Understanding the Effects of Hyper-Connections on Self-Attention Dynamics: A Bifurcation Analysis

Nhi Y. Pham
Center for AI Research
VinUniversity, Hanoi, Vietnam
nhi.py2@vinuni.edu.vn

Dung V. Nguyen
Department of Mathematics
National University of Singapore
dungnv@u.nus.edu

Trinh T. Nguyen
School of Mathematical Sciences
Shanghai Jiao Tong University
q0004349@sjtu.edu.cn

Thieu N. Vo[†]
Department of Computer Science
University of Bath
ntv22@bath.ac.uk

Tan M. Nguyen[†]
Department of Mathematics
National University of Singapore
tanmn@nus.edu.sg

Abstract

A central challenge in understanding deep sequence models is characterizing how token representations evolve across layers: whether they collapse, diverge, or converge to nontrivial structures. We study this question for Transformers with hyper-connections, whose learnable cross-layer coupling makes the stability of representation collapse analytically tractable. In contrast to standard residual connections, hyper-connections introduce additional parameters that directly control the hyperbolicity of the zero fixed point, where hyperbolicity allows the Jacobian to determine local stability, thereby enabling a rigorous and precise bifurcation analysis of the dynamical system underlying the model. In particular, we derive a continuous-depth ODE limit of self-attention with hyper-connections and use bifurcation theory to characterize its token dynamics. For the 1-stream case, we obtain a closed-form critical threshold and give conditions under which crossing this threshold produces a pitchfork bifurcation, separating representation local collapse from escape or convergence to nontrivial stable states. For the 2-stream case, the richer coupling structure yields two distinct types of stability boundaries: one where a real eigenvalue of the Jacobian reaches zero (static type), and another where a complex-conjugate pair reaches the imaginary axis at $\pm i\omega$, with $\omega > 0$ (oscillatory type). Experiments on pretrained hyper-connected language models provide qualitative comparisons with the ODE regimes.

1 Introduction

A central question in understanding modern deep sequence models, spanning architectures from self-attention in transformers [42, 36, 12, 19, 30, 6, 48] to state-space formulations [4, 9, 37, 23, 50, 3, 29], is a precise characterization of how token representations evolve across depth. Recent dynamical analyses [2, 8, 28, 34, 43, 22, 31, 1] reveal that these token dynamics organize into distinct regimes–collapse,

[†]Co-last authors.
Please correspond to: nhi.py2@vinuni.edu.vn and tanmn@nus.edu.sg.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

clustering, controlled divergence, and metastability–whose qualitative structure directly governs model expressivity and downstream performance [41, 6, 51, 55, 52, 17, 25, 38, 34]. A powerful technique for analyzing these phenomena is to formulate deep sequence models as continuous-depth dynamical systems, in which the layer index plays the role of time, and token representations interact and update one another as they propagate through the network [11, 2, 47, 26]. This viewpoint has yielded concrete insights into the asymptotic behavior of Transformers [16, 34, 7, 27] and selective state-space models such as Mamba [43, 37, 24, 20].

However, existing token dynamics analyses focus on standard residual architectures, where the strength of interaction between layers, i.e., the layer-to-layer coupling, is essentially fixed. As a result, there is no explicit parameter that controls whether representations collapse to a trivial state, remain stable, or become unstable, thereby precluding a precise bifurcation-theoretic characterization of when and how representations collapse. Recently, hyper-connections (HC) have replaced residual connections (RC) with a richer family of learnable couplings that control representation evolution [54, 48, 5].

1.1 Residual Connection vs. Hyper-Connection

Residual Connection. The classical Transformer [42] updates each layer k with an RC:

$$x^{(\ell)}(k+1) = x^{(\ell)}(k) + \mathrm{Layer}\big(x^{(\ell)}(k), \{x^{(i)}(k)\}_{i=1}^{L}\big). \tag{1}$$

Here, the superscript $\ell = 1,...,L$ indexes the tokens, Layer denotes any layer-level computation, such as self-attention and feed-forward layers, and the addition of the identity term $x^{(\ell)}(k)$ implements the residual connection. A key issue with residual connections is that they predefine the strength of connections between the output and input within a layer, thereby preventing the network from achieving optimal trade-offs between vanishing gradients and representation collapse.

Hyper-Connection. HC expands RC to N parallel streams per token and learns the optimal strength of connections to improve performance. The N-stream update in HC is given by:

$$x_i^{(\ell)}(k+1) = \sum_{j=1}^{N} h_{ij}^{(\ell)} x_j^{(\ell)}(k) + \beta_i^{(\ell)} \mathrm{Layer}\left(\sum_{j=1}^{N} \alpha_j^{(\ell)} x_j^{(\ell)}(k), \Big\{ \sum_{j=1}^{N} \alpha_j^{(l)} x_j^{(l)}(k) \Big\}_{l=1}^{L} \right), \tag{2}$$

where $i = 1,...,N$ indexes the streams and $\ell = 1,...,L$ indexes the tokens. Here, $H^{(\ell)} = (h_{ij}^{(\ell)}) \in \mathbb{R}^{N\times N}$ is a learnable intra-token connection matrix that replaces the fixed identity term in residual connections by combining information across streams. The coefficients $\alpha_j^{(\ell)}$ are input-scaling weights that aggregate the streams to form the input to the layer, and $\beta_i^{(\ell)}$ are output-scaling weights that distribute the layer output back across the streams. There are two HC variants: *Static HC* (SHC), where $H^{(\ell)}, \alpha^{(\ell)} = [\alpha_1^{(\ell)},...,\alpha_N^{(\ell)}]^\top, \beta^{(\ell)} = [\beta_1^{(\ell)},...,\beta_N^{(\ell)}]^\top$ are fixed learned parameters, and *Dynamic HC* (DHC), where they are additionally conditioned on the input via tanh-gated projections.

HC recovers RC, including both Pre-Norm and Post-Norm variants, as special cases with fixed, non-trainable parameters [54]. More importantly, they introduce additional learned parameters that render the hyperbolicity of the zero fixed point analytically tunable. A fixed point is hyperbolic when the Jacobian has no purely imaginary eigenvalues, under which linearisation fully determines local stability. For $L > 1$, our unmasked residual ODE has zero eigenvalues at the origin due to uniform token averaging. HC parameters shift the Jacobian spectrum, making hyperbolicity tunable and permitting simple stability crossings under the stated conditions. Despite this advantage, the dynamical effects of hyper-connections remain largely unexplored, motivating a systematic bifurcation analysis of deep sequence models such as self-attention under hyper-connected coupling.

1.2 Contribution

In this paper, we take a first-principles approach to understanding token dynamics in Transformers with HC. We formulate the forward pass as a continuous-depth dynamical system, thereby revealing the role of HC-induced architectural coupling in shaping the evolution of representations. This perspective shows that representation collapse is not merely an empirical phenomenon, but a question of the stability of a distinguished equilibrium. Building on this insight, we develop a bifurcation-theoretic framework that reveals how changes in coupling parameters lead to qualitatively distinct dynamical regimes. Our contribution is three-fold:

1. We derive continuous-depth ODE limits of self-attention with both 1-stream and 2-stream hyper-connections, and use these limits to characterize the local phase portrait around the collapsed (zero) state.

2. For the 1-stream case (SHC-1), we characterize the stability region of the collapsed state and establish sufficient conditions for a pitchfork bifurcation as a hyper-connection parameter passes through an explicitly determined critical value. This bifurcation separates regimes of collapse, convergence to nontrivial stable states, and local escape, with its precise nature determined by the spectrum of the value matrix.
3. For the 2-stream case (SHC-2), we show that the richer coupling structure introduces static and oscillatory stability boundaries; the oscillatory boundary has no analog in the 1-stream model under our assumptions. This demonstrates that adding parallel streams fundamentally reshapes the dynamical behavior, rather than merely increasing model capacity.

We compare the ODE predictions with pretrained hyper-connected language-model trajectories and perplexities. These comparisons are qualitative: the implementation and regime-label limitations in Appendix H prevent interpreting them as a direct validation of every ODE stability condition.

Scope of Our Work: We focus exclusively on SHC, since the time-independent parameters yield an *autonomous* ODE amenable to bifurcation analysis. In contrast, state-dependent weights in DHC change the nonlinear vector field; explicit depth dependence makes it non-autonomous, which falls outside the scope of our framework and requires fundamentally different techniques; we leave this direction for future work.

Within SHC, we study two cases:

(1) **SHC-1** ($N=1$): the 1-stream case, where $H^{(\ell)}$ degenerates to a learnable scalar $h^{(\ell)}$ and the pre/post mappings to scalars $\alpha^{(\ell)},\beta^{(\ell)}$:

$$x^{(\ell)}(k+1)=h^{(\ell)}x^{(\ell)}(k)+\beta^{(\ell)}\mathrm{Layer}\big(\alpha^{(\ell)}x^{(\ell)}(k),\{\alpha^{(i)}x^{(i)}(k)\}_{i=1}^{L}\big). \tag{3}$$

When $h^{(\ell)}=1$ and $\alpha^{(\ell)}=\beta^{(\ell)}=1$, (3) recovers (1).

(2) **SHC-2** ($N=2$, shared parameters): the simplest multi-stream case, analysed under the tractability assumption that $H^{(\ell)}=H\in\mathbb{R}^{2\times 2}$, $\alpha^{(\ell)}=\alpha\in\mathbb{R}^2$, and $\beta^{(\ell)}=\beta\in\mathbb{R}^2$ are shared across token positions. Substituting $N=2$ into (2) gives

$$x_i^{(\ell)}(k+1)=\sum_{j=1}^{2}h_{ij}x_j^{(\ell)}(k)+\beta_i\mathrm{Layer}\left(\sum_{j=1}^{2}\alpha_j x_j^{(\ell)}(k),\left\{\sum_{j=1}^{2}\alpha_j x_j^{(l)}(k)\right\}_{l=1}^{L}\right),\quad i=1,2. \tag{4}$$

We adopt token-wise hyperparameters for SHC-1 and shared ones for SHC-2 as complementary frameworks to isolate distinct architectural phenomena. A detailed justification is provided in Remark F.9.

2 Dynamics of Self-Attention with Hyper-Connections

We derive the continuous-depth limit of self-attention with hyper-connections. Following the setting in [34], we consider the case where $\mathrm{Layer}(\cdot)$ in (1) and (2) consists of **self-attention only** without a feed-forward sublayer, with weight matrices shared across all layers, i.e., the attention matrix $W\in\mathbb{R}^{d\times d}$ and the value matrix $V\in\mathbb{R}^{d\times d}$ are time-independent. We denote $Z^{(\ell)}(t)\in\mathbb{R}^d$ the representation of token ℓ at continuous depth t for the 1-stream case, and $Z_i^{(\ell)}(t)\in\mathbb{R}^d$ for the N-stream case (stream i).

The self-attention output for a query token u against context $\{v^{(i)}\}_{i=1}^{L}$ is given by:

$$\mathrm{Attn}(u;\{v^{(i)}\}):=V^\top\sum_{i=1}^{L}\mathrm{softmax}\big(u^\top Wv^{(i)}\big)v^{(i)}, \tag{5}$$

where $\mathrm{softmax}(u^\top Wv^{(i)})=e^{u^\top Wv^{(i)}}/\sum_j e^{u^\top Wv^{(j)}}$. Our ODE analysis considers unmasked self-attention: each token attends to all token positions. Hereafter, residual self-attention refers to this unmasked setting. Using the step-size scaling described in Appendix B, we obtain the following continuous-depth models for residual self-attention, SHC-1, and SHC-2, corresponding respectively to (1), (3), and (4). For each token $\ell=1,\dots,L$, the three models are

$$\dot{Z}^{(\ell)}=V^\top\sum_{i=1}^{L}\mathrm{softmax}\big({Z^{(\ell)}}^\top WZ^{(i)}\big)Z^{(i)}, \tag{ODE-1}$$

$$\dot{Z}^{(\ell)}=(h^{(\ell)}-1)Z^{(\ell)}+\beta^{(\ell)}V^\top\sum_{i=1}^{L}\mathrm{softmax}\big((\alpha^{(\ell)}Z^{(\ell)})^\top W(\alpha^{(i)}Z^{(i)})\big)\alpha^{(i)}Z^{(i)}, \tag{ODE-2}$$

$$\dot{Z}_i^{(\ell)}=\sum_{j=1}^{2}(h_{ij}-\delta_{ij})Z_j^{(\ell)}+\beta_iV^\top\sum_{h=1}^{L}\mathrm{softmax}\Big(\tilde{Z}^{(\ell)\top}W\tilde{Z}^{(h)}\Big)\tilde{Z}^{(h)},\qquad i=1,2, \tag{ODE-3}$$

where $\tilde{Z}^{(\ell)} := \sum_{j=1}^{2} \alpha_j Z_j^{(\ell)}$ is the weighted-sum attention input for token ℓ. Note that the dynamics in (ODE-1) follow the formulation of [34].

3 Assumptions, Fundamental Lemma, and Residual Connection's Limitation

In this section, we introduce the local stability framework used throughout the paper, discuss the standing assumptions, and identify a common fixed point shared by residual and hyper-connected dynamics. We then define the collapsed state, where all token representations vanish and explain how persistent zero eigenvalues complicate bifurcation analysis of the unmasked residual ODE.

3.1 Standing Assumptions and Common Fixed Point Lemma

We first state the following assumptions, which will be used throughout our theoretical analysis.

Assumption 3.1. *Consider the self-attention dynamics with hyper-connections introduced in Section 2. We assume the following:*

(A) The value matrix V is a real symmetric matrix, with nonzero eigenvalues listed with multiplicity as $\lambda_1,\dots,\lambda_r$ and corresponding eigenvectors $v_1,\dots,v_r \in \mathbb{R}^d$. For $r>0$, we denote $\lambda_{\max} := \max_k \lambda_k$ and $\lambda_{\min} := \min_k \lambda_k$.

(B) Input- and output-scaling weights are positive: $\alpha^{(\ell)},\beta^{(\ell)} > 0$ for SHC-1, and $\alpha_1,\alpha_2,\beta_1,\beta_2 > 0$ for SHC-2. For pitchfork sign claims only, assume $W = W^\top \succ 0$. Throughout, $L \geq 2$.

Discussion of Assumption 3.1. The symmetry of V ensures that all eigenvalues of $V^\top$ are real, which is the key technical condition underlying our analysis: the secular equation (7) and the stability thresholds in Theorems 4.2–5.2 are stated in terms of $\lambda_{\max}$ and $\lambda_{\min}$, and these are only well-defined as scalar quantities when the spectrum is real. Without symmetry, eigenvalues may be complex, and the stability characterization would require tracking complex spectral quantities, substantially complicating the analysis. Positivity of the input/output scaling weights is natural because they represent connection strengths and scaling weights, and it avoids degenerate sign cancellations unrelated to the stability mechanisms studied here.

Next, the following fundamental lemma motivates our analysis in this paper.

Lemma 3.2 (Common fixed point)**.** *The origin $Z=0$ is a fixed point of each of* (ODE-1)-(ODE-3).

This lemma shows that the fully collapsed representation is a common equilibrium of both residual and hyper-connected self-attention dynamics: at zero, the attention logits vanish, the attention weights become uniform, and the vector field is zero. The proof is elementary and given in Appendix C.

We now formalize the corresponding notion of collapse as follows.

Definition 3.3 (Representation collapse)**.** We say that the token dynamics undergo representation collapse if, along the continuous-depth trajectory, all token representations converge to the zero state, i.e.,

$$Z^{(\ell)}(t) \to 0 \qquad \text{for all } \ell = 1,\dots,L \qquad \text{as } t \to \infty.$$

Local asymptotic stability requires stability and attraction of all sufficiently close states. In this regime, all input-dependent information is lost. By contrast, instability of the origin indicates local escape for some perturbations, rather than necessarily global divergence.

Thus, our goal is to determine when the origin is stable or unstable as a function of the architectural parameters and the spectral properties of $V^\top$. This local stability question provides the foundation for the bifurcation analysis developed in the following sections.

3.2 Residual Connections: Persistent Non-Hyperbolicity Complicates Bifurcation Analysis

The residual connection (ODE-1) was studied in [34]. Two spectral scenarios arise:

(i) when $A := W(V^\top)^{-1} \prec 0$ and $W \succ 0$ (all eigenvalues of $V^\top$ are negative given V is symmetric, see Lemma C.3), the origin is a global attractor;

(ii) when V^T has at least one positive eigenvalue, the origin is unstable.

This dichotomy shows that residual self-attention can lead either to representation collapse or local instability, depending on the spectrum of the value matrix. For comparison, the hyper-connected systems studied below admit both locally exponentially stable and unstable regimes within each

spectral scenario, depending on the connection parameters (Theorems 4.2, 4.3, and 5.2). For the unmasked residual ODE, the persistent zero eigenvalues of the Jacobian at the origin require analysis beyond hyperbolic linearization.

The following proposition makes this degeneracy explicit.

Proposition 3.4 (Non-hyperbolicity of residual self-attention)**.** *For the residual self-attention dynamics* (ODE-1)*, the Jacobian at the origin is*

$$J_{\mathrm{RC}} = \frac{1}{L}\mathbf{1}_L\mathbf{1}_L^\top \otimes V^\top.$$

In particular, $\mathrm{rank}(J_{\mathrm{RC}}) \le d < dL$*, and therefore the origin is non-hyperbolic whenever* $L > 1$.

The proposition shows that the residual Jacobian retains zero eigenvalues as the value matrix varies. Persistent zero eigenvalues leave asymptotic stability undecided by linearization, and bifurcation analysis may require examining nonlinear terms. HC parameters can shift these eigenvalues, yielding regions of local exponential stability and tractable stability boundaries.

Remark 3.5 (Hyper-connections enable rigorous analysis)**.** The Jacobian matrix at the origin of (ODE-1) is rank-deficient. Setting $h^{(\ell)} \neq 1$ in (ODE-2) adds $(h^{(\ell)}-1)I_d$ to each diagonal block (see Lemma D.2), making hyperbolicity tunable away from the spectral boundaries and enabling a rigorous local bifurcation analysis. The condition $h^{(\ell)} \neq 1$ alone does not guarantee hyperbolicity.

4 One-Stream Static Hyper-Connection Dynamics: Bifurcation Analysis

We first analyze the 1-stream static hyper-connection self-attention model, SHC-1, where the residual identity coefficient is replaced by a learnable scalar parameter. We derive the eigenstructure of the linearization at the collapsed state, use it to characterize the stability region of the origin, and show that crossing a closed-form critical threshold induces a pitchfork bifurcation under some conditions. After rescaling $X^{(\ell)} = \alpha^{(\ell)} Z^{(\ell)}$, $\gamma^{(\ell)} = \alpha^{(\ell)}\beta^{(\ell)}$ (see Lemma D.1 in Appendix D), (ODE-2) becomes

$$\dot{X}^{(\ell)} = (h^{(\ell)}-1)X^{(\ell)} + \gamma^{(\ell)} V^\top \sum_{i=1}^{L} \mathrm{softmax}({X^{(\ell)}}^\top W X^{(i)}) X^{(i)}. \tag{6}$$

4.1 Eigenstructure of the Jacobian

We study the Jacobian matrix of (6) at the origin, specified by the following characteristic equation.

Theorem 4.1 (Characteristic equation for SHC-1)**.** *Let* $\lambda_1,\ldots,\lambda_r$ *be the nonzero eigenvalues of* $V^\top$*. The Jacobian* J *of* (6) *at* $X = 0$ *has the form*

$$J_{\mathrm{SHC1}} = H_{\mathrm{diag}} \otimes I_d + M \otimes V^\top,$$

with $H_{\mathrm{diag}} = \mathrm{diag}(h^{(1)}-1,\ldots,h^{(L)}-1)$ *and* $M = \frac{1}{L}\gamma\mathbf{1}_L^\top$ *where* $\gamma = (\gamma^{(1)},\gamma^{(2)},\ldots,\gamma^{(L)})^\top \in \mathbb{R}^L$*. For distinct* $h^{(i)}$*, its* dL *eigenvalues are (repeated parameters are treated in Theorem D.5):*

(a) the values $h^{(i)}-1$ *with multiplicity* $d-r$ *each; and*

(b) for each λ_k*, the* L *solutions of the* secular equation

$$f(\mu) := \sum_{i=1}^{L} \frac{\gamma^{(i)}}{\mu - (h^{(i)}-1)} = \frac{L}{\lambda_k}. \tag{7}$$

When all $h^{(i)}$ *are distinct, these solutions are strictly separated from the poles* $h^{(i)}-1$.

We first consider distinct carry-over coefficients $h^{(i)}$, which separate the secular poles and clarify the relation between carry-over coefficients and the stability spectrum. Repeated coefficients require grouping and additional eigenmodes for tokens sharing a coefficient. To focus the main text on the stability mechanism, we defer this general treatment, including the additional modes and their possible crossings, to Appendix D, along with all proofs. The function $f(\mu)$ is strictly decreasing between consecutive poles $p_i = h^{(i)}-1$, while the sign of L/λ_k determines the solutions' positions relative to the poles; this geometry drives the stability analysis. Full proofs are in Appendix D.

Theorem 4.1 reduces the dL-dimensional Jacobian spectrum to the scalar secular equation (7) for each nonzero eigenvalue λ_k of $V^\top$. The poles $h^{(i)}-1$ encode the hyper-connection architecture, while L/λ_k encodes the value-matrix spectrum, enabling closed-form stability and bifurcation analysis of SHC-1.

4.2 Stability Region

We use the eigenstructure above to characterize local stability of the collapsed state. Recall that $\lambda_{\max} := \max_k \lambda_k$ where $\{\lambda_k\}_{k=1}^r$ are nonzero eigenvalues of $V^\top$. We distinguish the cases $\lambda_{\max} < 0$ and $\lambda_{\max} > 0$, corresponding to contractive and amplifying value-matrix directions. Assuming $V^\top$ is full-rank and all $h^{(i)}$ are distinct (see Appendix D for the general case), we obtain the following results.

Theorem 4.2 (Stability: Scenario 1, $\lambda_{\max} < 0$)**.** *The origin is locally exponentially stable iff all $h^{(i)} \le 1$; or exactly one $h_{\max} > 1$ with all other $h^{(i)} < 1$ and $f(0) := \sum_i \frac{\gamma^{(i)}}{1-h^{(i)}} < \frac{L}{\lambda_{\max}}$. It is unstable if two or more $h^{(i)} \ge 1$, or if exactly one $h_{\max} > 1$ with all other $h^{(i)} < 1$ and $f(0) > \frac{L}{\lambda_{\max}}$.*

Theorem 4.3 (Stability: Scenario 2, $\lambda_{\max} > 0$)**.** *The origin is locally exponentially stable iff all $h^{(i)} < 1$ and $f(0) < \frac{L}{\lambda_{\max}}$. It is unstable if any $h^{(i)} \ge 1$, or all $h^{(i)} < 1$ with $f(0) > \frac{L}{\lambda_{\max}}$.*

Theorems 4.2 and 4.3 show that stability depends on the carry-over coefficients $h^{(i)}$, attention strengths $\gamma^{(i)}$, and the spectrum of $V^\top$. If $\lambda_{\max} < 0$, the value matrix is contractive, allowing limited expansion in a hyper-connection coefficient; if $\lambda_{\max} > 0$, an amplifying direction exists, requiring $h^{(i)} < 1$ and $f(0) < L/\lambda_{\max}$ for stability. Thus, $h^{(i)}$ can either help escape collapse or prevent divergence, depending on the spectral regime.

4.3 Pitchfork Bifurcation

The stability conditions above identify where the collapsed state changes stability. We now show that, under a one-parameter variation, this boundary corresponds to a pitchfork bifurcation, yielding a precise local phase portrait and explaining how nontrivial token representations emerge.

Fix $h^{(i)} < 1$ for $i \ge 2$ and vary $h^{(1)}$ as the bifurcation parameter. Define $m := \frac{L}{\lambda_{\max}} - \sum_{i=2}^{L} \frac{\gamma^{(i)}}{1-h^{(i)}}$.

For a scalar pitchfork classification, we consider a stability boundary at which exactly one eigenvalue reaches zero, with algebraic multiplicity one, while all remaining eigenvalues stay negative. We also require this eigenvalue to cross zero with nonzero speed as $h^{(1)}$ varies, and the cubic coefficient of the reduced equation to be nonzero. The first requirement ensures a genuine stability crossing; the second lets the cubic term determine the nearby equilibrium branches and their stability. The theorem below gives explicit parameter conditions ensuring these properties.

Theorem 4.4 (Pitchfork bifurcation of SHC-1)**.** *The stability boundary $f(0) = L/\lambda_{\max}$ is crossed at the closed-form critical value*

$$h_1^* := 1 - \gamma^{(1)}/m.$$

Assume the hidden dimension $d = 1$ and $W > 0$. Scenario 2 requires $m > 0$ and $u^* \neq \mathbf{1}_L/L$, with u^* defined in (28). Let c_η and c_3 be constants depending on h_1^* and the model parameters $(\gamma, \lambda_{\max}, W)$, defined explicitly via center manifold reduction (see (28)–(29) in the appendix), where $c_3 < 0$ corresponds to supercritical and $c_3 > 0$ to subcritical behaviour. Here $c_\eta > 0$ and $\operatorname{sign}(c_3) = \operatorname{sign}(\lambda_{\max})$. Locally, as $h^{(1)}$ crosses h_1^*, (6) undergoes a pitchfork bifurcation:

- *(i)* ***Stability switch.*** *For $h^{(1)}$ sufficiently close to h_1^*, the origin $X = 0$ is locally exponentially stable when $h^{(1)} < h_1^*$, and unstable when $h^{(1)} > h_1^*$.*
- *(ii)* ***Supercritical*** *($c_\eta c_3 < 0$)**.** For $h^{(1)} > h_1^*$ sufficiently close to the threshold, two stable fixed points emerge continuously from the origin with amplitude $\theta^* = \pm\sqrt{c_\eta (h^{(1)} - h_1^*)/|c_3|}(1 + o(1))$. Nearby trajectories off the origin's stable manifold converge to one of these states.*
- *(iii)* ***Subcritical*** *($c_\eta c_3 > 0$)**.** For $h^{(1)} > h_1^*$ sufficiently close to the threshold, no stable fixed point exists near the origin; and trajectories off the origin's local stable manifold leave a neighbourhood. For $h^{(1)} < h_1^*$ sufficiently close to the threshold, the origin is locally stable and two unstable nonzero branches exist at amplitudes $\pm\sqrt{-c_\eta (h^{(1)} - h_1^*)/c_3}(1 + o(1))$. These branches bound the local attracting interval in the reduced coordinate; their stable manifolds separate the local attracting and exiting regions.*

The general case of $d > 1$ is discussed in Appendix D.5. The proof proceeds via center manifold reduction. Let $\eta = h^{(1)} - h_1^*$ denote the signed distance from the critical value, and let $\theta \in \mathbb{R}$ be the amplitude of the token representation along the center direction. Reducing the full dL-dimensional system to

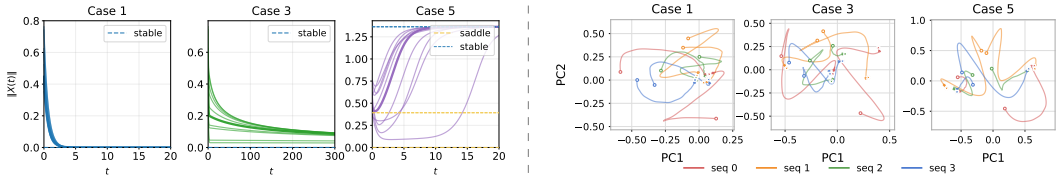


Figure 2: **SHC-1 Scenario 1** ($\lambda_{\max} < 0$). (Left) ODE token trajectories with $d = 1$, $L = 5$, $V = -1$, $W = 1$ and $\alpha_k = \beta_k = 1$. Cases 1, 3, and 5 illustrate configurations with a stable, critical, and unstable origin, respectively; their parameters are given in Appendix G. (Right) Shared-weight pretrained token trajectories projected by PCA. Circles and crosses denote initial and final layer positions. The pretrained trajectories provide a qualitative comparison.

