

On the Diverse Dynamical Behaviors Arising in Deep Linear Transformers

Sixu Li^{*1}, Thomas Jacob Maranzatto^{†2}, Jan Peszek^{‡3}, Trevor Teolis^{§4}, Semih Akkoc^{¶2},
Konstantin Riedl^{||5}, Sennur Ulukus^{**2}, and Nicolás García Trillos^{††1}

¹University of Wisconsin–Madison, Department of Statistics

²University of Maryland, Department of Electrical and Computer Engineering

³University of Warsaw, Institute of Applied Mathematics and Mechanics

⁴Rice University, Department of Computational and Applied Mathematics & Operations Research

⁵University of Oxford, Mathematical Institute

July 22, 2026

Abstract

We study the inference-time behavior of deep linear encoder-only transformers through the lens of interacting particle systems. In this perspective, tokens are modeled as particles that interact dynamically through successive linear self-attention layers. We show that in embedding dimension two, for any key, query, and value matrices, the dynamics can be reformulated as a generalized Kuramoto-type model with pure second-harmonic coupling. This formulation is amenable to Watanabe–Strogatz theory which reveals the dynamics are intrinsically low-dimensional regardless of the parameter matrices. For a class of token initializations associated with the Ott–Antonsen (OA) manifold, we show that the parameter matrices induce a diverse variety of long-time behaviors in linear transformers, including clustering, oscillations, and bifurcations. The oscillations and bifurcations are characterized by uncovering a hidden Hamiltonian structure in the dynamics. By establishing a structural stability result, we further show that dynamics initialized near the OA manifold exhibit the same long-time behavior as those initialized exactly on the manifold. Motivated by our theory in dimension two, we conduct numerical experiments for analogous parameter regimes in higher-dimensional transformers. Our numerical experiments suggest that the long-time behaviors characterized in our theoretical results persist in higher dimensions.

Keywords: transformers, linear self-attention, interacting particle systems, Kuramoto model, low-dimensional structure, Watanabe–Strogatz transformation, Ott–Antonsen ansatz.

AMS subject classifications: 34C15, 34D06, 35Q83, 82C22, 37C20

^{*}Email: sli739@wisc.edu

[†]Email: tmaran@umd.edu

[‡]Email: j.peszek@uw.edu.pl

[§]Email: tt111@rice.edu

[¶]Email: akkoc@umd.edu

^{||}Email: Konstantin.Riedl@maths.ox.ac.uk

^{**}Email: ulukus@umd.edu

^{††}Email: garciatrillo@wisc.edu

Contents

1	Introduction	2
1.1	Contributions	4
1.2	Related Works	6
1.3	Outline	7
2	Linear Self-Attention Model on the Circle and its Low-Dimensional Structure	8
2.1	Linear Self-Attention Model on the Circle	8
2.2	Watanabe–Strogatz Transformation and Ott–Antonsen Ansatz	10
2.2.1	Watanabe–Strogatz Transformation	10
2.2.2	Ott–Antonsen Ansatz	12
3	Long-Time Behavior Within the Ott–Antonsen Ansatz: Case Studies	14
3.1	Case 1: $V = I$, A Arbitrary	15
3.2	Case 2: $A = I$, V Symmetric	17
3.3	Case 3: Non-commuting A and V	18
3.4	Case 4: Hidden Hamiltonian Dynamics	20
4	Long-Time Behavior Beyond the Ott–Antonsen Ansatz: Stability Analysis	22
4.1	Stability of the WS Dynamics Beyond the OA-manifold	23
4.2	Long-Time Behavior	26
5	Experiments	29
5.1	Case 1: $V = I$, A Arbitrary	30
5.2	Case 2: $A = I$, V Symmetric	30
5.3	Case 3: Non-commuting A and V	32
5.4	Case 4: Hidden Hamiltonian Dynamics	34
6	Conclusion	36
	Appendix	45

1 Introduction

The transformer architecture, through its central role in modern large language and computer vision models [17, 37, 40, 99, 100] and, more broadly, in state-of-the-art foundation models [15], has become a fundamental component of contemporary machine learning and artificial intelligence. At its core lies the self-attention mechanism [103], which enables the model to capture long-range, context-aware dependencies within the input data. This mechanism distinguishes transformers from earlier neural network architectures such as feedforward [62] or recurrent neural networks [54], where semantic interactions within the data are either absent or propagated sequentially through a number of hidden states. While the capabilities of transformer models have experienced remarkable improvements in performance since their introduction, apart from phenomenological explanations it remains mostly unclear how these large-scale models process data, what the internal transformation of the data leads to, and, more generally, why these architectures turn out to be highly capable for the wide ranges of tasks they are employed for.

