \lambda_k \neq \mu}} (\lambda_k - \mu) v_k^2 - r, \qquad \boldsymbol{w}' = \boldsymbol{0}. \end{aligned} \tag{45}$$

For clarity, the first equation uses $\alpha(q)(\lambda - \mu)q + \lambda\chi(q) = G_\mu(q)/\mathcal{Z}(q)$. The last equation holds because the projection of the directional velocity onto H_μ is parallel to that projection of $\boldsymbol{x}$: $\boldsymbol{V}$ acts as $\mu\boldsymbol{I}$ on H_μ. Its normalized direction $\boldsymbol{w}$ is therefore constant for the limiting equation. This need not hold for the actual, time-dependent equation.

We now identify which coordinates expand near $\mathcal{K}_\mu$. The identity $G_\mu(q_*) = 0$ gives $\lambda\chi(q_*) = -\alpha(q_*)(\lambda - \mu)q_*$. Consequently $\psi_\mu(q_*) = \alpha(q_*)\mu = r_*$, as required for a raw equilibrium. Since the terms involving $\boldsymbol{v}$ in the equation for q are quadratic, the coefficient of $q - q_*$ in its linearization is

$$a_\mu := \frac{(1 - q_*^2)G_\mu'(q_*)}{r_*\mathcal{Z}(q_*)} > 0.$$

The inequality follows from (25). At the same equilibrium, the coefficient of v_k is $(\lambda_k - \mu)/\mu$. Finally, the linearized radial equation is $(r - r_*)' = \psi'_\mu(q_*)(q - q_*) - (r - r_*)$.

The radial coordinate is linearly coupled to $q - q_*$. Remove that coupling by setting

$$\ell := r - r_* - \frac{\psi'_\mu(q_*)}{a_\mu + 1}(q - q_*). \tag{46}$$

Indeed, substituting the two linearized equations into the derivative of this expression gives $\ell' = -\ell$ at linear order. Thus the linearized equations are $(q - q_*)' = a_\mu(q - q_*)$, $v'_k = ((\lambda_k - \mu)/\mu)v_k$, $\ell' = -\ell$, and $\boldsymbol{w}' = \mathbf{0}$.

Collect the v_k with $\lambda_k > \mu$ into $\boldsymbol{v}_>$ and those with $\lambda_k < \mu$ into $\boldsymbol{v}_<$. Define $\boldsymbol{p} = (q - q_*, \boldsymbol{v}_>)$ and $\boldsymbol{h} = (\ell, \boldsymbol{v}_<, \boldsymbol{w})$. The linearized equation then has the form (27). Its matrix $\boldsymbol{A}$ is diagonal with entries a_μ and $(\lambda_k - \mu)/\mu$ for $\lambda_k > \mu$; its matrix $\boldsymbol{C}$ has the entry -1, the entries $(\lambda_k - \mu)/\mu$ for $\lambda_k < \mu$, and zero on the ambient coordinates of $\boldsymbol{w}$. In particular, $\boldsymbol{A} \succeq \kappa_\mu \boldsymbol{I}$ for some $\kappa_\mu > 0$, whereas $\boldsymbol{C} \preceq 0$. The expanding vector $\boldsymbol{p}$ always has at least one coordinate, namely $q - q_*$.

We next check the difference estimate required by Lemma 5.4. Subtract the linear terms just described from the limiting equation. At $q = q_*$, $\boldsymbol{v} = \mathbf{0}$, and $\ell = 0$, the resulting remainder and its derivatives in the normal coordinates vanish. The limiting normal equations do not depend on $\boldsymbol{w}$, and its $\boldsymbol{w}$-equation is zero. Hence, by continuity on the compact unit sphere of H_μ, the derivatives of the limiting remainder can be made uniformly as small as desired by restricting $(q - q_*, \boldsymbol{v}, \ell)$ to a sufficiently small closed product neighborhood of zero.

Lemma 5.2 supplies the other part of the estimate. It says that the actual raw equation differs from the limiting one by a term tending to zero in C^1 on every compact annulus away from $\mathbf{0}$. The coordinate map above and its inverse have bounded derivatives of the orders needed on a slightly larger neighborhood of $\mathcal{K}_\mu$. The additional remainder in the $(\boldsymbol{p}, \boldsymbol{h})$ equation therefore also tends to zero in C^1, uniformly over the whole unit sphere of $\boldsymbol{w}$.

These derivative bounds give (28) for every pair of states, including pairs with different $\boldsymbol{w}$. To see this, first vary their normal coordinates along a line segment while holding $\boldsymbol{w}$ fixed. If $m \geq 2$, then vary $\boldsymbol{w}$ along a shortest spherical arc; its length is at most $\pi/2$ times the Euclidean distance between its endpoints. Integrating the derivative bounds along these paths controls the remainder difference. If $m = 1$, the sphere consists of two points. The limiting remainder agrees at both points, and the difference of the actual additional remainders is bounded by twice its uniform norm. Thus this case gives the same estimate. We may consequently choose the normal neighborhood small enough, and then a time S_μ late enough, that (28) holds with some $0 < \epsilon < \kappa_\mu/4$ for all $s \geq S_\mu$.

