First-Token Attraction in Mamba Dynamics

Trinh T. Nguyen[1*] **Minh-Khoi Nguyen-Nhat**[2*] **Duy-Tung Pham**[3]
Hoang-Son Do[4,5] **Tan Minh Nguyen**[2,†] **Thieu N. Vo**[6,†]

[1] School of Mathematical Sciences, Shanghai Jiao Tong University, China
[2] Department of Mathematics, National University of Singapore
[3] Khoury College of Computer Sciences, Northeastern University, USA
[4] Institute of Mathematics, Vietnam Academy of Science and Technology
[5] University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology
[6] Department of Computer Science, University of Bath, United Kingdom

[1]q0004349@sjtu.edu.cn [2]khoi.nguyen@u.nus.edu [3]pham.duy@northeastern.edu
[4]dhson@math.ac.vn [5]do-hoang.son@usth.edu.vn
[2]tanmn@nus.edu.sg [6]ntv22@bath.ac.uk

Abstract

We study token dynamics in deep Mamba models, with a particular focus on the attracting role of the first token. Unlike attention mechanisms, where token interactions are governed by positive scalar coefficients, Mamba dynamics involve full matrix-valued interaction coefficients with no positivity or definiteness guarantees. This structural difference makes Mamba dynamics substantially harder to analyze and prevents the direct application of existing techniques for attention-based models. To overcome this challenge, we introduce a new analytical framework based on a high-dimensional invariant box and a sharp exit-time argument, which together recover an effective form of sign-definiteness. Using this framework, we characterize the possible limit points of token trajectories and prove explicit exponential convergence rates. Our analysis reveals that the first token has privileged dynamics: its limiting direction attracts subsequent tokens, inducing token clustering and attention-sink phenomena in deep Mamba models. We further show that this attraction mechanism extends to broader sequence dynamics, including state-space models, causal attention, and classical time-series models. Experiments on Falcon-Mamba-7B validate our predictions, showing that concentration on the first token increases with depth and dominates in later layers.

1 Introduction

Modern sequence modeling has been dominated by transformers, while state space models (SSMs) have emerged as an important alternative. Transformers have set the standard for representation learning, yet their quadratic complexity with respect to sequence length poses challenges for efficiency and scalability [77]. In parallel, SSMs have reemerged as powerful alternatives that combine principles of recurrence and convolution with ideas from classical control theory [34]. Recent innovations in parameterization and implementation have made SSMs substantially more efficient [25, 26, 24, 29, 45, 36, 63, 72], enabling them to capture long-range dependencies effectively [23, 58] and to excel in benchmark tasks such as the Long Range Arena [74]. Beyond efficiency, these models have demonstrated strong empirical performance in domains as varied as audio, speech, and vision [21, 54, 70].

[*]Co-first authors
[†]Co-last authors. Correspondence to: q0004349@sjtu.edu.cn & ntv22@bath.ac.uk.

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

Within this line of development, the Mamba architecture has emerged as a selective SSM with striking empirical results. Its distinctive component, the selective state-space (S6) layer, introduces input-dependent parameters that grant the model content awareness [22]. This design has proven effective across a wide spectrum of applications, including language modeling [67, 41], vision and video analysis [48, 89, 85, 42], medical imaging [50, 80], tabular data [5], reinforcement learning [64], 3D perception [46], graph learning [79], and even N-dimensional sequence modeling [44]. The breadth of these successes has been linked to Mamba's ability to generate highly adaptive and expressive representations across modalities [68, 65, 84, 89].

The study of token dynamics–the evolution of token representations within a model–has proven to be a powerful framework for understanding transformers, offering insights into equilibrium structures, stability properties, and clustering phenomena. In contrast, the token dynamics of selective SSMs have received little attention, despite the strong empirical performance of architectures such as Mamba.

1.1 Our Contributions

In this paper, we investigate the dynamical properties of tokens in deep pretrained Mamba models. Prior theoretical studies often rely on restrictive conditions such as the absence of normalization and one-dimensional token embeddings (see [78]). We consider the continuous-time limit of pretrained Mamba models without these restrictive assumptions and systematically analyze the asymptotic behavior of token trajectories over time. Our main contributions are summarized as follows:

- We construct a new theoretical framework based on a high-dimensional invariant box for the Mamba dynamics and establish a sharp exit-time argument to determine when token trajectories remain confined to this domain. This framework overcomes the difficulties arising in previous approaches for attention settings.
- We characterize the set of limit points of token trajectories in Mamba models and establish explicit exponential convergence rates (Theorems 3.5, 3.3, 3.6 and 3.9. Our theoretical analysis further demonstrates that the first token exerts an attracting influence on the remaining tokens, a phenomenon that we empirically validate.
- We demonstrate a practical implication of our theoretical results to offer a principled explanation of the attention sink effect in Mamba models via the attraction of the first token.

To validate our theoretical findings and their applications, we conduct extensive experiments by visualizing attention concentration in Falcon-Mamba-7B across different depths. These experimental results confirm the predictive accuracy of our analysis and underscore its practical benefits in real-world scenarios.

Mamba dynamics is far more complicated than causal attention dynamics. Indeed, while our work is inspired by the recent analysis of Transformer dynamics in [35], the mathematical structure of Mamba is much more complex. In the attention setting, the effective interaction coefficients are positive scalars as they are normalized via softmax. The positivity of these interaction coefficients ensures sign-definiteness and enables the use of powerful tools from monotone and cooperative dynamical systems. In contrast, for Mamba, the effective interaction coefficients are full matrices with no positivity or definiteness guarantees. This lack of sign-definiteness is not a minor technical inconvenience; it is an intrinsic structural feature of practical Mamba architectures and invalidates essentially all positivity-based techniques that underpin the current theory of attention dynamics. As a consequence, techniques used for analyzing causal attention dynamics, including hemisphere-based Lyapunov methods [3] and Shub-type genericity arguments [35], completely fail in this architecture.

Our new analytical framework. To overcome the above challenge, we develop a new analytical framework based on a high-dimensional invariant box for the dynamical system governing the continuous-time limit of Mamba, together with a sharp exit-time argument that recovers effective sign-definiteness and determines when token trajectories remain confined to this domain. Within this framework, we characterize all possible token limit points and derive explicit exponential convergence rates toward them. This analysis yields a theoretical explanation for a widely observed empirical phenomenon: the limit point of the first token exerts an attracting influence on subsequent tokens, thereby inducing clustering and attention-sink phenomena. Finally, we extend our analysis of this

attracting effect to a broad class of autoregressive sequence models, including state-space models, causal Transformers, and classical time-series models.

Organization. Section 2 derives a dynamical system over the unit sphere arising from the continuous-time limit of Mamba. Section 3 analyzes the limitations of the first token and proves its attraction effect on subsequent tokens. Section 4.1 validates the theory in practical Mamba settings, while Section 5 applies the results to predict attention-sink phenomena in Mamba models.

2 Mamba Dynamics as a Flow Map on Sphere

2.1 Background – S6 layer

The building block of a selective state-space (Mamba) model is a Mamba block (see [13]). At the heart of the Mamba block is the S6 layer, which plays a key role in the success of deep Mamba models.

S6 **layer.** Formally, an S6 layer is defined as a function that maps a sequence of tokens $\mathbf{x} = (x_1, \dots, x_L) \in \mathbb{R}^{D\times L}$ to another sequence of tokens $\mathbf{y} = (y_1, \dots, y_L) \in \mathbb{R}^{D\times L}$ (with the same number of channels D). In this context, each vector $x_l \in \mathbb{R}^D$ (or $y_l \in \mathbb{R}^D$) represents a token/word, while the whole sequence $\mathbf{x} = (x_1, \dots, x_L)$ (or $\mathbf{y} = (y_1, \dots, y_L)$) is called a prompt/sentence.

For each $d = 1, \dots, D$, the d-th channel output $\hat{y}_d = (y_{d1}, \dots, y_{dL}) \in \mathbb{R}^L$ is determined recurrently from the corresponding d-th channel input $\hat{x}_d = (x_{d1}, \dots, x_{dL}) \in \mathbb{R}^L$ via a sequence of hidden states $h_{d1}, \dots, h_{dL} \in \mathbb{R}^N$ as follows:

$$\begin{cases} h_{dl} = \overline{A}_{dl} \cdot h_{d,l-1} + \overline{B}_{dl} \cdot x_{dl}, \quad h_{d0} = 0, \\ y_{dl} = C_l \cdot h_{dl}, \end{cases} \tag{1}$$

for $l = 1, \dots, L$. Here, the matrices $\overline{A}_{dl}$, $\overline{B}_{dl}$, and C_l are input-dependent and time-varying, determined by

$$\overline{A}_{dl} = e^{-a_d \Delta_d(x_l)}, \quad \overline{B}_{dl} = \Delta_d(x_l) S_B \cdot x_l, \quad \text{and} \quad C_l = (S_C \cdot x_l)^\top, \tag{2}$$

where $\Delta_d : \mathbb{R}^D \to \mathbb{R}$ is the step size function at the d-th channel and is defined for each $u \in \mathbb{R}^D$ by

$$\Delta_d(u) = \text{softplus}(S_{\Delta,d} \cdot u) = \ln\left(1 + e^{S_{\Delta,d}\cdot u}\right). \tag{3}$$

The hidden positive coefficients $a_d \in \mathbb{R}_{>0}$, and the step size vectors $S_{\Delta,d} \in \mathbb{R}^D$ for $d = 1, \dots, D$, as well as the input and output matrices $S_B, S_C \in \mathbb{R}^{N\times D}$, are learnable from input data.

By eliminating the hidden states h_{dl} in system (1) and recall that $y_l = (y_{1l}, \dots, y_{DL})$, we obtain the following form (see also [7]):

$$y_l = \sum_{j=1}^{l} A_{lj}(\mathbf{x}) x_j, \tag{4}$$

where $A_{lj} = \text{diag}(A_{1,lj}, \dots, A_{D,lj})$ and $S_{BC} = S_C^\top S_B$ as well as

$$A_{d,lj}(\mathbf{x}) = \begin{cases} x_l^\top S_{BC} x_l \cdot \Delta_d(x_l), & \text{if } l = j, \\ x_l^\top S_{BC} x_j \cdot \Delta_d(x_j) \cdot \exp\Big(-a_d \sum_{k=j+1}^{l} \Delta_d(x_k)\Big), & \text{if } l > j. \end{cases} \tag{5}$$

In our context, the matrix $S_{BC} = S_C^\top S_B \in \mathbb{R}^{D\times D}$, which we will refer to as the *input-output matrix*, plays an essential role in characterizing the dynamical properties of tokens.

2.2 Mamba Dynamic as Flow Map on Sphere

All practical implementations of Mamba models incorporate normalization [8], most commonly realized as root mean square (RMS) normalization [87]. In this approach, the sequence of tokens at each layer is normalized by dividing each token by its Euclidean norm. This operation enhances

stability by preventing divergence, thereby reducing rounding errors and avoiding numerical overflow. Consequently, the resulting dynamics can be viewed as evolving on the unit sphere $\mathbb{S}^{D-1}$ throughout.

Follow the framework of [35], a Mamba model including S6 layer and RMS normalization with skip connection can then be viewed as a flow map on $(\mathbb{S}^{D-1})^L$ of the dynamical system

$$\frac{d}{dt}x_l(t) = \mathbf{P}_{x_l(t)}\Big(\sum_{j=1}^{l} A_{lj}(\mathbf{x}(t),t)x_j(t)\Big), \quad l = 1,\dots,L, \tag{6}$$

$$\mathbf{x}(0) = (x_1(0),\dots,x_L(0)) \in (\mathbb{S}^{D-1})^L, \tag{7}$$

where

- for each $x \in \mathbb{S}^{D-1}$, the map $\mathbf{P}_x \colon \mathbb{R}^D \to T_x\mathbb{S}^{D-1}$ is the projection onto the tangent space at x of the sphere $\mathbb{S}^{D-1}$ and it is defined by $\mathbf{P}_x(y) = y - \langle y, x\rangle\, x$ for each $y \in \mathbb{R}^D$;
- the coefficient A_{lj} is an $D \times D$ diagonal matrix $A_{lj} = \mathrm{diag}(A_{1,lj},\dots,A_{D,lj})$ with

$$A_{d,lj}(\mathbf{x}(t),t) = \begin{cases} x_l(t)^\top S_{BC}(t)x_l(t)\cdot \Delta_d(x_l(t),t), & \text{if } l = j,\\ x_l(t)^\top S_{BC}(t)x_j\cdot \Delta_d(x_j(t),t)\cdot & \\ \quad \cdot \exp\Big(-a_d(t)\sum_{k=j+1}^{l}\Delta_d(x_k(t),t)\Big), & \text{if } l > j; \end{cases} \tag{8}$$

- the step size function at the d-th channel $\Delta_d : \mathbb{R}^D \times [0,+\infty) \to \mathbb{R}$ is defined by $\Delta_d(u,t) = \mathrm{softplus}(S_{\Delta,d}(t)u)$, for each $(u,t) \in \mathbb{R}^D \times [0,+\infty)$.

2.3 Mamba Interaction Coefficients Are More Complex Than Those in Causal Attention

In our continuous-time view for both Mamba and attention define dynamics, each $\frac{dx_l}{dt}$ is a linear combination of earlier tokens with interaction coefficients. For attention, the interaction coefficients are positive scalars as they are normalized by softmax. These positive scalar interaction coefficients makes the system comparatively structured and amenable to existing analysis (see [35]).

In contrast, the interaction coefficients of Mamba are far more complicated than those in the causal attention setting. Indeed, the Mamba interaction coefficients are $A_{lj}(\mathbf{x}(t),t)$ which are given in equation (8) are full matrices with no positivity or definiteness guarantees. As a consequence, almost all tools and conclusions from the attention setting, including hemisphere-based Lyapunov methods [3] and Shub-type genericity arguments [35], break down.

3 Theoretical Results for Mamba Dynamic

Our goal is to study the dynamical properties of tokens in the internal representations of deep Mamba models. In the continuous-time limit, this amounts to analyzing the solution trajectories of the system (6) and characterizing the limiting set $\lim_{t\to+\infty}\{x_1(t),\dots,x_L(t)\}$. In this section, we focus on a simplified Mamba setting in which the model parameters are assumed to be time-independent. An extension to the time-dependent parameter regime is discussed in Section 3.3. Intuitively, we will prove the following:

Observation 3.1. (i) The first token $x_1(t)$ has privileged dynamics: its trajectory determines a discrete set of possible limit points, primarily the coordinate unit vectors $\pm e_d$, which serve as attractors (Theorems 3.3 and 3.5). Here, e_d is the coordinate unit vector with 1 at the d-th coordinate and 0 elsewhere.

(ii) Subsequent tokens $x_l(t)$ with $2 \le l \le L$ are strongly influenced by the first token, often exponentially converging toward the same attractor, thereby creating clusters of tokens in the final layers (see Theorem 3.6 and Proposition 3.7 for the time-indepedent parameter setting and Theorem 3.9 for a time-dependent analogue).

The study of the convergence of all tokens based on the dynamical system (6) without any additional assumption on the model parameters is challenging. To make the proofs feasible while still maintaining the generality practicality of our theoretical findings, we will first assume that all parameters

$S_{BC}, S_{\Delta,d}$ and a_d are time-independent in Sections 3.1 and 3.2. Next, a generalization into the case of time-dependent parameters will be presented in Section 3.3. All proofs can be found in Appendix.

These theoretical observations will be empirically validate using practical Mamba architectures, considering both the continuous-time setting (in Subsection 4.1) and the discrete-time setting (in Subsection 4.2).

3.1 Possible Limit Points of the First Token

Let us consider the differential equation governing the trajectory of the first token, which is the first equation in the dynamical system (6), given by

$$\begin{aligned} &\frac{d}{dt}x_1(t) = \mathbf{P}_{x_1(t)}\big(A_{11}(\mathbf{x}(t))x_1(t)\big), \\ &x_1(0) \in \mathbb{S}^{D-1}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} A_{11} &= \operatorname{diag}(A_{1,11}, A_{2,11}, \cdots, A_{D,11}), \text{ and} \\ A_{d,11}(\mathbf{x}(t)) &= x_1(t)^\top S_{BC} x_1(t) \cdot \Delta_d(x_1(t)). \end{aligned}$$

A limit point must be an equilibrium point. We will determine the set of equilibrium points of differential equation (9) in the following proposition.