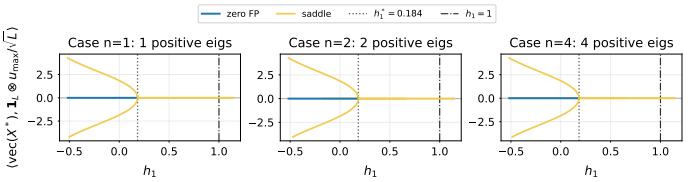


Figure 3: **ODE bifurcation diagram, SHC-1 Scenario 2** ($\lambda_{\max} > 0$)**.** Unstable branches (yellow) exist for $h^{(1)} < h_1^* = 0.184$, identical across $n = 1,2,4$ positive eigenvalues, confirming h_1^* depends only on $\lambda_{\max}$.

the center manifold yields the one-dimensional normal form $\dot{\theta} = c_\eta \eta\theta + c_3\theta^3 + O(|\theta|(\eta^2 + |\eta|\theta^2 + \theta^4))$, where c_η and c_3 are as defined in the theorem. Equation (30) is the standard normal form of the pitchfork bifurcation [46, Section 20.1e]. The full proof is in Appendix D.

Theorem 4.4 gives a complete local bifurcation picture for SHC-1: the explicit threshold $h^{(1)*}$ determines where stability is lost, while the sign of $c_\eta c_3$ decides whether stable nontrivial token states emerge or trajectories escape. This results in the following consequence.

Corollary 4.5 (Token fate near the threshold)**.** *Under Theorem 4.4, sufficiently close states collapse below h_1^*. Above it, states off the local stable manifold approach the two branches (supercritical) or leave a neighbourhood (subcritical).*

Corollary 4.5 provides an architectural interpretation of the carry-over parameter. The learnable carry-over parameter $h^{(\ell)}$ acts as an explicit stability knob with a closed-form critical value.

4.4 Numerical Validation

We study SHC-1 using ODE simulations and qualitative comparisons with pretrained language-model trajectories. For Scenario 1, we report three parameter configurations and a separate one-parameter bifurcation sweep (Section G). The trained networks have the limitations described in Appendix H.

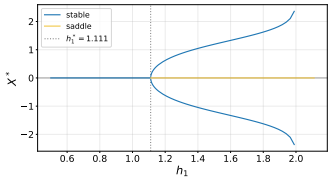


Figure 1: **ODE bifurcation diagram, SHC-1 Scenario 1** ($\lambda_{\max} < 0$)**.** Stable branches (blue) emerge at $h_1^* = 1.111$ with amplitude $\propto \sqrt{h^{(1)} - h_1^*}$. Dotted line: h_1^*.

Scenario 1 ($\lambda_{\max} < 0$**, supercritical).**The ODE norm plots (Figure 2) illustrate behavior across the configurations in Appendix G. The bifurcation diagram (Figure 1) illustrates the two sign-related equilibrium branches near h_1^*, consistent with Theorem 4.4. Pretrained trajectories provide a qualitative comparison. This is an empirical comparison, not a verification of Theorem 4.4: the trained model uses $V = -I$, whose leading eigenvalue is repeated d times, so the simple-mode hypothesis of the general-d pitchfork result does not hold, and the trained network differs from the ODE (Appendix H). Without weight sharing, this symmetry breaks and tokens diverge without settling (Figure 8 in Appendix I).

Scenario 2 ($\lambda_{\max} > 0$**, subcritical).** ODE trajectories (Figure 4) exhibit growth over the simulated interval, reaching norms of 10^{11}–10^{17}, while the bifurcation diagram (Figure 3) illustrates subcritical behaviour (Appendix D.5). The identical critical value $h_1^* = 0.184$ across different numbers of positive eigenvalues shows that h_1^* depends only on $\lambda_{\max}$, and pretrained trajectories (Figure 4) provide a qualitative comparison, subject to the implementation limitations in Appendix H.

5 Two-Stream Static Hyper-Connection Dynamics: Richer Bifurcations

We now study SHC-2, where a 2×2 coupling matrix H mixes two streams per token. Stability is characterized by trace and determinant conditions on the 2×2 spectral blocks, enabling new instabilities such as oscillatory transitions. As in SHC-1, we first locate the loss of exponential stability and then

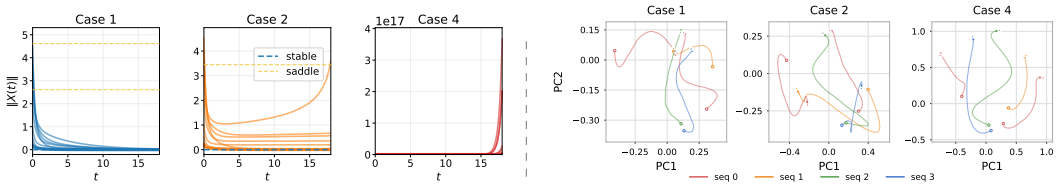


Figure 4: **SHC-1 Scenario 2** ($\lambda_{\max} > 0$). (Left) ODE trajectories. (Right) Pretrained model trajectories without weight sharing, projected onto the first two PCA components. Open and filled circles denote initial and final layer positions, respectively. Case 1, below the threshold, collapses; Case 2, at the threshold, diverges slowly; and Case 4, above the threshold, diverges more rapidly.

examine the nonlinear dynamics in the critical directions. We rewrite the 2-stream system in (4) as follows:

$$\dot{X}_1^{(\ell)} = (h_{11}-1)X_1^{(\ell)} + h_{12}X_2^{(\ell)} + \beta_1 V^\top \sum_k \mathrm{softmax}(\tilde{X}^{(\ell)\top} W \tilde{X}^{(k)})\tilde{X}^{(k)}, \tag{8}$$

$$\dot{X}_2^{(\ell)} = h_{21}X_1^{(\ell)} + (h_{22}-1)X_2^{(\ell)} + \beta_2 V^\top \sum_k \mathrm{softmax}(\tilde{X}^{(\ell)\top} W \tilde{X}^{(k)})\tilde{X}^{(k)}, \tag{9}$$

with $\tilde{X}^{(\ell)} = \alpha_1 X_1^{(\ell)} + \alpha_2 X_2^{(\ell)}$, $b = (\beta_1, \beta_2)^\top$, $\alpha = (\alpha_1, \alpha_2)^\top$, and $H = \left[\begin{smallmatrix} h_{11}-1 & h_{12} \\ h_{21} & h_{22}-1 \end{smallmatrix}\right]$. We assume H, α, β are shared across token positions (see Remark F.9).

5.1 Eigenstructure of the Jacobian

As in SHC-1, we linearize the dynamics at the collapsed state. The additional stream dimension combines the intra-stream coupling from H with the attention-mediated coupling from $V^\top$, yielding two spectral components: eigenvalues from H and quadratic roots coupled to the spectrum of $V^\top$.

Theorem 5.1 (Eigenvalues of J_{SHC2} for SHC-2)**.** *The $2dL$ eigenvalues of J_{SHC2} at $X = 0$ consist of:*

(a) eigenvalues of H (from token-contrast modes and uniform value-kernel modes, see Appendix E for further explanation of the terms), contributing $2dL - 2r$ eigenvalues in total; and

(b) for each nonzero eigenvalue λ_k of $V^\top$, two roots of

$$\mu^2 + B_k \mu + C_k = 0, \tag{10}$$

with

$$B_k = -\mathrm{tr}(H) - \lambda_k(\beta_1\alpha_1 + \beta_2\alpha_2), \qquad C_k = \det(H) + \lambda_k C_0(H),$$

where $C_0(H) = \beta_1\alpha_1(h_{22}-1) - \beta_2\alpha_1 h_{12} - \beta_1\alpha_2 h_{21} + \beta_2\alpha_2(h_{11}-1)$.

Theorem 5.1 shows that SHC-2 stability depends on $\mathrm{tr}(H)$, $\det(H)$, $C_0(H) = \alpha^\top \mathrm{adj}(H) b$ and the spectrum of $V^\top$. The high-multiplicity eigenvalues inherited from H allow stream-coupling instabilities to affect many token directions simultaneously.

5.2 Stability Region

We next characterize local stability of the collapsed state for SHC-2.

Theorem 5.2 (Stability of the origin for SHC-2)**.** *The origin is locally exponentially stable iff*

$$\mathrm{tr}(H) < 0, \qquad \det(H) > 0, \qquad B_k > 0, \qquad C_k > 0$$

for all k.

Corollary 5.3 (Binding constraints)**.** *Let $B := -\lambda_{\max}(\beta_1\alpha_1 + \beta_2\alpha_2)$, $\mathcal{D}_1 := \max(0, -\lambda_{\min} C_0(H))$, and $\mathcal{D}_2 := \max(0, -\lambda_{\max} C_0(H), -\lambda_{\min} C_0(H))$. Then:*

Scenario 1 *($\lambda_{\max} < 0$): exponentially stable iff $\mathrm{tr}(H) < 0$ and $\det(H) > \mathcal{D}_1$.*

Scenario 2 *($\lambda_{\max} > 0$): exponentially stable iff $\mathrm{tr}(H) < B$ and $\det(H) > \mathcal{D}_2$.*

Theorem 5.2 and Corollary 5.3 give two types of stability constraints: a trace upper bound and a determinant lower bound. When $\lambda_{\max} < 0$, these are $\mathrm{tr}(H) < 0$ and $\det(H) > \mathcal{D}_1$; when $\lambda_{\max} > 0$, they become $\mathrm{tr}(H) < B$ and $\det(H) > \mathcal{D}_2$. The determinant thresholds depend on H through $C_0(H) = \alpha^\top \mathrm{adj}(H) b$.

Table 1: Stability and bifurcation summary for SHC-2.

Scenario	Condition on H	Conclusion
$\lambda_{\max} < 0$	$\mathrm{tr}(H) < 0$ and $\det(H) > \mathcal{D}_1$	Origin **stable** (token collapse)
	$\det(H) = \mathcal{D}_1$, $\mathrm{tr}(H) < 0$	**Static boundary** (B1$^-$)
	$\mathrm{tr}(H) = 0$, $\det(H) > \mathcal{D}_1$	**Oscillatory boundary** (B2$^-$)
	$\mathrm{tr}(H) > 0$ or $\det(H) < \mathcal{D}_1$	Origin **unstable** (token escape)
$\lambda_{\max} > 0$	$\mathrm{tr}(H) < B$ and $\det(H) > \mathcal{D}_2$	Origin **stable** (token collapse)
	$\det(H) = \mathcal{D}_2$, $\mathrm{tr}(H) < B$	**Static boundary** (B1$^+$)
	$\mathrm{tr}(H) = B$, $\det(H) > \mathcal{D}_2$	**Oscillatory boundary** (B2$^+$)
	$\mathrm{tr}(H) > B$ or $\det(H) < \mathcal{D}_2$	Origin **unstable** (token escape)

5.3 Two Independent Bifurcation Mechanisms

Theorem 5.2 characterizes stability through $\mathrm{tr}(H)$, $\det(H)$, and B_k,C_k. When a stability constraint becomes tight, the boundary is *static* if a critical eigenvalue is zero (a zero-frequency linearized mode), or *oscillatory* if a nonzero imaginary conjugate pair is critical (oscillating linearized modes). These terms describe linear modes, not the emergence of equilibria or periodic solutions. An actual stability crossing requires eigenvalues to cross the imaginary axis; nonlinear conclusions are established separately in Theorem F.7.

Theorem 5.4 (Bifurcation types for SHC-2)**.** *On regular boundary portions of Corollary 5.3, transverse crossings require the active determinant/trace derivative to be nonzero. Two qualitatively distinct crossings arise:*

Scenario 1 *($\lambda_{\max} < 0$):*

(i) ***Static*** *(B1$^-$): $\det(H) = \mathcal{D}_1$ with $\mathrm{tr}(H) < 0$, a zero eigenvalue is carried by H if $C_0 < 0$, by $H + \lambda_{\min} b\alpha^\top$ if $C_0 > 0$, and by both if $C_0 = 0$;*

(ii) ***Oscillatory*** *(B2$^-$): $\mathrm{tr}(H) = 0$ with $\det(H) > \mathcal{D}_1$, a complex conjugate pair of eigenvalues of H reaches $\pm i\omega^*$, $\omega^* = \sqrt{\det(H)}$.*

Scenario 2 *($\lambda_{\max} > 0$):*

(i) ***Static*** *(B1$^+$): $\det(H) = \mathcal{D}_2$ with $\mathrm{tr}(H) < B$; active blocks satisfy $\det H = 0$ or $\det(H + \lambda_k b\alpha^\top) = 0$;*

(ii) ***Oscillatory*** *(B2$^+$): $\mathrm{tr}(H) = B$ with $\det(H) > \mathcal{D}_2$, giving $\omega^* = \sqrt{\det(H) + \lambda_{\max} C_0(H)}$.*

The nonlinear conclusions depend on the critical subspace. If it is entirely uniform, attention is exactly linear there, giving equilibria or constant-amplitude oscillations at criticality. Contrast modes can retain a cubic nonlinearity. Theorem F.7 classifies the critical subspaces, and Propositions F.5 and F.6 prove equilibrium and periodic branches in simple cases with $L = 2, d = 1$. The general multidimensional contrast dynamics are not classified here. Table 1 summarizes the linear boundaries.

5.4 Numerical Validation

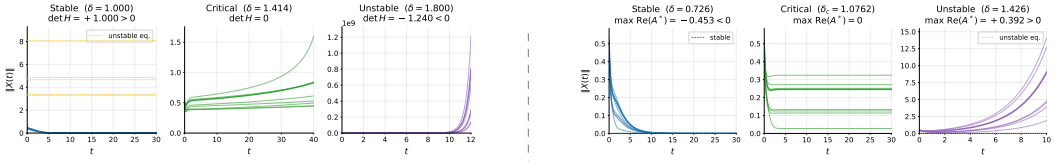
We validate the SHC-2 predictions with ODE simulations and pretrained 2-stream models, testing the four bifurcation regimes in Theorem 5.4; the setup follows Sections G and 6.

Scenario 1 ($\lambda_{\max} < 0$)**:** Static boundary (B1$^-$, Figure 5a): trajectories converge in the stable regime, exhibit IC-dependent critical slowing at $\det(H) = \mathcal{D}_1$, and grow beyond the threshold. Oscillatory crossing (B2$^-$, Figure 6a): trajectories show damped oscillations when stable, IC-dependent oscillations at $\mathrm{tr}(H) = 0$, and growing norms over the simulated interval after crossing; the finite-time plots do not establish periodic branches.

Scenario 2 ($\lambda_{\max} > 0$)**:** Static boundary (B1$^+$, Figure 5b): IC-stratified freezing at $\det(H) = \mathcal{D}_2$ is consistent with the conditional exact linearity in Lemma F.2. Oscillatory crossing (B2$^+$, Figure 6b): nearly flat oscillation envelopes at $\mathrm{tr}(H) = B$ reflect exact linearity of the uniform centre. Pretrained trajectories (Figures 7a, 7b) provide qualitative comparisons; their regime labels require the implementation check in Appendix H.

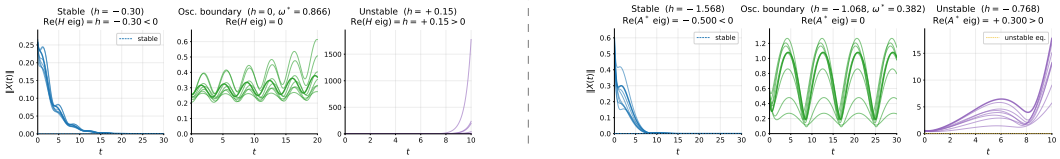
6 Additional Experimental Results

We further test the practical impact of the predicted regimes by training 1-stream and 2-stream GPT-2 [36] variants with different hyper-connection configurations. Tables 7 and 8 in Appendix H



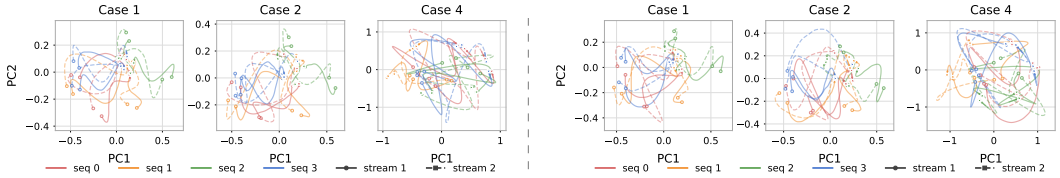
(a) Scenario 1 ($\lambda_{\max} < 0$). (b) Scenario 2 ($\lambda_{\max} > 0$).

Figure 5: **(a) SHC-2, S1: Static boundary** B1$^-$ (fix $\mathrm{tr}(H) < 0$). Left ($\det(H) > \mathcal{D}_1$): stable, convergence; dashed lines mark numerically found unstable equilibria (positive maximal real part; full saddle signatures not checked). Centre ($\det(H) = \mathcal{D}_1$): critical slowing. Right ($\det(H) < \mathcal{D}_1$): divergence. **(b) SHC-2, S2: Static boundary B1$^+$.** (fix $\mathrm{tr}(H) < B$). Left ($\det(H) > \mathcal{D}_2$): stable. Centre ($\det(H) = \mathcal{D}_2$, exact root δ_c): IC-stratified freezing suggests conditional uniform linearity; equilibria form a continuum there, so no isolated equilibria are marked. Right ($\det(H) < \mathcal{D}_2$): divergence.



(a) Scenario 1 ($\lambda_{\max} < 0$). (b) Scenario 2 ($\lambda_{\max} > 0$).

Figure 6: **(a) SHC-2, S1: oscillatory crossing B2$^-$** (fix $\det(H) > \mathcal{D}_1$). Left ($\mathrm{tr}(H) < 0$): damped oscillations. Centre ($\mathrm{tr}(H) = 0$): oscillations (contrast critical modes). Right ($\mathrm{tr}(H) > 0$): divergence. **(b) SHC-2, S2: oscillatory crossing B2$^+$** (fix $\det(H) > \mathcal{D}_2$). Left ($\mathrm{tr}(H) < B$): damped oscillations. Centre ($\mathrm{tr}(H) = B$): flat oscillation envelopes, consistent with exact uniform linearity. Right ($\mathrm{tr}(H) > B$): divergence, slower than Scenario 1.



(a) Scenario 1 ($\lambda_{\max} < 0$), PCA. (b) Scenario 2 ($\lambda_{\max} > 0$), PCA.

Figure 7: **Pretrained SHC-2 token trajectories.** Open/filled circles denote initial/final layers; solid/dashed curves denote streams 1/2. Stable cases converge, critical cases show threshold behavior, and unstable cases diverge with stream-specific patterns.

report the measured perplexities. Several configurations historically labelled unstable outperform the corresponding stable-labelled configurations, and the best reported Scenario 2 runs improve on their baselines. This is an empirical association, not a consequence of the local ODE theorems. In particular, the historical 2-stream Case 2 label and implementation conventions require the consistency checks described in Appendix H. The original measured values and historical labels are retained pending run-level revalidation; they should not be read as a certified stability classification of the trained model.

7 Concluding Remarks

We analyzed self-attention token dynamics with static hyper-connections through bifurcation theory. Hyper-connection parameters make the collapsed fixed point's hyperbolicity tunable, turning representation collapse into a tractable stability problem. For SHC-1, a closed-form threshold h_1^* gives a pitchfork under the stated conditions, separating local collapse from nontrivial equilibria or escape; for SHC-2, the coupling matrix H creates two stability boundaries and static/ oscillatory instabilities. Experiments explore related trajectory patterns and report perplexity improvements for some selected configurations. These observations do not prove a causal performance benefit from ODE instability; the training implementation and regime labels require the checks in Appendix H. Our analysis is limited to static hyper-connections and attention-only continuous-depth dynamics with shared weights, omitting dynamic hyper-connections, layer normalization, feed-forward networks, and positional encodings. In addition, the shared-parameter SHC-2 setting induces an exactly linear uniform center manifold when all critical modes are uniform, leaving general contrast-mode classification open for future work.

Acknowledgments and Disclosure of Funding

This paper is dedicated to Professor Richard G. Baraniuk on the occasion of his 60th birthday.

This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). This research is also supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023) (A-8002040-00-00, A-8002039-00-00), the NUS Presidential Young Professorship Award (A-0009807-01-00), the NUS Artificial Intelligence Institute–Seed Funding (A-8003062-00-00), and the Cross Faculty Grant 2025, CFG25-012 (A-8004460-00-00). T. T. Nguyen acknowledges support from the NSFC Excellent Young Scientists Fund (Overseas) and the Research Start-up Fund of Shanghai Jiao Tong University.

References

[1] Laziz Abdullaev and Tan Nguyen. Transformer meets twicing: Harnessing unattended residual information. In *International Conference on Learning Representations*, volume 2025, pages 57583–57606, 2025.

[2] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1285. URL https://doi.org/10.18653/v1/p19-1285.

[4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071. PMLR, 2024.

[5] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[6] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

[7] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

[8] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[10] Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

[11] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2018.

[12] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41, 2022.

[15] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024.

[16] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

[17] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *Advances in neural information processing systems*, 37:32653–32677, 2025.

[18] Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

[19] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[20] Fusheng Liu and Qianxiao Li. From generalization analysis to optimization designs for state space models, 2024. URL https://openreview.net/forum?id=EGjvMcKrrl.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2017.

[22] 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. *arXiv preprint arXiv:1906.02762*, 2019.

[23] Chao Ma and Ziyang Wang. Semi-mamba-unet: Pixel-level contrastive and cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300:112203, 2024.

[24] Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. *Advances in Neural Information Processing Systems*, 37:127226–127272, 2025.

[25] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

[26] Ho Huu Nghia Nguyen, Tan Nguyen, Huyen Vo, Stanley Osher, and Thieu Vo. Improving neural ordinary differential equations with nesterov's accelerated gradient method. *Advances in Neural Information Processing Systems*, 35:7712–7726, 2022.

[27] Tam Nguyen, Tan Nguyen, and Richard Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36: 80233–80256, 2023.

[28] Tam Minh Nguyen, César A Uribe, Tan Minh Nguyen, and Richard Baraniuk. Pidformer: Transformer meets control theory. In *Forty-first International Conference on Machine Learning*, 2024.

[29] Thai Khanh Nguyen, Vo Ngoc Bich Uyen, Thieu Vo, Tan Minh Nguyen, and Cuong Pham. MuonSSM: Orthogonalizing state space models for sequence modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GmP3VcfHi0.

[30] Stefan K Nielsen, Laziz U Abdullaev, Rachel S Teo, and Tan M Nguyen. Elliptical attention. *Advances in Neural Information Processing Systems*, 37:109748–109789, 2024.

[31] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[32] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112*, 2024.

[33] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

[34] Duy-Tung Pham, An Nguyen The, Viet-Hoang Tran, Nhan-Phu Chung, Xin T. Tong, Tan Minh Nguyen, and Thieu Vo. Dynamical properties of tokens in self-attention and effects of positional encoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=i9Vn8vV99g.

[35] Maciej Pióro, Kamil Ciebiera, Krystian Król, Jan Ludziejewski, and Sebastian Jaszczur. Moe-mamba: Efficient selective state space models with mixture of experts. *arXiv preprint arXiv:2401.04081*, 2024.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[37] Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. Mambalrp: Explaining selective state space sequence models. *Advances in Neural Information Processing Systems*, 37:118540–118570, 2025.

[38] George Saon, Ankit Gupta, and Xiaodong Cui. Diagonal state space augmented transformers for speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[39] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[40] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[43] 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. URL https://openreview.net/forum?id=qtTIP5Gjc5.

[44] Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660*, 2024.

[45] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.

[46] S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Texts in Applied Mathematics. Springer New York, 2003. ISBN 9780387001777. URL https://books.google.com.vn/books?id=RSI4RGdwnU4C.

[47] Hedi Xia, Vai Suliafu, Hangjie Ji, Tan Nguyen, Andrea Bertozzi, Stanley Osher, and Bao Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

[48] Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, et al. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

[49] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pages 10524–10533. PMLR, 2020.

[50] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[51] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, 34:30008–30022, 2021.

[52] Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024.

[53] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[54] 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. URL https://openreview.net/forum?id=9FqARW7dwB.

[55] Itamar Zimerman, Moran Baruch, Nir Drucker, Gilad Ezov, Omri Soceanu, and Lior Wolf. Converting transformers to polynomial form for secure inference over homomorphic encryption. *arXiv preprint arXiv:2311.08610*, 2023.

Supplement to "Understanding the Effects of Hyper-Connections on Self-Attention Dynamics: A Bifurcation Analysis"

Table of Contents

A Related Work **16**

B Setting and notation **16**

B.1 Hyper-connected self-attention and its continuous-depth limits 16

B.2 Standing assumptions 18

C Common fixed point and the residual-connection obstruction **18**

D SHC-1: one-stream static hyper-connections **19**

D.1 Rescaling and Jacobian matrix 19

D.2 Spectrum of Jacobian J_{SHC1} 20

D.2.1 Setup and Notation 20

D.2.2 Reduction and the block spectrum 20

D.2.3 Spectrum of J_{SHC1} 22

D.3 Stability region for fixed point 0 of SHC-1 22

D.4 Threshold and type of the crossing 25

D.5 SHC-1 in general hidden dimension d 30

E SHC-2: Spectrum and exponential stability **33**

E.1 Setting and notation 33

E.2 Jacobian at the collapsed state 33

E.3 Stability region 34

F SHC-2: nonlinear structure and crossings **36**

F.1 The nonlinearity vanishes when attention-input contrast vanishes 36

F.2 Exactly linear dynamics on uniform critical spaces 36

F.3 Contrast dynamics and simple bifurcations 37

F.4 Meaning of static and oscillatory stability boundaries 39

F.5 Which modes become critical at each boundary 39

F.6 Parameter sharing and its limits 40

G ODE Simulation Setup **41**

G.1 Hyper-Connection ODE System 41

G.2 Stability and Fixed Points 41

G.3 Bifurcation Along h_1 42

G.4 Simulation Regimes 42

G.5 Numerical Outputs 42

G.6 Scenario-2 SHC-1 simulation 43

G.7 SHC-2 ODE System 43

H Training Details **44**

H.1 1-stream Hyper-connection Models 44

H.2 2-stream Hyper-connection Models 45

I Additional Experiments **47**

J Hardware Settings **48**

K Broader Societal Impacts **49**

A Related Work

Token dynamics in deep sequence models. Recent works have studied deep sequence models by interpreting layer depth as a time variable and analyzing the induced dynamics of token representations [34, 43, 8]. For Transformers [42, 6, 19, 28, 51, 36, 11, 53, 39], this viewpoint models self-attention as an interacting particle system and has been used to characterize clustering, representation collapse, metastable behavior, and divergence of token embeddings. These analyses show that the qualitative regime of token trajectories can often be determined by model parameters, such as the attention and value matrices, and that convergent or collapsed regimes may negatively affect performance. Related dynamical analyses such as [43] have also been developed for selective state-space models such as Mamba [9, 45, 32, 44, 15, 10], where continuous-depth limits reveal convergence, divergence, token-order effects, and clustering phenomena. Together, these works suggest that token dynamics provide a useful theoretical lens for understanding representation quality in deep sequence models.