Let $\mathcal{U}_\mu$ be the raw image of this closed coordinate product, including the entire unit sphere of H_μ. It is compact, is contained in $\mathbb{R}^d \setminus \{\mathbf{0}\}$, and contains $\mathcal{K}_\mu$ in its interior. We have now verified all the local hypotheses of Lemma 5.4 for the actual equation at times $s \geq S_\mu$.

Denote by $\varphi(s, s_0, \boldsymbol{z}_0)$ the solution of the common raw equation at time s with state $\boldsymbol{z}_0$ at time s_0. This is the same equation for every choice of $\boldsymbol{u}_i$, because the predecessor trajectories were fixed at the start of the proof. Solutions from nonzero states exist for all later times and remain nonzero: the radial comparison of Lemma 4.2, applied from time s_0, bounds their radii below by a positive constant and above by a finite constant.

For an integer $n \geq S_\mu$, let $\mathcal{T}_{\mu,n}$ consist of the states at time n whose solutions remain in $\mathcal{U}_\mu$ for every $s \geq n$. This is a Borel set. Indeed, because $\mathcal{U}_\mu$ is closed and solutions are continuous in time, it is enough to require membership in $\mathcal{U}_\mu$ at every rational $s \geq n$. There are countably many such requirements, and each defines a Borel set by continuous dependence on the state at time n.

In the coordinates constructed above, Lemma 5.4 says that, for each fixed $\boldsymbol{h}$ at time n, at most one $\boldsymbol{p}$ belongs to $\mathcal{T}_{\mu,n}$. Tonelli's theorem therefore gives zero product volume to its coordinate image. Here the $\boldsymbol{w}$ factor carries the induced volume on the unit sphere of H_μ; when $m = 1$, that sphere has two points and carries its zero-dimensional volume. In local sphere charts, the coordinate map back to the raw state is smooth between spaces of dimension d and hence locally Lipschitz. A countable chart cover shows that $\mathcal{T}_{\mu,n}$ has zero d-dimensional Lebesgue measure in the raw state space.

The map $\boldsymbol{z}_0 \mapsto \varphi(n, 0, \boldsymbol{z}_0)$ is a local C^1 diffeomorphism on $\mathbb{R}^d \setminus \{\mathbf{0}\}$. The raw vector field is continuous in time and C^1 in the nonzero state. Its derivative flow along any finite solution segment

solves a linear variational equation with identity initial matrix, so it remains invertible. The inverse function theorem gives a local C^1 inverse. On sufficiently small neighborhoods these inverses are Lipschitz; a countable cover therefore shows that the preimage of the null set $\mathcal{T}_{\mu,n}$ is also null.

Finally, suppose $\operatorname{dist}(\boldsymbol{z}_i(t;\boldsymbol{u}_i),\mathcal{K}_\mu)\to 0$. Since $s=\Gamma_i(t)$ increases to infinity and $\mathcal{K}_\mu$ lies in the interior of $\mathcal{U}_\mu$, the solution eventually remains in $\mathcal{U}_\mu$. At some integer $n\geq S_\mu$ its state must therefore belong to $\mathcal{T}_{\mu,n}$. Its initial value belongs to the preimage of one of these null sets. Taking the countable union over integers $n\geq S_\mu$ proves that $\mathcal{E}_{i,\mu}$ has zero Lebesgue measure.

For completeness, $\mathcal{E}_{i,\mu}$ is Borel as well. At each fixed t, the map $\boldsymbol{u}_i\mapsto \boldsymbol{z}_i(t;\boldsymbol{u}_i)$ is continuous on the admissible initial states. Because $\mathcal{K}_\mu$ is compact, convergence of its distance to zero can be expressed using countable intersections and unions of the conditions $\operatorname{dist}(\boldsymbol{z}_i(t;\boldsymbol{u}_i),\mathcal{K}_\mu)\leq 1/k$ at rational times t beyond an integer threshold. Continuity in time extends the rational-time conditions to all later times. Intersecting this Borel convergence event with $\{\boldsymbol{u}_i:\langle \boldsymbol{u}_i,\boldsymbol{e}\rangle\neq 0\}$ gives exactly $\mathcal{E}_{i,\mu}$.