Proposition 3.2 (All possible equilibrium points of the first token)**.** *A vector $y \in \mathbb{S}^{D-1}$ is an equilibrium of equation (9) if and only if it belongs to one of the following types:*

> ***Type 1.*** *$y \in \bigcup_{A\subset\{1,2,\ldots,D\}} \mathcal{H}_{\Delta,A}$, where for each subset $A \subset \{1,\ldots,D\}$, we denote by $\mathcal{H}_{\Delta,A} := \langle \{\, S_{\Delta,j} - S_{\Delta,\min A},\, j \in A \,\} \cup \{e_k,\, k \in A\}\rangle^{\perp}$.*
>
> ***Type 2.*** *$y \in \mathcal{H}_{S_{BC}} := \{u \in \mathbb{R}^D \mid u^\top S_{BC}\, u = 0\}$, which is the null space of the quadratic surface defined by S_{BC}.*

If S_{BC} is definite, then $\mathcal{H}_{S_{BC}} = \emptyset$ and Type 2 equilibria do not exist. If, furthermore, the columns of S_Δ are linearly independent, then the Type 1 equilibria collapse to exactly two antipodal points $y = \pm\xi$, where ξ is the unique unit vector in the union $\bigcup_{A\subset\{1,2,\ldots,D\}} \mathcal{H}_{\Delta,A}$.

Not every equilibrium point is stable, and therefore not every equilibrium point can arise as a limit point. Even for the first token, analyzing stability near equilibria in order to characterize the limit points of equation (9) is delicate and technically involved. This difficulty already appears in the two-dimensional case $D = 2$: depending on the initial data, the first token may converge to different limiting states, as we see in the following theorem.

Theorem 3.3 (All possible limit points of the first token in two dimensions)**.** *Assume that $D = 2$ and let $y(t) = (y_1(t), y_2(t)) \in C^\infty([0,+\infty), \mathbb{S}^1)$ be a solution of differential equation (9). Assume further that*

(i) $S_{BC} = \operatorname{diag}(\lambda_1, \lambda_2)$ *with* $\lambda_1 \neq \lambda_2$ *or* $\lambda_1 = \lambda_2 \neq 0$*;*

(ii) $S_{\Delta,2} \neq S_{\Delta,1}$.

In this case, the circle $\mathbb{S}^1$ intersects the line $\langle S_{\Delta,2} - S_{\Delta,1}\rangle^{\perp}$ at two antipodal points, say $\pm\xi$. Then $\lim_{t\to+\infty} y(t)$ exists and belongs to the set

$$\begin{aligned} \Omega = &\{(0,\pm1),(\pm1,0)\} \cup \{\pm\xi\} \\ &\cup \left\{(\cos\varphi_\star, \sin\varphi_\star) : \varphi_\star \in [0,2\pi),\ \lambda_1\cos^2\varphi_\star + \lambda_2 \sin^2\varphi_\star = 0\right\}. \end{aligned}$$

Remark 3.4 (Possible limit points of the first token in practical setting)**.** Theorem 3.5 indicates that the first token converges to either of the coordinate unit vectors $\pm e_m$. In contrast, Theorem 3.3 reveals additional possible limit points, which are difficult to detect in simulations with randomly chosen model parameters. In practice, even when the assumptions of Theorems 3.5 and 3.3 do not strictly hold, we still observe that the first token typically converges to one of the coordinate unit vectors $\pm e_m$ (see Figure 1 below).

We leave the proof of the above proposition, as well as the study of the stability of these equilibrium points, to Appendix B.1.1. The following theorem proves that the first token $x_1(t)$ likely converges to a limit point of the form $\pm e_d$ for some $d = 1, \ldots, D$.

Theorem 3.5 (First token likely converges to $\pm e_d$ in general)**.** *Let $y(t) = (y_1(t), \ldots, y_D(t)) \in C^\infty([0, +\infty), \mathbb{S}^{D-1})$ be a solution of differential equation (9). Assume that the symmetric part of S_{BC} is positive definite, and $S_{\Delta,m} = \text{sign}(y_m(0)) \cdot e_m$ for each m such that $y_m(0) \neq 0$. Suppose that there exists $d_0 \in \{1, ..., D\}$ such that*

$$|y_{d_0}(0)| > |y_d(0)| > 0,$$

for every $d \neq d_0$. Then

$$\lim_{t \to \infty} y(t) = \pm e_{d_0},$$

where the sign is $\text{sign}(y_{d_0}(0))$.

3.2 All Other Tokens are Attracted by the First Token

Different from the first token, the problem of characterizing the possible limits points of the other token is far more challenge. In this subsection, to make the proofs feasible while still maintaining the generality and practicability of our findings, beside of the time-independent parameters assumption, we will assume further that:

$$S_{\Delta,1}(t) = \ldots = S_{\Delta,D}(t) =: \alpha \in \mathbb{R}^D, \text{ and}$$
$$a_1(t) = \ldots = a_D(t) =: a \in \mathbb{R},$$

for any time $t \in [0, +\infty)$. Under this assumption, we can rewrite dynamical system (6) as:

$$\frac{d}{dt} x_l(t) = \sum_{j=1}^{l-1} A_{lj}(\mathbf{x}(t)) \cdot \mathbf{P}_{x_l(t)}(x_j(t)), \quad l = 1, \ldots, L, \tag{10}$$

where the coefficients $A_{lj}(\mathbf{x}(t))$ are defined as

$$A_{lj}(\mathbf{x}(t)) = x_l(t)^\top S_{BC} x_j(t) \cdot \Delta(x_l(t)) e^{-a \sum_{k=j+1}^{l} \Delta(x_k(t))}, \tag{11}$$

with $\Delta(x_l(t)) = \ln(1 + e^{\alpha \cdot x_l(t)})$.

In this case, for the first token $x_1(t)$, we see that

$$\frac{d}{dt} x_1(t) = A_{11}(\mathbf{x}(t), t) \cdot \mathbf{P}_{x_1(t)}(x_1(t)) = 0.$$

This implies that $x_1(t) = x_1(0)$ is a constant function on $[0, +\infty)$. We will present all possible limit points of $x_i(t)$ for $i = 2, \ldots, L$ in Theorem 3.6 and Proposition 3.7 below.

Theorem 3.6 (All tokens are exponentially attracted by the first token)**.** *Let $\mathbf{x}(t) = (x_1(t), \ldots, x_L(t)) \in C^\infty([0, +\infty), (\mathbb{S}^{D-1})^L)$ be a solution of dynamical system (10). Assume that S_{BC} is symmetric and $\langle x_1(0), S_{BC} x_1(0) \rangle > 0$. Then, there exists a zero-measurable subset Ω of $(\mathbb{S}^{D-1})^{L-1}$ such that: if the initial point $(x_2(0), ..., x_L(0))$ does not belong in Ω then all tokens converge exponentially to $\pm x_1(0)$, i.e.,*

$$\sum_{i=1}^{L} |x_i(t) - s_i x_1(0)| \lesssim e^{-C_0 t},$$

for some constant $C_0 > 0$ and $s_i \in \{+1, -1\}$ for each i.

Proof strategy. The proof is organized into four steps.

Step 1 (two-token reduction). We first analyze the simplest nontrivial interaction between the first two tokens x_1 and x_2. The key scalar is $\langle x_2(t),\, S_{BC} x_1 \rangle$, whose sign controls the long-time behavior. A Grönwall-type estimate yields exponential convergence of $x_2(t)$ to $x_1(0)$.

Step 2 (global positivity regime). We then extend the two-token analysis to the full system under the global positivity assumption that $\langle x_i(0),\, S_{BC} x_j(0) \rangle > 0$ for all i, j. The main tool is a box-invariance argument for the interaction variables $\langle x_i(t),\, S_{BC} x_j(t) \rangle$, showing that trajectories starting

in the positive box $[0,1]^{L\times L}$ cannot exit it (by a contradiction argument based on uniqueness of ODE solutions). This yields exponential convergence of all downstream tokens to $\pm x_1(0)$.

Step 3 (removing positivity by induction). Next, we drop the positivity assumption and proceed inductively. Assuming that $x_1,\dots,x_{k-1}$ converge exponentially to x_1, we show that the interaction coefficient $\langle x_k(t),\, S_{BC}x_1\rangle$ decays exponentially, implying that $x_k(t)$ converges to an equilibrium orthogonal to $S_{BC}x_1$.

Step 4 (nongenericity of orthogonal equilibria). Finally, we prove that the orthogonal-drift scenario from Step 3 is nongeneric. For a candidate equilibrium $x_* = (x_1,\dots,x_1,\xi)$, with $\xi \perp S_{BC}x_1$, we linearize the flow at x_* and show that the Jacobian $J(x_*)$ admits a positive eigenvalue $\lambda_+ = a_{L1}(x_*) > 0$, so that x_* is a saddle point. The stable manifold theorem then implies that the set of initial conditions leading to the convergence of tokens to x_* lies in a lower-dimensional manifold of Lebesgue measure zero. This provides a complete proof of Theorem 3.6.

Under slightly stronger assumptions, we can apply Theorem 3.6 to recover the single-cluster case stated below.

Proposition 3.7 (Exponential alignment under positive angle conditions)**.** *Suppose the assumptions of Theorem 3.6 hold. In addition, assume that the initial configuration satisfies the positive angle condition:*

$$\langle x_i(0), S_{BC}x_j(0)\rangle > 0, \quad \forall\, i,j \in [1,L].$$

Then all tokens $x_l(t)$ converge exponentially to the aligned state $x_1(0)$, that is,

$$\sum_{i=1}^{L} \|x_i(t) - x_1(0)\| \;\lesssim\; e^{-C_0 t},$$

for some constant $C_0 > 0$. Moreover, the positivity of interactions is preserved for all time:

$$\langle x_i(t), S_{BC}x_j(t)\rangle > 0, \quad \forall\, t > 0.$$

Remark 3.8 (All token are exponentially attracted by the first token in practical settings)**.** Theorem 3.6 states that, under certain assumptions, each token converges exponentially either to $x_1(0)$ or to $-x_1(0)$, depending on the sign of the corresponding coefficient s_i. Thus, in this setting there are two possible clusters: one around $x_1(0)$ and one around $-x_1(0)$. Proposition 3.7 imposes stricter assumptions under which all tokens converge exponentially to $x_1(0)$. In this case, there is only one cluster.

In practice, when the assumptions of Theorem 3.6 and Proposition 3.7 do not fully hold, it is possible to observe more than two clusters at the end of the dynamics (see for examples in Figure 1 below). However, this occurs only rarely. Almost all trajectories still fall into one of the two dominant clusters predicted by Theorem 3.6.

3.3 A generalization to time-dependent regime

In this section, we discuss a generalization of Theorem 3.6 to the setting of time-dependent parameters. Specifically, we allow the parameters to be nonconstant functions of time, subject to additional mild conditions ensuring that the proof strategy from the previous section remains valid.

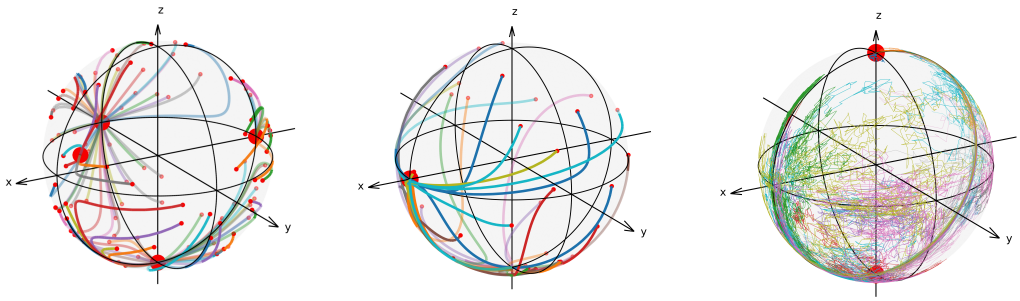
We borrow the notation from the previous subsection. To keep the proofs tractable while maintaining the generality and practical relevance of our results, we assume that the functions $S_{BC}(t)$, $\alpha(t)$, and $a(t)$ converge exponentially to S_∞, α_∞, and a_∞, respectively, when t tends to infinity. (In the previous section, these parameters are constant.) In this case, the interaction coefficients are the matrices

$$A_{lj}(t,x) = x_l^\top S_{BC}(t)x_j \cdot \Delta_t(x_l) e^{-a(t)\sum_{k=j+1}^{l}\Delta_t(x_k)},$$

where $\Delta_t(x_l) = \ln(1+e^{\alpha(t)\cdot x_l})$ for each $1 \le j < l \le L$ and $x = (x_1,...,x_L) \in (\mathbb{S}^{D-1})^L$. Under the above setting, we can generalize Theorem 3.6 to the time-dependent parameter regime as follows.

Theorem 3.9. *Let $\mathbf{x}(t) = (x_1(t),\dots,x_L(t))$ be a solution to the dynamical system*

$$\begin{cases} \frac{d}{dt}x_l(t) = \sum_{j=1}^{l-1} A_{lj}(t+c,\mathbf{x}(t)) \cdot \mathbf{P}_{x_l(t)}(x_j(t)), \\ \mathbf{x}(0) = (x_1(0),\dots,x_L(0)) \in (\mathbb{S}^{D-1})^L, \end{cases} \tag{12}$$



(a) First-token trajectories (100 random initializations; time-independent a_d, S_{BC}, S_Δ). 94 trajectories converge to some coordinate unit vector $\pm e_d$.

(b) A single sequence of 32 tokens (time-independent a_d, S_{BC}, S_Δ). Tokens converge toward a point at one coordinate unit vector $\pm e_d$ for some d.

(c) A single sequence of 16 tokens (time-dependent a_d, S_{BC}, S_Δ). Tokens converge to points at some coordinate unit vectors $\pm e_d$ for some $d = 1, \ldots, D$.

Figure 1: Simulated trajectories of dynamical system (6) with random parameters. The initial points (denoted by small red points in (a) and (b)) are sampled via a uniform distribution on the sphere. In these settings, the first token converges to discrete attractors at the big red points (mostly coordinate unit vectors $\pm e_d$) and draws subsequent tokens toward the same points, producing clusters and illustrating the attention sink effect.

for $l = 1, \ldots, L$, and $c \in \mathbb{R}$ is an arbitrary constant. Assume that $S_{BC}(t)$ is symmetric for every t and S_∞ is positive definite. Then, there exists a zero-measurable subset N_0 of $(0, \infty)$ satisfying the following property: if $c \in \mathbb{R} \setminus N_0$ then there exists a zero-measurable subset Ω of $(\mathbb{S}^{D-1})^{L-1}$ such that if the initial point $(x_2(0), ..., x_L(0))$ does not belong in Ω then all tokens converge exponentially to $\pm x_1(0)$.

Theorem 3.9 indicates that the downstream tokens $x_l(t)$, for $2 \leq l \leq L$, are strongly driven by the initial direction of the first token $x_1(0)$. In particular, for generic initializations, these tokens converge exponentially to the same attractor $\pm x_1(0)$, which leads to the emergence of token clusters in the final layers. This conclusion corroborates the phenomenon described at the beginning of Section 3. Further discussion and intuition for this result are provided in Appendix B.3.

4 Alignment with Observed Behavior in Practical Mamba Settings

We empirically validate the theoretical results established in the previous section using practical Mamba architectures, considering both the continuous-time setting (Subsection 4.1) and the discrete-time setting (Subsection 4.2). We study models with both time-dependent and time-independent parameters, and further investigate how practical architectural components of Mamba (such as pre-layer normalization and linear layers) impact the conclusions of our theory.

4.1 Simulations

To validate our theoretical results, we conduct extensive simulations of the dynamical system (6) under random choices of model parameters, demonstrating that the empirical behaviors observed in the simulations are consistent with our theoretical predictions. In particular, we numerically integrate the system to observe: (i) trajectories of first tokens across many random initializations with time-independent parameters a_d, S_{BC}, and S_Δ; (ii) trajectories of all tokens in a single sequence with time-independent parameters a_d, S_{BC}, and S_Δ; and (iii) trajectories of all tokens in a single sequence with time-dependent parameters a_d, S_{BC}, and S_Δ.