This paper is motivated by the general goal of developing a mathematical understanding of transformer architectures, and specifically the self-attention mechanism. Toward that goal, and as in recent theoretical works on the mathematical analysis of the inference dynamics of encoder-only transformers [5, 19–21, 23, 48, 92], we consider an idealized version of a deep encoder-only transformer and view the evolution of a sequence of input tokens (prompt) as a continuous-time process described by a system of coupled ODEs of the form

$$\dot{x}_k(t) = P_{x_k(t)}^\perp \left(\frac{1}{Z_k(t)} \sum_{j=1}^n h(\beta \langle Qx_k(t), Kx_j(t) \rangle) Vx_j(t) \right), \quad k = 1, \dots, n. \quad (1.1)$$

Here, $x_k(t) \in \mathbb{S}^{d-1}$ is interpreted as the output of the t -th self-attention layer for the k -th token representation, which we will thereafter also refer to as the k -th particle. The matrices $Q, K \in \mathbb{R}^{d_{\text{int}} \times d}$, and $V \in \mathbb{R}^{d \times d}$ (the query, key, and value parameters of the transformer) are real-valued matrices assumed to have been previously determined at the moment of training. Z_k is a normalization factor that in principle can be token dependent; different choices of the normalization factor lead to different, although in many cases qualitatively similar, models, see, e.g., [21, 48]. The operator $\mathbf{P}_x^\perp$ denotes the projection onto the tangent space $T_x \mathbb{S}^{d-1}$ of the sphere at a point $x \in \mathbb{S}^{d-1}$, i.e., $\mathbf{P}_x^\perp y := y - \langle x, y \rangle x$. This projection map ensures that the dynamics (1.1) remains on the sphere and can be regarded as a model for layer normalization (in particular, RMSNorm), see, e.g., [21, Section 1.2]. The function $h : \mathbb{R} \mapsto \mathbb{R}$ is a suitable attention kernel that may take a variety of forms. Most popularly, h takes exponential [103] and linear forms [59], which lead (provided suitable normalization) to softmax and linear self-attention, respectively. For simplicity, we use the notation $A = Q^\top K$ in the sequel.

After formulating the transformer dynamics (1.1), a central question is to characterize its long-time behavior. More specifically, we ask the following question:

For which choices of the weight matrices A and V does the associated dynamics converge to consensus and stabilize, and under which choices can other long-time behaviors emerge?

By answering this question, researchers hope to clarify the structure and qualitative properties of the token representations produced after repeated application of self-attention layers.

One class of settings in which this question is relatively well-understood is when (1.1) admits a Wasserstein gradient flow structure on the space of probability measures. In particular, [21, 48] study the case $h(y) = \exp(y)$ and shows that, if $A^\top = A$ and $V = cA$ with some constant $c \in \mathbb{R}$, then (1.1) can be written as a Wasserstein gradient flow for a suitable energy functional. A further analysis of the corresponding energy landscape shows that, under additional assumptions on A and V , for instance $A = V = I_d$, the dynamics exhibits clustering/synchronization behavior, in the sense that all token representations converge in the long-time limit to a single point.

In practical transformer models, however, the matrices A and V do not necessarily satisfy the structural assumptions required for the Wasserstein gradient flow formulation. This naturally leads to the following question:

What is the long-time behavior of the transformer dynamics (1.1) under more general choices of the weight matrices A and V ?

In the present work, we use a different analytical framework, distinct from the Wasserstein gradient flow approach, that enables us to study the transformer dynamics (1.1) for general matrices A and V , albeit with a different choice of attention kernel h . In particular, we focus on the *linear* self-attention model (LSA), corresponding to the choice $h(y) = y$. This model has received considerable attention in recent years given its advantage on fast inference [30, 59, 83, 93, 94]. It is also particularly well-suited for the analytical approach developed in this paper, for reasons that will become clear later. In this case, the model (1.1) takes the form

$$\dot{x}_k(t) = \mathbf{P}_{x_k(t)}^\perp \left(\frac{1}{n} \sum_{j=1}^n \beta \langle x_k(t), Ax_j(t) \rangle Vx_j(t) \right), \quad k = 1, \dots, n, \quad (\text{LSA})$$

where we have set the normalization factor Z_k equal to n for all tokens, following [48, Equation (USA)]. Without loss of generality, we henceforth set $\beta = 1$, since this parameter can be absorbed into either A or V .