C.12 Proof of Lemma 5.6

Fix the predecessor initial data as in the statement. By Lemma 5.3, every admissible solution for token i either converges directionally to $\boldsymbol{e}$ or $-\boldsymbol{e}$, or satisfies $\operatorname{dist}(\boldsymbol{z}_i(t),\mathcal{K}_\mu)\to 0$ for some lower eigenvalue $\mu<\lambda$ for which $\mathcal{K}_\mu$ is defined.

For each such μ, Lemma 5.5 gives a Borel null set $\mathcal{E}_{i,\mu}$ containing every admissible initial value that produces the corresponding limit that is not $\pm\boldsymbol{e}$. There are only finitely many distinct eigenvalues of $\boldsymbol{V}$. Hence

$$\mathcal{E}_i:=\bigcup_{\mu<\lambda}\mathcal{E}_{i,\mu}$$

is Borel and has zero d-dimensional Lebesgue measure.

C.13 Proof of Theorem 5.1

We work in the single-axis setting $\lambda_1>\lambda_2$, with the matrices and prescribed history rates fixed. The proof first establishes directional convergence outside a Borel null set by induction on the number of tokens. We then evaluate the limiting radial forcing and use the radial equation to obtain the remaining conclusions.

For $1\leq i\leq N$, let $\mathcal{D}_i:=\{(\boldsymbol{u}_1,\dots,\boldsymbol{u}_i)\in(\mathbb{R}^d)^i:\langle \boldsymbol{u}_j,\boldsymbol{e}\rangle\neq 0 \text{ for every } j\leq i\}$. Thus $\mathcal{D}_i=\mathcal{D}_1^i$ and $\mathcal{D}_N=\mathcal{D}$. Every initial token in these sets is nonzero. Moreover, because the leading eigenspace is $\operatorname{span}\{\boldsymbol{e}\}$, the direction $\boldsymbol{\xi}$ in Assumption 2(c) is either $\boldsymbol{e}$ or $-\boldsymbol{e}$. The nonzero leading projections therefore imply $|\langle \boldsymbol{u}_j,\boldsymbol{\xi}\rangle|/\|\boldsymbol{u}_j\|>0$ for every token. The initial-data conditions in Assumption 2(b),(c) are thus satisfied.

Lemma 4.2 gives a unique global solution for every such initial configuration, with radii bounded away from zero. The same conclusion holds for every prefix of the system: the first i equations involve only the first i tokens, and the radial estimates remain valid for those equations. Consequently, if $U^{(i)}=(\boldsymbol{u}_1,\dots,\boldsymbol{u}_i)\in\mathcal{D}_i$, the directions $\boldsymbol{x}_j(t;U^{(i)})$, $j\leq i$, are determined independently of any later initial tokens. This is the causal property that allows us to induct on i.

Let $\mathcal{G}_i\subset\mathcal{D}_i$ consist of the prefixes for which every one of the first i directions converges to $s_i\boldsymbol{e}$, and set $\mathcal{B}_i:=\mathcal{D}_i\setminus\mathcal{G}_i$. For each fixed finite t, the map $U^{(i)}\mapsto \boldsymbol{x}_j(t;U^{(i)})$ is continuous. This follows from continuous dependence for the raw ODE and the positivity of the radii, which makes normalization continuous. We now prove by induction that $\mathcal{B}_i$ has zero di-dimensional Lebesgue measure. For $i=1$, Lemma 4.3 applies to every $\boldsymbol{u}_1\in\mathcal{D}_1$ and gives convergence to $s_1\boldsymbol{e}$, where $s_1=\operatorname{sign}\langle \boldsymbol{u}_1,\boldsymbol{e}\rangle$. Thus $\mathcal{B}_1=\varnothing$.

Suppose the assertion holds for $i-1$. For each predecessor tuple $U^{(i-1)}\in\mathcal{D}_{i-1}$, define the section $\mathcal{B}_i(U^{(i-1)}):=\{\boldsymbol{u}_i\in\mathcal{D}_1:(U^{(i-1)},\boldsymbol{u}_i)\in\mathcal{B}_i\}$. If $U^{(i-1)}\in\mathcal{G}_{i-1}$, all its predecessor directions converge to poles. Their trajectories stay fixed as $\boldsymbol{u}_i$ varies, by causality. Lemma 5.6 therefore applies and shows that token i also converges to a pole outside a d-dimensional null set of choices of $\boldsymbol{u}_i$. Since the predecessors already converge, it follows that $\mathcal{B}_i(U^{(i-1)})$ has measure zero.