Figure 1, which demonstrates the expected results of the trajectories of the tokens under the above settings, reveals two robust patterns. First, the initial token exhibits privileged dynamics and converges to a discrete set of attractors, namely the coordinate vectors $\pm e_d$, in agreement with Theorems 3.5 and 3.3. Second, the downstream tokens are strongly driven by this first-token trajectory and typically converge to the same attractor, producing token clusters in the final layers, consistent

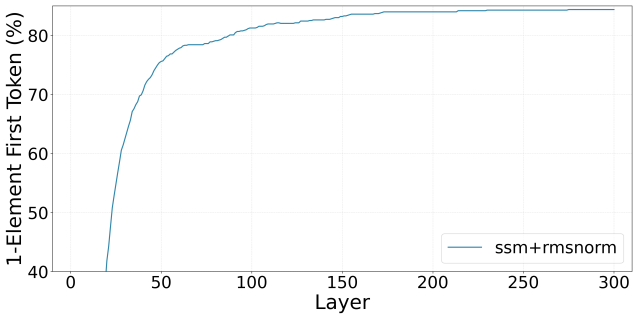
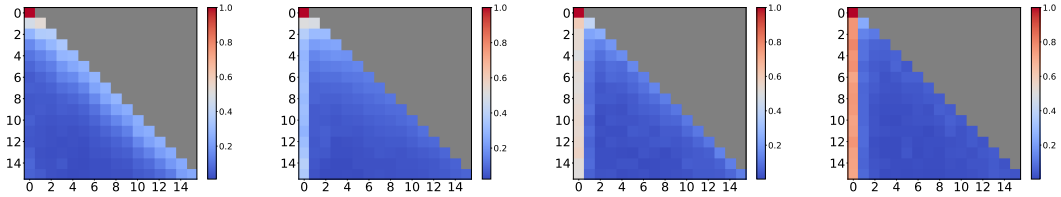


Figure 2: Percentage of first tokens exhibiting a single non-zero element (threshold 1×10^{-4}) across simplified Mamba-2 blocks with shared weights, after omitting causal convolution, input/output projections, activations, and inner residuals. Most tokens converged to 1-element tokens, consistent with simulations.



(a) Layer 1, Channel 8 (b) Layer 15, Channel 8 (c) Layer 46, Channel 8 (d) Layer 60, Channel 8

Figure 3: Attention concentration on the first token across different depths of the Falcon-Mamba-7B model [90]. Each subfigure shows the hidden attention matrix for a specific layer: (a) Layer 1, (b) Layer 15, (c) Layer 46, and (d) Layer 60. In early layers, tokens distribute attention more uniformly, but as depth increases, nearly all tokens concentrate disproportionately on the first token. This illustrates the emergence of an attention sink in state-space models.

with Theorem 3.6, Proposition 3.7, and Theorem 3.9. These robust patterns strongly support Observation 3.1.

4.2 Empirical Validation with Full Practical Setting

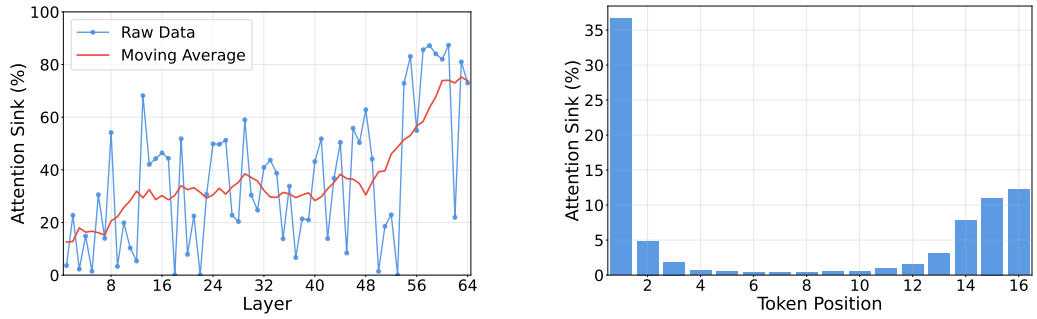
To validate our theoretical results on practical models, we conduct experiments on simplified versions of the Mamba-2 block. Specifically, we employ the configuration with shared-weight blocks. Within each block, we omit the causal convolution, input projection, and output projection.

We process a batch of 64 sequences through the model. We compute the percentage of sequences where the first token has only one non-near-zero element (threshold 1×10^{-4}). The results in Figure 2 reveal that most first tokens converge to a representation with only one non-zero element. This outcome aligns with our theoretical predictions and the simulation in Figure 1a.

5 An application – Predict Attention Sink

Observation 3.1, together with the theoretical justification provided in the previous section, suggests that the first token should be regarded with particular importance, as it serves as an attractor for subsequent tokens. This insight provides a principled explanation of the attention sink effect in Mamba models through the attracting role of the first token.

We analyze the hidden attention matrix in equation (4) for the Falcon-Mamba-7B model [90] to quantify how tokens attend to one another. The results in Figure 3 reveal that the concentration of attention on the first token consistently increases with depth, and in the final layers it dominates nearly all other tokens. This empirical trend validates our theoretical justification for token dynamics and, importantly, represents the first observation of the attention sink phenomenon in state-space models. While attention sink was originally identified in transformer-based architectures [83], we have showed that an analogous effect emerges in Mamba models as predicted from Observation 3.1.



(a) Attention sink (%) across layers for the first token.

(b) Attention sink (%) across all token positions.

Figure 4: Attention sink score [27] analysis in the Falcon-Mamba-7B model. Subfigure (a) shows the sink score of the first token and its moving average across layers; the score steadily increases with depth and dominates in the final layers. Subfigure (b) shows the average sink scores across token positions; the first token consistently receives the largest share.

To quantify the attention sink in state-space models, we employ the metrics Sink_ℓ and Sink (formally defined in Appendix C), adapted from [27] with modifications to account for the structural differences in our hidden attention representation. Specifically, while the original metric is computed from attention logits with shape $(\text{batch}, \text{head}, \text{seqlen}, \text{seqlen})$, our Mamba hidden attention has shape $(\text{batch}, \text{channel}, \text{seqlen}, \text{seqlen})$. Each channel plays an analogous role to an attention head, but given the large number of channels (4096), we uniformly sample 200 channels per layer for evaluation. We fix the threshold to $\varepsilon = 0.3$ to decide whether the first token exhibits sink in a given channel. The results, shown in Figure 4, reveal that the sink score of the first token increases consistently with depth and that, when averaged across token positions, the first token dominates all others, thereby extending the attention sink phenomenon to state-space models.

6 Conclusion and Limitation

In this work, we developed a theoretical framework for token dynamics in deep Mamba models under realistic assumptions. Modeling Mamba as flow maps on high-dimensional spheres, we characterized limit points of token trajectories, proved exponential convergence, and revealed the first token's role as a global attractor. We further extended our analysis to time-dependent parameters setting and to broader autoregressive models, as well as demonstrated a practical implication to explain the attention sink effect. These results advance the understanding of token dynamics and suggest new directions for improving Mamba models.

Limitation. Our analysis focuses on the continuous-time limit and asymptotic behavior, and future work is needed to fully bridge these insights with the discrete, large-scale implementations of Mamba models in practice.

Acknowledgments and Disclosure of Funding

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 Minh Nguyen. Transformer meets twicing: Harnessing unattended residual information. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Laziz U Abdullaev, Maksim Tkachenko, and Tan M Nguyen. Revisiting transformers with insights from image filtering and boosting. *SIAM Journal on Imaging Sciences*, 2025.

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

[4] Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137–185, 2005.

[5] Md Atik Ahamed and Qiang Cheng. Mambatab: A plug-and-play model for learning tabular data. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 369–375, 2024.

[6] Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *arXiv preprint arXiv:2407.01602*, 2024.

[7] Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. *arXiv preprint arXiv:2403.01590*, 2024.

[8] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[9] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. *arXiv preprint arXiv:2402.02098*, 2024.

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

[11] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12020–12030, 2022.

[12] Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

[13] 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.

[14] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

[15] Gbetondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

[16] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

[17] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[18] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

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

[20] 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.

[21] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *International Conference on Machine Learning*, pages 7616–7633. PMLR, 2022.

[22] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[23] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020.

[24] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

[25] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[26] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[27] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025.

[28] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

[29] 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.

[30] Seung-Yeal Ha, Jeongho Kim, Jinyeong Park, and Xiongtao Zhang. Complete cluster predictability of the cucker–smale flocking model on the real line. *Archive for Rational Mechanics and Analysis*, 231:319–365, 2019.

[31] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017.

[32] Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning*, pages 21502–21521. PMLR, 2024.

[33] Amir Joudaki, Hadi Daneshmand, and Francis Bach. On the impact of activation and normalization in obtaining isometric embeddings at initialization. *Advances in Neural Information Processing Systems*, 36:39855–39875, 2023.

[34] RE Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[35] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, 37:115652–115681, 2024.

[36] Chrysoula Kosma, Giannis Nikolentzos, and Michalis Vazirgiannis. Time-parameterized convolutional neural networks for irregularly sampled time series. *arXiv preprint arXiv:2308.03210*, 2023.

[37] Samory Kpotufe and Abdeslam Boularias. Gradient weights help nonparametric regressors. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[38] Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.

[39] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, pages 420–422. Springer, 1975.

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[41] Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] 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.

[43] 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.

[44] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.

[45] Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? In *The Eleventh International Conference on Learning Representations*, 2023.

[46] 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.

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

[48] Fusheng Liu and Qianxiao Li. From generalization analysis to optimization designs for state space models. *arXiv preprint arXiv:2405.02670*, 2024.

[49] 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.

[50] 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.

[51] William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning*, pages 35492–35506. PMLR, 2024.

[52] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM review*, 56(4):577–621, 2014.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Tuan Nguyen, Hirotada Honda, Takashi Sano, Vinh Nguyen, Shugo Nakamura, and Tan Minh Nguyen. From coupled oscillators to graph neural networks: Reducing over-smoothing via a kuramoto model-based approach. In *International Conference on Artificial Intelligence and Statistics*, pages 2710–2718. PMLR, 2024.

[60] 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.

[61] 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.

[62] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.

[63] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, pages 26670–26698. PMLR, 2023.

[64] Toshihiro Ota. Decision mamba: Reinforcement learning via sequence modeling with selective state spaces. *arXiv preprint arXiv:2403.19925*, 2024.

[65] 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.

[66] Duy-Tung Pham, An The Nguyen, Viet-Hoang Tran, Nhan-Phu Chung, Xin T Tong, Tan M Nguyen, and Thieu N Vo. Dynamical properties of tokens in self-attention and effects of positional encoding. *Advances in Neural Information Processing Systems*, 2026.

[67] 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.

[68] 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.

[69] Domenec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

[70] 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.

[71] 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.

[72] Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[73] Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2022.

[74] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021.

[75] Rachel S Teo and Tan M Nguyen. Momentumsmoe: Integrating momentum into sparse mixture of experts. *Advances in Neural Information Processing Systems*, 37:28965–29000, 2024.

[76] Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. Grand++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[78] Thieu N Vo, Tung D Pham, Xin T Tong, and Tan Minh Nguyen. Demystifying the token dynamics of deep selective state space models. *International Conference on Learning Representations (ICLR)*, 2025.

[79] Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024.

[80] 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.

[81] Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[82] 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.

[83] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

[84] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024.

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

[86] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

[87] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

[88] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural odes and invertible residual networks. In *International Conference on Machine Learning*, pages 11086–11095. PMLR, 2020.

[89] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, pages 62429–62442. PMLR, 2024.

[90] Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*, 2024.

Appendix of "First-Token Attraction in Mamba Dynamics"

Table of Contents

A Related Works **17**

B Theoretical Results **18**

B.1 Possible Limit Points of the First Token 19

B.1.1 Equilibrium Points According to the First Token 19

B.1.2 Limit points of the First Token 20

B.2 All Other Tokens are Attracted by the First Token 24

B.3 A Generalization to the Time-dependent Parameters Setting 36

C Experiment - Additional Details **37**

C.1 Experimental Setup 37

C.2 Evaluation 37

A Related Works

ResNets and Neural ODEs Our work connects to the literature that interprets deep neural networks (DNNs) through the lens of dynamical systems. Standard DNNs can be seen as discrete-time dynamical processes, where data propagate sequentially through layers [40]. Residual networks (ResNets) [10, 31, 37] exemplify this view, and their continuous counterparts, neural ordinary differential equations (neural ODEs) [47, 88, 82, 43, 55, 76, 69, 12], formalize the passage to continuous time. In this formulation, the residual block update $z = x + y(x, \theta)$ becomes a flow governed by a differential equation, mapping an initial input $x(0)$ to an output $x(T)$. A broad line of work has studied the approximation, interpolation, and controllability properties of these architectures [47, 88, 43, 73, 69, 12, 75].

Oversmoothing and Rank Collapse Transformers face two critical degeneracies: oversmoothing and rank collapse. Oversmoothing causes token representations to become indistinguishable, leading to a loss of expressive power. Nguyen et al. [56, 59], Wu et al. [81], Dovonon et al. [15], and Chen et al. [11] proved that self-attention alone cannot prevent this collapse. Remedies include *ContraNorm*, a normalization method inspired by contrastive learning [28], and the use of residual connections with normalization, which have been shown to mitigate oversmoothing [71, 1, 2].

Rank collapse, on the other hand, occurs when the hidden representations degenerate into a low-dimensional subspace. Prior studies [14, 17, 61, 33, 86, 57, 62, 60, 9] identified this as a major bottleneck and proposed stabilization strategies to maintain diversity and improve performance.

Theoretical Analyses of Mamba Recent studies have begun to explore Mamba from a theoretical standpoint. Ali et al. [7] interpreted the S6 layer as a linear-time, softmax-free variant of attention, proving that Mamba is more expressive than transformers. In contrast, Muca Cirone et al. [53] applied Rough Path Theory to show that diagonal selective state space models (SSMs) are less expressive than non-diagonal ones. Merrill et al. [51] used circuit complexity theory to argue that both SSMs and transformers belong to the class TC^0, implying computational equivalence. Jelassi et al. [32] provided both theoretical and empirical evidence that transformers surpass state space models in copying tasks. Together, these results form a theoretical basis for understanding the expressive and generalization limits of Mamba.

Token Dynamics and Clustering Effects Another perspective interprets sequence models, such as Transformers and Mamba, as interacting particle systems. Lu et al. [49] formulated self-attention as a numerical solver for a convection–diffusion equation, while Dutta et al. [16] proposed *TransEvolve*, a particle-inspired scheme that reduces complexity. These approaches connect Transformer dynamics

to broader nonlinear systems, including flocking [30], synchronization [39, 4], and consensus models [38, 52].

From this dynamical viewpoint, the long-term behavior of tokens has been studied via clustering. Geshkovski et al. [20] proved convergence to limit objects influenced by initial conditions, while follow-up work [19, 18] showed metastable multi-cluster states that persist for long times before collapsing into a single cluster. Alcalde et al. [6] further confirmed asymptotic convergence to clustered equilibria in hardmax Transformers. Pham et al. [66] investigated how different forms of positional encoding - specifically absolute and rotary - affect these dynamical regimes. In contrast, the token dynamics of Mamba models remain largely unexplored. Vo et al. [78] studied the dynamical properties of tokens in a deep Mamba model under a certain restrictive assumptions, and based on that, suggested improvements to the model by excluding negatively impactful scenarios and reordering tokens using importance scores.

In this paper, we take one step further by studying the token dynamics for Mamba without the restrictive assumptions.