Residual connections and hyper-connections. Residual connections [13, 18, 39] are central to the trainability of deep architectures [35, 23, 25, 15, 52, 14, 40], but different residual designs can induce different trade-offs between optimization stability and representation collapse. In Transformers, Pre-Norm [49] variants help alleviate gradient instability but may lead to highly similar hidden states across layers, while Post-Norm variants can reduce collapse at the cost of more difficult optimization [54]. Hyper-connections [54, 48, 5] were proposed as a learnable alternative to standard residual connections, introducing trainable depth-wise and width-wise couplings that mix hidden states across layers and streams. Empirically, hyper-connected models reduce adjacent-layer similarity and improve pretraining performance, indicating that cross-layer coupling plays an important role in preserving representation diversity.

B Setting and notation

B.1 Hyper-connected self-attention and its continuous-depth limits

Throughout, $d \geq 1$ and $L \geq 2$. Tokens are indexed by $\ell = 1,\ldots,L$ and layers by k. The classical residual Transformer update is

$$x^{(\ell)}(k+1) = x^{(\ell)}(k) + \mathrm{Layer}\big(x^{(\ell)}(k), \{x^{(i)}(k)\}_{i=1}^{L}\big). \tag{11}$$

Hyper-connections (HC) expand this to N parallel streams per token with learnable couplings:

$$x_i^{(\ell)}(k+1) = \sum_{j=1}^{N} h_{ij}^{(\ell)} x_j^{(\ell)}(k) + \beta_i^{(\ell)} \mathrm{Layer}\Big(\sum_{j=1}^{N} \alpha_j^{(\ell)} x_j^{(\ell)}(k), \Big\{\sum_{j=1}^{N} \alpha_j^{(l)} x_j^{(l)}(k)\Big\}_{l=1}^{L}\Big), \qquad i = 1,\ldots,N, \tag{12}$$

with intra-token connection matrix $H^{(\ell)} = (h_{ij}^{(\ell)}) \in \mathbb{R}^{N\times N}$, input-scaling weights $\alpha_j^{(\ell)}$ and output-scaling weights $\beta_i^{(\ell)}$. The paper studies the *static* variant (SHC), for which the parameters are input-independent, in two cases:

- **SHC-1** ($N = 1$, token-dependent parameters $h^{(\ell)}, \alpha^{(\ell)}, \beta^{(\ell)}$):

$$x^{(\ell)}(k+1) = h^{(\ell)} x^{(\ell)}(k) + \beta^{(\ell)} \mathrm{Layer}\big(\alpha^{(\ell)} x^{(\ell)}(k), \{\alpha^{(i)} x^{(i)}(k)\}_{i=1}^{L}\big); \tag{13}$$

- **SHC-2** ($N=2$, parameters $H\in\mathbb{R}^{2\times 2}$, $\alpha,\beta\in\mathbb{R}^2$ shared across token positions).

$$x_i^{(\ell)}(k+1)=\sum_{j=1}^{2}h_{ij}x_j^{(\ell)}(k)+\beta_i\mathrm{Layer}\Big(\sum_{j=1}^{2}\alpha_j x_j^{(\ell)}(k),\Big\{\sum_{j=1}^{2}\alpha_j x_j^{(l)}(k)\Big\}_{l=1}^{L}\Big),\qquad i=1,2. \tag{14}$$

Token-dependence in SHC-1 and parameter sharing in SHC-2 are complementary tractability choices; see Remark F.9 (Section F.6) for the discussion of what happens under the opposite choices.

Following Pham et al. [34], the layer computation is self-attention only, with layer-independent attention matrix $W\in\mathbb{R}^{d\times d}$ and value matrix $V\in\mathbb{R}^{d\times d}$:

$$\mathrm{Attn}(u;\{v^{(i)}\}):=V^\top\sum_{i=1}^{L}\mathrm{softmax}\big(u^\top Wv^{(i)}\big)v^{(i)},\qquad \mathrm{softmax}\big(u^\top Wv^{(i)}\big)=\frac{e^{u^\top Wv^{(i)}}}{\sum_j e^{u^\top Wv^{(j)}}}. \tag{15}$$

Writing $Z^{(\ell)}(t)\in\mathbb{R}^d$ (resp. $Z_i^{(\ell)}(t)$) for the representation of token ℓ at continuous depth t, for each token $\ell=1,\dots,L$ the three models are

$$\dot{Z}^{(\ell)}=V^\top\sum_{i=1}^{L}\mathrm{softmax}\big({Z^{(\ell)}}^\top WZ^{(i)}\big)Z^{(i)}, \tag{ODE-1}$$

$$\dot{Z}^{(\ell)}=(h^{(\ell)}-1)Z^{(\ell)}+\beta^{(\ell)}V^\top\sum_{i=1}^{L}\mathrm{softmax}\big((\alpha^{(\ell)}Z^{(\ell)})^\top W(\alpha^{(i)}Z^{(i)})\big)\alpha^{(i)}Z^{(i)}, \tag{ODE-2}$$

$$\dot{Z}_i^{(\ell)}=\sum_{j=1}^{2}(h_{ij}-\delta_{ij})Z_j^{(\ell)}+\beta_iV^\top\sum_{h=1}^{L}\mathrm{softmax}\Big(\tilde{Z}^{(\ell)\top}W\tilde{Z}^{(h)}\Big)\tilde{Z}^{(h)},\qquad i=1,2, \tag{ODE-3}$$

where $\tilde{Z}^{(\ell)}:=\sum_{j=1}^{2}\alpha_jZ_j^{(\ell)}$ is the weighted-sum attention input for token ℓ.

Here matrix indices use the output-row/input-column convention: h_{ij} maps stream j to stream i. δ_{ij} is the Kronecker delta, equal to 1 if $i=j$ and 0 otherwise.

From discrete hyper-connections to continuous depth We start from the HC update (12), with $\mathrm{Layer}=\mathrm{Attn}$ as defined in (15), and additionally assume that all parameters are shared across layers. For brevity, write

$$\tilde{x}^{(\ell)}(k):=\sum_{j=1}^{N}\alpha_j^{(\ell)}x_j^{(\ell)}(k).$$

For K layers over a fixed depth interval $[0,T]$, set $\Delta t=T/K$. A finite continuous-depth limit requires the layer increment to be of order Δt; this does not generally hold for the unscaled update above. Following the step-size construction for residual networks in Haber and Ruthotto [11, Eq. (2.1)], we introduce a family of HC updates whose effective discrete coefficients are

$$\begin{aligned}h_{ij,\Delta t}^{(\ell)}&:=\delta_{ij}+\Delta t\big(h_{ij}^{(\ell)}-\delta_{ij}\big),\\ \beta_{i,\Delta t}^{(\ell)}&:=\Delta t\beta_i^{(\ell)},\qquad \alpha_{j,\Delta t}^{(\ell)}:=\alpha_j^{(\ell)}.\end{aligned} \tag{16}$$

The unsubscripted parameters are fixed as Δt varies, whereas the subscripted coefficients are those used in each discrete update. At $\Delta t=1$, they coincide with the original coefficients. This construction preserves the HC form, but the step-size scaling is an additional modelling choice rather than a requirement of the original architecture.

Writing $x_{i,\Delta t}^{(\ell)}(k)$ for the resulting discrete states and $\tilde{x}_{\Delta t}^{(\ell)}(k)=\sum_j\alpha_j^{(\ell)}x_{j,\Delta t}^{(\ell)}(k)$, the update gives the exact difference quotient

$$\begin{aligned}&\frac{x_{i,\Delta t}^{(\ell)}(k+1)-x_{i,\Delta t}^{(\ell)}(k)}{\Delta t}\\ &\quad=\sum_{j=1}^{N}\big(h_{ij}^{(\ell)}-\delta_{ij}\big)x_{j,\Delta t}^{(\ell)}(k)+\beta_i^{(\ell)}\mathrm{Attn}\Big(\tilde{x}_{\Delta t}^{(\ell)}(k);\{\tilde{x}_{\Delta t}^{(q)}(k)\}_{q=1}^{L}\Big).\end{aligned} \tag{17}$$

To make the passage to continuous depth explicit, set $t_k=k\Delta t$ and introduce smooth interpolating curves $Z_{i,\Delta t}^{(\ell)}:[0,T]\to\mathbb{R}^d$ satisfying

$$Z_{i,\Delta t}^{(\ell)}(t_k)=x_{i,\Delta t}^{(\ell)}(k),\qquad Z_{i,\Delta t}^{(\ell)}(t_k+\Delta t)=x_{i,\Delta t}^{(\ell)}(k+1).$$

The subscript Δt indicates that the interpolating curve depends on the discrete step size. We reserve $Z_i^{(\ell)}(t)$ for its continuous-depth limit as $\Delta t \to 0$.

To identify the limiting equation, suppose these interpolants converge in C^1 to smooth curves $Z_i^{(\ell)}$ as $\Delta t \to 0$. First-order expansion then gives, uniformly over the grid points,

$$Z_{i,\Delta t}^{(\ell)}(t_k+\Delta t) = Z_{i,\Delta t}^{(\ell)}(t_k) + \Delta t \dot{Z}_{i,\Delta t}^{(\ell)}(t_k) + o(\Delta t).$$

Substituting this expression into (17) yields

$$\dot{Z}_{i,\Delta t}^{(\ell)}(t_k) + o(1) = \sum_{j=1}^{N} \big(h_{ij}^{(\ell)} - \delta_{ij}\big) Z_{j,\Delta t}^{(\ell)}(t_k) + \beta_i^{(\ell)} \mathrm{Attn}\Big(\tilde{Z}_{\Delta t}^{(\ell)}(t_k); \{\tilde{Z}_{\Delta t}^{(q)}(t_k)\}_{q=1}^{L}\Big),$$

where $\tilde{Z}_{\Delta t}^{(\ell)} := \sum_j \alpha_j^{(\ell)} Z_{j,\Delta t}^{(\ell)}$. Letting $\Delta t \to 0$, we obtain

$$\dot{Z}_i^{(\ell)}(t) = \sum_{j=1}^{N} \big(h_{ij}^{(\ell)} - \delta_{ij}\big) Z_j^{(\ell)}(t) + \beta_i^{(\ell)} \mathrm{Attn}\Big(\tilde{Z}^{(\ell)}(t); \{\tilde{Z}^{(q)}(t)\}_{q=1}^{L}\Big), \qquad \tilde{Z}^{(\ell)}(t) := \sum_{j=1}^{N} \alpha_j^{(\ell)} Z_j^{(\ell)}(t). \tag{18}$$

For fixed parameters, the vector field F is smooth and satisfies $\|F(X)\| \le C\|X\|$, since attention weights are nonnegative and sum to one. Gronwall's inequality bounds exact and Euler trajectories on each fixed finite interval, and standard Euler convergence then applies. This also excludes finite-time blow-up, but does not establish uniform-in-time approximation or stability of the unscaled discrete network. The three models follow as special cases: (ODE-1) takes $N=1$ with $h_{11}^{(\ell)} = \alpha_1^{(\ell)} = \beta_1^{(\ell)} = 1$ for every token; (ODE-2) takes $N=1$ with token-dependent parameters $h_{11}^{(\ell)} = h^{(\ell)}$, $\alpha_1^{(\ell)} = \alpha^{(\ell)}$, and $\beta_1^{(\ell)} = \beta^{(\ell)}$; and (ODE-3) takes $N=2$ with parameters shared across token positions. We then analyse the long-time dynamics of the resulting continuous-depth model. Its parameter conditions guide the experimental configurations, allowing us to test whether the predicted qualitative regimes are also observed in finite-depth pretrained models. Agreement provides empirical support for the relevance of the ODE analysis beyond its idealised setting; an exact correspondence of stability regimes additionally requires checking the discrete update and the model assumptions.

B.2 Standing assumptions

Assumption B.1. *Consider the self-attention dynamics with hyper-connections introduced in Section 2. We assume the following:*

(A) The value matrix V is a real symmetric matrix, with nonzero eigenvalues listed with multiplicity as $\lambda_1, \ldots, \lambda_r$ and corresponding eigenvectors $v_1, \ldots, v_r \in \mathbb{R}^d$. For $r > 0$, we denote $\lambda_{\max} := \max_k \lambda_k$ and $\lambda_{\min} := \min_k \lambda_k$.

(B) Input- and output- scaling weights are positive: $\alpha^{(\ell)}, \beta^{(\ell)} > 0$ for SHC-1, and $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ for SHC-2. For pitchfork sign claims only, assume $W = W^\top \succ 0$. Throughout, $L \ge 2$.

C Common fixed point and the residual-connection obstruction

Lemma C.1 (Common fixed point)**.** *The origin $Z = 0$ is a fixed point of each of* (ODE-1)–(ODE-3).

Proof. At $Z = 0$ all attention logits vanish, so $\mathrm{softmax}(0) = \frac{1}{L}\mathbf{1}_L$. Hence each right-hand side evaluates to 0, confirming that $Z = 0$ is a fixed point. □

Definition C.2 (Zero-state collapse and local stability)**.** A trajectory collapses to zero if every token and stream representation tends to zero as $t \to \infty$. We say that the origin is locally asymptotically stable when it is Lyapunov stable and attracts all initial conditions in some neighbourhood. Local exponential stability additionally requires a uniform exponential decay estimate. Convergence of one trajectory is not equivalent to stability of the equilibrium. Instability implies that some arbitrarily small perturbations leave a neighbourhood; it does not imply that every trajectory escapes, that norms diverge, or that performance improves.

Lemma C.3 (Stable $V^\top$)**.** *Let $W, V \in \mathbb{R}^{n\times n}$ be symmetric with $W \succ 0$. If V is invertible and $W(V^\top)^{-1} \prec 0$, then all eigenvalues of $V^\top$ are strictly negative.*

Proof. Put $A = W(V^\top)^{-1}$. For any eigenpair (λ, x) of V we have $x^\top Ax = \frac{1}{\lambda} x^\top Wx$. Since $A \prec 0$ implies $x^\top Ax < 0$ and $W \succ 0$ implies $x^\top Wx > 0$, we get $1/\lambda < 0$, hence $\lambda < 0$. □

Proposition C.4 (Non-hyperbolicity of residual self-attention)**.** *For the residual self-attention dynamics* (ODE-1)*, the Jacobian at the origin is*

$$J_{\mathrm{RC}}(0) = \frac{1}{L} \mathbf{1}_L \mathbf{1}_L^\top \otimes V^\top \in \mathbb{R}^{dL \times dL}.$$

In particular $\operatorname{rank}(J_{\mathrm{RC}}) = \operatorname{rank}(V) \le d < dL$*, and therefore the origin is non-hyperbolic whenever* $L > 1$.

Proof. Write $s_{\ell k} := Z^{(\ell)\top} W Z^{(k)}$ and $p_k^{(\ell)} := \mathrm{softmax}_k(s_{\ell \cdot}) = e^{s_{\ell k}} / \sum_{h=1}^{L} e^{s_{\ell h}}$, so that

$$F^{(\ell)}(Z) = V^\top \sum_{k=1}^{L} p_k^{(\ell)} Z^{(k)}, \qquad F_a^{(\ell)} = \sum_{c=1}^{d} V_{ca} \sum_{k=1}^{L} p_k^{(\ell)} Z_c^{(k)}, \qquad a = 1, \dots, d.$$

By the product rule, for $j = 1, \dots, L$ and $b = 1, \dots, d$,

$$\frac{\partial F_a^{(\ell)}}{\partial Z_b^{(j)}} = \sum_{c=1}^{d} V_{ca} \Big(\sum_{k=1}^{L} Z_c^{(k)} \frac{\partial p_k^{(\ell)}}{\partial Z_b^{(j)}} + p_j^{(\ell)} \delta_{cb} \Big).$$

At $Z = 0$: every $Z_c^{(k)} = 0$ while $\partial p_k^{(\ell)} / \partial Z_b^{(j)}$ is finite (the softmax is smooth), so the first sum vanishes; and $s_{\ell k} = 0$ gives $p_j^{(\ell)}(0) = \frac{1}{L}$. Hence

$$\frac{\partial F_a^{(\ell)}}{\partial Z_b^{(j)}}\Big|_{Z=0} = \frac{1}{L} V_{ba} = \frac{1}{L} \big(V^\top\big)_{ab} \Longrightarrow \frac{\partial F^{(\ell)}}{\partial Z^{(j)}}\Big|_{Z=0} = \frac{1}{L} V^\top \quad \forall \ell, j,$$

$$J_{\mathrm{RC}}(0) = \Big(\frac{\partial F^{(\ell)}}{\partial Z^{(j)}}\Big|_0 \Big)_{\ell,j=1}^{L} = \frac{1}{L} \begin{pmatrix} V^\top & \cdots & V^\top \\ \vdots & & \vdots \\ V^\top & \cdots & V^\top \end{pmatrix} = \frac{1}{L} \mathbf{1}_L \mathbf{1}_L^\top \otimes V^\top.$$

Finally,

$$\operatorname{rank} J_{\mathrm{RC}}(0) = \operatorname{rank}\big(\mathbf{1}_L \mathbf{1}_L^\top\big) \cdot \operatorname{rank} V^\top = \operatorname{rank} V \le d < dL \quad (L > 1),$$

so $\dim(\ker J_{\mathrm{RC}})(0) \ge d(L-1) \ge 1$: $J_{\mathrm{RC}}(0)$ has the eigenvalue 0, and the origin is non-hyperbolic. □

D SHC-1: one-stream static hyper-connections

D.1 Rescaling and Jacobian matrix

Lemma D.1 (Rescaled coordinates)**.** *Define the rescaled variables*

$$X^{(\ell)}(t) := \alpha^{(\ell)} Z^{(\ell)}(t), \qquad \gamma^{(\ell)} := \alpha^{(\ell)} \beta^{(\ell)}, \qquad \ell = 1, \dots, L.$$

Under this substitution the vector field (ODE-2) *transforms to*

$$\begin{aligned} \dot{X}^{(\ell)} &= (h^{(\ell)} - 1) X^{(\ell)} + \gamma^{(\ell)} V^\top \sum_{i=1}^{L} \mathrm{softmax}\big(X^{(\ell)\top} W X^{(i)}\big) X^{(i)} \\ &=: F^{(\ell)}(X), \qquad \ell = 1, \dots, L. \end{aligned} \tag{19}$$

We henceforth work in the X-coordinates.

Proof. Multiplying both sides of (ODE-2) for token ℓ by $\alpha^{(\ell)}$ and substituting $X^{(\ell)} = \alpha^{(\ell)} Z^{(\ell)}$, $X^{(i)} = \alpha^{(i)} Z^{(i)}$:

$$\frac{dX^{(\ell)}}{dt} = \alpha^{(\ell)} \frac{dZ^{(\ell)}}{dt} = \alpha^{(\ell)} (h^{(\ell)} - 1) Z^{(\ell)} + \alpha^{(\ell)} \beta^{(\ell)} V^\top \sum_{i=1}^{L} \mathrm{softmax}\big(\alpha^{(\ell)} Z^{(\ell)\top} W \alpha^{(i)} Z^{(i)}\big) \alpha^{(i)} Z^{(i)}.$$

Using $\alpha^{(\ell)} Z^{(\ell)} = X^{(\ell)}$, $\alpha^{(i)} Z^{(i)} = X^{(i)}$ and $\alpha^{(\ell)} \beta^{(\ell)} = \gamma^{(\ell)}$ yields (19). □

Lemma D.2 (Jacobian structure of SHC-1)**.** *The Jacobian of system* (19) *at $X = 0$ is*

$$J_{\mathrm{SHC1}} = H_{\mathrm{diag}} \otimes I_d + M \otimes V^\top \in \mathbb{R}^{dL \times dL}, \qquad H_{\mathrm{diag}} = \mathrm{diag}\big(h^{(1)} - 1, \dots, h^{(L)} - 1\big) \in \mathbb{R}^{L \times L}, \tag{20}$$

where $M = \frac{1}{L} \gamma \mathbf{1}_L^\top \in \mathbb{R}^{L \times L}$ with $\gamma = (\gamma^{(1)}, \dots, \gamma^{(L)})^\top$.

Proof. Write $p_k^{(\ell)} := \mathrm{softmax}_k\big(X^{(\ell)\top} W X^{(\cdot)}\big)$, so that

$$F^{(\ell)}(X) = (h^{(\ell)}-1)X^{(\ell)} + \gamma^{(\ell)} V^\top \sum_{k=1}^{L} p_k^{(\ell)} X^{(k)} \in \mathbb{R}^d,$$

of which each component

$$F_a^{(\ell)} = (h^{(\ell)}-1)X_a^{(\ell)} + \gamma^{(\ell)} \sum_{c=1}^{d} V_{ca} \sum_{k=1}^{L} p_k^{(\ell)} X_c^{(k)} \in \mathbb{R}.$$

For $j = 1,\dots,L$ and $a,b = 1,\dots,d$,

$$\frac{\partial F_a^{(\ell)}}{\partial X_b^{(j)}} = (h^{(\ell)}-1)\delta_{\ell j}\delta_{ab} + \gamma^{(\ell)} \sum_{c=1}^{d} V_{ca} \Big(\sum_{k=1}^{L} X_c^{(k)} \frac{\partial p_k^{(\ell)}}{\partial X_b^{(j)}} + p_j^{(\ell)} \delta_{cb} \Big).$$

At $X = 0$ we have $X_c^{(k)} = 0$ and $p_j^{(\ell)}(0) = \frac{1}{L}$, hence

$$\frac{\partial F_a^{(\ell)}}{\partial X_b^{(j)}}\bigg|_{X=0} = (h^{(\ell)}-1)\delta_{\ell j}\delta_{ab} + \frac{\gamma^{(\ell)}}{L}\big(V^\top\big)_{ab}$$

$$\Longrightarrow (J_{\mathrm{SHC1}})_{\ell j} := \frac{\partial F^{(\ell)}}{\partial X^{(j)}}\bigg|_{X=0} = (h^{(\ell)}-1)\delta_{\ell j} I_d + \frac{\gamma^{(\ell)}}{L} V^\top.$$

Since the (ℓ,j) block of $A \otimes B$ is $A_{\ell j} B$, and $(H_{\mathrm{diag}})_{\ell j} = (h^{(\ell)}-1)\delta_{\ell j}$, $M_{\ell j} = \frac{1}{L}\gamma^{(\ell)}$,

$$J_{\mathrm{SHC1}} = \big((J_{\mathrm{SHC1}})_{\ell j}\big)_{\ell,j=1}^{L} = \big((H_{\mathrm{diag}})_{\ell j} I_d + M_{\ell j} V^\top\big)_{\ell,j=1}^{L} = H_{\mathrm{diag}} \otimes I_d + M \otimes V^\top. \qquad \square$$

D.2 Spectrum of Jacobian J_{SHC1}

D.2.1 Setup and Notation

Let $P_1,\dots,P_s$ be the *distinct* values taken by $h^{(1)},\dots,h^{(L)}$, and set

$$\mathcal{G}_g := \{i : h^{(i)} = P_g\}, \qquad q_g := |\mathcal{G}_g|,$$

so $\sum_g q_g = L$, and $s = L$ exactly when all $h^{(i)}$ are distinct. Define the secular function

$$f(\mu) := \sum_{g=1}^{s} \frac{\sum_{i\in\mathcal{G}_g} \gamma^{(i)}}{\mu - (P_g - 1)}. \tag{21}$$

By Assumption B.1(A), V is real symmetric; fix an orthonormal eigenbasis of $\mathbb{R}^d$. Let Λ be the set of distinct nonzero eigenvalues of $V^\top$. For each $\lambda \in \Lambda$, define its eigenspace $E_\lambda := \ker(V^\top - \lambda I)$ of dimension n_λ, so $\sum_{\lambda\in\Lambda} n_\lambda = r = \mathrm{rank}(V)$ and $\dim(\ker V^\top) = d - r$. Write $A_\lambda := H_{\mathrm{diag}} + \frac{\lambda}{L}\gamma \mathbf{1}_L^\top$.

D.2.2 Reduction and the block spectrum

Lemma D.3 (Invariant splitting)**.** *The decomposition*

$$\mathbb{R}^{dL} = \Big(\mathbb{R}^L \otimes \ker V^\top\Big) \oplus \bigoplus_{\lambda\in\Lambda} \Big(\mathbb{R}^L \otimes E_\lambda\Big)$$

is J_{SHC1}-invariant, and

$$J_{\mathrm{SHC1}}\big|_{\mathbb{R}^L\otimes\ker V^\top} = H_{\mathrm{diag}} \otimes I_{d-r}, \qquad J_{\mathrm{SHC1}}\big|_{\mathbb{R}^L\otimes E_\lambda} = A_\lambda \otimes I_{n_\lambda}.$$

Proof. The splitting. Since V is symmetric, the spectral theorem gives the orthogonal decomposition $\mathbb{R}^d = \ker V^\top \oplus \bigoplus_{\lambda\in\Lambda} E_\lambda$. Choosing a basis of each summand and taking the union produces a basis of $\mathbb{R}^d$; the corresponding vectors $e_i \otimes v_a$ then form a basis of $\mathbb{R}^{dL}$, partitioned according to the summand containing v_a. This is the displayed decomposition, and the dimensions add: $L(d-r) + \sum_\lambda L n_\lambda = L(d-r) + Lr = dL$.