As the (USA) model studied in [21, 48], that is, model (1.1) with $h(y) = \exp(y)$, the model (LSA) also admits a Wasserstein gradient flow formulation when $A^\top = A$ and $V = cA$ for some constant c . A rigorous justification is given in Lemma B.1. However, our goal is to go beyond the restrictive structural assumptions required by the Wasserstein gradient flow framework. The key observation is that, at least in the case of dimension $d = 2$, the Fourier representation of (LSA) takes the form of a *pure* second-harmonic; see (2.4) and (2.5). This type of interaction has an intrinsic low-dimensional structure, which in turn allows us to apply dimensionality-reduction techniques developed in the study of the Kuramoto model [79, 96], a classical mathematical model for synchronization. By “low-dimensional structure” we mean that both the full n -particle system and the mean-field PDE which will be introduced shortly are completely governed by a small number of effective ODEs.

To make this idea more precise, we represent each token x_k as a complex number on the unit circle, written as $e^{i\theta_k}$. In these variables, the equations of motion (LSA) become

$$\dot{\theta}_k(t) = b(\theta_k(t), \mathcal{R}_2^n(t)), \quad k = 1, \dots, n, \quad (1.2)$$

for a function b that depends implicitly on the matrices A and V ; see Theorem 2.1. Importantly, the particles interact only through the *second order parameter* $\mathcal{R}_2^n(t)$ of the empirical distribution of tokens f_t^n ,

$$f_t^n = f^n(t, \bullet) := \frac{1}{n} \sum_{k=1}^n \delta_{\theta_k(t)}$$

and

$$\mathcal{R}_2^n(t) := \int_0^{2\pi} e^{i2\theta} f_t^n(d\theta) = \frac{1}{n} \sum_{j=1}^n e^{i2\theta_j(t)}. \quad (1.3)$$

An analogous characterization continues to hold for the dynamics of a “typical” particle in the mean-field limit, that is, as $n \rightarrow \infty$ and under the assumption that the initial tokens are drawn from a fixed distribution f_0 . More precisely, one may consider the characteristic flow

$$\dot{\Theta}(t) = b(\Theta(t), \mathcal{R}_2(t))$$

to describe the evolution of this typical particle, where $f := \text{Law}(\Theta)$ solves the continuity equation

$$\partial_t f_t + \partial_\theta (b(\bullet, \mathcal{R}_2(t)) f_t) = 0, \quad \text{with } \mathcal{R}_2(t) = \int_0^{2\pi} e^{i2\theta} f_t(d\theta).$$

The above formulation shows that each token is influenced by the others only through the order parameter $\mathcal{R}_2$, highlighting the central role of this quantity in the analysis of transformer dynamics. Moreover, $\mathcal{R}_2$ quantifies the degree of synchronization of the system. Indeed, the modulus $|\mathcal{R}_2(f)|$ is close to 1 if and only if the distribution f is concentrated near two antipodal points; see Figure 1 for an illustration. This observation motivates tracking the evolution of the order parameter $\mathcal{R}_2$.

The behavior of $\mathcal{R}_2$ becomes considerably more tractable for a special class of initialization of the angle Θ . More precisely, we prove that there exists a *finite-dimensional* submanifold $\mathcal{M}_{\text{OA}}$ of the space of probability measures over $\mathbb{T}^1 := \mathbb{R}/2\pi\mathbb{Z}$, referred to as the *Ott–Antonsen (OA) manifold* [79], that is *invariant* under the dynamics of the double-angle variable $\Xi := 2\Theta$, interpreted modulo 2π . Furthermore, within this submanifold, the dynamics of Ξ are completely determined by a single ODE of the order parameter $\mathcal{R}_2$ in the complex plane. This ODE takes the form

$$\dot{\mathcal{R}}_2(t) = 2i \left(B(\mathcal{R}_2(t)) (\mathcal{R}_2(t))^2 + C(\mathcal{R}_2(t)) \mathcal{R}_2(t) + \overline{B}(\mathcal{R}_2(t)) \right), \quad (1.4)$$

where the functions B and C are linear in $\mathcal{R}_2$ and its complex conjugate, with coefficients depending on the matrices A and V ; see (A.4).