B Theoretical Results

Recall that a pre-trained Mamba model can be viewed as a flow map on $(\mathbb{S}^{D-1})^L$ defined by the following dynamical system:

$$\frac{d}{dt}x_l(t) = \mathbf{P}_{x_l(t)}\left(\sum_{j=1}^{l} A_{lj}(\mathbf{x}(t))x_j(t)\right), \quad t > 0,\, l = 1, \dots, L, \tag{13}$$

$$\mathbf{x}(0) = (x_1(0), \dots, x_L(0)) \in (\mathbb{S}^{d-1})^L, \tag{14}$$

where

- for each $x \in \mathbb{S}^{D-1}$, the map $\mathbf{P}_x \colon \mathbb{R}^D \to T_x\mathbb{S}^{D-1}$ is the projection onto the tangent space at x of the sphere $\mathbb{S}^{D-1}$ and it is defined by
$$\mathbf{P}_x(y) = y - \langle y, x\rangle\, x, \quad y \in \mathbb{R}^D;$$
- the coefficient matrices A_{lj} is an $D \times D$ diagonal matrix $A_{lj} = \mathrm{diag}(A_{1,lj}, \dots, A_{D,lj})$ with
$$A_{d,lj}(\mathbf{x}(t)) = \begin{cases} x_l(t)^\top S_{BC} x_l(t) \cdot \Delta_d(x_l(t)), & \text{if } l = j, \\ x_l(t)^\top S_{BC} x_j \cdot \Delta_d(x_j(t)) \cdot \exp\left(-a_d(t) \sum_{k=j+1}^{l} \Delta_d(x_k(t))\right), & \text{if } l > j; \end{cases} \tag{15}$$
- the step size function at the d-th channel $\Delta_d : \mathbb{R}^D \to \mathbb{R}$ is defined by
$$\Delta_d(u) = \mathrm{softplus}(S_{\Delta,d}(t) \cdot u) = \ln\left(1 + e^{S_{\Delta,d}\cdot u}\right), \qquad u \in \mathbb{R}^D. \tag{16}$$

The dynamical system (6) is parameterized by the input-output matrix

$$S_{BC} = S_C^\top S_B \in \mathbb{R}^{D\times D},$$

the step-size matrix

$$S_\Delta = [S_{\Delta,1}, \dots, S_{\Delta,D}] \in \mathbb{R}^{D\times D},$$

and the state coefficient vector

$$a = (a_1, \dots, a_D) \in \mathbb{R}^D_{>0}.$$

Our main problem is studying the dynamical properties of the solutions of system (6). In particular, we want to understand how the set $\lim_{t\to+\infty}\{x_1(t), \dots, x_L(t)\}$ looks like and when it reduces to a set of a single point.

Notation. Throughout this section, the notation $O(1)$ denotes a universal positive constant. We write $A \lesssim B$ to mean that there exists a constant $C_0 > 0$, independent of time, such that $|A| \le C_0|B|$.

B.1 Possible Limit Points of the First Token

In this subsection, we will characterize all possible limit points for the first token. In order to do that, let us consider the differential equation governing the trajectory of the first equation, which is the first equation in the dynamical system (13), given by

$$\frac{d}{dt}x_1 = \mathbf{P}_{x_1}\big(A_{11}(x)x_1\big) \tag{17}$$

where

$$A_{11}(x) = \operatorname{diag}(A_{1,11}, A_{2,11}, \cdots, A_{d,11})$$

and $A_{d,11} = \langle x_1, S_{BC}x_1\rangle\Delta_d(x_1)$. We will determine the set of equilibrium points of this differential equation in Subsection B.1.1. However not every equilibrium point is a limit point. The set of possible actual limit points of the first token is then presented in Subsection B.1.2. Moreover, since

$$\langle S_{BC}x_1, x_1\rangle = \langle Sx_1, x_1\rangle,$$

where

$$S = \tfrac{1}{2}\big(S_{BC} + S_{BC}^\top\big)$$

denotes the symmetric part of S_{BC}, it follows that only the symmetric component contributes to the quadratic form. Throughout this section, we may therefore assume without loss of generality that S_{BC} is symmetric.

B.1.1 Equilibrium Points According to the First Token

Theorem B.1 (Equilibria of the first token)**.** *A vector $y \in \mathbb{S}^{D-1}$ is an equilibrium of equation (17) if and only if it belongs to one of the following types:*

Type 1. *(Alignment equilibria) The vector y may have some zero coordinates, but restricted to its nonzero entries we obtain*

$$y^{(\neq 0)} = (y_{k_1}, \ldots, y_{k_\ell}) \in \mathbb{S}^{\ell-1},$$

which lies in the common kernel of the column differences

$$S_{\Delta,k_j} - S_{\Delta,k_1}, \qquad j = 1, \ldots, \ell,$$

for the active indices $\{k_1, \ldots, k_\ell\}$.

Type 2. *(Quadratic constraint equilibria) The vector y satisfies*

$$y^\top S_{BC}\, y = 0.$$

If S_{BC} is definite, then Type 2 equilibria do not exist. Furthermore, if the columns of S_Δ are linearly independent, then the Type 1 equilibria collapse to exactly two antipodal points

$$y^{(\neq 0)} = \pm\xi,$$

where ξ is the unique unit vector in the one-dimensional intersection of the kernels above.

Proof. By definition,

$$A_{11}(x) = \langle x_1, S_{BC}x_1\rangle \operatorname{diag}\big(\Delta_1(x_1), \Delta_2(x_1), \ldots, \Delta_D(x_1)\big).$$

Substituting $x_1 = y = (y_1, \ldots, y_D) \in \mathbb{S}^{D-1}$, the dynamics take the form

$$y' = \mathbf{P}_y\big(A_{11}(y)y\big) = \langle S_{BC}y, y\rangle\, \mathbf{P}_y\big(\Delta_1(y)y_1, \ldots, \Delta_D(y)y_D\big),$$

where $\mathbf{P}_y$ denotes the orthogonal projection onto $T_y\mathbb{S}^{D-1}$.

For each component this yields

$$y_i' = \langle S_{BC}y, y\rangle\, \big\{\Delta_i(y) - \Delta(y)\big\}y_i, \tag{18}$$

with

$$\Delta(y) = \sum_{k=1}^{D} \Delta_k(y)y_k^2, \qquad \Delta_i(y) = \ln\big(1 + e^{S_{\Delta,i}\cdot y}\big). \tag{19}$$

Equation (18) shows that equilibria arise in two distinct cases:

(i) either $\langle S_{BC}y, y\rangle = 0$, corresponding to Type 2 equilibria,

(ii) or else $\Delta_i(y) = \Delta(y)$ for every index i with $y_i \neq 0$.

Since $|y| = 1$, there exists at least one index i_0 with $y_{i_0} \neq 0$. For any i with $y_i \neq 0$, condition (ii) implies

$$\Delta_i(y) = \Delta_{i_0}(y),$$

which is equivalent to

$$(S_{\Delta,i} - S_{\Delta,i_0}) \cdot y = 0. \tag{20}$$

Let $\{k_1, \dots, k_\ell\}$ denote the indices of the nonzero entries of y, and write

$$y^{(\neq 0)} = (y_{k_1}, \dots, y_{k_\ell})$$

for the restriction of y to these coordinates. Then (20) becomes

$$(S_{\Delta,k_j} - S_{\Delta,k_1}) \cdot y^{(\neq 0)} = 0, \qquad j = 1, \dots, \ell,$$

which precisely characterizes Type 1 equilibria.

Finally, if S_{BC} is definite, then $\langle S_{BC}y, y\rangle > 0$ for all $y \neq 0$, so no Type 2 equilibria can exist. Moreover, if the columns $\{S_{\Delta,k_j}\}_{j=1}^{\ell}$ are linearly independent, the intersection

$$\bigcap_{j=1}^{\ell} \langle S_{\Delta,k_j} - S_{\Delta,k_1}\rangle^{\perp}$$

is one-dimensional in $\mathbb{R}^\ell$. Its intersection with the unit sphere $\mathbb{S}^{\ell-1}$ therefore consists of exactly two antipodal points $\pm\xi$, which are the only possible equilibria. □

B.1.2 Limit points of the First Token

The analysis of stability near equilibrium points, in order to characterize the limit sets of equation (17), is in general quite delicate. Thanks to Theorem B.1, under the assumption that S_{BC} is definite and the columns of S_Δ are linearly independent, the nonzero-part of equilibria are the two antipodal points

$$y^{(\neq 0)} = (y_{k_1}, \cdots, y_{k_\ell}) = \pm\xi,$$

where ξ is the unique intersection of $\mathbb{S}^{\ell-1}$ with the one-dimensional subspace

$$\bigcap_{j=1}^{\ell} \ker\big((S_{\Delta,k_j} - S_{\Delta,k_1})^\top\big).$$

To avoid unnecessary technical complications, we shall henceforth restrict attention to this case. In particular, we consider the high-dimensional setting in which

$$S_{\Delta,m} = \pm e_m,$$

with e_m denoting the m-th coordinate vector (see Theorem B.5). Apart from the special case $D = 2$, where the structure can be worked out explicitly (see Theorem B.7), this reduction allows us to describe all possible limit points of the first token.

We start with the following auxiliary lemma.

Lemma B.2 (Sign Preservation)**.** *Let $y(t) = (y_1(t), \dots, y_D(t)) \in C^\infty([0, +\infty), \mathbb{R}^D)$ be a solution of equation (17). If $y_i(0) < 0$ then $y_i(t) < 0$ for all $t > 0$. Similarly, if $y_i(0) > 0$ then $y_i(t) > 0$ for all $t > 0$.*

Proof. Suppose $y_i(0) < 0$ but $y_i(t) \geq 0$ for some $t > 0$. Then there exists a minimal $T > 0$ such that

$$y_i(T) = 0, \qquad y_i(t) < 0 \quad \text{for all } t \in [0, T).$$

By continuity of y_i, we have

$$\lim_{t \to T^-} y_i(t) = 0.$$

For $t \in (0, T)$ the evolution equation reads

$$\frac{d}{dt} y_i(t) = (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta(y)\big)\, y_i(t).$$

Dividing by $y_i(t)$ yields

$$\frac{d}{dt}\big(\ln |y_i(t)|\big) = (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta(y)\big).$$

Let $\{t_n\}_{n=1}^{\infty} \subset (0, T)$ be an increasing sequence with $t_n \to T$. Then

$$\ln |y_i(t_n)| - \ln |y_i(0)| = \int_0^{t_n} (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta(y)\big)\, ds.$$

The right-hand side is bounded, hence the left-hand side is also bounded. But since $|y_i(t_n)| \to 0$, we have $\ln |y_i(t_n)| \to -\infty$, a contradiction. Thus $y_i(t) < 0$ for all $t > 0$.

The case $y_i(0) > 0$ is analogous. □

Lemma B.3. *Let $y(t) = (y_1(t), \dots, y_D(t)) \in C^{\infty}([0, +\infty), \mathbb{R}^D)$ be a solution of equation (17). Assume that $S_{\Delta,m} = \pm e_m$ and $y^{\top} S_{BC} y \geq 0$ for all $y \neq 0$. Here, the sign $\pm$ in $\pm e_m$ is the sign of $y_m(0)$. Suppose that there are $i, j \in \{1, ..., D\}$ such that*

$$0 < |y_j(0)| < |y_i(0)|.$$

Then

$$0 < |y_j(t)| < |y_i(t)|,$$

for every $t > 0$.

Proof. By Lemma B.2, we have $y_j(t), y_i(t) \neq 0$ for every $t > 0$. It remains to prove that

$$|y_j(t)| < |y_i(t)|,$$

for every $t > 0$.

The same computation as in the proof of Lemma B.2 gives

$$\frac{d}{dt}\big(\ln |y_i(t)|\big) = (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta(y)\big),$$

and

$$\frac{d}{dt}\left(\ln |y_i(t)|\right) = (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta(y)\big).$$

Hence, we have

$$\frac{d}{dt}\Big(\ln \frac{|y_i(t)|}{|y_j(t)|}\Big) = (S_{BC} y \cdot y)\big(\Delta_i(y) - \Delta_j(y)\big). \tag{21}$$

Note that, since y_j and y_i are continuous and nonvanishing on $[0, \infty)$, they have constant sign on $[0, \infty)$. Therefore,

$$S_{\Delta,i}(y(t)) = \text{sign}(y_i(0)) y_i(t) = |y_i(t)| \text{ and } S_{\Delta,j}(y(t)) = \text{sign}(y_j(0)) y_j(t) = |y_j(t)|,$$

for every $t > 0$. Consequently, we get

$$\Delta_d(y(t)) = \ln\left(1 + e^{|y_d(t)|}\right),$$

for $d = i, j$. Combining this with (21), we obtain

$$\frac{d}{dt}\Big(\ln \frac{|y_i(t)|}{|y_j(t)|}\Big) = (S_{BC} y \cdot y)\left(\ln\left(1 + e^{|y_i(t)|}\right) - \ln\left(1 + e^{|y_j(t)|}\right)\right). \tag{22}$$

Now, suppose to the contrary that $|y_j(t)| \geq |y_i(t)|$ for some $t > 0$. Then there is $T > 0$ such that $|y_j(t)| < |y_i(t)|$ for every $0 \leq t < T$ and $|y_j(T)| = |y_i(T)|$. Integrating both side of (21) and using the fact $S_{BC} y \cdot y \geq 0$, we get

$$\ln \frac{|y_i(T)|}{|y_j(T)|} - \ln \frac{|y_i(0)|}{|y_j(0)|} = \int_0^T (S_{BC} y \cdot y)\left(\ln\left(1 + e^{|y_i(t)|}\right) - \ln\left(1 + e^{|y_j(t)|}\right)\right) ds \geq 0.$$

Then

$$0 = \ln \frac{|y_i(T)|}{|y_j(T)|} \geq \ln \frac{|y_i(0)|}{|y_j(0)|} > 0.$$

We get a contradiction. Thus, $|y_j(t)| < |y_i(t)|$ for every $t > 0$.

The proof is complete. □

Lemma B.4. *Let $y(t) = (y_1(t), \dots, y_D(t)) \in C^\infty([0, +\infty), \mathbb{R}^D)$ be a solution of equation (17). Assume that the symmetric part*

$$S = \tfrac{1}{2}(S_{BC} + S_{BC}^\top)$$

of S_{BC} is positive definite, and that $S_{\Delta,m} = \pm e_m$, where the sign $\pm$ agrees with the sign of $y_m(0)$. Suppose that there are $i, j \in \{1, ..., D\}$ such that

$$0 < |y_j(0)| < |y_i(0)|.$$

Then

$$\lim_{t\to\infty} y_j(t) = 0.$$

Proof. Define

$$c(t) = \frac{|y_i(t)|}{|y_j(t)|}.$$

By Lemma B.3, we have $c(t) > 1$ for all $t \geq 0$. Moreover, it follows from (22) that

$$\frac{d}{dt} \ln c(t) = (Sy \cdot y) \left(\ln \left(1 + e^{|y_i(t)|}\right) - \ln \left(1 + e^{|y_j(t)|}\right) \right) \geq 0. \tag{23}$$

In particular, $c(t)$ is increasing.

We distinguish two cases.

Case 1: $c(t) \to \infty$ as $t \to \infty$.
In this case we have

$$|y_j(t)| \leq \tfrac{1}{c(t)} \to 0 \text{ as } t \to \infty.$$

Case 2: $c(t)$ is bounded on $(0, \infty)$.
Let $f(t) = \ln(1 + e^t)$. By the mean value theorem, there exists $u \in (|y_j|, |y_i|)$ such that

$$f(|y_i|) - f(|y_j|) = f'(u)\,(|y_i| - |y_j|) = \frac{e^u}{1 + e^u}\,(c(t) - 1)|y_j(t)| \geq \frac{(c(0) - 1)|y_j(t)|}{2}.$$

Combining this with (23) gives

$$|y_j(t)| \leq C \frac{d}{dt} \ln c(t), \tag{24}$$

where $C = \frac{2}{(c(0)-1) \min_{y \in \mathbb{S}^{D-1}} (Sy \cdot y)} > 0$. Integrating both sides of (24), we get

$$\int_0^\infty |y_j(t)| dt \leq C(\lim_{T\to\infty} \ln c(T) - \ln c(0)) < \infty.$$

As a consequence, we have

$$\lim_{n\to\infty} \int_n^{n+1} |y_j(t)| dt = 0.$$

Since $y_j(t)$ is Lipschitz continuous, it follows that

$$\lim_{t\to\infty} |y_j(t)| = 0.$$

This completes the proof. □

Theorem B.5 (Convergence of the first tokens in high dimension)**.** *Assume that the symmetric part of S_{BC} is positive definite. Let $y(t) = (y_1(t), \dots, y_D(t)) \in C^\infty([0,\infty), \mathbb{R}^D)$ be a solution of equation (17). Assume that $S_{\Delta,m} = \pm e_m$, where the sign is chosen according to $\operatorname{sgn}(y_m(0)) \neq 0$, for all $m = 1, \dots, D$. Suppose that there exists $d_0 \in \{1, ..., D\}$ such that*

$$|y_{d_0}(0)| > |y_d(0)| > 0,$$

for every $d \neq d_0$. Then

$$\lim_{t\to\infty} y(t) = \pm e_{d_0},$$

where the sign is $\operatorname{sign}(y_{d_0}(0))$.