Invariance and the restrictions. By the mixed-product rule, for $u \in \mathbb{R}^L$, $v \in \mathbb{R}^d$,

$$J_{\mathrm{SHC1}}(u \otimes v) = (H_{\mathrm{diag}} u) \otimes v + (Mu) \otimes (V^\top v). \tag{$*$}$$

Let $v_1,\dots,v_{d-r}$ be a basis of $\ker V^\top$. Every $X\in\mathbb{R}^L\otimes\ker V^\top$ has a unique expansion $X=\sum_a u_a\otimes v_a$ with $u_a\in\mathbb{R}^L$, and $V^\top v_a=0$ gives

$$J_{\mathrm{SHC1}}X=\sum_{a=1}^{d-r}(H_{\mathrm{diag}}u_a)\otimes v_a\in\mathbb{R}^L\otimes\ker V^\top.$$

Similarly, let $v_1,\dots,v_{n_\lambda}$ be a basis of E_λ. For $X=\sum_a u_a\otimes v_a\in\mathbb{R}^L\otimes E_\lambda$, $V^\top v_a=\lambda v_a$ gives

$$J_{\mathrm{SHC1}}X=\sum_{a=1}^{n_\lambda}\big[(H_{\mathrm{diag}}+\lambda M)u_a\big]\otimes v_a=\sum_{a=1}^{n_\lambda}(A_\lambda u_a)\otimes v_a\in\mathbb{R}^L\otimes E_\lambda.$$

Hence every summand is J_{SHC1}-invariant. In the bases $\{e_i\otimes v_a\}$, i.e. identifying $\sum_a u_a\otimes v_a$ with $\sum_a u_a\otimes e_a$, the two displays read $J_{\mathrm{SHC1}}|_{\mathbb{R}^L\otimes\ker V^\top}=H_{\mathrm{diag}}\otimes I_{d-r}$ and $J_{\mathrm{SHC1}}|_{\mathbb{R}^L\otimes E_\lambda}=A_\lambda\otimes I_{n_\lambda}$. □

Lemma D.4 (Spectrum of A_λ)**.** *Fix $\lambda\neq0$. The eigenvalues of $A_\lambda\in\mathbb{R}^{L\times L}$ are:*

(A) P_g-1 for each g with $q_g\geq2$, with eigenspace

$$\mathcal{C}_g:=\Big\{u\in\mathbb{R}^L:\mathrm{supp}(u)\subseteq\mathcal{G}_g,\ \textstyle\sum_{i\in\mathcal{G}_g}u_i=0\Big\},\qquad \dim\mathcal{C}_g=q_g-1;$$

For a vector $u=(u_1,\dots,u_L)^\top\in\mathbb{R}^L$, its support is $\mathrm{supp}(u):=\{i\in\{1,\dots,L\}:u_i\neq0\}$, the set of indices of its nonzero entries.

(B) the s roots $\mu_1^\lambda<\cdots<\mu_s^\lambda$ of $f(\mu)=L/\lambda$, which are real, simple, strictly interlaced with the poles, and distinct from every P_g-1; the eigenvector for μ_t^λ is

$$u_j^{(t,\lambda)}=\frac{\lambda}{L}\cdot\frac{\gamma^{(j)}}{\mu_t^\lambda-\big(h^{(j)}-1\big)},\qquad j=1,\dots,L,\tag{22}$$

which is already normalised by $\mathbf{1}_L^\top u^{(t,\lambda)}=\frac{\lambda}{L}f(\mu_t^\lambda)=1$.

In particular A_λ is diagonalizable, with $(L-s)+s=L$ eigenvalues counted with multiplicity, and $P_g-1\in\mathrm{spec}(A_\lambda)$ if and only if $q_g\geq2$.

Proof. Let $A_\lambda u=\mu u$. Explicitly,

$$(H_{\mathrm{diag}}-\mu I_L)u=-\frac{\lambda}{L}\gamma(\mathbf{1}_L^\top u)\tag{23}$$

Componentwise,

$$\big(h^{(j)}-1-\mu\big)u_j=-\frac{\lambda\gamma^{(j)}}{L}\mathbf{1}_L^\top u,\qquad j=1,\dots,L.\tag{24}$$

Case $\mu=P_g-1$ for some g. Taking $j\in\mathcal{G}_g$ in (24) makes the left-hand side vanish, so $0=-\frac{\lambda\gamma^{(j)}}{L}\mathbf{1}_L^\top u$; since $\lambda\neq0$ and $\gamma^{(j)}>0$, this forces $\mathbf{1}_L^\top u=0$. With this identity, (24) reads $(h^{(j)}-1-\mu)u_j=0$ for all j, so $u_j=0$ whenever $h^{(j)}\neq P_g$, i.e. $\mathrm{supp}(u)\subseteq\mathcal{G}_g$; together with $\mathbf{1}_L^\top u=0$ this says exactly $u\in\mathcal{C}_g$. Conversely every $u\in\mathcal{C}_g$ satisfies $A_\lambda u=H_{\mathrm{diag}}u=(P_g-1)u$, the rank-one term being annihilated by $\mathbf{1}_L^\top u=0$. The space $\mathcal{C}_g$ is the kernel of the nonzero linear functional $u\mapsto\sum_{i\in\mathcal{G}_g}u_i$ on the q_g-dimensional coordinate subspace $\mathrm{span}\{e_i:i\in\mathcal{G}_g\}$, so $\dim\mathcal{C}_g=q_g-1$ by rank-nullity; it is nonzero iff $q_g\geq2$. (In particular, a singleton group contributes no pole eigenvector; if all groups are singletons, there are no pole eigenvalues.)

Case $\mu\notin\{P_1-1,\dots,P_s-1\}$. Then $h^{(j)}-1-\mu\neq0$ for all j. If $\mathbf{1}_L^\top u=0$ then (24) gives $(h^{(j)}-1-\mu)u_j=0$ for all j, hence $u_j=0$ for all j, contradicting $u\neq0$. Thus $\mathbf{1}_L^\top u\neq0$, and replacing u by $\frac{u}{\mathbf{1}_L^\top u}$ we may normalise $\mathbf{1}_L^\top u=1$. Solving (24) gives (22), and applying $\mathbf{1}_L^\top$ yields

$$1=\frac{\lambda}{L}\sum_{j=1}^{L}\frac{\gamma^{(j)}}{\mu-(h^{(j)}-1)}=\frac{\lambda}{L}\sum_{g=1}^{s}\frac{\sum_{j\in\mathcal{G}_g}\gamma^{(j)}}{\mu-(P_g-1)},\tag{25}$$

equivalently,

$$f(\mu)=\frac{L}{\lambda}.\tag{26}$$

Since $\sum_{j\in\mathcal{G}_g}\gamma^{(j)}>0$, $f'(\mu)=-\sum_{g=1}^{s}\frac{\sum_{j\in\mathcal{G}_g}\gamma^{(j)}}{(\mu-P_g+1)^2}<0$: f is strictly decreasing on each of the $s+1$ intervals determined by the poles, sweeping $0^-\to-\infty$ on $(-\infty,P_{\min}-1)$, $+\infty\to-\infty$ on each of the $s-1$ interior intervals, and $+\infty\to 0^+$ on $(P_{\max}-1,\infty)$. Hence the horizontal line L/λ is met exactly s times, transversally, with the roots interlacing the poles. Each such μ has a one-dimensional eigenspace, since (22) determines u from $\mathbf{1}_L^\top u$.

Counting, $\sum_g(q_g-1)+s=(L-s)+s=L$ independent eigenvectors, so A_λ is diagonalizable. □

D.2.3 Spectrum of J_{SHC1}

Theorem D.5 (Spectrum of the SHC-1 Jacobian; general form)**.** *Under Assumption B.1, with the notation above, the dL eigenvalues of J_{SHC1} are exactly the following two families, which are pairwise distinct.*

Family A (pole modes). *For each $g=1,\dots,s$ the value P_g-1, with geometric multiplicity*

$$\mathrm{mult}(p_g)=q_g d-r$$

and eigenspace

$$\Big(\mathrm{span}\{e_i : i\in\mathcal{G}_g\}\otimes\ker V^\top\Big)\oplus\Big(\mathcal{C}_g\otimes\mathrm{range}V\Big),$$

of dimensions $q_g(d-r)$ and $(q_g-1)r$ respectively. The value P_g-1 is an eigenvalue of J_{SHC1} if and only if $q_g d>r$.

Family B (secular modes). *For each $\lambda\in\Lambda$ and each of the s roots μ_t^λ of $f(\mu)=L/\lambda$, the value μ_t^λ, with geometric multiplicity*

$$\mathrm{mult}(\mu_t^\lambda)=n_\lambda$$

and eigenspace $u^{(t,\lambda)}\otimes E_\lambda$, where $u^{(t,\lambda)}$ is given by (22). There are rs such eigenvalues counted with multiplicity.

The two families together account for $\sum_g(q_g d-r)+rs=dL$ eigenvalues, all real; J_{SHC1} is diagonalizable.

Proof. By Lemma D.3 it suffices to collect the spectra of the restrictions. On $\mathbb{R}^L\otimes\ker V^\top$ the operator is $H_{\mathrm{diag}}\otimes I_{d-r}$, whose eigenvalue P_g-1 has eigenspace $\mathrm{span}\{e_i : i\in\mathcal{G}_g\}\otimes\ker V^\top$ of dimension $q_g(d-r)$. On $\mathbb{R}^L\otimes E_\lambda$ the operator is $A_\lambda\otimes I_{n_\lambda}$, so by Lemma D.4 its spectrum is P_g-1 with multiplicity $(q_g-1)n_\lambda$ and each μ_t^λ with multiplicity n_λ. Summing the pole contributions over λ gives $(q_g-1)r$, and $q_g(d-r)+(q_g-1)r=q_g d-r$.

The families are disjoint: no μ_t^λ equals a pole (Lemma D.4), and $\mu_t^\lambda=\mu_{t'}^{\lambda'}$ would give $L/\lambda=f(\mu)=L/\lambda'$, hence $\lambda=\lambda'$ by injectivity of $\lambda\mapsto L/\lambda$, and then $t=t'$ by simplicity. Distinct p_g are distinct by construction. Finally $\sum_g(q_g d-r)+\sum_\lambda s\,n_\lambda=dL-sr+sr=dL$, so the listed geometric multiplicities exhaust the dimension and J_{SHC1} is diagonalizable. □

Corollary D.6 (Shared parameters)**.** *If $h^{(\ell)}\equiv h$ and $\gamma^{(\ell)}\equiv\gamma$ then $s=1$, $q_1=L$, $\Gamma_1=L\gamma$, and $f(\mu)=L\gamma/(\mu-(h-1))$. Family A is the single value $h-1$ with multiplicity $Ld-r$; Family B is $\mu^\lambda=(h-1)+\gamma\lambda$, one per distinct λ, with multiplicity n_λ. This is the shared-parameter spectrum of Remark F.9.*

Proof. For shared coefficients, $H_{\mathrm{diag}}=(h-1)I_L$ and $M=(\gamma/L)\mathbf{1}_L\mathbf{1}_L^\top$. On $\mathbf{1}_L^\perp$, $M=0$; on $\mathrm{span}\{\mathbf{1}_L\}$ it acts as multiplication by γ. Combining this with the value eigenspaces proves the formula and multiplicities. □

D.3 Stability region for fixed point 0 of SHC-1

Theorem D.5 describes the whole spectrum of the Jacobian J_{SHC1} at $X=0$. In this section we use it to decide when the collapsed state is stable. Since J_{SHC1} has real spectrum, its largest eigenvalue is negative, zero, or positive. The origin is locally exponentially stable in the first case and unstable in the third. In the second case linearization alone is inconclusive, and the nonlinear analysis of Section D.4 takes over. Our task is therefore to locate the eigenvalues of J_{SHC1} relative to the origin.

Notation. We keep the notation of Section D.2.1: $P_1,\dots,P_s$ are the distinct values among $h^{(1)},\dots,h^{(L)}$, $\mathcal{G}_g=\{i:h^{(i)}=P_g\}$ and $q_g=|\mathcal{G}_g|$. We assume $V\neq 0$, so that $\lambda_{\max}$ is defined, and we set

$$\Gamma_g:=\sum_{j\in\mathcal{G}_g}\gamma^{(j)}>0, \qquad \text{so that} \qquad f(\mu)=\sum_{g=1}^{s}\frac{\Gamma_g}{\mu-(P_g-1)}.$$

The poles of f are the numbers P_g-1. The origin $\mu=0$ is a pole exactly when $P_g=1$ for some g; otherwise

$$f(0)=\sum_{g=1}^{s}\frac{\Gamma_g}{1-P_g}=\sum_{i=1}^{L}\frac{\gamma^{(i)}}{1-h^{(i)}}.$$

Finally, let

$$n_+:=\#\{g:P_g>1\}$$

denote the number of positive poles of f.

The argument is graphical. By Theorem D.5, the secular eigenvalues attached to $\lambda\in\Lambda$ are the abscissae at which the graph of f meets the horizontal line $y=L/\lambda$. Recall from the proof of Lemma D.4 that f is strictly decreasing on each of the $s+1$ intervals cut out by its poles. It decreases from 0^- to $-\infty$ on the leftmost interval, from $+\infty$ to $-\infty$ on every interior interval, and from $+\infty$ to 0^+ on the rightmost one. The following lemma turns this picture into a count.

Lemma D.7 (Nonnegative secular roots)**.** *Assume that $P_g\neq 1$ for every g, and let $c\in\mathbb{R}\setminus\{0\}$. Then:*

(i) *$\mu=0$ solves $f(\mu)=c$ if and only if $f(0)=c$;*

(ii) *the number of solutions of $f(\mu)=c$ in $(0,\infty)$ is*

$$N_+(c)=n_+-1+\mathbb{I}\{f(0)>c\}+\mathbb{I}\{c>0\};$$

where $\mathbb{I}\{\cdot\}$ equals 1 if the condition holds and 0 otherwise;

(iii) *consequently, $f(\mu)=c$ has no solution in $[0,\infty)$ if and only if*

- *for $c<0$: either $n_+=0$, or $n_+=1$ and $f(0)<c$;*
- *for $c>0$: $n_+=0$ and $f(0)<c$.*

Proof. Part (i) is immediate, since f is defined and continuous at 0.

For (ii), suppose first that $n_+\geq 1$ and let $0<\pi_1<\cdots<\pi_{n_+}$ be the positive poles. They split $(0,\infty)$ into the open intervals

$$I_0=(0,\pi_1), \qquad I_j=(\pi_j,\pi_{j+1})\ (1\leq j\leq n_+-1), \qquad I_{n_+}=(\pi_{n_+},\infty).$$

Because f is strictly decreasing on each of them, the equation $f(\mu)=c$ has at most one solution in each, and it has one exactly when c lies strictly between the two end values of f. On I_0 the function decreases from $f(0)$ to $-\infty$, so there is a solution if and only if $f(0)>c$. On each of the n_+-1 intervals I_j it decreases from $+\infty$ to $-\infty$, so there is always exactly one solution. On I_{n_+} it decreases from $+\infty$ to 0^+, so there is a solution if and only if $c>0$. Adding up gives the formula.

If $n_+=0$, all poles are negative, hence every term $\Gamma_g/(1-P_g)$ is positive and $f(0)>0$. On $(0,\infty)$ the function decreases from $f(0)$ to 0^+, so $f(\mu)=c$ has a solution there if and only if $0<c<f(0)$. For $c<0$ this never happens, while $\mathbb{I}\{f(0)>c\}=1$ and $\mathbb{I}\{c>0\}=0$; for $c>0$ it happens if and only if $f(0)>c$. In both cases the count agrees with the formula for $n_+=0$.

Part (iii) follows by requiring $N_+(c)=0$ and $f(0)\neq c$. For $c<0$ this reads $n_++\mathbb{I}\{f(0)>c\}=1$; when $n_+=0$ it holds automatically because $f(0)>0>c$, and when $n_+=1$ it holds if and only if $f(0)<c$. For $c>0$ it reads $n_++\mathbb{I}\{f(0)>c\}=0$, that is, $n_+=0$ and $f(0)<c$. □

We can now state the stability region. Recall that Scenario 1 refers to $\lambda_{\max}<0$ and Scenario 2 to $\lambda_{\max}>0$.

Theorem D.8 (Stability region of the origin)**.** *Let Assumption B.1 hold, with the notation above.*

(i) Stability. *All eigenvalues of J_{SHC1} are negative if and only if the following two conditions hold.*

(S1) (pole modes) $P_g<1$ for every g with $q_g d>r$.

(S2) (secular modes)

- *Scenario 1: either (a) $P_g \le 1$ for all g; or (b) there is exactly one index g_0 with $P_{g_0} > 1$, all other groups satisfy $P_g < 1$, and $f(0) < L/\lambda_{\max}$.*
- *Scenario 2: $P_g < 1$ for all g, and $f(0) < L/\lambda_{\max}$.*

In this case the origin is locally exponentially stable. Moreover, in case (b) of Scenario 1 the exceptional group must satisfy $q_{g_0} d = r$; that is, $\mathcal{G}_{g_0}$ is a single token and $V^\top$ is invertible.

(ii) Zero eigenvalues (not necessarily a stability boundary). *J_{SHC1} has the eigenvalue 0 if and only if either $P_g = 1$ for some g with $q_g d > r$, or $P_g \neq 1$ for all g and $f(0) = L/\lambda$ for some $\lambda \in \Lambda$.*

(iii) Positive eigenvalues and the actual boundary. *Put $D_\gamma = \mathrm{diag}(\gamma^{(1)},\dots,\gamma^{(L)})$, $g = D_\gamma^{1/2}\mathbf{1}_L$, and $\widehat{\lambda} = \mathrm{maxspec}(V)$ (including zero if present). Then*

$$a(J_{\mathrm{SHC1}}) := \mathrm{maxspec} J_{\mathrm{SHC1}} = \lambda_{\mathrm{largest}}\left(H_{\mathrm{diag}} + \frac{\widehat{\lambda}}{L} g g^\top\right).$$

If $a(J_{\mathrm{SHC1}}) > 0$, the origin is unstable even if J_{SHC1} also has zero eigenvalues. The boundary of the Hurwitz region is $a(J_{\mathrm{SHC1}}) = 0$; this condition, rather than singularity alone, identifies the points where nonlinear analysis is needed. In particular, if neither (i) nor (ii) applies, the origin is unstable.

Proof. By Theorem D.5, the spectrum of J_{SHC1} is the union of the pole family (Family A) and the secular family (Family B), and it is real. We treat the two families separately and write $c_\lambda := L/\lambda$ for $\lambda \in \Lambda$.

Pole modes. The number $P_g - 1$ is an eigenvalue of J_{SHC1} exactly when $q_g d > r$. Hence Family A lies in $(-\infty, 0)$ if and only if (S1) holds.

Secular modes in Scenario 1. Here every $\lambda \in \Lambda$ is negative, so every line $y = c_\lambda$ lies below the axis. Since $\lambda \mapsto L/\lambda$ is decreasing on $(-\infty, 0)$, the lowest of these lines is $c_{\lambda_{\max}} = L/\lambda_{\max}$, i.e. $c_{\lambda_{\max}} \le c_\lambda$ for all $\lambda \in \Lambda$.

Sufficiency of (a). If $P_g \le 1$ for all g, every pole is nonpositive. Fix $\lambda \in \Lambda$. The solution on the leftmost interval lies below the smallest pole, and the solution on each interior interval lies strictly below the right endpoint of that interval, which is at most the largest pole and hence at most 0. On the rightmost interval $f > 0 > c_\lambda$, so there is no solution there. All s secular roots are therefore negative.

Sufficiency of (b). Now 0 is not a pole and $n_+ = 1$. For every $\lambda \in \Lambda$ we have $f(0) < c_{\lambda_{\max}} \le c_\lambda < 0$, so Lemma D.7(iii) rules out solutions in $[0, \infty)$.

Necessity. Assume all secular roots are negative. If two poles were positive, f would take every real value on the interval between them, and the line $y = c_{\lambda_{\max}}$ would meet it at a positive abscissa. Hence $n_+ \le 1$. If $n_+ = 0$ we are in case (a). If $n_+ = 1$, let $\pi > 0$ be the positive pole. Were some other P_g equal to 1, the origin would be a pole and f would sweep $(-\infty, +\infty)$ on $(0, \pi)$, again producing a positive root; so all other groups satisfy $P_g < 1$. Then 0 is not a pole, and Lemma D.7(iii) with $c = c_{\lambda_{\max}}$ gives $f(0) < L/\lambda_{\max}$, which is case (b).

Secular modes in Scenario 2. Now $c_{\lambda_{\max}} > 0$, and since $\lambda \mapsto L/\lambda$ is decreasing on $(0, \infty)$, $c_{\lambda_{\max}} \le c_\lambda$ for every positive $\lambda \in \Lambda$.

Sufficiency. If $P_g < 1$ for all g, then $n_+ = 0$ and 0 is not a pole. For positive λ we have $f(0) < c_{\lambda_{\max}} \le c_\lambda$, and for negative λ we have $c_\lambda < 0$; in both cases Lemma D.7(iii) excludes roots in $[0, \infty)$.

Necessity. If $P_g \ge 1$ for some g, the largest pole is nonnegative. On the rightmost interval f decreases from $+\infty$ to 0^+, so it meets the positive line $y = c_{\lambda_{\max}}$ to the right of that pole, at a positive abscissa. Hence $P_g < 1$ for all g, and Lemma D.7(iii) with $c = c_{\lambda_{\max}}$ yields $f(0) < L/\lambda_{\max}$.

The exceptional group. In case (b) of Scenario 1 we have $P_{g_0} > 1$, so (S1) forces $q_{g_0} d \le r$. As $q_{g_0} \ge 1$ and $r \le d$, this is possible only if $q_{g_0} = 1$ and $r = d$.

Parts (ii) and (iii). Family A contains 0 exactly when $P_g = 1$ for some g with $q_g d > r$. A secular root is never a pole (Lemma D.4), so Family B can contain 0 only if 0 is not a pole, i.e. $P_g \neq 1$ for all g; in that

case, by Lemma D.7(i), 0 is a secular root for λ if and only if $f(0) = L/\lambda$. This proves (ii). For (iii), $D_\gamma^{-1/2} A_\lambda D_\gamma^{1/2} = H_{\text{diag}} + (\lambda/L) g g^\top$, a real symmetric matrix. Its largest eigenvalue is nondecreasing in λ by the Rayleigh quotient, and the zero-value block is H_{diag}. Thus the largest full eigenvalue is the displayed one. Adding the same ε to all $h^{(i)}$ adds εI to J_{SHC1}, so every point with $a(J_{\text{SHC1}}) = 0$ is approached from both sides of the Hurwitz region. Points with $a(J_{\text{SHC1}}) > 0$ are unstable by the indirect method.

The dynamical conclusions follow from Lyapunov's indirect method. □

In the setting when all $h^{(i)}$ are distinct and $V^\top$ is invertible, every group is a single token. (S1) is then empty, and Theorem D.8 takes the following explicit form in each scenario.

Remark D.9 (Interpretation)**.** Stability is governed by three ingredients: the carry-over coefficients $h^{(i)}$, the attention strengths $\gamma^{(i)}$, and the spectrum of $V^\top$. Only the largest eigenvalue $\lambda_{\max}$ enters the secular condition, since the line $y = L/\lambda_{\max}$ is always the binding one. This is why the critical value in Figure 3 is the same for one, two and four positive eigenvalues. When $\lambda_{\max} < 0$, the value matrix is contractive and the model can afford one token with $h^{(i)} > 1$, provided the weighted sum $f(0)$ stays below $L/\lambda_{\max}$. Raising that single coefficient is therefore a way out of collapse. When $\lambda_{\max} > 0$ there is an amplifying direction, and stability requires every $h^{(i)} < 1$ together with $f(0) < L/\lambda_{\max}$. Here lowering the carry-over coefficients is what prevents divergence. On the one-token parameter paths admitting a secular first crossing, the boundary is the hypersurface $f(0) = L/\lambda_{\max}$, which Section D.4 crosses along a one-parameter family.

D.4 Threshold and type of the crossing

Setting. Throughout this subsection $h^{(1)}$ is the bifurcation parameter. All other parameters are fixed, and the remaining carry-over coefficients satisfy $h^{(i)} < 1$ for $i \geq 2$; ties among them are allowed. We write

$$S := \sum_{i=2}^{L} \frac{\gamma^{(i)}}{1 - h^{(i)}} > 0, \qquad m := \frac{L}{\lambda_{\max}} - S, \qquad F(h^{(1)}) := \frac{\gamma^{(1)}}{1 - h^{(1)}} + S \quad (h^{(1)} \neq 1),$$

so that $F(h^{(1)})$ is the value $f(0)$ of Section D.3 along the parameter path. We denote by $n_{\max}$ the multiplicity of $\lambda_{\max}$ as an eigenvalue of $V^\top$.

Lemma D.10 (Critical value)**.** *The function F is strictly increasing on $(-\infty, 1)$, with range (S, ∞), and on $(1, \infty)$, with range $(-\infty, S)$. The equation $F(h^{(1)}) = L/\lambda_{\max}$ has a solution if and only if $m \neq 0$; the solution is then unique and equals*

$$h_1^* = 1 - \frac{\gamma^{(1)}}{m}. \tag{27}$$

In particular $h_1^* < 1$ if $m > 0$ and $h_1^* > 1$ if $m < 0$. In Scenario 1 one always has $m < 0$, hence $h_1^* > 1$; in Scenario 2 both signs of m are possible.

Proof. The monotonicity and the ranges follow from $\gamma^{(1)} > 0$. The equation $F(h^{(1)}) = L/\lambda_{\max}$ is equivalent to $\gamma^{(1)}/(1 - h^{(1)}) = m$, which has no solution for $m = 0$ and the unique solution (27) otherwise. In Scenario 1, m is the sum of the negative number $L/\lambda_{\max}$ and the negative number $-S$. □

Recall from Theorem D.5 that the eigenvalues of J_{SHC1} are of two kinds: the values $P_g - 1$ (Family A) and the roots of the secular equations $f(\mu) = L/\lambda$, $\lambda \in \Lambda$ (Family B). As $h^{(1)}$ increases from a value at which the origin is exponentially stable, the Hurwitz region ends when some eigenvalue reaches 0. The next proposition shows that this eigenvalue is either a secular root or the Family A eigenvalue $h^{(1)} - 1$, and says which one.

Proposition D.11 (First loss of exponential stability)**.** *Let Assumption B.1 hold, and let $h^{(1)}$ increase from a value at which the origin is exponentially stable. The first loss of exponential stability happens in one of two ways.*

- *(i)* A secular root crosses 0 first. *This happens if either $r = d$ in Scenario 1, or Scenario 2 with $m > 0$ and any r. Before the crossing $h^{(1)} < h_1^*$, and the crossing happens at $h^{(1)} = h_1^*$. There, $\mu = 0$ is a root of $f(\mu) = L/\lambda_{\max}$ and of no other secular equation, so it is an eigenvalue of J_{SHC1} of multiplicity $n_{\max}$.*

(ii) The eigenvalue $h^{(1)}-1$ crosses 0 first. *This happens if $r<d$ in Scenario 1. Before the crossing $h^{(1)}<1$, and the crossing happens at $h^{(1)}=1$, with eigenspace $E_c:=\mathrm{span}\{e_1\}\otimes\ker V^\top$ of dimension $d-r$.*

In both cases all other eigenvalues of J_{SHC1} are negative at the crossing. In Scenario 2 with $m\le 0$ the origin is unstable for every $h^{(1)}$, so no loss-of-stability crossing occurs (other eigenvalues may still cross zero).

Proof. Along the path only the group of token 1 changes, and $f(0)=F(h^{(1)})$.

Where the origin is stable. We apply Theorem D.8(i). Consider first (S1). Groups not containing token 1 have value <1, and the group of token 1 has value $h^{(1)}$. If this group contains another token, then $h^{(1)}<1$; otherwise it is a singleton, and $qd>r$ holds exactly when $r<d$. Hence (S1) holds if and only if $r=d$ or $h^{(1)}<1$.