For predecessor tuples in $\mathcal{B}_{i-1}$, no restriction on the size of this section is needed: those tuples themselves form a null set by the induction hypothesis. To apply Tonelli without any issue from unbounded sections, fix an integer $L \geq 1$ and restrict all raw coordinates to $[-L, L]$. Writing vol_k for k-dimensional Lebesgue measure, Tonelli gives

$$\mathrm{vol}_{di}\big(\mathcal{B}_i \cap [-L,L]^{di}\big) = \int_{\mathcal{D}_{i-1}\cap[-L,L]^{d(i-1)}} \mathrm{vol}_d\big(\mathcal{B}_i(U^{(i-1)}) \cap [-L,L]^d\big)\, dU^{(i-1)} = 0.$$

The integrand vanishes on $\mathcal{G}_{i-1}$ and is at most $(2L)^d$ on the null set $\mathcal{B}_{i-1}$. Its measurability follows from the Borel property of $\mathcal{B}_i$ and Tonelli's theorem. Thus the displayed integral is zero. Letting L range over the positive integers covers $\mathcal{B}_i$ by countably many null sets and completes the induction. In particular, this argument does not require choosing the conditional exceptional sets in Lemma 5.6 measurably as the predecessors vary.

Set $\mathcal{E} := \mathcal{B}_N$. It is Borel and has zero Nd-dimensional Lebesgue measure. For every $U \in \mathcal{D} \setminus \mathcal{E}$, all directions converge to one of two unit leading eigenvectors of $\boldsymbol{V}$, proving the desired convergence $\boldsymbol{x}_i(t) \to s_i \boldsymbol{e}$.

Fix such a configuration U, and write $\boldsymbol{X}_\infty := (s_1\boldsymbol{e}, \ldots, s_N\boldsymbol{e})$ for its limiting directional configuration. For each token, define its radial forcing by $f_i(t) := \langle \boldsymbol{x}_i(t), \boldsymbol{F}_i(\boldsymbol{X}(t))\rangle$. The softmax denominator is strictly positive, so the attention field is continuous in the directions. Consequently, $f_i(t) \to \langle s_i\boldsymbol{e}, \boldsymbol{F}_i(\boldsymbol{X}_\infty)\rangle$. We compute this limit directly. Since $\boldsymbol{W}\boldsymbol{e} = m_0\boldsymbol{e}$ and $\boldsymbol{V}\boldsymbol{e} = \lambda\boldsymbol{e}$, the limiting score and radial value for a source $j \leq i$ are

$$\frac{(s_i\boldsymbol{e})^\top \boldsymbol{W}(s_j\boldsymbol{e})}{\sqrt{d}} = \frac{m_0 s_i s_j}{\sqrt{d}}, \qquad \langle s_i\boldsymbol{e}, \boldsymbol{V}(s_j\boldsymbol{e})\rangle = \lambda s_i s_j.$$

There are $1 + n_{i,s_i}$ sources with the same sign as token i, including the self term, and $n_{i,-s_i}$ sources with the opposite sign. Their scores are respectively $m_0/\sqrt{d}$ and $-m_0/\sqrt{d}$, and their radial values are respectively λ and $-\lambda$. Therefore

$$\begin{aligned}\lim_{t\to\infty} f_i(t) &= \lambda \frac{(1+n_{i,s_i})e^{m_0/\sqrt{d}} - n_{i,-s_i}e^{-m_0/\sqrt{d}}}{(1+n_{i,s_i})e^{m_0/\sqrt{d}} + n_{i,-s_i}e^{-m_0/\sqrt{d}}} \\ &= \lambda \frac{(1+n_{i,s_i})e^{2m_0/\sqrt{d}} - n_{i,-s_i}}{(1+n_{i,s_i})e^{2m_0/\sqrt{d}} + n_{i,-s_i}} = R_i.\end{aligned}$$

The second equality multiplies numerator and denominator by $e^{m_0/\sqrt{d}}$. Lemma 4.1 applies to the limiting configuration $\boldsymbol{X}_\infty$ itself, so $R_i = \langle s_i\boldsymbol{e}, \boldsymbol{F}_i(\boldsymbol{X}_\infty)\rangle \geq \eta > 0$.

Finally, the exact radial equation is $\dot{r}_i = \gamma_i(t)(f_i - r_i)$. For any fixed $T \geq 0$, the estimate (42), with $L_\mu = R_i$, gives

$$|r_i(t) - R_i| \leq e^{-\int_T^t \gamma_i(v)\,dv}|r_i(T) - R_i| + \sup_{v\geq T}|f_i(v) - R_i|, \qquad t \geq T.$$

Assumption 2(a) implies $\int_T^\infty \gamma_i(v)\,dv = \infty$, since the integral over $[0,T]$ is finite. Thus the exponential term tends to zero for fixed T. Taking the upper limit as $t \to \infty$ and then letting $T \to \infty$ proves $r_i(t) \to R_i$, because $f_i(t) \to R_i$. Multiplying this radial limit by $\boldsymbol{x}_i(t) \to s_i\boldsymbol{e}$ yields $\boldsymbol{z}_i(t) = r_i(t)\boldsymbol{x}_i(t) \to R_i s_i \boldsymbol{e}$. This proves (20) and completes the theorem.