Proof. By Lemma B.4, we have

$$\lim_{t\to\infty} y_d(t) = 0, \tag{25}$$

for every $d \neq d_0$. Then

$$\lim_{t\to\infty} |y_{d_0}(t)| = 1.$$

Since $y_{d_0}(t)$ has constant sign on $[0,\infty)$, it follows that

$$\lim_{t\to\infty} y_{d_0}(t) = \operatorname{sign}(y_{d_0}(0)). \tag{26}$$

Combining (25) and (26), we get

$$\lim_{t\to\infty} y(t) = \operatorname{sign}(y_{d_0}(0)) e_{d_0},$$

as desired. □

Remark B.6. The conclusion of Theorem B.5 highlights a striking selection mechanism: under the assumption that the columns of S_Δ are linearly independent, the long-time dynamics do not converge to a large collection of equilibria lying in a lower-dimensional sphere in Theorem B.1. Instead, the trajectory is forced to pick out one of the most "extreme" coordinate directions, namely the basis vector $\pm e_d$ corresponding to the largest initial coordinate in magnitude.

In the two-dimensional case, we can actually determine all possible limit points of the first token in the following theorem.

Theorem B.7 (Convergence of the first token in two dimensions)**.** *Let $D = 2$ and let $y(t) = (y_1(t), y_2(t)) \in C^\infty([0,+\infty), \mathbb{S}^1)$ be a solution of (17). Assume that*

- *$S_{BC} = \operatorname{diag}(\lambda_1, \lambda_2)$ with either $\lambda_1 \neq \lambda_2$ or $\lambda_1 = \lambda_2 \neq 0$;*
- *$S_{\Delta,2} \neq S_{\Delta,1}$.*

In this setting, the unit circle $\mathbb{S}^1$ intersects the line $\langle S_{\Delta,2} - S_{\Delta,1}\rangle^\perp$ at two antipodal points $\pm\xi$. Then the limit $\lim_{t\to+\infty} y(t)$ exists and belongs to the set

$$\Omega = \{(0,\pm1), (\pm1,0)\} \cup \{\pm\xi\} \cup \left\{(\cos\varphi_\star, \sin\varphi_\star) : \lambda_1 \cos^2\varphi_\star + \lambda_2 \sin^2\varphi_\star = 0\right\}.$$

Proof. Since $D = 2$, we may parametrize

$$y(t) = (\cos\varphi(t),\ \sin\varphi(t)), \qquad \varphi(t) \in [0, 2\pi].$$

Define $y^\perp = (-y_2, y_1)$. Then

$$\varphi'(t) y^\perp = y'(t) = A_{11} y - (A_{11} y \cdot y)\, y.$$

Taking the inner product with $y^\perp$, and using $|y| = 1$ and $y \cdot y^\perp = 0$, yields

$$\varphi'(t) = A_{11} y \cdot y^\perp = (\lambda_1 y_1^2 + \lambda_2 y_2^2) y_1 y_2 \big(\Delta_2(y) - \Delta_1(y)\big).$$

Substituting $y = (\cos\varphi, \sin\varphi)$, we obtain

$$\varphi'(t) = \big[(\lambda_1 + \lambda_2) + (\lambda_1 - \lambda_2)\cos(2\varphi)\big] \sin(2\varphi) \big[h(R_2, \varphi - \beta) - h(R_1, \varphi - \alpha)\big],$$

where

$$h(R,u) = \tfrac{1}{4} \ln\big(1 + e^{R\cos u}\big), \qquad S_{\Delta,2} = R_2(\cos\beta, \sin\beta), \qquad S_{\Delta,1} = R_1(\cos\alpha, \sin\alpha).$$

Thus φ satisfies the autonomous ODE

$$\varphi' = F(\varphi),$$

where the points $(\cos\varphi, \sin\varphi)$ corresponding to zeros of F in $[0, 2\pi]$ form precisely Ω. Since Ω is finite, the zeros of F partition $(0, 2\pi)$ into intervals on each of which F has a definite sign. By the phase portrait of φ, every trajectory converges to one of the equilibria of F, and hence $y(t)$ converges to a point in Ω. This proves the claim. □

B.2 All Other Tokens are Attracted by the First Token

Different from the first token, the study of the convergence of all tokens based on the dynamical system (13) is much more challenging. To make the proofs feasible while maintaining the generality of our theoretical findings, in this whole subsection, we will consider the dynamical system (13) under the following assumptions:

$$\textit{Assumption: } S_{\Delta,1} = \ldots = S_{\Delta,D} =: \alpha \in \mathbb{R}^D \text{ and } a_1 = \ldots = a_D =: a \in \mathbb{R}.$$

In this case we can rewrite dynamical system (13) as:

$$\frac{d}{dt}x_i = \sum_{j=1}^{i-1} A_{ij}\mathbf{P}_{x_i}(x_j) \tag{27}$$

where A_{ij} is defined as

$$A_{ij} = (x_i \cdot S_{BC}x_j)a_{ij}, \qquad a_{ij} = \Delta(x_i)e^{-a\sum_{k=j+1}^{i}\Delta(x_k)}, \qquad \text{and} \qquad \Delta(x_i) = \ln(1+e^{\alpha\cdot x_i}).$$

Throughout this section, we will assume that $x(t) = (x_1(t), x_2(t), \ldots, x_L(t)) \in C^\infty([0,+\infty);\mathbb{R}^D)^L$ is a solution of the dynamical system (27).

For the first token $x_1(t)$, we see that

$$\frac{d}{dt}x_1 = A_{11}\mathbf{P}_{x_1}(x_1) = 0$$

This implies that $x_1(t) = x_1$ is a constant for all $t \geq 0$. In the following, we will present all possible limit points of $x_l(t)$ with $l = 2, \ldots, L$.

Define the consensus set

$$\Omega_0 = \{(\pm x_1, \ldots, \pm x_1)(0) \in (\mathbb{S}^{D-1})^{L-1}\}.$$

This set has exactly 2^{L-1} distinct elements, corresponding to the situation where every token is either aligned or anti-aligned with x_1.

Our main theorem of this section is stated as follows:

Theorem B.8. *Consider L trajectories $x_1(t), \ldots, x_L(t)$ evolving on the unit sphere $\mathbb{S}^{D-1}$ under the dynamics (27). Assume that S_{BC} is a $D \times D$ symmetric matrix, and that the initial reference token $x_1 = x_1(0)$ satisfies*

$$\langle x_1, S_{BC}x_1\rangle > 0.$$

Suppose further that for each $i \geq 2$,

$$\langle x_i(0), S_{BC}x_1\rangle \notin \{0, -1, 1\}, \qquad \langle x_i(0), x_1\rangle \notin \{-1, 1\}.$$

Then there exists a set of measure zero $N \subset (\mathbb{S}^{D-1})^{L-1}$, depending only on x_1 and $S_{BC}x_1$, such that for any initialization $(x_2, \ldots, x_L)(0) \in (\mathbb{S}^{D-1})^{L-1} \setminus N$, precisely one of the trajectory $x(t) = (x_2, \ldots, x_L)(t)$ converges to some $\omega_0 \in \Omega_0$ as $t \to \infty$. Moreover, there exists $C_0 > 0$ such that

$$\|x(t) - \omega_0\| \lesssim e^{-C_0 t}, \qquad t \geq 0.$$

Outline of the proof. The structure of this section is as follows. We first establish an affirmative result: under a favorable positive-angle condition on the initial token configuration, the system converges to a state where all tokens align with x_1. This is proved in Proposition B.9, with the base case $L = 2$ treated separately in Proposition B.15. We then proceed by induction: assuming exponential convergence of the first $k-1$ tokens to $\pm x_1$, we show that the k-th token either converges exponentially to $\pm x_1$ or to a point in the orthogonal complement $\langle S_{BC}x_1\rangle^\perp$. Finally, Proposition B.13 demonstrates that the latter scenario occurs only for a set of initial conditions of measure zero. The main difficulty arises from the fact that the interaction terms $\langle x_i, S_{BC}x_j\rangle$ are not sign-definite—a feature that distinguishes the model from classical consensus or synchronization dynamics. This necessitates a careful combination of nonlinear analysis and linear stability techniques to exclude nongeneric limiting behaviors.

Proposition B.9. *Assume that $\langle x_1(0), S_{BC}x_1(0)\rangle > 0$ and that:*

- $\langle x_i(0), S_{BC} x_j(0)\rangle > 0$ *for all* $i, j \in [1, L]$, *and*
- $\langle x_i(0), x_1(0)\rangle \notin \{-1, 1\}$ *for all* $i \in [2, L]$.

Then,

$$\sum_{i=1}^{L} |x_i(t) - x_1(0)| \lesssim e^{-O(1)t}.$$

Moreover, we have

$$\langle x_i(t), S_{BC} x_j(t)\rangle > 0 \quad \text{for all } t > 0.$$

The proof of Proposition B.9 will be postponed to the end of this section, as it requires several preparatory steps. To analyze the token dynamics, we first encode the pairwise interactions between tokens through the scalar quantities

$$X_{ij}(t) = \langle x_i(t), x_j(t)\rangle, \qquad U_{ij}(t) = \langle x_i(t), S_{BC} x_j(t)\rangle, \qquad i, j \in [1, L].$$

Here X_{ij} measures the relative alignment between tokens i and j on the sphere $\mathbb{S}^{d-1}$, while U_{ij} captures their interaction through the transformation S_{BC}. These variables will serve as the fundamental building blocks in our nonlinear analysis and will play a central role in the proof of Theorem B.8.

Our first step is to establish their evolution equations.

Lemma B.10. *Assume that S_{BC} is symmetric. For $i > 1$, we have*

$$\begin{aligned} X'_{i1}(t) &= \sum_{j=1}^{i-1} a_{ij} U_{ij} \left[X_{j1} - X_{ij} X_{i1}\right], \\ U'_{i1}(t) &= \sum_{j=1}^{i-1} a_{ij} U_{ij} \left[U_{j1} - X_{ij} U_{i1}\right]. \end{aligned} \tag{28}$$

If $i \geq j > 1$, then

$$\begin{aligned} X'_{ij}(t) &= \sum_{k=1}^{i-1} a_{ik} U_{ik} \left[X_{kj} - X_{ik} X_{ij}\right] + \sum_{k=1}^{j-1} a_{jk} U_{jk} \left[X_{ki} - X_{jk} X_{ji}\right], \\ U'_{ij}(t) &= \sum_{k=1}^{i-1} a_{ik} U_{ik} \left[U_{kj} - X_{ik} U_{ij}\right] + \sum_{k=1}^{j-1} a_{jk} U_{jk} \left[U_{ki} - X_{jk} U_{ji}\right]. \end{aligned}$$

Proof. Recalling (27), we have

$$\frac{d}{dt} x_i = \sum_{j=1}^{i-1} a_{ij} U_{ij} \left[x_j - X_{ij} x_i\right].$$

Dotting both sides of the above equation with x_1, we have

$$X'_{i1}(t) = \sum_{j=1}^{i-1} a_{ij} U_{ij} \left[X_{j1} - X_{ij} X_{i1}\right].$$

Next for $i \geq j > 1$, we compute

$$\begin{aligned} X'_{ij} &= \frac{d}{dt}\langle x_i(t), x_j(t)\rangle = \frac{d}{dt} x_i \cdot x_j + \frac{d}{dt} x_j \cdot x_i \\ &= \sum_{k=1}^{i-1} a_{ik} U_{ik} (x_k \cdot x_j - X_{ik} x_i \cdot x_j) + \sum_{k=1}^{j-1} a_{jk} U_{jk} (x_k \cdot x_i - X_{jk} x_j \cdot x_i) \\ &= \sum_{k=1}^{i-1} a_{ik} U_{ik} (X_{kj} - X_{ik} X_{ij}) + \sum_{k=1}^{j-1} a_{jk} U_{jk} (X_{ki} - X_{jk} X_{ji}). \end{aligned}$$

We also have

$$\frac{d}{dt}U_{i1} = \frac{d}{dt}x_i \cdot S_{BC}x_1 = \sum_{j=1}^{i-1} a_{ij}U_{ij}\left[U_{j1} - X_{ij}U_{i1}\right],$$

and similarly

$$\begin{aligned}\frac{d}{dt}U_{ij} &= \frac{d}{dt}x_i \cdot S_{BC}x_j + \frac{d}{dt}x_j \cdot S_{BC}x_i \\ &= \sum_{k=1}^{i-1} a_{ik}U_{ik}\left[U_{kj} - X_{ik}U_{ij}\right] + \sum_{k=1}^{j-1} a_{jk}U_{jk}\left[U_{ki} - X_{jk}U_{ji}\right].\end{aligned}$$

The proof is complete. □

Proposition B.11. *Assume that S_{BC} is symmetric and that the leading token $x_1 = x_1(0)$ satisfies*

$$\langle x_i(0),\, S_{BC}x_1\rangle \notin \{-1, 0, 1\}, \qquad \text{for all } i \in [2, L].$$

Then there exists a minimal index $k \in [3, L]$ such that exactly one of the following holds:

(i) Every token up to level k converges exponentially to the expected limit direction:

$$x_i(t) \to \pm x_1(0) \quad \text{as} \quad t \to \infty, \qquad i \le k,$$

with uniform exponential rate

$$\sum_{i=1}^{k} \left|x_i(t) - (\pm x_1(0))\right| \lesssim e^{-O(1)t}.$$

(ii) All earlier tokens align exponentially:

$$x_i(t) \to \pm x_1(0), \qquad i \le k-1,$$

while the k-th token loses its component along the direction $S_{BC}x_1(0)$:

$$\lim_{t\to\infty} \langle x_k(t),\, S_{BC}x_1(0)\rangle = 0, \qquad \left|\langle x_k(t),\, S_{BC}x_1(0)\rangle\right| \lesssim e^{-O(1)t}. \tag{29}$$

and $x_k(t)$ converges to some unit vector $\xi \in \langle S_{BC}x_1(0)\rangle^{\perp}$.

Moreover, after possibly replacing each x_i by its reflection $-x_i$, we may assume without loss of generality that

$$\lim_{t\to\infty} x_i(t) = x_1(0), \qquad i \le k-1.$$

Proof. For simplicity of notation, we may assume $k = L$ and proceed by induction on L. Our main tool is Proposition B.9.