Consider next (S2). In Scenario 1, case (a) means $h^{(1)}\le 1$, and case (b) means $h^{(1)}>1$ and $F(h^{(1)})<L/\lambda_{\max}$, that is, $1<h^{(1)}<h_1^*$ by Lemma D.10; hence (S2) holds if and only if $h^{(1)}<h_1^*$. In Scenario 2, (S2) requires $h^{(1)}<1$ and $F(h^{(1)})<L/\lambda_{\max}$. Since $F(h^{(1)})>S$ for $h^{(1)}<1$, this never holds when $m\le 0$, that is, when $L/\lambda_{\max}\le S$. The origin is then unstable: for $h^{(1)}\ge 1$ by the Scenario 2 necessity argument in Theorem D.8, and for $h^{(1)}<1$ by Lemma D.7(ii), which gives $N_+(L/\lambda_{\max})=1$. When $m>0$, (S2) holds if and only if $h^{(1)}<h_1^*$, and (S1) then holds automatically because $h_1^*<1$.

Combining the two conditions, the origin is exponentially stable exactly for $h^{(1)}<h_1^*$ in the cases of (i), and exactly for $h^{(1)}<1$ in case (ii), where $h_1^*>1$.

Which eigenvalue crosses. At the right endpoint J_{SHC1} has the eigenvalue 0, and Theorem D.8(ii) tells us where it comes from. In case (i) we have $h_1^*\neq 1$, so no group has value 1, while $f(0)=L/\lambda_{\max}$. Thus 0 is a root of $f(\mu)=L/\lambda_{\max}$, and of no other secular equation, since $f(0)\neq L/\lambda$ for $\lambda\neq\lambda_{\max}$; by Theorem D.5 it is an eigenvalue of multiplicity $n_{\max}$. In case (ii), token 1 forms a singleton group with value 1 and $d>r$, so 0 is the Family A eigenvalue $h^{(1)}-1$, with eigenspace E_c of dimension $d-r$ (Theorem D.5); it is not a secular root, because $\mu=0$ is then a pole of f.

Finally, the eigenvalues of J_{SHC1} depend continuously on $h^{(1)}$ and are negative before the crossing. So at the crossing they are all ≤ 0, and all except the zero eigenvalues just identified are negative. □

Proposition D.11 identifies where the origin first loses exponential stability and which directions become critical. To determine the local dynamics at this threshold, we must examine the full system along those directions: do nearby nonzero equilibria appear, and what is their stability? We treat the two crossings separately. We call case (i) a *secular crossing* and case (ii) a *null-space crossing*, since in case (ii) the critical eigenvectors are $e_1\otimes v$ with v in the null space of $V^\top$. The two crossings have very different characters.

For the null-space crossing, the critical subspace is exactly invariant, and the attention contribution vanishes there. This allows us to determine the restricted dynamics directly. For the secular crossing, we use a centre-manifold reduction to account for the nonlinear terms.

Lemma D.12 (Exact linearity on the kernel subspace)**.** *The subspace*

$$\mathcal{K}:=\mathbb{R}^L\otimes\ker V^\top=\{X: X^{(\ell)}\in\ker V^\top \text{ for all } \ell\}$$

is invariant under the full system (19). On $\mathcal{K}$ the dynamics are linear and decoupled,

$$\dot{X}^{(\ell)}=(h^{(\ell)}-1)X^{(\ell)},\qquad \ell=1,\dots,L,$$

at every amplitude.

Proof. For $X\in\mathcal{K}$ the attention term at token ℓ is

$$\gamma^{(\ell)}\sum_{i=1}^{L}\mathrm{softmax}\big(X^{(\ell)\top}WX^{(i)}\big)V^\top X^{(i)}=0,$$

because the softmax weights are scalars and $V^\top X^{(i)}=0$. The remaining local term $(h^{(\ell)}-1)X^{(\ell)}$ preserves $\ker V^\top$. □

Proposition D.13 (Exact linear dynamics at the null-space crossing)**.** *Assume Scenario 1 with $r<d$, and let*

$$E_c := \{X : X^{(1)} \in \ker V^\top, X^{(\ell)} = 0 \text{ for } \ell = 2,\dots,L\}$$

be the critical subspace identified in Proposition D.11(ii). Then the following statements hold.

(i) For every $h^{(1)}$, E_c is invariant under (19), and the flow on it is $\dot{\xi} = (h^{(1)} - 1)\xi$ with $\xi := X^{(1)} \in \ker V^\top$. At $h^{(1)} = 1$, E_c is a centre manifold of the origin, and the reduced equation has vanishing coefficients at every nonlinear order.

(ii) Near $(X, h^{(1)}) = (0, 1)$ the equilibria are exactly $X = 0$ for $h^{(1)} \neq 1$. At $h^{(1)} = 1$, the equilibria in this neighbourhood are precisely the points of E_c. Thus there is a continuum of equilibria at the critical parameter, but no nearby nonzero equilibrium branch extending to either side of the crossing.

(iii) At $h^{(1)} = 1$ the origin is Lyapunov stable but not asymptotically stable. For $h^{(1)} > 1$, every solution starting in $E_c \setminus \{0\}$ satisfies $X^{(1)}(t) = e^{(h^{(1)}-1)t} X^{(1)}(0)$ and its norm tends to infinity as $t \to \infty$.

Thus, the origin loses stability through exactly linear dynamics on the critical invariant subspace, with no nonzero equilibrium branch extending to either side of the threshold.

Proof. We first identify the dynamics on the critical subspace, then characterize the nearby equilibria and their stability.

(i) Since $E_c \subset \mathcal{K}$, Lemma D.12 applies. The decoupled flow keeps $X^{(\ell)} = 0$ for $\ell \geq 2$ and acts on $X^{(1)}$ by $\dot{\xi} = (h^{(1)} - 1)\xi$. At $h^{(1)} = 1$, E_c is an invariant manifold equal to the centre eigenspace (Proposition D.11(ii)), hence a centre manifold.

(ii) Regard $h^{(1)}$ as an additional state variable satisfying $\dot{h}^{(1)} = 0$. Near $(0, 1)$, the set $E_c \times \mathbb{R}$ is an invariant manifold tangent to the centre subspace of this augmented system, and hence is a local centre manifold. Every sufficiently nearby equilibrium lies on this manifold. Its equilibrium equation is exactly

$$(h^{(1)} - 1)\xi = 0.$$

If $h^{(1)} \neq 1$, this equation forces $\xi = 0$, so $X = 0$. If $h^{(1)} = 1$, every $\xi \in \ker V^\top$ satisfies it. This gives the asserted local equilibrium set and rules out a nearby nonzero equilibrium branch for parameters on either side of the crossing.

(iii) At $h^{(1)} = 1$ every solution of the reduced equation is constant, so the origin is Lyapunov stable within the centre manifold. All transverse eigenvalues have negative real part; the centre-manifold reduction principle therefore gives Lyapunov stability for the full system. However, $r < d$ implies that $\ker V^\top$ contains nonzero vectors. There are consequently nonzero equilibria in E_c arbitrarily close to the origin, and these equilibria do not converge to it. The origin is therefore not asymptotically stable. Finally, for $h^{(1)} > 1$, solving the exact equation from (i) gives $X^{(1)}(t) = e^{(h^{(1)}-1)t} X^{(1)}(0)$. Every nonzero initial condition in E_c thus grows exponentially. □

We now turn to the secular crossing. Unlike the null-space case, the critical eigenspace need not be invariant under the nonlinear system. We therefore construct a local centre manifold tangent to it and study the dynamics on this manifold. We first consider $d = 1$, for which the critical eigenvalue is simple and the reduced equation is scalar. The sign symmetry forces this equation to be odd; computing its cubic coefficient then determines whether the crossing produces a supercritical or subcritical pitchfork, provided that coefficient is nonzero. The extension to higher hidden dimension is discussed afterward.

Theorem D.14 (Pitchfork bifurcation of SHC-1)**.** *Let $d = 1$, write $W = \omega$ for the scalar score coefficient, let $L \geq 2$, and assume Assumption B.1(B), together with either Scenario 1 or Scenario 2 with $m > 0$. Let h_1^* be the critical value (27), write $h_1^* := h_1^*$ and $h_j^* := h^{(j)}$ for $j \geq 2$, and define*

$$u_j^* := \frac{\lambda_{\max}}{L} \frac{\gamma^{(j)}}{1 - h_j^*}, \qquad S_2 := \sum_{j=1}^{L} \frac{\gamma^{(j)}}{(1 - h_j^*)^2}, \qquad c_\eta := \frac{\gamma^{(1)}}{S_2 (1 - h_1^*)^2} > 0. \tag{28}$$

In Scenario 2 assume in addition $u^* \neq \mathbf{1}_L/L$. Set

$$c_3 := \frac{\lambda_{\max}\omega}{L}\Big(\|u^*\|^2 - \frac{1}{L}\Big)\frac{1}{S_2}\sum_{j=1}^{L}\frac{(\gamma^{(j)})^2}{(1-h_j^*)^2}. \tag{29}$$

Then $\operatorname{sign} c_3 = \operatorname{sign}\lambda_{\max}$*, and system* (19) *undergoes a pitchfork bifurcation at $h^{(1)} = h_1^*$. Write $\eta := h^{(1)} - h_1^*$. For $|\eta|$ small and initial data sufficiently close to the origin:*

(i) Stability switch. *The origin is locally exponentially stable for $\eta < 0$ and unstable for $\eta > 0$. At $\eta = 0$, it is locally asymptotically stable, but not exponentially stable, in Scenario 1 and unstable in Scenario 2.*

(ii) Scenario 1 is supercritical ($c_3 < 0$). *For $\eta > 0$, two stable equilibria $X_\pm^* = \theta_\pm^* u^* + O(|\theta^*|^3)$ emerge from the origin, with*

$$\theta_\pm^* = \pm\sqrt{c_\eta \eta/|c_3|}\big(1+o(1)\big).$$

Every trajectory off the codimension-one local stable manifold of the origin converges to one of them. Trajectories on that manifold converge to the origin.

(iii) Scenario 2 is subcritical ($c_3 > 0$). *For $\eta > 0$ there is no stable equilibrium near the origin, and every trajectory off the local stable manifold of the origin leaves a fixed neighbourhood of it. For $\eta < 0$ the origin is stable, but the width of its local attracting region in the reduced centre coordinate is $O(\sqrt{|\eta|})$ as $\eta \to 0^-$, bounded by two unstable equilibria at distance $O(\sqrt{|\eta|})$. Their local stable manifolds separate the local attracting and exiting regions. This gives no assertion that the full global basin shrinks to the origin.*

Proof. We separate the critical (centre) and stable directions, construct a local centre manifold, and compute the scalar dynamics on it.

Linearization and nonlinear terms. Set $\eta := h^{(1)} - h_1^*$ and let A^* be the Jacobian at the critical parameter. The system can be written as

$$\dot{X} = A^* X + \eta e_1 e_1^\top X + N(X),$$

where e_1 is the first standard basis vector and

$$N_\ell(X) = \gamma^{(\ell)}\lambda_{\max}\sum_{i=1}^{L}\Big[\operatorname{softmax}_i(\omega X_\ell X_i) - \frac{1}{L}\Big]X_i.$$

Since $d = 1$, $V = \lambda_{\max}$ here. The attention scores are unchanged under $X \mapsto -X$, so N is odd. Expanding the softmax at zero gives

$$N_\ell(X) = \frac{\lambda_{\max}\omega}{L}\gamma^{(\ell)} X_\ell\big(\|X\|^2 - L\bar{X}^2\big) + O(\|X\|^5), \qquad \bar{X} := \frac{1}{L}\sum_i X_i.$$

Thus the nonlinearity starts at cubic order, and $DN(X) = O(\|X\|^2)$.

Centre and stable coordinates. By Proposition D.11, A^* has a simple zero eigenvalue and all its other eigenvalues are negative. Let u^* and w^* be right and left zero eigenvectors, normalized by $\mathbf{1}_L^\top u^* = 1$ and $w^{*\top} u^* = 1$. Solving the eigenvector equations yields

$$u_j^* = \frac{\lambda_{\max}\gamma^{(j)}}{L(1-h_j^*)}, \qquad w_j^* = \frac{L}{\lambda_{\max} S_2 (1-h_j^*)}, \qquad S_2 = \sum_j \frac{\gamma^{(j)}}{(1-h_j^*)^2}.$$

The normalization of u^* follows from the secular condition $\sum_j \gamma^{(j)}/(1-h_j^*) = L/\lambda_{\max}$.

Choose a matrix P_s whose columns form a basis for the stable subspace, and define the stable coordinate map Q_s by

$$\begin{pmatrix} w^{*\top} \\ Q_s \end{pmatrix} = \begin{pmatrix} u^* & P_s \end{pmatrix}^{-1}.$$

Writing $X = \theta u^* + P_s\zeta$, we have $\theta = w^{*\top} X$ and $\zeta = Q_s X$. With $B := Q_s A^* P_s$, whose eigenvalues are negative, the transformed equations are

$$\begin{aligned} \dot{\theta} &= c_\eta \eta\theta + \eta w_1^* (P_s\zeta)_1 + w^{*\top} N(\theta u^* + P_s\zeta), \\ \dot{\zeta} &= B\zeta + \eta Q_s e_1\big(\theta u_1^* + (P_s\zeta)_1\big) + Q_s N(\theta u^* + P_s\zeta), \end{aligned}$$

where

$$c_\eta = w_1^* u_1^* = \frac{\gamma^{(1)}}{S_2(1-h_1^*)^2} > 0.$$

This is also the derivative of the critical eigenvalue with respect to η at the crossing.

The invariant graph. Adjoin $\dot\eta = 0$. The centre-manifold theorem gives a sufficiently smooth local invariant graph $\zeta = \Psi(\theta,\eta)$, with $\Psi(0,0) = 0$ and $D\Psi(0,0) = 0$. We choose it to respect the sign symmetry: $\Psi(-\theta,\eta) = -\Psi(\theta,\eta)$.

For clarity, if F_c and F_s denote the right-hand sides of the transformed centre and stable equations, invariance means

$$\partial_\theta \Psi(\theta,\eta) F_c(\theta,\Psi(\theta,\eta),\eta) = F_s(\theta,\Psi(\theta,\eta),\eta).$$

There is no parameter-derivative term because $\dot\eta = 0$. Although this equation can be used to compute the coefficients of Ψ, we only need its order of vanishing. Smoothness and oddness give $\Psi(\theta,\eta) = \partial_\theta \Psi(0,\eta)\theta + O(|\theta|^3)$, while tangency gives $\partial_\theta \Psi(0,\eta) = O(|\eta|)$. Hence

$$\Psi(\theta,\eta) = O\big(|\theta|(|\eta|+\theta^2)\big).$$

Dynamics on the centre manifold. Substitute $\zeta = \Psi(\theta,\eta)$ into the centre equation. The graph estimate implies

$$\eta w_1^* (P_s \Psi)_1 = O\big(|\theta|(\eta^2 + |\eta|\theta^2)\big).$$

Since $DN(X) = O(\|X\|^2)$, it also implies

$$N(\theta u^* + P_s\Psi) - N(\theta u^*) = O\big(|\theta|^3(|\eta|+\theta^2)\big).$$

Consequently, the graph correction does not affect the pure cubic coefficient. On $X = \theta u^*$, the normalization $\mathbf{1}_L^\top u^* = 1$ gives $\|X\|^2 - L\bar X^2 = \theta^2(\|u^*\|^2 - 1/L)$. Projecting the cubic expansion with $w^{*\top}$ therefore yields

$$\dot\theta = c_\eta \eta\theta + c_3\theta^3 + O\big(|\theta|(\eta^2 + |\eta|\theta^2 + \theta^4)\big), \tag{30}$$

where

$$c_3 = \frac{\lambda_{\max}\omega}{L}\left(\|u^*\|^2 - \frac{1}{L}\right)\sum_j w_j^* \gamma^{(j)} u_j^* = \frac{\lambda_{\max}\omega}{L}\left(\|u^*\|^2 - \frac{1}{L}\right)\frac{1}{S_2}\sum_j \frac{(\gamma^{(j)})^2}{(1-h_j^*)^2}.$$

Sign of the cubic coefficient. Since $\sum_j u_j^* = 1$,

$$\|u^*\|^2 - \frac{1}{L} = \sum_j \left(u_j^* - \frac{1}{L}\right)^2.$$

This factor vanishes exactly when $u^* = \mathbf{1}_L/L$, which is excluded in Scenario 2. In Scenario 1, such equality would imply $1 - h_j^* = \lambda_{\max}\gamma^{(j)} < 0$ for $j \geq 2$, contrary to $h^{(j)} < 1$. Thus this factor is positive in both cases. The remaining sum is positive, and $\omega > 0$, so $c_3 \neq 0$ and $\operatorname{sign}(c_3) = \operatorname{sign}(\lambda_{\max})$.

Bifurcation and stability. The reduced equation is odd in θ, so it can be written as $\dot\theta = \theta G(\theta,\eta)$, where G is smooth, even in θ, and

$$G(\theta,\eta) = c_\eta\eta + c_3\theta^2 + O(\eta^2 + |\eta|\theta^2 + \theta^4).$$

Because $\partial_\eta G(0,0) = c_\eta > 0$, the implicit function theorem expresses the nonzero-equilibrium condition as

$$\eta = -\frac{c_3}{c_\eta}\theta^2 + O(\theta^4).$$

It follows that two opposite branches exist on the side $\eta c_3 < 0$, with

$$\theta_\pm^* = \pm\sqrt{-\frac{c_\eta\eta}{c_3}}\big(1 + O(|\eta|)\big).$$

The derivative of the reduced vector field at these equilibria is $-2c_\eta\eta + O(\eta^2)$. They are therefore stable for $\eta > 0$ and unstable for $\eta < 0$. At the origin, the derivative is $c_\eta\eta + O(\eta^2)$, so the origin changes from stable to unstable as η increases through zero. This gives a supercritical pitchfork in Scenario 1 and a subcritical pitchfork in Scenario 2.

At $\eta = 0$, the reduced equation is $\dot\theta = c_3\theta^3 + O(\theta^5)$. The origin is therefore asymptotically stable in Scenario 1 and unstable in Scenario 2. Its zero linear eigenvalue rules out exponential stability in Scenario 1.

The transverse directions remain stable near the crossing, so centre-manifold reduction transfers these local conclusions to the full system. All sufficiently nearby equilibria lie on the invariant graph. Along the bifurcating branches, $\eta = O(\theta^2)$, and hence

$$X_\pm^* = \theta_\pm^* u^* + P_s \Psi(\theta_\pm^*, \eta) = \theta_\pm^* u^* + O(|\theta_\pm^*|^3).$$

For $\eta > 0$, the origin has one unstable direction and a codimension-one local stable manifold. In Scenario 1, nearby trajectories outside that manifold converge to one of the two stable equilibria. In Scenario 2, they leave a sufficiently small fixed neighbourhood of the origin. For $\eta < 0$ in Scenario 2, the two unstable reduced equilibria bound the local attracting interval, whose width is $O(\sqrt{|\eta|})$; their stable manifolds provide the corresponding local boundaries in the full system. These statements concern local dynamics and make no claim about the global basin of attraction or divergence to infinite norm. □

Corollary D.15 (Token fate near the threshold)**.** *Under the hypotheses of Theorem D.14, the single number h_1^* determines the local behaviour near the collapsed state. Below it, all nearby representations converge to zero. Above it, apart from the measure-zero local stable manifold of the origin, they either converge to one of two nontrivial states (Scenario 1) or leave a neighbourhood of the origin (Scenario 2). This local result does not prove divergence to infinite norm.*

Remark D.16 (Architectural interpretation)**.** Within the stated heterogeneous SHC-1 parameter path, raising $h^{(1)}$ through $h_1^* > 1$ in Scenario 1 creates two attracting nonzero equilibria; in Scenario 2, lowering it below $h_1^* < 1$ restores local exponential stability. These statements concern the frozen ODE. They do not establish global collapse or divergence for residual attention, nor a performance ordering for trained Transformers. Proposition C.4 only establishes non-hyperbolicity of residual attention at zero.

Remark D.17 (Role of Assumption B.1(B))**.** The query–key matrix does not enter the linear analysis: at $X = 0$ the attention weights are uniform, so J_{SHC1}, the thresholds h_1^* and 1, and Propositions D.11 and D.13 are independent of W. It enters only through the cubic coefficient, where c_3 carries the factor $\lambda_{\max}\omega$. Assumption B.1(B) fixes $\omega > 0$ and thereby ties the direction of the bifurcation to the sign of $\lambda_{\max}$ alone. Without it, the same proof shows that the direction is decided by the sign of $\lambda_{\max}\omega$, so a negative ω would exchange the two cases.

Remark D.18 (Summary)**.** With $h^{(1)}$ as the bifurcation parameter and $h^{(i)} < 1$ fixed for $i \geq 2$, the first crossing is:

Condition	Rank	Threshold	Critical space	Character
$\lambda_{\max} < 0$	$r = d$	$h_1^* > 1$	$u^* \otimes E_{\lambda_{\max}}$	Supercritical pitchfork†
$\lambda_{\max} < 0$	$r < d$	$h^{(1)} = 1$	$e_1 \otimes \ker V^\top$	Exactly linear (Prop. D.13)
$\lambda_{\max} > 0$, $m > 0$	Any	$h_1^* < 1$	$u^* \otimes E_{\lambda_{\max}}$	Subcritical pitchfork†
$\lambda_{\max} > 0$, $m \leq 0$	Any	None	—	Unstable for all $h^{(1)}$

†Requires $W \succ 0$, simple $\lambda_{\max}$, and $u^* \neq \mathbf{1}_L/L$; see Theorem D.14 and Proposition D.21. Repeated value eigenvalues and the uniform exception are discussed below.

D.5 SHC-1 in general hidden dimension d

We now extend the scalar pitchfork result to vector-valued tokens. The essential requirement is that the full Jacobian have a simple zero eigenvalue at the secular crossing, with all remaining eigenvalues negative. Under this condition, the centre manifold at the crossing is still one-dimensional, even when $d > 1$. We first identify the critical space and then show that the scalar calculation applies on an invariant subspace of the full system.

Throughout this subsection, we consider a secular first crossing as in Proposition D.11(i). All starred coefficients are evaluated at $h^{(1)} = h_1^*$. Write

$$A_\lambda^* := \mathrm{diag}(h_1^* - 1, \ldots, h_L^* - 1) + \frac{\lambda}{L}\gamma \mathbf{1}_L^\top, \qquad A^* := A_{\lambda_{\max}}^*.$$

We use the critical token vector u^* and the normalization from the scalar analysis:

$$u_j^* = \frac{\lambda_{\max}\gamma^{(j)}}{L(1-h_j^*)}, \qquad S_2 = \sum_j \frac{\gamma^{(j)}}{(1-h_j^*)^2}, \qquad w_j^* = \frac{L}{\lambda_{\max} S_2 (1-h_j^*)}.$$

Thus $A^*u^* = 0$, $w^{*\top}A^* = 0$, $\mathbf{1}_L^\top u^* = 1$, and $w^{*\top}u^* = 1$.

Lemma D.19 (Critical space at a secular first crossing)**.** *At a secular first crossing, zero is a simple eigenvalue of A^*. The kernel of the full Jacobian is*

$$\ker J_{\mathrm{SHC1}} = \{u^* \otimes v : v \in E_{\lambda_{\max}}\},$$

and its dimension equals the multiplicity $n_{\max}$ of $\lambda_{\max}$ as an eigenvalue of $V^\top$. All remaining eigenvalues of J_{SHC1} are negative. In particular, the full Jacobian has a simple zero eigenvalue if and only if $\lambda_{\max}$ is simple.

Proof. No pole equals zero at a secular first crossing. Hence $f(0) = L/\lambda$ can hold for only one distinct nonzero value eigenvalue, namely $\lambda_{\max}$. The determinant identity gives

$$\frac{d}{d\mu}\det(\mu I - A^*)\Big|_{\mu=0} = -\det(-H_{\mathrm{diag}}^*)\frac{\lambda_{\max}}{L} f'(0) \neq 0,$$

where $H_{\mathrm{diag}}^* = \mathrm{diag}(h_1^* - 1, \ldots, h_L^* - 1)$. Indeed, this diagonal matrix is invertible and $f'(0) = -S_2 < 0$. Thus zero is simple in A^*.

For each $v \in E_\lambda$, the Jacobian acts on token-feature products by $J_{\mathrm{SHC1}}(u \otimes v) = (A_\lambda^* u) \otimes v$. The spectral decomposition therefore gives the stated kernel, with one independent critical direction for each independent vector in $E_{\lambda_{\max}}$. Negativity of all remaining eigenvalues follows from Proposition D.11. □

The preceding lemma separates two notions of simplicity. A secular root may be simple in the token matrix A^* while appearing several times in the full Jacobian because $\lambda_{\max}$ is repeated. Distinct carry-over coefficients $h^{(i)}$ do not remove this multiplicity. For $d = 1$, however, the value matrix is scalar, so its nonzero eigenvalue is necessarily simple.

The next lemma connects the vector-valued system to the scalar problem. It concerns an entire subspace of token configurations, rather than only the critical line.

Lemma D.20 (Invariant value-eigendirection subspaces)**.** *Let $v \neq 0$ satisfy $V^\top v = \lambda v$. Then*

$$\mathcal{V}_v := \{X : X^{(\ell)} = s_\ell v,\ \ell = 1, \ldots, L, \text{ for some } s \in \mathbb{R}^L\}$$

is invariant under the full SHC-1 system. With $\omega := v^\top W v$, its restricted dynamics are

$$\dot{s}_\ell = (h^{(\ell)} - 1)s_\ell + \gamma^{(\ell)}\lambda \sum_{i=1}^{L} \mathrm{softmax}_i(\omega s_\ell s_i) s_i, \qquad \ell = 1, \ldots, L.$$

Proof. For $X^{(\ell)} = s_\ell v$, the attention scores are $X^{(\ell)\top} W X^{(i)} = \omega s_\ell s_i$. Every value vector is $V^\top X^{(i)} = \lambda s_i v$, so the attention contribution is parallel to v. The local term is also parallel to v. Thus the vector field preserves $\mathcal{V}_v$, and comparison of the coefficients of v gives the stated scalar equations. □

We can now extend the pitchfork theorem. Simplicity of $\lambda_{\max}$ supplies a single critical feature direction, and the invariant subspace above allows the centre-manifold calculation to be carried out using scalar token amplitudes.

Proposition D.21 (Pitchfork for arbitrary hidden dimension)**.** *Assume Assumption B.1, including $W = W^\top \succ 0$, and let $L \geq 2$. Vary $h^{(1)}$ while keeping $h^{(i)} < 1$ fixed for $i \geq 2$; these fixed coefficients need not be distinct. Suppose either:*

- *(i) Scenario 1 holds and $r = d$; or*
- *(ii) Scenario 2 holds and $m > 0$.*

Assume that $\lambda_{\max}$ is simple, with unit eigenvector v^*, and that $u^* \neq \mathbf{1}_L/L$. The latter condition is automatic in Scenario 1.