We first consider the case $L = 2$. If $\langle x_2(0), S_{BC}x_1\rangle > 0$, then the result follows directly from Proposition B.9. If instead $\langle x_2(0), S_{BC}x_1\rangle < 0$, we define

$$\widetilde{x}_2(t) = -x_2(t).$$

Since $\mathbf{P}_{-y}(x) = \mathbf{P}_y(x)$ for all $x, y \in \mathbb{S}^{D-1}$, we obtain

$$\frac{d}{dt}\widetilde{x}_2 = -\frac{d}{dt}x_2 = -a_{21}\langle x_2, S_{BC}x_1\rangle \mathbf{P}_{x_2}(x_1) = a_{21}\langle \widetilde{x}_2, S_{BC}x_1\rangle \mathbf{P}_{\widetilde{x}_2}(x_1).$$

Hence, by Proposition B.9 with $\langle \widetilde{x}_2(0), S_{BC}x_1\rangle > 0$, we deduce

$$|\widetilde{x}_2(t) - x_1| \lesssim e^{-O(1)t}.$$

From this point onward, starting from the third token, we replace (x_1, x_2) by $(x_1, \widetilde{x}_2)$. We call this a positive angle rearrangement. By abuse of notation, $\widetilde{x}_i$ may denote either x_i or $-x_i$, depending on whether $x_i \to x_1$ or $x_i \to -x_1$. This completes the case $L = 2$. Now assume we have the rearrangement $(x_1, \widetilde{x}_2, \cdots, \widetilde{x}_{L-1})$ such that

$$\sum_{i=2}^{L-1} |\widetilde{x}_i(t) - x_1| \lesssim e^{-O(1)t}.$$

We claim that: if $\langle x_L(t), S_{BC}x_1\rangle \not\to 0$ as $t\to\infty$, then $x_L(t)\to \pm x_1$ as $t\to\infty$. Indeed, since $\langle x_L(t), S_{BC}x_1\rangle \not\to 0$ as $t\to\infty$, there exists $\epsilon_0>0$ such that either the set

$$\{t>0: \quad \langle x_L(t), S_{BC}x_1\rangle \geq \epsilon_0\},$$

or

$$\{t>0: \quad \langle x_L(t), S_{BC}x_1\rangle \leq -\epsilon_0\}$$

contains an increasing sequence of times $\{t_n\}_{n=1}^{\infty}$. We will show that there exists n such that

$$\langle x_L(t_n), S_{BC}x_1\rangle, \cdots, \langle x_L(t_n), S_{BC}\widetilde{x}_{L-1}(t_n)\rangle$$

all have the same sign (positive or negative). If this is the case, we can perform the reflection $\widetilde{x}_L$, apply Proposition B.9, and conclude $\widetilde{x}_L\to x_1$ as $t\to\infty$.

Since $\widetilde{x}_k\to x_1$ for all $k\leq L-1$ as $t\to\infty$, there exists $T>0$ such that for any $k\in[2,L-1]$,

$$|\langle x_L(t), S_{BC}x_1\rangle - \langle x_L(t), S_{BC}\widetilde{x}_k(t)\rangle| \leq 0.1\epsilon_0 \quad \text{for all} \quad t\geq T.$$

This implies

$$\langle x_L(t), S_{BC}x_1\rangle - 0.1\epsilon_0 \leq \langle x_L(t), S_{BC}x_k(t)\rangle \leq \langle x_L(t), S_{BC}x_1\rangle + 0.1\epsilon_0 \quad \forall t\geq T.$$

- If there is an increasing sequence of times $\{t_n\}$ such that $\langle x_L(t_n), S_{BC}x_1\rangle\geq\epsilon_0$, then we can find $t_n\geq T$ such that

  $$\langle x_L(t_n), S_{BC}x_k(t_n)\rangle \geq 0.9\epsilon_0.$$

  Hence Proposition B.9 applied to $(x_1,\widetilde{x}_2,\cdots,x_L)$ at time $t=t_n$ yields $x_L(t)\to x_1$ as $t\to\infty$.

- If there is an increasing sequence of times $\{t_n\}$ such that $\langle x_L(t_n), S_{BC}x_1\rangle\leq-\epsilon_0$, then we can find $t_n\geq T$ such that

  $$\langle x_L(t_n), S_{BC}x_k(t_n)\rangle \leq -0.9\epsilon_0.$$

  Hence Proposition B.9 applied to $(x_1,\widetilde{x}_2,\cdots,\widetilde{x}_L)$ at time $t=t_n$ yields $\widetilde{x}_L(t)\to x_1$ as $t\to\infty$.

The claim is proved.

Now if $\langle x_L(t), S_{BC}x_1\rangle\to 0$, we show that

$$|\langle x_L(t), S_{BC}x_1\rangle| \lesssim e^{-O(1)t}.$$

Recalling that $U_{L1}(t)=\langle x_L(t), S_{BC}x_1\rangle$, the above is equivalent to

$$U_{L1}^2(t)\to 0 \quad \text{as} \quad t\to\infty.$$

Using the evolution equation of U_{L1} in B.10 and the fact that all previous tokens converge to $\pm x_1(0)$, there exists a time $T>0$ and positive constants $c_T,\sigma,C_\sigma>0$ such that

$$\frac{d}{dt}U_{L1}^2 \geq c_T U_{L1}^2 - C_\sigma e^{-\sigma t} \quad \text{for all} \quad t\geq T. \tag{30}$$

Hence for any time $T_1\geq T$, we obtain

$$U_{L1}^2(t)\geq e^{c_0(t-T_1)}\left[U_{L1}(T_1)^2 - \frac{C_\sigma}{c_T+\sigma}e^{-\sigma T_1}\right] + \frac{C_\sigma}{c_T+\sigma}e^{-\sigma t}, \qquad t\geq T_1.$$

Since $U_{L1}^2(t)\in[0,1]$ for all $t\geq 0$, by letting $t\to\infty$ it follows that

$$U_{L1}(T_1)^2 \leq \frac{C_\sigma}{c_T+\sigma}e^{-\sigma T_1}.$$

Since the above inequality is true for any time $T_1\geq T$, we get (29). Now we have

$$\left|\frac{dx_k}{dt}\right| \lesssim \left|\langle x_k(t), S_{BC}x_1(0)\rangle\right| \lesssim e^{-O(1)t}.$$

This implies that $\{x_k(t)\}_{t>0}$ is Cauchy and hence $x_k(t)$ converges to some unit vector $\xi\in\langle S_{BC}x_1(0)\rangle^{\perp}$ as $t\to\infty$. The proof is complete.

□

Remark B.12. Proposition B.11 shows that, starting from the third token onward, each trajectory either converges to a point in $\{\pm x_1\}$ or else its component in the $S_{BC}x_1(0)$ direction vanishes asymptotically.

In the next proposition, we establish that if all preceding tokens converge to $\{\pm x_1\}$, then the probability that the subsequent token converges to any point outside $\{\pm x_1\}$ is zero.

Proposition B.13. *Let $L \geq 3$ and assume that $\langle S_{BC}x_1(0), x_1(0)\rangle > 0$. Denote by K the set of all points $\xi \in \mathbb{S}^{D-1}$ such that*

$$\langle \xi, S_{BC}x_1(0)\rangle = 0.$$

Then the set of initial configurations $(x_2(0), \dots, x_L(0)) \in (\mathbb{S}^{D-1})^{L-1}$ for which the corresponding trajectory satisfies

$$(x_2(t), \dots, x_{L-1}(t), x_L(t)) \longrightarrow (\pm x_1, \dots, \pm x_1, \xi) \quad \textbf{\textit{for some}}\ \xi \in K \text{ as } t \to \infty$$

has measure zero in $(\mathbb{S}^{D-1})^{L-1}$.

Proof. Without loss of generality, we may assume that $x_k \to x_1$ for all $k \in [2, L-1]$. Our first goal is to show that, for each $\xi \in K$ the set of initial conditions leading to $x_L \to \xi$ has measure zero in $(\mathbb{S}^{D-1})^{L-1}$.

Recall the notation

$$X_{ij} = x_i \cdot x_j, \qquad U_{ij} = x_i \cdot S_{BC}x_j.$$

For convenience, we define the limiting configuration associated with ξ by

$$x_\star = x_\star(\xi) := (x_1, \dots, x_1, \xi) \in (\mathbb{S}^{D-1})^L.$$

We perturb the configuration $(x_i)_{i=1}^L$ around $x_\star$ by setting

$$\begin{cases} \widetilde{x}_1 = 0, \quad x_i = x_1 + \delta\widetilde{x}_i + O(\delta^2) & \text{for } 2 \leq i \leq L-1, \\ x_L = \xi + \delta\widetilde{x}_L + O(\delta^2), \\ (x_1, \dots, x_L) = x_\star + \delta\,(0, \widetilde{x}_2, \dots, \widetilde{x}_L) + O(\delta^2). \end{cases}$$

Here

$$x_1 \cdot \widetilde{x}_i = 0 \quad \text{for } 2 \leq i \leq L-1, \qquad \xi \cdot \widetilde{x}_L = 0.$$

For $i, j \leq L$, we introduce the first variations

$$\widetilde{X}_{ij} = \widetilde{x}_i \cdot x_{\star,j} + x_{\star,i} \cdot \widetilde{x}_j, \qquad \widetilde{U}_{ij} = \widetilde{x}_i \cdot S_{BC}x_{\star,j} + x_{\star,i} \cdot S_{BC}\widetilde{x}_j,$$

and we set

$$\gamma = S_{BC}x_1 \cdot x_1 > 0.$$

We obtain the following first-order expansions:

$$\begin{cases} X_{Li} &= \xi \cdot x_1 + \delta\big[\widetilde{X}_{L1} + \xi \cdot \widetilde{x}_i\big] + O(\delta^2), \qquad 1 \leq i \leq L-1, \\ U_{Li} &= \delta\big[\xi \cdot S_{BC}\widetilde{x}_i + \widetilde{U}_{L1}\big] + O(\delta^2), \qquad 1 \leq i \leq L-1, \\ X_{ij} &= 1 + \delta\widetilde{X}_{ij} + O(\delta^2) = 1 + O(\delta^2), \qquad 1 \leq i, j \leq L-1, \\ U_{ij} &= \gamma + \delta\, S_{BC}x_1 \cdot (\widetilde{x}_i + \widetilde{x}_j) + O(\delta^2), \qquad 1 \leq i, j \leq L-1. \end{cases}$$

Substituting the above expansions into the equations of Lemma B.10 and collecting the $O(\delta)$ terms, we obtain:

(i) Linearized equation for U_{L1}:

$$\gamma^{-1}\,\partial_t\widetilde{U}_{L1} = a_{L1}(x_\star)\widetilde{U}_{L1} + \sum_{j=2}^{L-1} a_{Lj}(x_\star)\big[\widetilde{U}_{L1} + \xi \cdot S_{BC}\widetilde{x}_j\big].$$

(ii) For $2 \le j \le L-1$, we have

$$\begin{aligned}\gamma^{-1}\,\partial_t\big[\xi\cdot S_{BC}\widetilde{x}_j+\widetilde{U}_{L1}\big] &= \big(a_{L1}(x_\star)+a_{j1}(x_\star)\big)\widetilde{U}_{L1}\\ &\quad+\Big(a_{Lj}(x_\star)-\sum_{k=1}^{j-1}a_{jk}(x_\star)\Big)\big[\xi\cdot S_{BC}\widetilde{x}_j+\widetilde{U}_{L1}\big]\\ &\quad+\sum_{k=2}^{j-1}a_{jk}(x_\star)\big[\xi\cdot S_{BC}\widetilde{x}_k+\widetilde{U}_{L1}\big]\\ &\quad+\sum_{\substack{2\le k\le L-1\\ k\neq j}}a_{Lk}(x_\star)\big[\xi\cdot S_{BC}\widetilde{x}_k+\widetilde{U}_{L1}\big].\end{aligned}$$

We now introduce the vector $\vec{X}\in\mathbb{R}^{L-1}$ defined by

$$\vec{X}(t)=\begin{bmatrix}\widetilde{U}_{L1}\\ \xi\cdot S_{BC}\widetilde{x}_2+\widetilde{U}_{L1}\\ \vdots\\ \xi\cdot S_{BC}\widetilde{x}_{L-1}+\widetilde{U}_{L1}\end{bmatrix}=[X_1,X_2,\dots,X_{L-1}]^T.$$

The vector $\vec{X}$ collects the first variations of the interactions $U_{L1},\dots,U_{L,L-1}$ relative to the limiting configuration $x_\star\in(\mathbb{S}^{D-1})^L$. From the previous computations we obtain

$$\gamma^{-1}\,\partial_t X_1=a_{L1}X_1+\sum_{j=2}^{L-1}a_{Lj}X_j,$$

$$\begin{aligned}\gamma^{-1}\,\partial_t X_j &= (a_{L1}+a_{j1})X_1+\sum_{k=2}^{j-1}(a_{jk}+a_{Lk})X_k\\ &\quad+\Big(a_{Lj}-\sum_{k=1}^{j-1}a_{jk}\Big)X_j+\sum_{k=j+1}^{L-1}a_{Lk}X_k,\qquad 2\le j\le L-1.\end{aligned}$$

Hence the vector $\vec{X}$ satisfies the linear system

$$\gamma^{-1}\,\partial_t\vec{X}(t)=J(x_\star)\vec{X}(t),$$

where $J(x_\star)\in\mathbb{R}^{(L-1)\times(L-1)}$ is the coefficient matrix of the reduced linearized system with rows $R_i=(J(x_\star))_{i,:}$ defined as follows (all coefficients a_{ij} are evaluated at $x_\star=x_\star(\xi)$):

$$\begin{aligned}R_1&=[a_{L1}\quad a_{L2}\quad a_{L3}\quad\cdots\quad a_{L,L-1}],\\ R_2&=[a_{L1}+a_{21}\quad a_{L2}-a_{21}\quad a_{L3}\quad\cdots\quad a_{L,L-1}],\\ R_3&=[a_{L1}+a_{31}\quad a_{32}+a_{L2}\quad a_{L3}-(a_{31}+a_{32})\quad\cdots\quad a_{L,L-1}],\\ &\vdots\\ R_{L-2}&=\big[a_{L1}+a_{L-2,1}\quad a_{L-2,2}+a_{L2}\quad a_{L-2,3}+a_{L3}\quad\cdots\quad a_{L-2,L-3}+a_{L,L-3}\quad a_{L,L-2}-\textstyle\sum_{k=1}^{L-3}a_{L-2,k}\quad a_{L,L-1}\big],\\ R_{L-1}&=\big[a_{L1}+a_{L-1,1}\quad a_{L-1,2}+a_{L2}\quad\cdots\quad a_{L-1,L-2}+a_{L,L-2}\quad a_{L,L-1}-\textstyle\sum_{k=1}^{L-2}a_{L-1,k}\big].\end{aligned}$$

We now show that the matrix $J(x_\star)$ always possesses at least one positive eigenvalue. Every row of $J(x_\star)$ has sum $\sum_{j=1}^{L-1}a_{Lj}(x_\star)$. Hence

$$J(x_\star)\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}=\left(\sum_{j=1}^{L-1}a_{Lj}(x_\star)\right)\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}.$$

To verify that this eigenvalue corresponds to a tangent direction of the original system, take

$$\widetilde{x}_i = 0 \quad (1 \le i \le L-1), \qquad \widetilde{x}_L = P_\xi(x_1).$$

Since $\xi \cdot S_{BC} x_1 = 0$, we have

$$\widetilde{x}_L \cdot S_{BC} x_1 = P_\xi(x_1) \cdot S_{BC} x_1 = \gamma.$$

The earlier perturbations remain zero, and the linearized last-token equation gives

$$\begin{aligned}\partial_t \widetilde{x}_L &= \left(\sum_{j=1}^{L-1} a_{Lj}(x_\star) \right) (\widetilde{x}_L \cdot S_{BC} x_1) P_\xi(x_1) \\ &= \gamma \left(\sum_{j=1}^{L-1} a_{Lj}(x_\star) \right) \widetilde{x}_L.\end{aligned}$$

Consequently, the tangent linearization of the original system admits the positive eigenvalue

$$\lambda_+ = \gamma \sum_{j=1}^{L-1} a_{Lj}(x_\star) > 0.$$

We now prove that, for almost every initial configuration

$$\vec{x}_0 = (x_2(0), \dots, x_L(0)) \in (\mathbb{S}^{D-1})^{L-1},$$

the trajectory converges with limits in $\{\pm x_1\}$. For $t \ge 0$, set

$$\vec{x}(t) = (x_2(t), x_3(t), \dots, x_L(t)) \in (\mathbb{S}^{D-1})^{L-1}.$$

Define the flow map $\Phi : (\mathbb{S}^{D-1})^{L-1} \to (\mathbb{S}^{D-1})^{L-1}$ by

$$\vec{x}_0 = (x_2, \dots, x_L)(0) \;\mapsto\; \vec{x}(t) = (x_2, \dots, x_L)(t).$$

If $\vec{x}(t) = \Phi^t(\vec{x}_0)$ does not converge to $(x_1, \dots, x_1)$ (up to a reflection of tokens), then by Proposition B.11, we have

$$\langle x_L(t), S_{BC} x_1 \rangle \to 0$$

as $t \to \infty$. Now fix any $\xi \in \mathbb{S}^{D-1}$ such that $\langle \xi, S_{BC} x_1(0) \rangle = 0$. If $\Phi^t(\vec{x}_0) \to (\pm x_1, \cdots, \pm x_1, \xi) = x_\star(\xi)$ as $t \to \infty$, then for any $t \in \mathbb{Q}^+$, we have

$$\Phi^t(\vec{x}_0) \in W^{\text{sc}}_{\text{loc}}(x_\star(\xi)), \tag{31}$$

where $W^{\text{sc}}_{\text{loc}}(x_\star(\xi))$ denotes the local center-stable manifold at $x_\star(\xi)$. Since the linearization at $x_\star(\xi) \in (\mathbb{S}^{d-1})^{L-1}$ admits at least one positive eigenvalue,

$$\lambda_+(\xi) = \gamma \sum_{j=1}^{L-1} a_{Lj}(x_\star(\xi)) > 0,$$

in the tangent direction $(0, \dots, 0, P_\xi(x_1))$, the local stable–center manifold $W^{\text{sc}}_{\text{loc}}(x_\star(\xi))$ has dimension at most $(d-1)(L-1) - 1$. Consequently, it has measure zero in $(\mathbb{S}^{d-1})^{L-1}$. From (31) it follows that

$$\vec{x}_0 \in \bigcup_{t \in \mathbb{Q}^+} \Phi^{-t}(W^{\text{sc}}_{\text{loc}}(x_\star(\xi))),$$

Hence $\vec{x}_0$ belongs to a set of measure zero, namely $N_\xi := \bigcup_{t \in \mathbb{Q}^+} \Phi^{-t}(W^{\text{sc}}_{\text{loc}}(x_\star(\xi)))$. This completes the first goal of the proof.

We now note that for each $\xi \in K$, there exists an open ball $B_\xi = B(x_\star(\xi), r_\xi)$ such that if $\vec{x}(t_1) \in B_\xi \setminus W^{\text{sc}}_{\text{loc}}(x_\star(\xi))$ for some $t_1 > 0$, then there exists $t_2 > t_1$ such that $\vec{x}(t_2) \notin B_\xi$. As a consequence, if $\vec{x}_0 \notin N_\xi$ then $\vec{x}(t)$ does not converge to any point in B_ξ. Moreover, since K is compact, there exist $\xi_1, ..., \xi_m \in K$ such that

$$\{(x_1, ..., x_1, \xi) \in (\mathbb{S}^{D-1})^{L-1} : \xi \in K\} \subset \bigcup_{j=1}^{m} B_{\xi_j}.$$

Hence, if $\vec{x}_0 \notin N := \bigcup_{j=1}^m N_{\xi_j}$ then $\vec{x}(t)$ does not converge to any point of the form $(x_1, ..., x_1, \xi)$, $\xi \in K$.

The proof is complete. □

Before proving Proposition B.9, we establish a useful lemma.

Lemma B.14. *Let $g : [t_0, \infty) \to \mathbb{R}$ be a differentiable function such that g' is continuous. Assume $g(t_0) > \delta$ and*

$$g'(T) > 0 \quad \textit{whenever } g(T) = \delta,$$

then $g(t) \geq \delta$ for all $t \geq t_0$.

Proof. Assume this is not true, since $g(t_0) > \delta$, there exists a smallest time $T > t_0$ such that $g(T) = \delta$. Since g' is continuous, there exists $\epsilon_0 > 0$ such that

$$g'(t) > 0 \quad \text{for all} \quad t \in (T - \epsilon_0, T + \epsilon_0).$$

This implies $g(t) < g(T) = \delta$ for all $t \in (T - \epsilon_0, T)$. By continuity of g and the fact that $g(t_0) > \delta$, there exists $T_1 \in (t_0, T)$ such that $g(T_1) = \delta$, violating the minimality of T. The proof is complete, and hence we have

$$g(t) > \delta \quad \text{for all} \quad t > t_0.$$

□

The following establishes Proposition B.9 in the special case $L = 2$.

Proposition B.15 (Proposition B.9 for $L = 2$)**.** *Assume that $\langle x_1(0), S_{BC} x_1(0) \rangle > 0$. Then*

$$\lim_{t \to \infty} x_2(t) = x_1(0) \quad \textit{whenever} \quad \langle x_2(0), S_{BC} x_1(0) \rangle > 0.$$

Moreover, the following hold:

- *Exponential convergence:*

$$|x_2(t) - x_1(0)| \;\leq\; |x_2(0) - x_1(0)|\, e^{-O(1)\, t},$$

- *Positivity preservation:*

$$\langle x_2(t), S_{BC} x_1(0) \rangle > 0, \qquad \textit{for all } t > 0.$$

Proof. Since

$$|x_2(t) - x_1|^2 = 2\big[1 - \langle x_1, x_2(t) \rangle\big],$$

it suffices to prove that

$$\lim_{t \to \infty} \langle x_2(t), x_1 \rangle = 1.$$

From the dynamical system (27), we have

$$\frac{d}{dt} x_2 = \langle x_2, S_{BC} x_1 \rangle\, a_{21} \big(x_1 - (x_1 \cdot x_2) x_2\big),$$

where $a_{21} > 0$ for all $t > 0$. Define

$$f(t) = \langle x_2(t), x_1 \rangle, \qquad g(t) = \langle x_2(t), S_{BC} x_1 \rangle.$$

A direct computation yields

$$f' = a_{21} g (1 - f^2),$$
$$g' = a_{21} g\, (x_1 \cdot S_{BC} x_1 - f g).$$

Step 1: Positivity of $g(t)$. We claim that

$$g(t) \;\geq\; \min\{\, S_{BC} x_1 \cdot x_1,\; g(0)\}, \quad \text{for all} \quad t > 0. \tag{32}$$

At $t = 0$, we have $g(0) = \langle x_2(0), S_{BC} x_1 \rangle > 0$. Fix any $\delta \in (0, \min\{S_{BC} x_1 \cdot x_1, g(0)\})$. If $g(T) = \delta$ at some time T, then

$$g'(T) = a_{21} \delta \big(x_1 \cdot S_{BC} x_1 - \delta f(T)\big) \;\geq\; a_{21} \delta \big(S_{BC} x_1 \cdot x_1 - \delta\big) > 0.$$

By Lemma B.14, this implies $g(t) > \delta$ for all $t > 0$, proving (32).

Step 2: Convergence of $f(t)$. From (33),

$$f' = a_{21} g (1 - f^2),$$

with $a_{21}, g > 0$. Hence $f'(t) > 0$ whenever $f(t) \in (-1, 1)$, so f is strictly increasing. Therefore, there exists $\theta = \sup_{t\geq 0} f(t) \in [-1, 1]$, with necessarily $\theta > -1$. *Step 3: Exponential convergence.* For large t, we estimate

$$\frac{d}{dt}(1 - f) = -f' = -a_{21}g(1 - f)(1 + f) \lesssim -(1 + \theta)(1 - f(t)).$$

Integrating, we obtain

$$1 - f(t) \leq (1 - f(0))e^{-O(1)\,t}.$$

Thus $f(t) \to 1$ exponentially fast as $t \to \infty$.

Combining Steps 1–3, we conclude that $\lim_{t\to\infty} x_2(t) = x_1(0)$, which completes the proof. □

We are ready to provide the proof of Proposition B.9.

Proof of Proposition B.9. We proceed by induction on L. The case $L = 2$ has already been established in Proposition B.15. Assume now that $L \geq 3$ and that the proposition holds for $L - 1$, i.e.,

$$\sum_{i=1}^{L-1} |x_i(t) - x_1| \lesssim e^{-O(1)t},$$

and moreover $\langle x_i(t), S_{BC}x_j(t)\rangle > 0$ for all $i, j = 1, \ldots, L - 1$. We will show that $x_L(t) \to x_1$ with exponential decay rate. The proof of Proposition B.9 is divided into several claims below.

Claim 1. *For any i, j, we have $U_{ij}(t) = \langle x_i(t), S_{BC}x_j(t)\rangle > 0$ for all $t > 0$.*

Proof of Claim 1. Indeed, by the induction hypothesis in Proposition B.9, it suffices to consider U_{Lj} with $1 \leq j \leq L$. We define the set

$$\mathcal{S} = [0, \infty)^L \subset \mathbb{R}^L.$$

By the assumption of Proposition B.9, we have

$$(U_{L1}(0), U_{L2}(0), \ldots, U_{LL}(0)) \in \mathcal{S}.$$

We now show that the trajectory

$$(U_{L1}(t), U_{L2}(t), \ldots, U_{LL}(t))$$

remains inside $\mathcal{S}$ for all $t > 0$. Suppose, for contradiction, that the trajectory leaves $\mathcal{S}$. Since it is bounded within a box, it must cross the boundary of $\mathcal{S}$ at some finite time. There are two possible scenarios:

- There exists a smallest exit time $T > 0$ and a unique index $i_\star \in [1, L]$ such that
$$U_{Li_\star}(T) = 0, \qquad U_{Li_\star}(t) > 0 \quad \text{for all } t \in (0, T),$$
while
$$\sum_{k\neq i_\star} U_{Lk}^2(T) > 0.$$
That is, the trajectory exits $\mathcal{S}$ through the $i_\star^{\text{th}}$ component $U_{Li_\star}$, rather than exiting all components simultaneously at time T.
- There exists a smallest exit time $T > 0$ such that $U_{Li}(T) = 0$ for all i, with $U_{Li}(t) > 0$ for all $t \in (0, T)$. In this case, the trajectory exits $\mathcal{S}$ all at once at time T.

We now consider the first case, when the trajectory exits $\mathcal{S}$ only at the component $i_\star$. By Lemma B.10, for $i_\star > 1$ we have

$$\begin{aligned} U'_{Li_\star}(T) &= \sum_{k\leq L-1} a_{Lk}U_{Lk}(T) + \sum_{k\leq i_\star-1} a_{jk}U_{jk}(T)\left[U_{Lk}(T) - X_{jk}(T)U_{Li_\star}(T)\right] \\ &\geq \sum_{k\leq L-1} a_{Lk}U_{Lk}(T) \\ &> 0, \end{aligned}$$

since at least one of the $U_{Lk}(T)$ remains strictly positive.

For $i_\star = 1$, we obtain

$$\begin{aligned} U'_{L1}(T) &= \sum_{j \le L-1} a_{Lj} U_{Lj}(T) \left[U_{j1}(T) - X_{Lj}(T) U_{L1}(T) \right] \\ &= \sum_{j \le L-1} a_{Lj} U_{Lj}(T) U_{j1}(T) \\ &> 0. \end{aligned}$$

Thus, in both cases we conclude that $U'_{Li_\star}(T) > 0$.

By continuity of $U'_{Li_\star}$, there exists $\epsilon_0 > 0$ such that

$$U'_{Li_\star}(t) > 0 \quad \text{for all } t \in (T - \epsilon_0, T + \epsilon_0).$$

Hence $U_{Li_\star}(t)$ is strictly increasing on $(T - \epsilon_0, T + \epsilon_0)$, which implies

$$U_{Li_\star}(t) < U_{Li_\star}(T) = 0 \quad \forall\, t \in (T - \epsilon_0, T).$$

Since $U_{Li_\star}(0) > 0$, there must exist some $T_1 \in (0, T - \epsilon_0)$ such that $U_{Li_\star}(T_1) = 0$. Moreover, $U_{Li}(t) > 0$ for all $t \in [0, T_1] \subset [0, T]$. This contradicts the minimality of T. Therefore, the first case cannot occur.

Now we consider the second case, when the trajectory

$$(U_{L1}(t), U_{L2}(t), \dots, U_{LL}(t))$$

exits $\mathcal{S}$ simultaneously at some time T. In this situation, Lemma B.10 implies that

$$U'_{Lk}(T) = 0 \qquad \text{for all } k \in [1, L].$$

Thus the trajectory passes through the equilibrium point $0 \in \mathbb{R}^L$ at time $t = T$. Choose $\epsilon_0 > 0$ sufficiently small, and consider the flow ODE for $(U_{L1}(t), U_{L2}(t), \dots, U_{LL}(t))$ on the interval $I = (T - \epsilon_0, T + \epsilon_0)$. On this interval we observe two distinct flows satisfying the same system with the same initial data at $t = T$: the trivial solution $(U_{L1}(t), \dots, U_{LL}(t)) \equiv 0$ and the trajectory $(U_{L1}(t), \dots, U_{LL}(t))$ with $U_{Li}(t) > 0$ for all $t \in (T - \epsilon_0, T)$. This contradicts the uniqueness of solutions to the differential equations on I.

Hence the trajectory cannot exit the boundary of $\mathcal{S}$ simultaneously at all components. This completes the proof for the second case and concludes the proof of Claim 1.

□

Claim 2. *There exist a time $T > 0$ and a constant $c_T > 0$ such that $U_{ij}(t) \ge c_T$ for all $t \ge T$ and all $1 \le i, j \le L$.*

Proof of Claim 2. For $i, j \le L - 1$, we observe that

$$\lim_{t \to \infty} U_{ij}(t) = \lim_{t \to \infty} \langle x_i(t), S_{BC} x_j(t) \rangle = \langle x_1, S_{BC} x_1 \rangle > 0.$$

Therefore, the claim holds for such i and j. It remains to consider the case of $U_{Lj}(t)$ for $1 \le j \le L$. Now, for any $i, j \le L - 1$, there exists a time $T > 0$ such that

$$U_{ij}(t) \ge 0.9 \left(S_{BC} x_1 \cdot x_1 \right), \quad \forall t \ge T.$$

We aim to show that for $2 \le j \le L$,

$$U_{Lj}(t) \ge \min \Big\{ 0.8 \left(S_{BC} x_1 \cdot x_1 \right), U_{L1}(T), \dots, U_{Lj}(T) \Big\}, \quad \forall t \ge T.$$

This will be established by induction on j.

For the base case $j = 1$, using Lemma B.10, we have

$$U'_{L1}(t) = \sum_{k \le L-1} a_{Lk} U_{Lk} \big[U_{k1} - X_{Lk} U_{L1} \big].$$

Choose any $\delta > 0$ satisfying

$$\delta < \min\left\{0.8\left(S_{BC}x_1 \cdot x_1\right),\, U_{L1}(T)\right\}.$$

We claim that

$$U_{L1}(t) \geq \delta, \quad \forall t \geq T. \tag{33}$$

Indeed, by Lemma B.14, it suffices to check that whenever $U_{L1}(T_\star) = \delta$, we have

$$\begin{aligned} U'_{L1}(T_\star) &= \sum_{k \leq L-1} a_{Lk} U_{Lk}(T_\star)\left[U_{k1}(T_\star) - X_{Lk}(T_\star)\,\delta\right] \\ &\geq \sum_{k \leq L-1} a_{Lk} U_{Lk}(T_\star)\left[0.9\left(S_{BC}x_1 \cdot x_1\right) - \delta\right] > 0. \end{aligned}$$

This establishes (33) for $j = 1$. Now, assume that for all $k \leq j-1$,

$$\inf_{t \geq T} U_{Lk}(t) \geq \min\left\{0.8\left\langle S_{BC}x_1, x_1\right\rangle,\, U_{L1}(T), \ldots, U_{Lk}(T)\right\}.$$

We will show that

$$\inf_{t \geq T} U_{Lj}(t) \geq \min\left\{0.8\left\langle S_{BC}x_1, x_1\right\rangle,\, U_{L1}(T), \ldots, U_{Lj}(T)\right\}.$$

To this end, fix any δ satisfying

$$0 < \delta < \min\left\{0.8\left\langle S_{BC}x_1, x_1\right\rangle,\, U_{L1}(T), \ldots, U_{Lj}(T)\right\}.$$

We will prove that $U_{Lj}(t) \geq \delta$ for all $t \geq T$.