Then the system undergoes a pitchfork bifurcation at $h^{(1)} = h_1^*$. With $\eta := h^{(1)} - h_1^*$, the equation on a local centre manifold is

$$\dot{\theta} = c_\eta \eta\theta + c_3\theta^3 + O\big(|\theta|(\eta^2 + |\eta|\theta^2 + \theta^4)\big),$$

where

$$c_\eta = \frac{\gamma^{(1)}}{S_2(1-h_1^*)^2} > 0, \qquad c_3 = \frac{\lambda_{\max}(v^{*\top} W v^*)}{L}\left(\|u^*\|^2 - \frac{1}{L}\right)\langle w^*, \gamma \odot u^*\rangle.$$

Here $\odot$ denotes componentwise multiplication.

The local stability and bifurcation conclusions of Theorem D.14 hold: the pitchfork is supercritical in Scenario 1 and subcritical in Scenario 2. The two nonzero equilibrium branches satisfy

$$X_\pm^* = \theta_\pm^*(u^* \otimes v^*) + O(|\theta_\pm^*|^3), \qquad \theta_\pm^* = \pm\sqrt{-\frac{c_\eta \eta}{c_3}}\big(1 + O(|\eta|)\big),$$

on the side $\eta c_3 < 0$. Every token on these branches is exactly parallel to v^*.

Proof. By Lemma D.19, the full Jacobian has a simple zero eigenvalue with eigenvector $u^* \otimes v^*$, and all its other eigenvalues are negative. In Scenario 1, the full-rank assumption excludes an earlier null-space crossing. In Scenario 2, all $h_j^* < 1$, so any null-space modes remain stable at the secular threshold.

By Lemma D.20, the subspace $\mathcal{V}_{v^*}$ is exactly invariant. Its dynamics are the scalar-token system with value coefficient $\lambda_{\max}$ and attention coefficient $\omega = v^{*\top} W v^* > 0$. We may therefore apply the centre-manifold construction from Theorem D.14 within this subspace.

After adjoining $\dot{\eta} = 0$, the resulting manifold is invariant and tangent to the full centre subspace, consisting of the direction $u^* \otimes v^*$ and the parameter direction. It is consequently a centre manifold for the full system. The scalar calculation gives the displayed reduced equation and coefficients.

To determine the cubic sign, observe that

$$\langle w^*, \gamma \odot u^*\rangle = \frac{1}{S_2}\sum_j \frac{(\gamma^{(j)})^2}{(1-h_j^*)^2} > 0, \qquad \|u^*\|^2 - \frac{1}{L} = \sum_j \left(u_j^* - \frac{1}{L}\right)^2 > 0.$$

Hence $c_3 \neq 0$ and $\operatorname{sign}(c_3) = \operatorname{sign}(\lambda_{\max})$. The scalar pitchfork classification applies, and the stable complement gives the same local stability conclusions in the full state space.

All sufficiently nearby equilibria lie on this centre manifold, which is contained in $\mathcal{V}_{v^*}$. Their tokens are therefore exactly parallel to v^*. Finally, the centre-manifold correction is $O(|\theta|(|\eta| + \theta^2))$ and the equilibrium branches satisfy $\eta = O(\theta^2)$, giving the stated expansion for $X_\pm^*$. □

The positive-definiteness assumption on W fixes the sign of the effective scalar coefficient $v^{*\top} W v^*$. If it is replaced by $v^{*\top} W v^* \neq 0$, the same reduction holds, but the pitchfork direction is determined by $\operatorname{sign}(\lambda_{\max} v^{*\top} W v^*)$: a negative sign gives the supercritical case and a positive sign gives the subcritical case.

Corollary D.22 (The uniform critical profile)**.** *Retain the hypotheses of the preceding proposition except the condition $u^* \neq \mathbf{1}_L/L$. The cubic coefficient vanishes if and only if the critical token profile is uniform. Equivalently,*

$$u^* = \frac{\mathbf{1}_L}{L} \quad \Longleftrightarrow \quad h_j^* = 1 - \lambda_{\max}\gamma^{(j)} \quad \textit{for every } j.$$

This case cannot occur in Scenario 1. When it occurs in Scenario 2, the full system has the exact line of equilibria

$$\{X : X^{(j)} = a v^* \textit{ for every } j,\ a \in \mathbb{R}\}$$

at the critical parameter. The restricted vector field vanishes identically on this line, so higher-order terms do not restore a nonzero reduced dynamics there.

Proof. All factors in c_3 except $\|u^*\|^2 - 1/L$ are nonzero, and this factor vanishes exactly when $u^* = \mathbf{1}_L/L$. Substituting the formula for u_j^* gives the stated conditions on h_j^*. In Scenario 1, these conditions would imply $h_j^* > 1$ for $j \geq 2$, contrary to the hypotheses.

For a uniform state $X^{(j)} = av^*$, all attention scores in each row are equal, so the attention weights are uniform. The vector field at token j is therefore

$$\dot{X}^{(j)} = \big(h_j^* - 1 + \lambda_{\max}\gamma^{(j)}\big)av^* = 0.$$

This identity holds for every a, proving the final claim. □

Remark D.23 (Repeated leading value eigenvalues)**.** If $\lambda_{\max}$ has multiplicity $q > 1$, the secular threshold is still covered by the preceding spectral analysis, but the critical space has dimension q. The centre-manifold equation is then a system of q equations, so the scalar pitchfork classification does not apply to the full system.

Each fixed value eigendirection still defines an invariant scalar-token subsystem. However, a scalar calculation within one such subsystem does not determine stability with respect to the other critical directions. Our pitchfork classification is therefore restricted to simple $\lambda_{\max}$, which is automatic for $d = 1$. A full nonlinear classification of repeated leading eigenvalues is left for future work.

E SHC-2: Spectrum and exponential stability

E.1 Setting and notation

The 2-stream system (from (ODE-3) with $N = 2$ and shared parameters) reads

$$\dot{X}_1^{(\ell)} = (h_{11} - 1)X_1^{(\ell)} + h_{12}X_2^{(\ell)} + \beta_1 V^\top \sum_k \mathrm{softmax}\big(\tilde{X}^{(\ell)\top} W \tilde{X}^{(k)}\big)\tilde{X}^{(k)}, \tag{31}$$

$$\dot{X}_2^{(\ell)} = h_{21}X_1^{(\ell)} + (h_{22} - 1)X_2^{(\ell)} + \beta_2 V^\top \sum_k \mathrm{softmax}\big(\tilde{X}^{(\ell)\top} W \tilde{X}^{(k)}\big)\tilde{X}^{(k)}, \tag{32}$$

where $\tilde{X}^{(\ell)} = \alpha_1 X_1^{(\ell)} + \alpha_2 X_2^{(\ell)}$, $\alpha = (\alpha_1, \alpha_2)^\top$, $b = (\beta_1, \beta_2)^\top$, and $H = \left[\begin{smallmatrix} h_{11}-1 & h_{12} \\ h_{21} & h_{22}-1 \end{smallmatrix}\right]$. The matrix H is the continuous-depth generator, including the diagonal shifts, and H, α, b are shared across token positions. We assume the standing symmetry and positivity conditions on V, α, b; positive definiteness of W is not needed for the linear results.

Write $X^{(\ell)} = (X_1^{(\ell)}, X_2^{(\ell)}) \in \mathbb{R}^{2d}$ for the two streams at token ℓ. Define the uniform and contrast subspaces by

$$U = \{X : X^{(\ell)} = X^{(1)} \text{ for every } \ell\}, \qquad U^\perp = \{X : \textstyle\sum_\ell X^{(\ell)} = 0\}.$$

Every state splits uniquely into its token mean and its zero-sum contrasts, so $\mathbb{R}^{2dL} = U \oplus U^\perp$, with $\dim U = 2d$. Uniformity refers to equality across tokens, not equality between streams.

We follow the same sequence as for SHC-1: identify the critical linear modes, locate the stability boundary, and then study the full system in the critical directions.

E.2 Jacobian at the collapsed state

Lemma E.1 (Linearisation of the attention term)**.** *The origin $X = 0$ is a fixed point of (31)–(32). The partial derivative of the nonlinear attention term $F_i^{(l)}(X) = \beta_i V^\top \sum_{k=1}^L \mathrm{softmax}\big(\tilde{X}^{(l)\top} W \tilde{X}^{(k)}\big)\tilde{X}^{(k)}$ with respect to $X_j^{(k)}$, evaluated at $X = 0$, is*

$$\left.\frac{\partial F_i^{(l)}}{\partial X_j^{(k)}}\right|_{X=0} = \frac{\beta_i}{L}\alpha_j V^\top, \qquad i,j = 1,2, \quad k = 1,\dots,L.$$

Proof. At $X = 0$ every attention input $\tilde{X}^{(k)} = 0$, so every logit $\tilde{X}^{(l)\top} W \tilde{X}^{(k)} = 0$ and $\mathrm{softmax}(0) = \frac{1}{L}\mathbf{1}$, giving $F_i^{(l)}(0) = 0$; hence $X = 0$ is a fixed point. For the Jacobian, each logit is $O(\|X\|^2)$, so to first order

$$\begin{aligned} F_i^{(l)}(X) &= \frac{\beta_i}{L} V^\top \sum_{k=1}^L \tilde{X}^{(k)} + O(\|X\|^3) \\ &= \frac{\beta_i}{L} V^\top \sum_{k=1}^L \big(\alpha_1 X_1^{(k)} + \alpha_2 X_2^{(k)}\big) + O(\|X\|^3). \end{aligned}$$

Differentiating with respect to $X_j^{(k)}$ at $X = 0$ gives the stated formula. □

Lemma E.2 (Jacobian structure for the 2-stream system)**.** *The Jacobian of* (31)–(32) *at* $X=0$ *is* $J_{\mathrm{SHC2}}=J_{\mathrm{loc}}+J_{\mathrm{attn}}$*, where* $J_{\mathrm{loc}}=\mathrm{diag}(H\otimes I_d,\ldots,H\otimes I_d)$ *(*L *identical blocks) and every* (l,k)*-th* $2d\times 2d$ *block of* J_{attn} *equals* $\frac{1}{L}b\alpha^\top\otimes V^\top$*. Equivalently,*

$$J_{\mathrm{SHC2}}=I_L\otimes H\otimes I_d+\frac{\mathbf{1}_L\mathbf{1}_L^\top}{L}\otimes b\alpha^\top\otimes V^\top.$$

Proof. The local linear terms contribute $H\otimes I_d$ in each diagonal block. By Lemma E.1, the derivative of $F_i^{(l)}$ with respect to $(X_1^{(k)},X_2^{(k)})$ is $\frac{1}{L}(\beta_i\alpha_1,\beta_i\alpha_2)\otimes V^\top$, which is the i-th row of $\frac{1}{L}b\alpha^\top\otimes V^\top$. Since b and α are independent of both l and k, every block of J_{attn} is identical. □

Proposition E.3 (Uniform and contrast decomposition)**.** *The Jacobian preserves* U *and* $U^\perp$*. Identifying a uniform state with its common token vector gives*

$$J_{\mathrm{SHC2}}|_U=H\otimes I_d+b\alpha^\top\otimes V^\top,\qquad J_{\mathrm{SHC2}}|_{U^\perp}=I_{L-1}\otimes H\otimes I_d,$$

where the second formula uses an orthonormal basis of the zero-sum token space. Along a value eigendirection with eigenvalue λ*, the uniform stream matrix is* $H+\lambda b\alpha^\top$*.*

Proof. The token-averaging matrix $\mathbf{1}_L\mathbf{1}_L^\top/L$ is the identity on constant token vectors and zero on zero-sum token vectors. Applying the Jacobian formula to these two subspaces gives the restrictions. For $V^\top v=\lambda v$, the uniform restriction sends $s\otimes v$ to $[(H+\lambda b\alpha^\top)s]\otimes v$, proving the last assertion. □

This is a decomposition of the linearized system. The full nonlinear flow also preserves U, as we prove below, but it need not preserve $U^\perp$ for general L.

Theorem E.4 (Eigenvalues of J_{SHC2})**.** *The spectrum consists of the spectrum of* H *repeated* $Ld-r$ *times and, for each nonzero value eigenvalue* λ_k *listed with multiplicity, the two roots of*

$$\mu^2+B_k\mu+C_k=0, \tag{33}$$

where

$$B_k=-\mathrm{tr}(H)-\lambda_k\alpha^\top b,\qquad C_k=\det(H)+\lambda_k C_0(H),$$

and

$$\begin{aligned}C_0(H)&=\alpha^\top\mathrm{adj}(H)b\\&=\beta_1\alpha_1(h_{22}-1)+\beta_2\alpha_2(h_{11}-1)-\beta_2\alpha_1h_{12}-\beta_1\alpha_2h_{21}.\end{aligned}$$

All multiplicities are algebraic and add when roots from different blocks coincide. No diagonalizability assumption on H *is required.*

Proof. The contrast subspace contributes $(L-1)d$ copies of H. The uniform value-kernel directions contribute another $d-r$ copies, while each nonzero value eigendirection contributes $H+\lambda_k b\alpha^\top$. This accounts for $2(L-1)d+2(d-r)+2r=2Ld$ eigenvalues.

The trace of the latter block is $\mathrm{tr}(H)+\lambda_k\alpha^\top b$. The rank-one determinant identity gives

$$\det(H+\lambda_k b\alpha^\top)=\det(H)+\lambda_k\alpha^\top\mathrm{adj}(H)b.$$

Its characteristic polynomial is therefore (33). Equivalently, for $\mu\notin\mathrm{spec}H$, the determinant lemma gives

$$\alpha^\top(\mu I_2-H)^{-1}b=\frac{1}{\lambda_k}. \tag{34}$$

The polynomial formula remains valid at eigenvalues of H, so no roots are lost when the blocks have coincident eigenvalues. □

E.3 Stability region

Having identified all linear modes, we require each block to be stable. This determines exponential stability; behaviour on the boundary requires the nonlinear analysis in the next section.

Theorem E.5 (Exponential stability of the origin)**.** *For* $L\geq 2$*, the origin is locally exponentially stable if and only if*

$$\mathrm{tr}(H)<0,\qquad \det(H)>0,\qquad B_k>0,\qquad C_k>0\quad \textit{for all } k.$$

Proof. A real quadratic $\mu^2+b\mu+c$ has both roots in the open left half-plane exactly when $b,c>0$. Apply this to the characteristic polynomial of H and to (33). The Jacobian is therefore Hurwitz exactly under the stated conditions, which are equivalent to local exponential stability for this smooth system. □

Corollary E.6 (Binding constraints)**.** *For $V\neq 0$, write $S:=\alpha^\top b>0$, $B:=-\lambda_{\max}S$, and*

$$D_1:=\max\bigl(0,-\lambda_{\min}C_0(H)\bigr),$$
$$D_2:=\max\bigl(0,-\lambda_{\max}C_0(H),-\lambda_{\min}C_0(H)\bigr).$$

In Scenario 1, exponential stability is equivalent to $\operatorname{tr}(H)<0$ and $\det(H)>D_1$. In Scenario 2, it is equivalent to $\operatorname{tr}(H)<B$ and $\det(H)>D_2$.

Proof. By Theorem E.5, local exponential stability is equivalent to: (a) $\operatorname{tr}(H)<0$ and $\det(H)>0$; and (b) $B_k>0$ and $C_k>0$ for every $k=1,\dots,r$. We reduce (a)–(b) to the stated scalar inequalities in each scenario.

Scenario 1 ($\lambda_{\max}<0$, so all $\lambda_k<0$).

Condition on B_k. $B_k=-\operatorname{tr}(H)-\lambda_k(\beta_1\alpha_1+\beta_2\alpha_2)$. Since $\lambda_k<0$ and $S=\beta_1\alpha_1+\beta_2\alpha_2>0$, we have $-\lambda_k S>0$ for every k, hence $B_k>-\operatorname{tr}(H)$. Given condition (a) ($\operatorname{tr}(H)<0$), $B_k>0$ holds automatically for every k; conversely, condition (a) is required in its own right by the H-modes. So no constraint beyond $\operatorname{tr}(H)<0$ arises.

Condition on C_k. $C_k=\det(H)+\lambda_k C_0(H)$; we need $C_k>0$ for every k. *Case $C_0(H)\le 0$:* then $\lambda_k C_0(H)\ge 0$ for every $\lambda_k<0$, so $C_k\ge\det(H)$, and $C_k>0$ for all k whenever $\det(H)>0$ — already required by (a). Here the constraint imposes nothing beyond $\det(H)>0=\max(0,-\lambda_{\min}C_0(H))$. *Case $C_0(H)>0$:* then $\lambda_k C_0(H)<0$ for every k, and the binding constraint is at $\lambda_{\min}$ (the most negative eigenvalue), since $\lambda_{\min}C_0(H)$ is the most negative contribution:

$$C_k\ge\det(H)+\lambda_{\min}C_0(H)\quad\text{for all } k,$$

with equality at $k=\operatorname{argmin}_k\lambda_k$. Hence $C_k>0$ for all k iff $\det(H)>-\lambda_{\min}C_0(H)$. Combining both cases with $\det H>0$: $C_k>0$ for all k iff $\det(H)>\max\bigl(0,-\lambda_{\min}C_0(H)\bigr)=D_1$.

Conclusion. Conditions (a)–(b) reduce to $\operatorname{tr}(H)<0$ (which ensures $B_k>0$ for all k and, with $\det(H)>0$, the stability of the H-modes) and $\det(H)>D_1$ (which ensures $C_k>0$ for all k and implies $\det(H)>0$). These are necessary and sufficient.

Scenario 2 ($\lambda_{\max}>0$, remaining λ_k of either sign).

Condition on B_k. Since $S>0$, the term $-\lambda_k S$ is decreasing in λ_k, so $\min_k B_k$ is attained at $\lambda_k=\lambda_{\max}$:

$$\min_k B_k=-\operatorname{tr}(H)-\lambda_{\max}S=-\operatorname{tr}(H)+B,$$

recalling $B=-\lambda_{\max}S<0$. Hence $B_k>0$ for all k iff $-\operatorname{tr}(H)+B>0$, i.e. $\operatorname{tr}(H)<B$.

Condition on C_k. The minimum of $C_k=\det(H)+\lambda_k C_0(H)$ over k depends on the sign of $C_0(H)$. *Case $C_0(H)>0$:* C_k is decreasing in $-\lambda_k$, i.e. the binding constraint is at $\lambda_{\min}$ (most negative): $C_k>0$ for all k iff $\det(H)>-\lambda_{\min}C_0(H)$. *Case $C_0(H)<0$:* C_k is increasing in $-\lambda_k$; the binding constraint is at $\lambda_{\max}$ (largest, positive): $C_k>0$ for all k iff $\det(H)>-\lambda_{\max}C_0(H)$. *Case $C_0(H)=0$:* $C_k=\det(H)>0$ for all k, already required by (a). Combining with $\det H>0$: $C_k>0$ for all k iff $\det(H)>\max\bigl(0,-\lambda_{\max}C_0(H),-\lambda_{\min}C_0(H)\bigr)=D_2$; the 0 in the maximum accounts for (a).

Condition (a) is implied. Since $\lambda_{\max}>0$ and $S>0$, $B=-\lambda_{\max}S<0$, so $\operatorname{tr}(H)<B<0$; and $D_2\ge 0$, so $\det(H)>D_2\ge 0$. Hence (a) is automatic, and conditions (a)–(b) reduce to $\operatorname{tr}(H)<B$ and $\det(H)>D_2$. □

These thresholds depend on H through $C_0(H)$. They must therefore be recomputed when H varies, rather than treated as fixed levels in an unrestricted trace–determinant plane. When $V=0$, attention vanishes and exponential stability is determined by H alone.

F SHC-2: nonlinear structure and crossings

The spectral conditions locate the loss of exponential stability, but do not determine whether nonzero equilibria or periodic solutions appear. As in SHC-1, we must examine the full vector field in the critical directions. We first identify an exact invariant subspace on which the dynamics are linear. We then compute the cubic terms for contrast modes and prove bifurcation results when the critical dynamics have one real amplitude or one oscillatory pair.

F.1 The nonlinearity vanishes when attention-input contrast vanishes

Define the mean attention input and the token contrast

$$\bar{\tilde{X}} := \frac{1}{L}\sum_j \tilde{X}^{(j)}, \qquad \delta^{(\ell)} := \tilde{X}^{(\ell)} - \bar{\tilde{X}},$$

and write $p_{\ell k} := \mathrm{softmax}_k\big(\tilde{X}^{(\ell)\top} W \tilde{X}^{(k)}\big)$ for the attention weights (so $\sum_k p_{\ell k} = 1$). The following identity shows why uniform critical modes differ from contrast modes.

Lemma F.1 (Exact decomposition and contrast dependence)**.** *The system* (31)–(32) *can be written* exactly *as*

$$\dot{X}^{(\ell)} = (J_{\mathrm{SHC2}}X)^{(\ell)} + N^{(\ell)}, \qquad N^{(\ell)} = b \otimes \Big(V^\top \sum_k \big(p_{\ell k} - \tfrac{1}{L}\big)\delta^{(k)}\Big),$$

with

$$p_{\ell k} - \frac{1}{L} = \frac{1}{L}\tilde{X}^{(\ell)\top} W \delta^{(k)} + \textit{h.o.t.}, \qquad N = O\big(|X||\delta|^2\big) \textit{ locally.}$$

The weights still depend on the query $\tilde{X}^{(\ell)}$, and hence generally on the mean as well as the contrasts. In particular $N \equiv 0$ on the token-uniform subspace U, exactly and at every amplitude. *Only row-stochasticity $\sum_k p_{\ell k} = 1$ is used for the linearity on U, so this conclusion survives positional bias and causal masking (the invariant splitting of Proposition E.3 does not).*

Proof. Row-stochasticity and $\sum_k \delta^{(k)} = 0$ give the decomposition: writing $\tilde{X}^{(k)} = \bar{\tilde{X}} + \delta^{(k)}$,

$$\sum_k p_{\ell k}\tilde{X}^{(k)} = \bar{\tilde{X}} + \sum_k p_{\ell k}\delta^{(k)} = \underbrace{\frac{1}{L}\sum_k \tilde{X}^{(k)}}_{\text{attention part of } J_{\mathrm{SHC2}}X} + \sum_k \big(p_{\ell k} - \tfrac{1}{L}\big)\delta^{(k)},$$

where the term $b \otimes (V^\top \bar{\tilde{X}})$ is exactly the attention part of $(J_{\mathrm{SHC2}}X)^{(\ell)}$ (Lemma E.2). For the estimate, recentre using the translation invariance of the softmax: $p_{\ell k} = \mathrm{softmax}_k(z_{\ell k})$ with $z_{\ell k} = \tilde{X}^{(\ell)\top} W \delta^{(k)}$ (subtracting the k-independent quantity $\tilde{X}^{(\ell)\top} W \bar{\tilde{X}}$ from every logit leaves the softmax unchanged); a Taylor expansion in z gives $p_{\ell k} - \frac{1}{L} = \frac{1}{L} z_{\ell k} + O(|z|^2)$. Since $z = O(|X||\delta|)$ and N carries one further factor of δ, $N = O(|X||\delta|^2)$: both factors vanish with δ, hence $N \equiv 0$ on U exactly.

For the linear invariance caveat, when $V \neq 0$: $U^\perp$ is preserved by the Jacobian if and only if the zero-state weight matrix P_0 is doubly stochastic, $\mathbf{1}^\top P_0 = \mathbf{1}^\top$ (for $V = 0$ invariance holds for every P_0); it fails for the causal mask (column sums $\sum_{\ell \geq k} 1/\ell \neq 1$ for $L > 1$) and for a key-only bias $P_0 = \mathbf{1}\pi^\top$ exactly when π is nonuniform. These are statements about the Jacobian: for $L > 2$, $U^\perp$ is generally not invariant under the full nonlinear flow even for unmasked attention. □

The nonlinearity is therefore controlled by differences between attention inputs, although the weights also depend on their mean through the query. The exact linearity on U uses only that the weights in each row sum to one. It survives row-normalized positional bias or causal masking, but the spectral decomposition above need not survive those changes: the zero-state attention matrix is then generally not uniform.

F.2 Exactly linear dynamics on uniform critical spaces

Suppose all critical directions are uniform. The preceding lemma then allows us to construct a centre manifold directly, without solving for a nonlinear graph over the critical eigenspace.

Lemma F.2 (An exactly linear centre manifold)**.** *Suppose the entire centre spectral subspace E^c of J_{SHC2} is contained in U, with all complementary modes hyperbolic. Then E^c is an invariant centre*

manifold at criticality. For a smooth parameter path ε with a spectral gap, it continues to an invariant subspace $E^c(\varepsilon) \subset U$ with exactly linear restricted dynamics

$$\dot{z} = A_c(\varepsilon) z.$$

Here $A_c(\varepsilon)$ represents the restriction to the continued critical subspace. At a static boundary with semisimple zero eigenvalues, every point of E^c is an equilibrium. If the critical spectrum consists only of a semisimple pair $\pm i\Omega$, possibly repeated, every nonzero orbit in E^c is periodic with period $2\pi/\Omega$. These orbits form a continuum, not isolated limit cycles.

Proof. On U, the vector field is exactly the linear map $H \otimes I_d + b\alpha^\top \otimes V^\top$. Its critical spectral subspace is therefore invariant for the full system and tangent to the centre subspace, so it is a centre manifold. A spectral gap allows a smooth continuation of this invariant subspace along the parameter path. After adjoining $\dot{\varepsilon} = 0$, the resulting family is an invariant manifold tangent to the augmented centre subspace. Its fibre dynamics are the displayed linear equation.

At a semisimple zero eigenvalue, the critical restriction is zero. At a semisimple imaginary pair with common frequency Ω, the restriction is a direct sum of rotation blocks of that frequency. These give the asserted equilibria and periodic orbits. □

This lemma constructs an exact reduction in linear coordinates on the invariant subspace. There is consequently no nonlinear amplitude term to select an isolated branch there. Its hypothesis is essential: uniform critical directions alone do not suffice if contrast directions are critical at the same parameter.

F.3 Contrast dynamics and simple bifurcations

We next consider critical modes involving differences between tokens. The attention nonlinearity need not vanish on these modes. We first count the critical directions and compute their cubic terms; a scalar classification will then be possible in the simple cases identified below.

Proposition F.3 (Scenario-1 oscillatory critical space)**.** *In Scenario 1, at $\mathrm{tr}H = 0$ and $\det H > D_1$, the critical eigenvalues are $\pm i\sqrt{\det H}$, each with algebraic multiplicity $Ld - r$. The real centre space is*

$$E^c = U^\perp \oplus \big(\mathrm{span}\{\mathbf{1}_L\} \otimes \mathbb{R}^2 \otimes \ker V^\top\big).$$

All complementary eigenvalues have negative real parts. In particular, for full-rank V, $E^c = U^\perp$ and each imaginary eigenvalue has multiplicity $(L-1)d$. Along a smooth path, $(\mathrm{Re}\mu)' = (\mathrm{tr}H)'/2$ for these eigenvalues at the boundary.

Proof. Every contrast block and every uniform value-kernel block is a copy of H. At zero trace and positive determinant, its eigenvalues are $\pm i\sqrt{\det H}$. There are $(L-1)d + (d-r) = Ld - r$ such copies. For each nonzero $\lambda_k < 0$, the uniform block has negative trace $\lambda_k S$ and positive determinant by $\det H > D_1$, so it is stable. Near the boundary, the real part of the eigenvalues of H equals $\mathrm{tr}H/2$, giving the derivative formula. □

Unlike in the scalar SHC-1 theorem, the critical space can be large even when the value eigenvalues are simple. To see what the nonlinearity does, we evaluate it on a token contrast, a stream vector, and a value eigendirection. This gives a cubic term in the centre equation, but not yet a classification of a multidimensional crossing.

Proposition F.4 (The contrast cubic and its projection)**.** *Let $u \neq 0$ satisfy $\mathbf{1}_L^\top u = 0$, let $s \in \mathbb{R}^2$, and let $V^\top v = \lambda v$ with $\|v\| = 1$. Then*

$$N(\theta u \otimes s \otimes v) = \frac{\lambda (v^\top W v)\|u\|^2 (\alpha^\top s)^3}{L} \theta^3 (u \otimes b \otimes v) + O(\theta^5).$$

The cubic is nonzero exactly when $\lambda(v^\top W v)(\alpha^\top s) \neq 0$. At the oscillatory boundary of Proposition F.3, this is also the centre-manifold cubic along the stated direction in spectral-projection coordinates; it is not a radial normal-form coefficient.

At a static boundary with $\det H = 0$, $T := \mathrm{tr}H < 0$, and all nonzero-value uniform blocks stable, choose $Hs_0 = 0$, $w_0^\top H = 0$, and $w_0^\top s_0 = 1$. The centre projection along $u \otimes s_0 \otimes v$ has cubic coefficient

$$c(u,v) = \frac{\lambda (v^\top W v)\|u\|^2}{L} (\alpha^\top s_0)^3 (w_0^\top b), \qquad C_0(H) = T(\alpha^\top s_0)(w_0^\top b).$$

In particular, $C_0(H) \neq 0$ ensures that both stream factors are nonzero.

Proof. Along this direction the mean attention input is zero, and $\tilde{X}^{(\ell)} = \theta u_\ell(\alpha^\top s)v$. The softmax expansion therefore gives

$$p_{\ell j} - \frac{1}{L} = \frac{\theta^2}{L} u_\ell u_j (\alpha^\top s)^2 (v^\top W v) + O(\theta^4).$$

Multiplying by $\tilde{X}^{(j)}$, summing over j, and applying $b \otimes V^\top$ proves the first formula.

Choose a centre manifold respecting the odd symmetry of the vector field. At criticality its graph over the centre space starts at cubic order. Its linear contribution vanishes under the centre spectral projection, while its substitution into N changes only terms of fifth order or higher. Thus the projected cubic can be computed directly on the centre subspace, as in the SHC-1 proof. At the oscillatory boundary, $u \otimes b \otimes v$ already lies in that subspace. At the static boundary, the stream projection is $s_0 w_0^\top$, giving $c(u,v)$. Finally, H has eigenvalues $0, T$, so $\mathrm{adj} H = T I_2 - H = T s_0 w_0^\top$. Multiplication by $\alpha^\top$ and b gives the identity for $C_0(H)$. □

For $L = 2, d = 1$, the contrast token space has dimension one. A static contrast crossing can therefore have a single critical amplitude, and an oscillatory contrast crossing can have a single critical pair. These are the cases in which the preceding calculation yields a full local bifurcation result.

Proposition F.5 (A simple contrast pitchfork)**.** *Let $L = 2$, $d = 1$, $V = \lambda \neq 0$, and $W = \omega \neq 0$. Suppose $\det H = 0$, $T := \mathrm{tr}\, H < 0$, and $H + \lambda b \alpha^\top$ is Hurwitz at the critical parameter. Assume the zero eigenvalue of H crosses zero with nonzero derivative along a smooth parameter path. Using this eigenvalue as the local parameter, the reduced equation has a nonzero cubic coefficient of sign $-\mathrm{sign}\omega$. The pitchfork is supercritical for $\omega > 0$ and subcritical for $\omega < 0$.*

Proof. The full Jacobian has one zero eigenvalue and three stable eigenvalues. Take $u = (1,-1)^\top/\sqrt{2}$ and use the normalized stream eigenvectors from the preceding proposition. Stability of the uniform block implies $\lambda C_0(H) > 0$. Its projected cubic is therefore

$$c_3 = \frac{\omega}{2} (\alpha^\top s_0)^2 \frac{\lambda C_0(H)}{T}, \qquad \mathrm{sign} c_3 = -\mathrm{sign}\omega.$$

The identity for $C_0(H)$ shows that $\alpha^\top s_0 \neq 0$, so $c_3 \neq 0$. The centre manifold is one-dimensional and its equation is odd. The same invariant-graph and equilibrium-branch argument as in Theorem D.14 applies, with the crossing eigenvalue as parameter. The other directions remain stable, giving the stated classification in the full system. □

Proposition F.6 (A simple contrast oscillatory bifurcation)**.** *Let $L = 2$, $d = 1$, $V = \lambda < 0$, and $W = \omega \neq 0$. Suppose $\mathrm{tr} H = 0$, $\det H > D_1$, and $(\mathrm{tr} H)' \neq 0$ along a smooth parameter path. A branch of small periodic orbits emerges at this crossing. Oriented towards increasing trace, it is supercritical with stable periodic orbits for $\omega > 0$, and subcritical with unstable periodic orbits for $\omega < 0$.*

Proof. The subspace $X^{(1)} = y$, $X^{(2)} = -y$, with $y \in \mathbb{R}^2$, is invariant under the full system: it is the fixed subspace of token exchange followed by sign reversal. Its exact equation is

$$\dot{y} = Hy + \lambda b (\alpha^\top y) \tanh\big(\omega (\alpha^\top y)^2\big) = Hy + \lambda \omega b (\alpha^\top y)^3 + O(\|y\|^7).$$

At criticality this plane is a centre manifold, and the uniform complement is stable by Proposition F.3.

Let H_0 be the critical matrix and $\Omega = \sqrt{\det H_0} > 0$. Since H_0 is real-similar to a rotation generator, choose a symmetric matrix $Q \succ 0$ with $H_0^\top Q + Q H_0 = 0$. In the amplitude $\rho^2 = y^\top Q y$, the linear orbits can be written $y = \rho Q^{-1/2} e(\phi)$, where $e(\phi) = (\cos\phi, \sin\phi)^\top$ and the angular speed has magnitude Ω. The cubic contribution to $\dot{\rho}$ is

$$\lambda \omega \rho^3 (e^\top Q^{1/2} b)(\alpha^\top Q^{-1/2} e)^3.$$

Averaging over one rotation gives the radial cubic coefficient

$$\ell = \frac{3\lambda\omega}{8} (\alpha^\top b)(\alpha^\top Q^{-1} \alpha).$$

Indeed, for planar vectors c, d, the angular average of $(d \cdot e)(c \cdot e)^3$ is $3(d \cdot c)\|c\|^2/8$. There are no quadratic terms to contribute further corrections at cubic order. Since $\alpha^\top b > 0$ and $\lambda < 0$, $\ell \neq 0$ and its sign is $-\mathrm{sign}\omega$.

The linear radial growth rate is $\mathrm{tr} H/2$, with nonzero derivative along the path. The bifurcation theorem for a simple imaginary pair therefore gives a stable small periodic orbit on the positive-trace side when $\ell < 0$, and an unstable one on the negative-trace side when $\ell > 0$. The complementary uniform modes remain stable, so these stability conclusions also hold in the full system. □

F.4 Meaning of static and oscillatory stability boundaries

The terminology refers to the time dependence of the critical linearized modes. For a real eigenvalue μ with eigenvector v, the solution along that eigendirection is $x(t)=ce^{\mu t}v$, where c is its initial amplitude. There is no oscillatory factor, and the mode is constant at $\mu=0$. This explains the name *static* or zero-frequency boundary; it does not imply that the nonlinear system has a new branch of equilibria.

For a conjugate pair $a\pm i\omega$, with $\omega>0$, choose real coordinates (y_1,y_2) on an invariant plane in which the linearized equation is $\dot{y}_1=ay_1-\omega y_2$, $\dot{y}_2=\omega y_1+ay_2$. Writing $y_1=r\cos\phi$ and $y_2=r\sin\phi$ gives

$$\dot{r}=ar, \qquad \dot{\phi}=\omega, \qquad r(t)=r(0)e^{at}, \qquad \phi(t)=\phi(0)+\omega t.$$

Thus the mode oscillates while its amplitude decays for $a<0$ or grows for $a>0$; at $a=0$ the linearized oscillations have constant amplitude. This explains the name *oscillatory* or nonzero-frequency boundary. These statements apply to the invariant planes of the semisimple imaginary pairs considered here, including repeated pairs.

A boundary describes parameter values, whereas a crossing describes a parameter path passing through it. The derivative conditions in the following theorem ensure that the critical real part changes sign with nonzero speed along the path. Neither a static nor an oscillatory crossing alone determines the nonlinear branches. In particular, periodic orbits at the critical parameter on an exactly linear uniform centre are not isolated cycles born on one side of the boundary. Proposition F.6 proves a periodic branch in the simple contrast case; the general multidimensional contrast case is not classified here.

F.5 Which modes become critical at each boundary

We now combine the spectral conditions with the preceding nonlinear results. At a static boundary, we identify the blocks with zero determinant; at an oscillatory boundary, we identify those with zero trace and positive determinant. This determines whether the critical space is uniform, contains contrast directions, or includes both. The theorem states the nonlinear conclusion only where the corresponding hypotheses hold.

Theorem F.7 (Boundary carriers and proved nonlinear conclusions)**.** *Assume $L\geq 2$ and $V\neq 0$. Use the exponential-stability region of Corollary E.6. The following describes its regular boundary portions; trace/determinant intersections are excluded. A crossing along a parameter path additionally requires the transversality condition stated below.*

Define the zero space contributed by H at $\det H=0$ by

$$E_H=(\mathbf{1}_L^{\perp}\otimes\ker H\otimes\mathbb{R}^d)\oplus(\operatorname{span}\{\mathbf{1}_L\}\otimes\ker H\otimes\ker V^{\top}), \qquad \dim E_H=Ld-r.$$

The corresponding uniform zero space for a nonzero value eigenvalue λ is $\operatorname{span}\{\mathbf{1}_L\}\otimes\ker(H+\lambda b\alpha^{\top})\otimes E_\lambda$.

(i) ***Scenario 1, static*** *$(B1-)$: $\det H=D_1$, $\operatorname{tr}H<0$. If $C_0<0$, then $D_1=0$, the centre space is E_H, and every nonzero-value uniform block is stable. If $C_0>0$, then $D_1=-\lambda_{\min}C_0>0$; only the uniform blocks for $\lambda_{\min}$ have zero eigenvalues, and Lemma F.2 gives an exactly linear reduction. If $C_0=0$, then $\det H=0$ and every uniform block also has a zero eigenvalue: both carriers are critical, and Lemma F.2 does not classify their coupled centre dynamics.*

(ii) ***Scenario 1, oscillatory*** *$(B2-)$: $\operatorname{tr}H=0$, $\det H>D_1$. The frequency, centre space, multiplicity, and stable complement are given by Proposition F.3. For full-rank V the centre space consists entirely of contrast modes; otherwise it also contains uniform kernel modes. Proposition F.4 gives a nonzero contrast cubic on rays with $\lambda(v^{\top}Wv)(\alpha^{\top}s)\neq 0$. This is a proved oscillatory crossing with a contrast nonlinearity, not a general proof of periodic branches. The simple case $L=2,d=1$ is the oscillatory bifurcation proved in Proposition F.6.*

(iii) ***Scenario 2, static*** *$(B1+)$: $\det H=D_2$, $\operatorname{tr}H<B$. If $D_2>0$, H is stable and the centre space is the direct sum of the uniform zero spaces for all nonzero λ satisfying $D_2+\lambda C_0=0$. Lemma F.2 applies, giving a linear subspace of exact equilibria. If $D_2=0$ and $C_0>0$, all nonzero λ_k must be positive; the centre space is E_H, and all nonzero-value uniform blocks are stable. If $D_2=0$ and $C_0=0$, both carriers are critical as in (i). There is no case $D_2=0$, $C_0<0$ in Scenario 2.*

(iv) ***Scenario 2, oscillatory*** *$(B2+)$: $\operatorname{tr}H=B$, $\det H>D_2$. The frequency is $\Omega=\sqrt{\det H+\lambda_{\max}C_0}>0$. The pair is carried by the uniform $\lambda_{\max}$ blocks, with complex multiplicity $n_{\max}$ and real centre dimension $2n_{\max}$; all other modes are stable. Lemma F.2 gives an exactly linear centre with a continuum of periodic orbits, rather than a branch of isolated periodic orbits away from criticality.*

At a static boundary a carrying block has trace $T<0$ and a simple zero; its eigenvalue derivative is $\mu' = D'/T$, where D is the determinant of that block along the path. Thus $D' \neq 0$ is the transversality condition. For (ii) it is $(\operatorname{tr} H)' \neq 0$, and for (iv) it is $[\operatorname{tr} H + \lambda_{\max} S]' \neq 0$ (with the parameters in this expression varied along the path). The real-part derivative is half this trace derivative. These statements include the variation of $C_0(H)$; treating D_1, D_2 as constants while varying H is generally incorrect.

Proof. For the H blocks, stability requires $\operatorname{tr} H < 0$, $\det H > 0$. For a nonzero-value uniform block, the trace is $\operatorname{tr} H + \lambda S$ and the determinant is $\det H + \lambda C_0$. These are the full set of blocks by Proposition E.3, including the uniform kernel copies of H.

In (i), all nonzero-value uniform traces are negative. If $C_0 < 0$, their determinants at $\det H = 0$ are $\lambda C_0 > 0$. If $C_0 > 0$, the least determinant occurs at $\lambda_{\min}$ and first vanishes at $\det H = -\lambda_{\min} C_0 > 0$, while H remains Hurwitz. If $C_0 = 0$, all determinants equal $\det H$ and vanish together. The simple zero of a 2×2 matrix with negative trace gives exactly the eigenspaces stated. Part (ii) is Proposition F.3.

In (iii), $\operatorname{tr} H < B$ makes every block trace strictly negative. If $D_2 > 0$, H is Hurwitz and only the listed uniform determinants vanish. If $D_2 = 0$, its definition requires $\lambda_k C_0 \geq 0$ for every nonzero λ_k. Because $\lambda_{\max} > 0$, necessarily $C_0 \geq 0$. For $C_0 > 0$ all nonzero eigenvalues must be positive and the uniform determinants are strictly positive, leaving exactly E_H critical. For $C_0 = 0$ every determinant vanishes. This proves all cases of (iii).

In (iv), $\operatorname{tr} H = B < 0$ and $\det H > D_2 \geq 0$ make H Hurwitz. All uniform determinants are positive. Their traces are $(\lambda_k - \lambda_{\max}) S$, zero exactly for eigenvalues equal to $\lambda_{\max}$ and negative for the others. Hence the critical pair is semisimple, with the stated multiplicity. The entire centre space is in U, so Lemma F.2 applies. The same argument proves its use in the static uniform-only cases, whose zero eigenvalues are semisimple.

Finally, differentiating $\mu^2 - T\mu + D = 0$ at $\mu = 0$ gives $-T\mu' + D' = 0$. Near a nonzero imaginary pair the real part is $T/2$. These identities give the transversality statements. □

Remark F.8 (Scope, degeneracies, and a missing boundary case)**.** For a mixed-sign nonzero value spectrum, $D_2 = 0$ is equivalent to $C_0 = 0$. When all nonzero value eigenvalues are positive, however, $D_2 = 0$ is equivalent to $C_0 \geq 0$. For example,

$$H = \begin{pmatrix} -1 & -4 \\ -1/2 & -2 \end{pmatrix}, \qquad \alpha = b = (1,1)^\top, \qquad V = 1$$

has $\operatorname{tr} H = -3 < B = -2$, $\det H = 0$, $C_0 = 3/2 > 0$, and $\operatorname{tr}(H + b\alpha^\top) = -1$, $\det(H + b\alpha^\top) = 3/2$. Thus the Scenario 2 static crossing is a contrast crossing even for $d = 1$. This case cannot be labelled an exactly linear uniform crossing.

For $r = d$, the multiplicity of an H eigenvalue is $(L-1)d$. Only the token factor $\mathbf{1}_L^\perp$ is the standard irreducible S_L representation; the full centre representation includes stream and value factors and is not generally irreducible. Except for a simple critical mode/pair as in Propositions F.5–F.6, a scalar pitchfork calculation or a radial coefficient for one mode cannot classify the entire centre dynamics. Multidimensional branch existence, selection and stability remain open here. A nonzero projected cubic is evidence of a nonlinear centre reduction, not a proof of such a classification.

At $C_0 = 0$, simultaneous uniform and contrast criticality needs its own analysis; exact linearity of the entire centre manifold is not asserted. For singular V, $\mathbb{R}^L \otimes \mathbb{R}^2 \otimes \ker V^\top$ carries exactly the linear dynamics $I_L \otimes H$. If $V = 0$, the full system is linear.

A real 2×2 carrying block A has nonreal eigenvalues precisely if $(A_{11} - A_{22})^2 + 4A_{12}A_{21} < 0$. A negative off-diagonal product is necessary but is not sufficient by itself. At an oscillatory boundary $\operatorname{tr} A = 0$, the additional condition $\det A > 0$ supplies this inequality. Positive off-diagonal entries rule out the corresponding pair for H and for $H + \lambda_{\max} b\alpha^\top$ when $\lambda_{\max} > 0$.

F.6 Parameter sharing and its limits

Remark F.9 (Shared-parameter SHC-1)**.** For $h^{(\ell)} = h$ and $\gamma^{(\ell)} = \gamma$, the uniform eigenvalues are $(h-1) + \gamma\lambda$ for every eigenvalue λ of V, including zero; contrast eigenvalues are $h - 1$, with multiplicity $(L-1)d$. Thus local exponential stability requires $h < 1$ and, if $V \neq 0$, $h < 1 - \gamma\lambda_{\max}$. Repeated poles must be included as pole modes; the secular root is not repeated L times.

In Scenario 1 with $r = d$, the first crossing is at $h = 1$ with $(L-1)d$ contrast zero modes. Along $u \otimes v$, $u \perp \mathbf{1}_L$, $V^\top v = \lambda v$, the cubic is $[\gamma\lambda(v^\top W v)\|u\|^2/L]u \otimes v$ for unit v; this is a ray evaluation of a vector cubic, not generally a scalar normal form. If $r < d$, uniform kernel modes also cross, giving total centre dimension $Ld - r$. In Scenario 2 the first crossing is at $h = 1 - \gamma\lambda_{\max}$; the critical space is $\mathrm{span}\{\mathbf{1}_L\} \otimes E_{\lambda_{\max}}$ and its dynamics are exactly linear. Token-dependence can remove these degeneracies, but does not guarantee a simple critical value eigenvalue or exclude the uniform exception of Corollary D.22. These are the reasons for the explicit pitchfork hypotheses.

Token-dependent SHC-2 generally loses the simple uniform/contrast block decomposition and needs a different analysis; it does not automatically produce a high-dimensional or degenerate centre. Likewise, state-dependent HC coefficients need not make an ODE non-autonomous: dependence on the current state alone gives an autonomous nonlinear vector field, while explicit depth/time dependence gives a non-autonomous one.

G ODE Simulation Setup

G.1 Hyper-Connection ODE System

We consider a continuous-depth Hyper-Connection dynamical system with $L = 5$ tokens. The state is written as

$$X(t) = \begin{bmatrix} X_1(t) \\ \vdots \\ X_5(t) \end{bmatrix},$$

The Scenario 1 figures use `hc_simulation.py` with $L = 5$, $d = 1$, $V = -1$, $W = 1$, and $\alpha_k = \beta_k = 1$. The three ODE panels in Figure 2 use the carry-over vectors

$$\begin{aligned} &\text{Case 1:} \quad (-0.5,-0.5,-0.5,-0.5,-0.5), \\ &\text{Case 3:} \quad (10/9,0,0,0,0), \\ &\text{Case 5:} \quad (1.5,1.2,0,0,0). \end{aligned}$$

These are separate parameter configurations, not three points on a single h_1 sweep. The separate bifurcation diagram in Figure 1 fixes $h_2 = \cdots = h_5 = 0$ and varies h_1, with critical value $h_1^* = 10/9$. Here X denotes the unrescaled state of (ODE-2); with these unit input weights it coincides with the coordinates of (6). The variable $X_k(t)$ denotes the state of token k.

Each token evolves according to

$$\frac{dX_k}{dt} = (h_k - 1)X_k + \beta_k V^\top \sum_{i=1}^{5} a_{ki}(X)\alpha_i X_i, \qquad k = 1,\ldots,5, \tag{35}$$

where the first term is a token-wise linear contribution and the second term is an attention-weighted Hyper-Connection coupling across tokens.

The attention weights are defined by

$$a_{ki}(X) = \frac{\exp\left(\alpha_k X_k^\top W \alpha_i X_i\right)}{\sum_{j=1}^{5} \exp\left(\alpha_k X_k^\top W \alpha_j X_j\right)}, \tag{36}$$

so that the softmax is applied over the token index i for each fixed query token k.

- $h_k \in \mathbb{R}$ controls the linear decay or growth of token k.
- $\alpha_k, \beta_k \in \mathbb{R}$ are token-dependent coupling parameters.
- $V^\top$ is chosen so that all eigenvalues are negative.
- W is chosen positive definite, denoted $W > 0$.

G.2 Stability and Fixed Points

A fixed point X^* satisfies

$$0 = (h_k - 1)X_k^* + \beta_k V^\top \sum_{i=1}^{5} a_{ki}(X^*)\alpha_i X_i^*, \qquad k = 1,\ldots,5. \tag{37}$$

The origin $X=0$ is always a fixed point. To identify additional fixed points, the simulation solves the nonlinear system

$$\frac{dX}{dt}=0$$

from multiple initial guesses. Each fixed point is classified by the eigenvalues of the Jacobian of the full stacked system: it is stable if all eigenvalues have negative real part and unstable if at least one eigenvalue has positive real part.

G.3 Bifurcation Along h_1

For the bifurcation experiment, the parameters $\{h_i\}_{i=2}^5$, $\{\alpha_i\}_{i=1}^5$, and $\{\beta_i\}_{i=1}^5$ are fixed, while only h_1 is varied.

Let

$$\lambda_{\max}=\max_j \lambda_j(V^\top)$$

denote the largest eigenvalue of $V^\top$. Since $V^\top$ is negative definite, $\lambda_{\max}<0$. The critical value is determined by

$$\sum_{i=1}^{5}\frac{\alpha_i\beta_i}{1-h_i}=\frac{5}{\lambda_{\max}}. \tag{38}$$

Equivalently, solving for the bifurcation parameter h_1 gives

$$h_1^*=1-\frac{\alpha_1\beta_1}{\frac{5}{\lambda_{\max}}-\sum_{i=2}^{5}\frac{\alpha_i\beta_i}{1-h_i}}. \tag{39}$$

G.4 Simulation Regimes

The simulations verify five parameter regimes.

Case 1. All eigenvalues of $V^\top$ are negative, $W>0$, and $h_i<0$ for all i. The origin is locally exponentially stable; the plotted trajectories converge to 0. A finite fixed-point search does not establish uniqueness globally.

Case 2. All eigenvalues of $V^\top$ are negative, $W>0$, $h_1>1$, and $h_i<1$ for $i=2,...,5$, with

$$\sum_{i=1}^{5}\frac{\alpha_i\beta_i}{1-h_i}<\frac{5}{\lambda_{\max}}.$$

The origin is locally exponentially stable and sufficiently small initial states converge to 0.

Case 3. The same conditions as Case 2 hold, but with equality:

$$\sum_{i=1}^{5}\frac{\alpha_i\beta_i}{1-h_i}=\frac{5}{\lambda_{\max}}.$$

This is the critical transition point.

Case 4. The same conditions as Case 2 hold, but

$$\sum_{i=1}^{5}\frac{\alpha_i\beta_i}{1-h_i}>\frac{5}{\lambda_{\max}}.$$

The origin becomes unstable. Under the simple-mode pitchfork hypotheses, two stable nonzero equilibria exist sufficiently close to the threshold;

Case 5. All eigenvalues of $V^\top$ are negative, $W>0$, and at least two token parameters satisfy $h_i,h_j>1$. The origin is unstable. The one-parameter local pitchfork theorem does not establish the number or stability of nonzero equilibria in this regime.

G.5 Numerical Outputs

For each regime, the simulation plots the total norm

$$\|X(t)\|$$

as well as the per-token norms

$$\|X_k(t)\|,\qquad k=1,...,5.$$

The bifurcation diagram is drawn with h_1 on the horizontal axis and the fixed-point value, or fixed-point norm, on the vertical axis. Stable fixed points are shown in blue and unstable fixed points are shown in red.

G.6 Scenario-2 SHC-1 simulation

The Scenario 2 scripts `hc_simulation_d5.py` and `hc_simulation_d5_setting2.py` use $L=5, \alpha_i=\beta_i=1, W=0.2I_5$, and $(h_2,h_3,h_4,h_5)=(-1.6,-2.0,-2.4,-2.8)$. The symmetric value matrix has one, two, or four positive eigenvalues, with the simple maximum $\lambda_{\max}=2$. The threshold is $h_1^*=1-[5/2-\sum_{i=2}^{5}(1-h_i)^{-1}]^{-1}\simeq 0.184$.

G.7 SHC-2 ODE System

The SHC-2 simulation equations are implemented in `shc2_ode_simulation_final.py`. They use $L=5, d=2, V^\top=\lambda I_2, W=-I_2$ and the system (8)–(9). The feature dimension and the number of streams both equal two; these are different indices.

$$\tilde{X}^{(\ell)}=\alpha_1 X_1^{(\ell)}+\alpha_2 X_2^{(\ell)}, \tag{40}$$

and the shared attention output at position ℓ,

$$\mathcal{A}^{(\ell)}(X)=V^\top\sum_{k=1}^{L}\mathrm{softmax}_k\big(\tilde{X}^{(\ell)\top}W\tilde{X}^{(k)}\big)\tilde{X}^{(k)}, \tag{41}$$

where the softmax is taken over all $k=1,\dots,L$. The two sub-streams evolve as

$$\frac{dX_1^{(\ell)}}{dt}=(h_{11}-1)X_1^{(\ell)}+h_{12}X_2^{(\ell)}+\beta_1\mathcal{A}^{(\ell)}(X), \tag{42}$$

$$\frac{dX_2^{(\ell)}}{dt}=h_{21}X_1^{(\ell)}+(h_{22}-1)X_2^{(\ell)}+\beta_2\mathcal{A}^{(\ell)}(X). \tag{43}$$

Fixed parameters used in all experiments are listed in Table 2.

Parameter	Value
d	2
L	5
β_1,β_2	$1.0, 0.8$
α_1,α_2	$0.9, 1.1$
W	$-I_2$
h_{12}	1.0
h_{21}	-1.0

Table 2: Fixed parameters of the SHC-2 system.