Recall from Lemma B.10 that

$$U'_{Lj} = \sum_{k \leq L-1} a_{Lk} U_{Lk}\left[U_{kj} - X_{Lk}U_{Lj}\right] + \sum_{k \leq j-1} a_{jk} U_{jk}\left[U_{Lk} - X_{jk}U_{Lj}\right].$$

Suppose $U_{Lj}(T_\star) = \delta$ for some $T_\star \geq T$. Then, we have

$$\begin{aligned} U'_{Lj}(T_\star) &\geq \sum_{k \leq L-1} a_{Lk} U_{Lk}(T_\star)\left[U_{kj}(T_\star) - \delta\right] + \sum_{k \leq j-1} a_{jk} U_{jk}(T_\star)\left[U_{Lk}(T_\star) - \delta\right] \\ &\geq \sum_{k \leq j-1} a_{jk} U_{jk}\left[U_{Lk}(T_\star) - \delta\right] > 0. \end{aligned}$$

The claim then follows by an application of Lemma B.14. □

Claim 3. *There exist a time $T > 0$ and a constant $c_T > 0$ such that*

$$1 + X_{L1}(t) \geq c_T \quad \textit{for all } t \geq T. \tag{34}$$

Proof of Claim 3. From Lemma B.10, we have

$$\begin{aligned} X'_{L1}(t) = a_{L1}U_{L1}(1 - X_{L1}^2) \\ + \sum_{j=2}^{L-1} a_{Lj} U_{Lj}\left[X_{j1} - X_{Lj}X_{L1}\right]. \end{aligned}$$

Fix $\varepsilon \in (0,1)$ sufficiently small. Since $X_{j1}(t) \to 1$ as $t \to \infty$ for all $j \leq L-1$, there exists $T > 0$ such that

$$X_{j1}(t) > 1 - \varepsilon \quad \text{for all } t \geq T.$$

If $X_{L1}(t) \to 1$ as $t \to \infty$, then the claim holds trivially. Suppose, to the contrary, that $X_{L1}(t) \not\to 1$ as $t \to \infty$. Then there exists an increasing sequence $\{t_n\}_{n=1}^{\infty} \subset [T, \infty)$ such that

$$X_{L1}(t_n) \leq 1 - \varepsilon.$$

We now show that not all t_n can satisfy $X_{L1}(t_n) = -1$. Assume for contradiction that

$$X_{L1}(t_n) = \left\langle x_L(t_n), x_1\right\rangle = -1 \quad \text{for all } n \geq 1.$$

This implies $x_L(t_n) = -x_1$ for all $n \geq 1$. Consequently,

$$U_{L1}(t_n) = \langle x_L(t_n), S_{BC} x_1 \rangle = -\langle x_1, S_{BC} x_1 \rangle < 0,$$

which contradicts the fact that $U_{L1}(t) > 0$ for all $t > 0$.

Therefore, there exists a time $t_\star \geq T$ such that

$$X_{L1}(t_\star) \leq 1 - \varepsilon, \quad \text{and} \quad 1 + X_{L1}(t_\star) > 0.$$

We now prove that

$$X_{L1}(t) \geq \min\{X_{L1}(t_\star),\, 1 - \varepsilon\} \quad \text{for all } t \geq t_\star,$$

which will imply inequality (34).

To this end, fix any $\delta \in (-1, \min\{X_{L1}(t_\star),\, 1 - \varepsilon\})$. We will show that $X_{L1}(t) \geq \delta$ for all $t \geq t_\star$ by applying Lemma B.14.

Indeed, suppose $X_{L1}(t) = \delta$ for some $t \geq t_\star$. Then, using the earlier lower bounds and the fact that $U_{ij}(t) > 0$, we obtain

$$X'_{L1}(t) \geq a_{L1} U_{L1}(1 - \delta^2) + \sum_{j=2}^{L-1} a_{Lj} U_{Lj} \left[(1 - \varepsilon) - \delta\right] > 0.$$

This verifies the condition of Lemma B.14, completing the proof of Claim 3. □

Claim 4. *There exist a time $T > 0$ and a constant $c_T > 0$ such that for all $t \geq T$,*

$$1 - X_{L1}(t) \lesssim e^{-c_T t}.$$

Proof of Claim 4. By the preceding claims and the induction hypothesis, there exist $T > 0$ and $c_T > 0$ such that for all $t \geq T$,

$$1 + X_{L1}(t) \geq c_T, \quad U_{ij}(t) \geq c_T, \quad \max_{2 \leq j \leq L-1} (1 - X_{j1}(t)) \lesssim e^{-c_T t}.$$

For $t \geq T$, we compute

$$\begin{aligned}
\frac{d}{dt}\left[1 - X_{L1}(t)\right] &= -X'_{L1}(t) = -a_{L1} U_{L1}(1 - X_{L1}^2) - \sum_{j=2}^{L-1} a_{Lj} U_{Lj} \left[X_{j1} - X_{Lj} X_{L1}\right] \\
&= -a_{L1} U_{L1}(1 - X_{L1}^2) - \sum_{j=2}^{L-1} a_{Lj} U_{Lj}(X_{j1} - 1) - \sum_{j=2}^{L-1} a_{Lj} U_{Lj}(1 - X_{Lj} X_{L1}) \\
&= -a_{L1} U_{L1}(1 - X_{L1})(1 + X_{L1}) + \sum_{j=2}^{L-1} a_{Lj} U_{Lj}(1 - X_{j1}) - \sum_{j=2}^{L-1} a_{Lj} U_{Lj}(1 - X_{Lj} X_{L1}).
\end{aligned}$$

Note that the third term is non-positive, since $|X_{Lj} X_{L1}| \leq 1$, so we may drop it to obtain an upper bound:

$$\begin{aligned}
\frac{d}{dt}\left[1 - X_{L1}(t)\right] &\leq -a_{L1} U_{L1}(1 + X_{L1})(1 - X_{L1}) + \sum_{j=2}^{L-1} a_{Lj} U_{Lj}(1 - X_{j1}) \\
&\leq -c_1(1 - X_{L1}(t)) + C_0 e^{-c_T t},
\end{aligned}$$

for some positive constants $c_1, C_0 > 0$, where we have used the lower bounds $U_{L1}(t) \geq c_T$, $1 + X_{L1}(t) \geq c_T$, and the exponential decay $1 - X_{j1}(t) \lesssim e^{-c_T t}$.

Applying the Gronwall inequality to this differential inequality yields

$$1 - X_{L1}(t) \lesssim e^{-\min\{c_1, c_T\} t},$$

which completes the proof of Claim 4. □

Proposition B.9 now follows from Claims 1–4. □

We end this subsection by the proof of Theorem B.8.

Proof of Theorem B.8. By Proposition B.15 and Proposition B.11, the second token converges to $\pm x_1(0)$. Moreover, by Proposition B.11, there exists a minimal integer $k \in [3, L]$ such that, after possibly replacing each x_i by its reflection $\widetilde{x}_i = -x_i$ for $i \leq k-1$, one of the following two cases holds:

$$\text{(a)} \sum_{i=2}^{k-1} |x_i(t) - x_1(0)| + \left|\langle x_k(t), S_{BC} x_1(0)\rangle\right| \lesssim e^{-O(1)t},$$

$$\text{(b)} \sum_{i=2}^{k-1} |x_i(t) - x_1(0)| + |x_k(t) - x_1(0)| \lesssim e^{-O(1)t}.$$

If the case (a) holds for some $3 \leq k \leq L$ then we have

$$\left|\frac{dx_k}{dt}\right| \lesssim \left|\langle x_k(t), S_{BC} x_1(0)\rangle\right| \lesssim e^{-O(1)t}.$$

This implies that $x_k(t)$ converges to some unit vector $\xi \in \langle S_{BC} x_1(0)\rangle^{\perp}$. By Proposition B.13, the set N_k of initial configurations for which such a limit ξ occurs is of measure zero in the phase space $(\mathbb{S}^{D-1})^{k-1}$. Denote

$$N = \bigcap_{k=3}^{L} N_k \times (\mathbb{S}^{D-1})^{L-k}.$$

We have N is a set of measure zero in $(\mathbb{S}^{D-1})^{L-1}$. Moreover, the case (b) holds for $k = L$ if $(x_2(0), ..., x_L(0) \notin N$.

This completes the proof. □

B.3 A Generalization to the Time-dependent Parameters Setting

In this section, we consider the problem (10) in the case where S_{BC}, α and a are time-dependent. Moreover, we assume that $S_{BC}(t), \alpha(t)$ and $a(t)$ converge exponentially to S_∞, α_∞ and a_∞, respectively. As a sequence, there exists a constant $c_0 > 0$ such that the functions

$$\tilde{S}(s) = \begin{cases} S_{BC}\left(-\frac{\ln|s|}{c_0}\right), s \neq 0, \\ S_\infty, \quad s = 0, \end{cases} \qquad \tilde{\alpha}(s) = \begin{cases} \alpha\left(-\frac{\ln|s|}{c_0}\right), s \neq 0, \\ \alpha_\infty, \quad s = 0, \end{cases} \quad \text{and}$$

$$\tilde{a}(s) = \begin{cases} a\left(-\frac{\ln|s|}{c}\right), s \neq 0, \\ a_\infty, \quad s = 0, \end{cases},$$

are differentiable on $\mathbb{R}$. In addition, the gradients of these functions vanish at $t = 0$.

For every $1 \leq j < l \leq L$, $x = (x_1, ..., x_L) \in (\mathbb{S}^{D-1})^L$ and $s \in \mathbb{R}$, we denote

$$\tilde{A}_{lj}(s, x) = x_l^\top \tilde{S}(s) x_j \cdot \tilde{\Delta}_s(x_l) e^{-\tilde{a}(s)\sum_{k=j+1}^{l} \tilde{\Delta}_s(x_k)},$$

where $\tilde{\Delta}_s(x_l) = \ln(1 + e^{\tilde{\alpha}(s)\cdot x_l})$.

By the same argument as for Theorem 3.6, we obtain the following result, whose proof we omit.

Theorem B.16. *Consider the problem*

$$\begin{cases} \frac{d}{dt}x_0(t) = -c_0 x_0(t), \\ \frac{d}{dt}x_l(t) = \sum_{j=1}^{l-1} \tilde{A}_{lj}(x_0(t), \mathbf{x}(t)) \cdot \mathbf{P}_{x_l(t)}(x_j(t)), \quad l = 1, \ldots, L, \\ x_0(0) > 0, \mathbf{x}(0) = (x_1(0), \ldots, x_L(0)) \in (\mathbb{S}^{D-1})^L. \end{cases} \tag{35}$$

Let $(x_0(t), \mathbf{x}(t) = (x_1(t), \dots, x_L(t))$ be a solution of this problem. Assume that $\tilde{S}(t)$ is symmetric for every t and S_∞ is positive definite. Then, there exists a zero-measurable subset N_0 of $(0, \infty)$ satisfying the following property: if $x_0(0) \in (0, \infty) \setminus N_0$ and $x_1(0) \in \mathbb{S}^{D-1}$ then there exists a zero-measurable subset Ω of $(\mathbb{S}^{D-1})^{L-1}$ such that if the initial point $(x_2(0), ..., x_L(0))$ does not belong in Ω then all tokens converge exponentially to $\pm x_1(0)$.

For every $1 \le j < l \le L$, $x = (x_1, ..., x_L) \in (\mathbb{S}^{D-1})^L$ and $t \ge 0$, we denote

$$A_{lj}(t, x) = x_l^\top S_{BC}(t) x_j \cdot \Delta_t(x_l) e^{-a(t) \sum_{k=j+1}^{l} \Delta_t(x_k)},$$

where $\Delta_t(x_l) = \ln(1 + e^{\alpha(t) \cdot x_l})$. Note that if $\frac{d}{dt} x_0(t) = -c_0 x_0(t)$ then $x_0(t) = c\, e^{-c_0 t}$ with $c = x_0(0) > 0$, and it follows that

$$\tilde{A}_{lj}(x_0(t), x) = A_{lj}(t - \frac{\ln c}{c_0}, x).$$

The following result is a direct consequence of Theorem B.16.

Theorem B.17. *For each constant $c \in \mathbb{R}$, we consider the following problem:*

$$\begin{cases} \frac{d}{dt} x_l(t) = \sum_{j=1}^{l-1} A_{lj}(t + c, \mathbf{x}(t)) \cdot \mathbf{P}_{x_l(t)}(x_j(t)), \quad l = 1, \dots, L, \\ \mathbf{x}(0) = (x_1(0), \dots, x_L(0)) \in (\mathbb{S}^{D-1})^L. \end{cases} \tag{36}$$

Let $(\mathbf{x}(t) = (x_1(t), \dots, x_L(t))$ be a solution of this problem. Assume that $S_{BC}(t)$ is symmetric for every t and S_∞ is positive definite. Then, there exists a zero-measurable subset N_0 of $(0, \infty)$ satisfying the following property: if $c \in \mathbb{R} \setminus N_0$ and $x_1(0) \in \mathbb{S}^{D-1}$ then there exists a zero-measurable subset Ω of $(\mathbb{S}^{D-1})^{L-1}$ such that if the initial point $(x_2(0), ..., x_L(0))$ does not belong in Ω then all tokens converge exponentially to $\pm x_1(0)$.

C Experiment - Additional Details

C.1 Experimental Setup

We adopt the threshold-based metric of [83], adapted to the hidden attention structure of SSMs. For a given layer $\ell \in \{1, \dots, L\}$ and channel c, let $A^{\ell,c} \in \mathbb{R}^{T \times T}$ denote the hidden attention matrix, where T is the sequence length. The importance of token k in (ℓ, c) is defined as

$$\alpha_k^{\ell,c} = \frac{1}{T - k + 1} \sum_{i=k}^{T} A_{i,k}^{\ell,c}. \tag{37}$$

We say that channel (ℓ, c) exhibits an attention sink at the first token if

$$\alpha_1^{\ell,c} > \varepsilon, \tag{38}$$

where ε is a fixed threshold (we set $\varepsilon = 0.3$ in our experiments).

The *layer-wise sink rate* is given by

$$\mathrm{Sink}_\ell = \frac{1}{H_s} \sum_{c \in \mathcal{C}_\ell} \mathbf{1}\left(\alpha_1^{\ell,c} > \varepsilon\right), \tag{39}$$

where $\mathcal{C}_\ell$ is a subset of channels sampled from layer ℓ, and $H_s = |\mathcal{C}_\ell|$. Finally, the *model-level sink score* is

$$\mathrm{Sink} = \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{Sink}_\ell. \tag{40}$$

C.2 Evaluation

To empirically estimate these quantities, we evaluate the Falcon-Mamba-7B model [90] on the dataset of [83], which consists of 170 samples of length $T = 16$. Since each layer contains 4096 channels, we uniformly sample $H_s = 200$ channels per layer to estimate $\mathrm{Sink}\ell$ and the overall Sink. We then report $\mathrm{Sink}\ell$ as a function of layer depth and the aggregate Sink across tokens in Figure 4. For visualization, we additionally average the softmax-normalized attention matrices $\mathrm{softmax}(A^{\ell,c})$ across the samples to obtain per-layer attention maps, as shown in Figure 3.

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 claims made in the abstract and introduction accurately reflect the paper's theoretical contributions and scope. They are supported by the main theorems (Theorems 3.3, 3.5, 3.6, and 3.9) and are stated under explicit assumptions.

   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: Limitations of this work is discussed in the Conclusion and Limitations section.

   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 do provide the full set of assumptions and complete proofs for the main theorems.

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 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 conclusions of the paper.

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 provide open access to the data and code in Supplementary Material, with sufficient instructions to faithfully reproduce the main experimental results.

   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 specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results in Section 5.

   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: [Yes]

   Justification: The paper reports suitably and correctly defined error bars.

   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 provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   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 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: [N/A]

    Justification: This work has no societal impacts.

    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: The paper does not release any data, pre-trained models, image generators, or scraped datasets. Therefore, safeguards for responsible release are not applicable.

    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 is primarily theoretical and does not rely on external datasets or pre-trained models. All prior works and relevant external resources used in the paper are properly cited, and no restricted assets are 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: [N/A]

    Justification: The paper does not introduce new assets such as datasets, trained models, or software packages. Therefore, asset documentation is not applicable.

    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 experiments or research with human subjects. Therefore, participant instructions, screenshots, and compensation details are not applicable.

    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 study participants or human-subject research. Therefore, participant risks, risk disclosure, and IRB approval are 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 are not used as an important, original, or non-standard component of the core methodology. Any use of LLMs is limited to writing, editing, or formatting assistance and does not affect the scientific content, rigor, or originality of the research.

    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.