The value matrix V and the connection matrix H vary by scenario.

With h_{12} and h_{21} fixed, we parametrise the diagonal of H by two scalar degrees of freedom:

$$H=\begin{pmatrix} h+\delta & h_{12} \\ h_{21} & h-\delta \end{pmatrix}, \tag{44}$$

setting $h_{11}=h+\delta+1$ and $h_{22}=h-\delta+1$. Here h controls $\mathrm{tr}(H)=2h$ while δ controls $\det(H)=h^2-\delta^2-h_{12}h_{21}$ independently of the trace.

Define the coupling scalar

$$C_0(H)=\beta_1\alpha_1(h_{22}-1)-\beta_2\alpha_1 h_{12}-\beta_1\alpha_2 h_{21}+\beta_2\alpha_2(h_{11}-1), \tag{45}$$

and the static bifurcation threshold

$$\mathcal{D}(\delta,h)=\max\big(0,-\lambda C_0(H)\big), \tag{46}$$

where $\lambda=\lambda_{\max}(V)$. The origin loses stability across a static boundary when

$$\det(H)=\mathcal{D}(\delta,h), \qquad \mathrm{tr}H<\min(0,-\lambda\alpha^\top b). \tag{47}$$

This is a static stability boundary, not by itself a pitchfork theorem. The oscillatory boundary and its linear frequency are

$$\mathrm{tr}H=\begin{cases} 0, & \lambda<0, \\ B:=-\lambda(\beta_1\alpha_1+\beta_2\alpha_2), & \lambda>0, \end{cases} \qquad \det H>\mathcal{D}(h,\delta), \tag{48}$$

$$\omega^*=\begin{cases} \sqrt{\det H}, & \lambda<0, \\ \sqrt{\det H+\lambda C_0(H)}, & \lambda>0. \end{cases} \tag{49}$$

For $\lambda<0$ the frequency comes from the contrast blocks H; it does not require $C_0=0$.

We study two scenarios distinguished by the sign of λ.

Scenario 1 ($\lambda=-1.5$). The static experiment fixes $h=-1$ and uses $\delta=1,\sqrt{2},1.8$. The oscillatory experiment fixes $\delta=0.5$ and uses $h=-0.30,0,0.15$. At $h=0$, $C_0=0.37$, $\mathcal{D}=0.555<\det H=0.75$, and $\omega^*=\sqrt{0.75}\simeq 0.866$. Each critical imaginary eigenvalue has multiplicity $(L-1)d=8$. This is a multidimensional contrast crossing, outside the simple periodic-branch theorem.

Scenario 2 ($\lambda=1.2$). The static experiment fixes $h=-2$. Here $\det(H+\lambda b\alpha^\top)=1.184-0.024\delta-\delta^2$, so the exact positive threshold is $\delta_c=-0.012+\sqrt{1.184144}$. The figure uses this exact root ($\delta_c\approx 1.07618$), with offsets $-0.35,0,+0.35$. At δ_c the equilibria form a continuum, so the middle panel marks no isolated equilibria. Equilibria marked as unstable in Figures 5 and 6 have a Jacobian eigenvalue with positive real part; their full saddle signatures were not checked. The oscillatory experiment fixes $\delta=0.4$ and varies $h=-1.568,-1.068,-0.768$. At $h_c=-1.068$, $\det H=1.980624$, $C_0=-1.52904$, and $\omega^*=\sqrt{0.145776}\simeq 0.382$. The critical space is uniform, of real dimension four because $V=1.2I_2$. Its flow is exactly linear at every amplitude within this space; there is no isolated nonlinear limit cycle selected there.

Initial conditions and observable. The plotted observable is the full Euclidean norm,

$$\|X(t)\|_2=\left(\sum_{\ell=1}^{L}\sum_{i=1}^{2}\|X_i^{(\ell)}(t)\|^2\right)^{1/2}, \tag{50}$$

not the mean of stream 1. Each panel uses eight random initial states uniformly sampled from $[-s,s]^{20}$: $s=0.1$ in the Scenario 1 oscillatory panels and $s=0.2$ in the other panels. The code uses seed 42 except for Scenario 2 oscillatory panels, which use $13+\lfloor|10h|\rfloor$. It does not initialize those states on the critical eigenspace. Integration uses RK45 with relative tolerance 10^{-8} and absolute tolerance 10^{-10}. The displayed frequency is the linear state-space frequency; a norm need not oscillate at that frequency. A numerical root search and finite-time trajectories do not establish a complete equilibrium classification or existence of periodic branches.

H Training Details

All experiments train a 96-layer GPT-style language model on FineWeb-10B [33]. Each run shares a common base configuration and differs only in the choice of token-wise hyper-connection coefficients and the spectral regime of W_V. Our experiments are conducted on a server with 2 H200 GPUs.

Model architecture and optimiser. The model uses 6 attention heads, embedding dimension 150, context length $L=1024$, no dropout, and no bias terms, giving approximately 30M parameters. We train with AdamW [21] ($\beta_1=0.9$, $\beta_2=0.95$, weight decay 0.1, gradient clip 1.0) at peak learning rate 6×10^{-4} with cosine decay to 6×10^{-5} over 20 000 iterations and a 2 000-iteration linear warm-up.

Token budget. Two batch configurations are used. Full-sequence fixed variants use batch size 64, context length 1 024, and gradient accumulation over 1 step, giving $\approx$1.31B tokens in total, assuming the accumulation count is the global count. Partial-constrained configurations set gradient accumulation to two, corresponding to $\approx 2.62B$ tokens under the same convention; the distributed launch configuration must be checked against each reported run.

H.1 1-stream Hyper-connection Models

Stream architecture. The 1-stream hyper-connection path is active with scalars $\alpha_\ell=\beta_\ell=1$ fixed across all token positions and layers. The token-wise hyper-connection coefficients $h=(h_1,\ldots,h_L)$ are initialized per case; in full-sequence fixed variants they are held non-learnable, while in partial-constrained variants only the first 10 positions are subject to a learnable stability projection. The supplied 1-stream configuration ties queries and keys, giving a positive-semidefinite score matrix; strict positive definiteness is not guaranteed.

Value matrix spectral constraint. In Setting 1 all eigenvalues of W_V are negative ($\lambda_{\max}=-1$), implemented by fixing $W_V=-I$. In Setting 2 exactly one eigenvalue of W_V is positive ($\lambda_{\max}=+1$), implemented by constraining W_V to a fixed eigenvalue spectrum with one positive eigenvalue while leaving the eigenvectors free; for partial-constrained variants this projection is enforced only over the first 10 token positions, matching the window over which h is projected.

Hyper-connection coefficients. With $\alpha_\ell = \beta_\ell = 1$, the stability sum reduces to $\Sigma = \sum_{\ell=1}^{n} (1-h_\ell)^{-1}$ and the threshold to $n/\lambda_{\max}$, where n is the number of constrained positions. Tables 3 and 4 list the h initialisation for each case.

Table 3: Hyper-connection coefficients for Setting 1 ($\lambda_{\max} = -1$).

Case	Regime	h initialisation	Stability condition
`s1_case1`	Convergence	$h_\ell = 0.9\ \forall \ell$	All $h_\ell < 1$
`s1_case2`	Convergence (one supercritical)	$h_1 > 1$ chosen s.t. $\Sigma < -n$; $h_{\ell \geq 2} = 0.9$	$\Sigma < n/\lambda_{\max}$
`s1_case3`	Critical	$h_1 > 1$ chosen s.t. $\Sigma = -n$; $h_{\ell \geq 2} = 0.9$	$\Sigma = n/\lambda_{\max}$
`s1_case4`	Divergence	$h_1 = 1.1$; $h_{\ell \geq 2} = 0.9$	$\Sigma > n/\lambda_{\max}$
`s1_case5`	Divergence (two supercritical)	$h_1 = 1.1$, $h_2 = 1.05$; $h_{\ell \geq 3} = 0.9$	$\#\{h_\ell > 1\} = 2$

Table 4: Hyper-connection coefficients for Setting 2 ($\lambda_{\max} = +1$).

Case	Regime	h initialisation	Stability condition
`s2_case1`	Convergence	$h_\ell = -0.01\ \forall \ell$	$\Sigma \approx 0.99n < n$
`s2_case2`	Critical*	h_1 chosen s.t. $\Sigma = n$; $h_{\ell \geq 2} = 0.0$	$\Sigma = n/\lambda_{\max}$
`s2_case3`	Divergence	$h_\ell = 0.9\ \forall \ell$	$\Sigma = 10n > n$
`s2_case4`	Divergence (one subcritical)	$h_1 = 0.5$; $h_{\ell \geq 2} = 1.1$	$\#\{h_\ell < 1\} = 1$

*With $\alpha_\ell = \beta_\ell = 1$, $\lambda_{\max} = 1$ and $h_{\ell \geq 2} = 0$, the condition $\Sigma = n$ forces $h_1 = 0$, so all constrained coefficients are equal. In the unmasked ODE surrogate this is the uniform critical-profile exception (Corollary D.22): the centre dynamics are exactly linear, not a generic subcritical pitchfork.

Constraint verification. At construction the per-token scalar $c_\ell = \alpha_\ell \beta_\ell / (1-h_\ell)$ is computed for each constrained position, and the sum $\Sigma = \sum_\ell c_\ell$ is checked against the threshold $n/\lambda_{\max}$. This arithmetic check does not establish a stability theorem for the causal, multi-head, finite-depth trained network.

H.2 2-stream Hyper-connection Models

All experiments train a 96-layer GPT-style language model on FineWeb-10B. Each run shares a common base configuration and differs only in the choice of H and the spectral regime of W_V.

Stream architecture. We activate the 2-stream SHC residual path with two residual streams and a single hyper-connection fraction. The mixing and output scaling scalars are fixed to $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0.5$. The hyper-connection matrix H is initialized per case, held non-learnable, and shared identically across all layers and token positions. The supplied 2-stream base configuration sets `positive_symmetric_qk=False`; it does not impose the positive-definiteness hypothesis used for pitchfork sign conclusions.

Value matrix spectral constraint. The stability conditions depend on $\lambda_{\max}(V^\top)$, which is fixed deterministically at construction. In Setting 1 we pin $W_V = -I$, giving $\lambda_{\max} = -1$. In Setting 2 we constrain W_V to a fixed eigenvalue spectrum with exactly one positive eigenvalue, giving $\lambda_{\max} = +1$.

Hyper-connection matrices. Tables 5 and 6 list the matrix H assigned to each case in Settings 1 and 2, together with the resulting $\mathrm{tr}(H)$ and $\det(H)$.

Consistency limitations of the training comparison. †Under the ODE convention of Section 5, the listed Setting 2 Case 2 has $H = -I_2$, $\alpha = b = (1/2, 1/2)^\top$ and $\lambda_{\max} = 1$. Thus $C_0 = -1/2$, $D_2 = 1/2$, and $\det H = 1 > D_2$: the origin is exponentially stable in this surrogate, not on a static boundary. Its leading uniform block has eigenvalues $-1, -1/2$. The historical "critical" label in the perplexity table must therefore be revalidated against the actual run.

There is also a convention mismatch in the supplied implementation: `SHC2StreamAttention.forward` multiplies the streams by the configured matrix directly, whereas H in Section 5 denotes the shifted continuous-depth generator. If the configured matrix is a discrete skip matrix, the corresponding near-identity generator is that matrix minus I_2. The configuration validator used for the reported runs computed C_0 incorrectly and did not implement

Table 5: Hyper-connection matrices for Setting 1 ($\lambda_{\max} = -1$).

Case	Regime	H	$\mathrm{tr}(H)$	$\det(H)$
`s1_case1`	Asymptotically stable	$\begin{pmatrix} -0.6 & 0.1 \\ 0.1 & -0.5 \end{pmatrix}$	-1.1	0.29
`s1_case2`	Static boundary	$\begin{pmatrix} -0.5 & 0.0 \\ 0.0 & 0.0 \end{pmatrix}$	-0.5	0.00
`s1_case3`	Oscillatory boundary	$\begin{pmatrix} 0.0 & 0.5 \\ -0.5 & 0.0 \end{pmatrix}$	0.0	0.25
`s1_case4`	Unstable	$\begin{pmatrix} 0.2 & 0.0 \\ 0.0 & 0.1 \end{pmatrix}$	0.3	0.02

Table 6: Hyper-connection matrices for Setting 2 ($\lambda_{\max} = +1$).

Case	Regime	H	$\mathrm{tr}(H)$	$\det(H)$
`s2_case1`	Asymptotically stable	$\begin{pmatrix} -1.1 & 0.0 \\ 0.0 & -1.1 \end{pmatrix}$	-2.2	1.21
`s2_case2`	Stable ODE surrogate†	$\begin{pmatrix} -1.0 & 0.0 \\ 0.0 & -1.0 \end{pmatrix}$	-2.0	1.00
`s2_case3`	Oscillatory boundary	$\begin{pmatrix} -0.25 & -0.8 \\ 0.8 & -0.25 \end{pmatrix}$	-0.5	0.70
`s2_case4`	Unstable	$\begin{pmatrix} 0.1 & 0.8 \\ 0.8 & 0.1 \end{pmatrix}$	0.2	-0.63

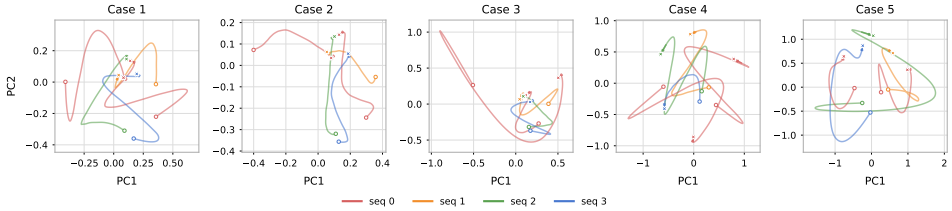


Figure 8: **Pretrained SHC-1 token trajectories, Scenario 1 ($\lambda_{\max} < 0$), PCA projection.** Open and filled circles denote the initial and final layer positions, respectively. Cases 1–2, which are before the threshold, exhibit collapse. Case 3, at the threshold, shows slower collapse. Cases 4–5, after the threshold, diverge.

the full endpoint conditions of Corollary 5.3. The released code replaces it with a validator that uses $C_0 = \alpha^{\top} \mathrm{adj}(G) b$, checks every block G and $G + \lambda_k b \alpha^{\top}$ for $\lambda_k \in \{\lambda_{\min}, \lambda_{\max}\}$, distinguishes stable, unstable and non-hyperbolic boundary cases, and makes the generator/skip-matrix convention an explicit option. Under this validator, the configurations in Tables 5 and 6 have the listed regimes except `s2_case2`, which is rejected as stable. The tables are retained as records of the configurations used for the reported runs; their regime labels do not certify the trained networks' dynamics.

The implemented attention is causal, multi-head, and followed by an output projection; the network also includes normalization and feed-forward layers. Its local map is not the unmasked attention-only ODE studied here. Constraining the first ten positions does not make the full-context system satisfy the theorem with $L = 10$. The empirical comparison is therefore qualitative, and an exact correspondence requires a separate analysis of the implemented model and verification of each run.

Remark H.1. The hyper-connection coefficients in Tables 3 and 4 are chosen to place each run in a precisely controlled stability regime, not to optimise perplexity. This paper is theory-oriented: our theoretical claims concern local ODE stability and bifurcations, while the experiments explore associations between the selected configurations and perplexity rather than propose a practical training recipe. A systematic search over h to minimise perplexity is left for future work.

Table 7: Perplexity results of 1-stream GPT-2 for different configurations.

Config	Property	# Params	PPL
`baseline`	-	30M	32.16
`s1-case1`	Converge	30M	38.09
`s1-case2`	Converge	30M	38.03
`s1-case3`	Critical	30M	34.46
`s1-case4`	Diverge	30M	33.78
`s1-case5`	Diverge	30M	33.62
`s2-case1`	Converge	30M	33.15
`s2-case2`	Critical	30M	33.07
`s2-case3`	Diverge	30M	31.59
`s2-case4`	Diverge	30M	**31.56**

Table 8: Perplexity results of 2-stream GPT-2 for different configurations.

Config	Property	# Params	PPL
`baseline`	-	30M	29.27
`s1-case1`	Converge	30M	32.79
`s1-case2`	Critical	30M	32.46
`s1-case3`	Diverge‡	30M	32.14
`s1-case4`	Diverge	30M	32.16
`s2-case1`	Converge	30M	30.47
`s2-case2`	Critical†	30M	27.94
`s2-case3`	Diverge‡	30M	27.11
`s2-case4`	Diverge	30M	**26.76**

Labels are the historical run labels. †Under the ODE convention of Section 5 this configuration is exponentially stable, not critical (Appendix H). ‡Tables 5 and 6 list this configuration as an oscillatory boundary. All labels require run-level revalidation.

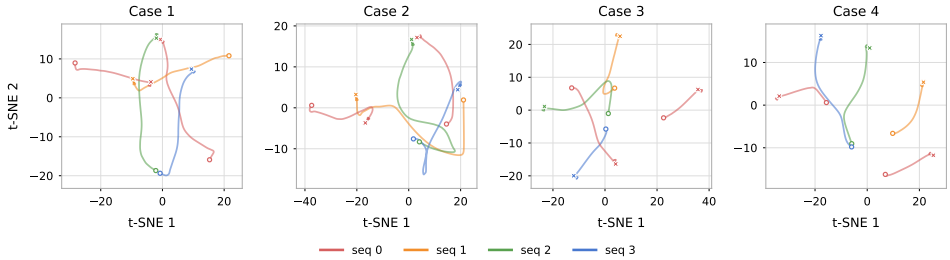


Figure 9: **Pretrained SHC-1 token trajectories, Scenario 2, t-SNE projection.** The trajectories show the qualitative transition predicted by the Scenario 2 dynamics.

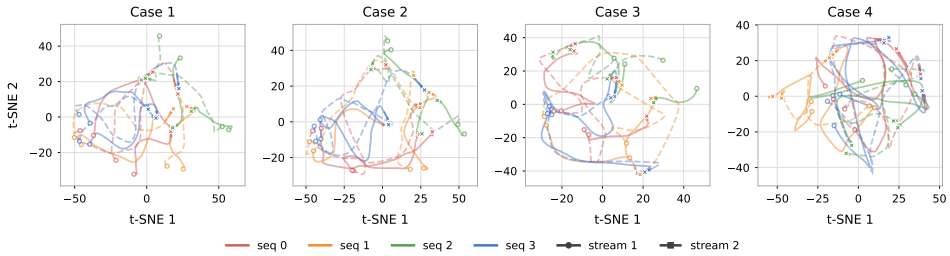


Figure 10: **Pretrained SHC-2 token trajectories, Setting 1, t-SNE projection.** The figure visualizes the evolution of token representations across attention layers in a GPT-2 model with 2-stream hyper-connections.

I Additional Experiments

Model and Dataset. To empirically validate our theoretical findings, we follow the experimental paradigm of [34]. Specifically, we pretrain a variant of nanoGPT-2 with 96 layers on the FineWeb10B dataset [33]. Our objectives are two-fold: (1) to demonstrate that convergence behavior can adversely affect the performance of Hyper-Connection Transformers and therefore should be mitigated; and (2) to examine whether configurations labelled as divergent are associated with better perplexity. These comparisons are empirical associations; they do not show that ODE instability causes better performance (Appendix H). Throughout the experiments, we compare our modified Transformer variants against baseline Hyper-Connection Transformers with the same configuration. Detailed information about the model architecture, hyperparameters, and training procedures has been provided in Appendix H.

Configurations. Tables 7 and 8 report the perplexity results for the 1-stream and 2-stream GPT-2 variants, respectively. For the 1-stream setting, we evaluate ten configurations in total: one baseline,

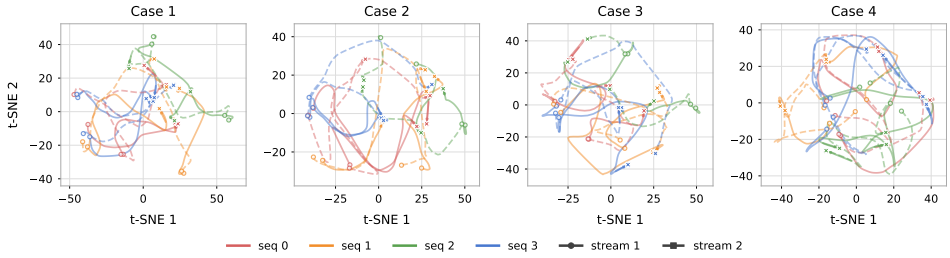


Figure 11: **Pretrained SHC-2 token trajectories, Setting 2, t-SNE projection.** The trajectories illustrate the qualitative behavior of token representations under the second 2-stream hyper-connection setting.

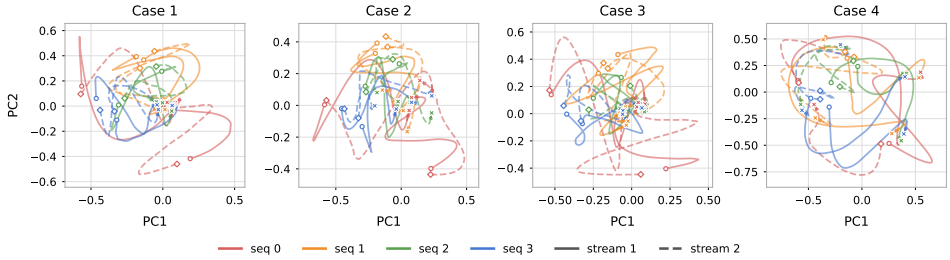


Figure 12: **Shared-weight SHC-2 token trajectories, Setting 1.** Cases before the transition collapse toward the origin, while cases after the transition separate into nonzero symmetric clusters.

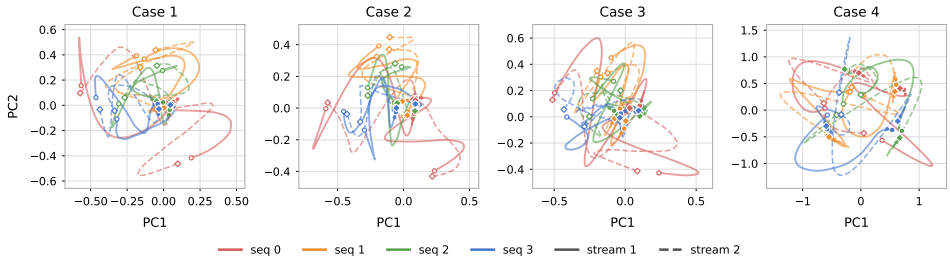


Figure 13: **Shared-weight SHC-2 token trajectories, Setting 2.** The first case collapses toward the origin, the second case shows transitional behavior, and later cases diverge.

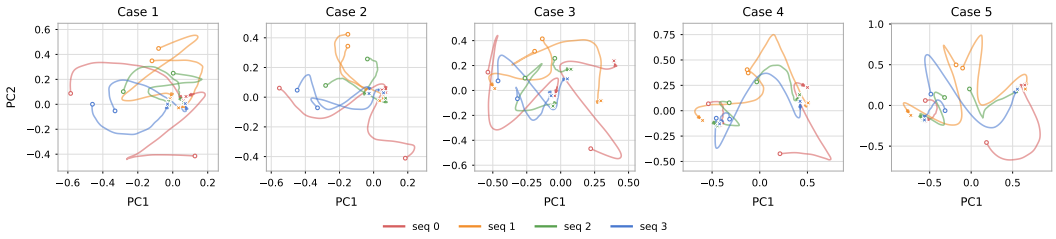


Figure 14: **Shared-weight SHC-1 token trajectories, Scenario 1.** Cases before the threshold collapse toward the origin; the threshold case exhibits slower collapse; cases after the threshold diverge.

five cases under Scenario 1, and four cases under Scenario 2. For the 2-stream setting, we evaluate nine configurations in total: one baseline, four cases under Scenario 1, and four cases under Scenario 2. Here, `s1` and `s2` denote Scenario 1 and Scenario 2, while `case1`–`case5` denote the corresponding cases within each scenario.

J Hardware Settings

We conducted all experiments on 2 NVIDIA H200 141GB GPUs.

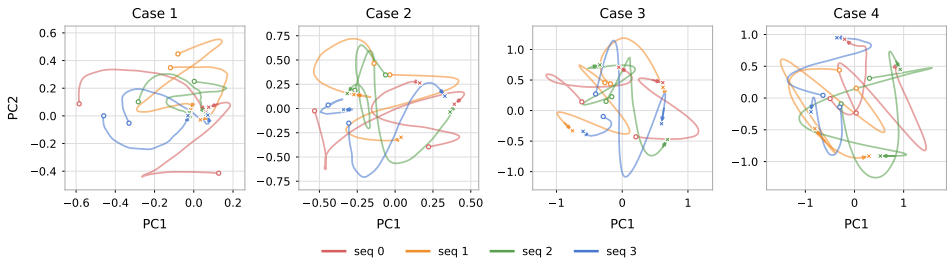


Figure 15: **Shared-weight SHC-1 token trajectories, Scenario 2.** The trajectories show the transition from collapse to instability as the control parameter crosses the critical regime.

K Broader Societal Impacts

This work is primarily theoretical: we study token dynamics in self-attention with static hyper-connections and show how architectural coupling governs representation collapse through stability and bifurcation mechanisms. Its societal impact is therefore indirect, through improving our understanding of how Transformer design choices affect stability, representation quality, and downstream performance. A more principled understanding of collapse may help researchers design more reliable and efficient models, reduce trial-and-error experimentation, and better diagnose failure modes such as collapse, divergence, or sensitivity to architectural parameters. At the same time, stronger Transformer architectures may also amplify broader risks associated with capable language models, including misinformation, misuse, and deployment in high-stakes settings. Our work does not directly address alignment, fairness, privacy, robustness, or factual reliability, and it focuses on static, attention-only, continuous-depth dynamics with shared weights; thus, the results should be viewed as a theoretical foundation rather than a safety guarantee for deployed systems.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately reflect the scope and contributions of the paper. In particular, Contributions 1, 2, and 3 in Section 1.2 are substantiated in Sections 2, 4, and 5, respectively.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Concluding Remarks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided the full set of assumptions and complete proofs of the theoretical results in Section 3 and the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided all experimental setups in Appendices G, H, and I. Besides, we also provide the code to reproduce the results in the paper, which can be found in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have included the code in the supplemental material, together with detailed instructions for reproducing the experimental results. The datasets used in our experiments consist either of synthetic data generated by our code or publicly available benchmark datasets explicitly specified in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the training and test details, which can be found in Appendix G, H, and I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our work is primarily theoretical, focusing on the dynamics of token representations in deep sequence models. Accordingly, the experiments are designed to visualize token trajectories under different settings and verify their convergence or divergence behavior. Since the objective is qualitative dynamical verification rather than statistical performance evaluation, error bars are neither necessary nor applicable in this context.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided information about computing resources needed in Appendix J.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts of the work performed in Appendix K.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper has credited all code and data that have been used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The source code of the paper is based on existing source (which has been credited properly). We also provide a README file to run the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only as a writing assistant for editing and polishing the manuscript (grammar, clarity, concision).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.