This reduction allows us to study the influence of the matrices A and V on the behavior of the mean-field PDE through the simpler ODE (1.4) for the order parameter $\mathcal{R}_2$, when the dynamics is initialized on the OA manifold. On top of that, we further establish structural stability results showing that, for certain choices of A and V , dynamics initialized near the OA manifold exhibit the same long-term behavior as those initialized exactly on it.

1.1 Contributions

The main contributions of this paper are the following.

- (i) **A pure second-harmonics Kuramoto formulation of linear self-attention.** We show that, in dimension $d = 2$, the linear self-attention model (LSA) can be formulated as a generalized Kuramoto model with pure second harmonics; see Theorem 2.1. This formulation reveals an intrinsic low-dimensional structure generated by the Watanabe–Strogatz transformation. Importantly, this property holds *independently* of the matrices A and V , and forms the basis of an analytical framework that differs fundamentally from the Wasserstein gradient flow approach.

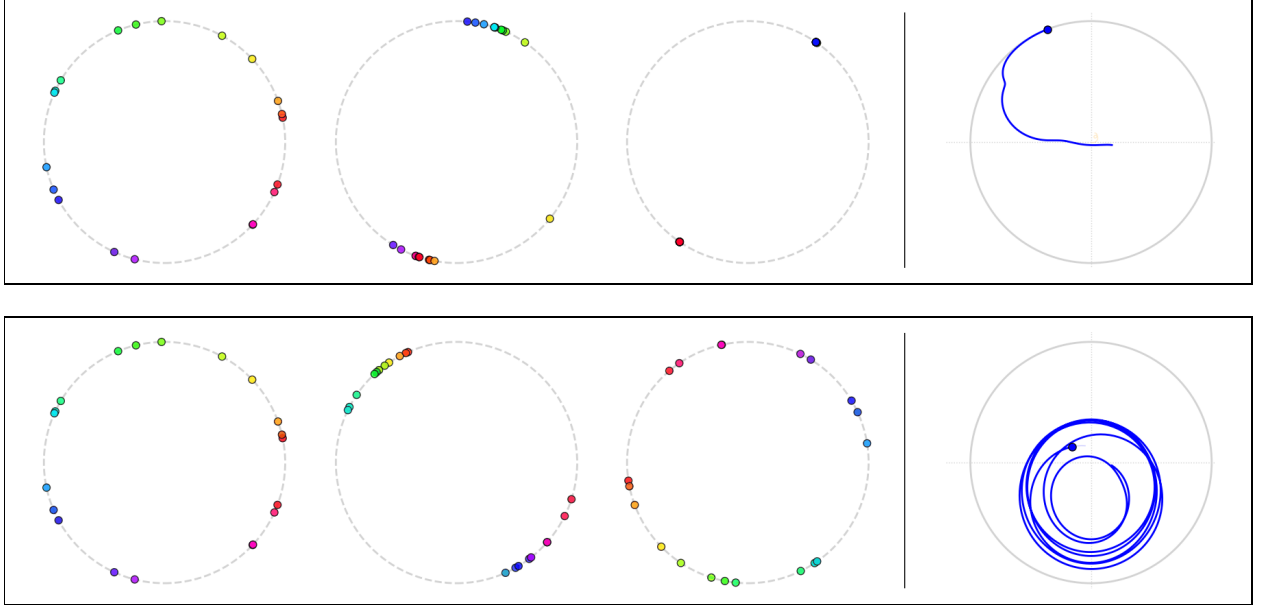


Figure 1: Snapshots of two systems with 20 particles evolving according to (LSA) under different choices of the matrices A and V . The left three columns show the particle positions at three points in time, and the right-most column shows the corresponding time evolution of the order parameter $\mathcal{R}_2^n(t)$. In the system at the top, the dynamics form two antipodal clusters, and the associated quantity $|\mathcal{R}_2^n(t)|$ converges to 1. In the system at the bottom, the particles neither cluster nor converge, and the associated quantity $|\mathcal{R}_2^n(t)|$ does not tend to 1.

- (ii) **Diverse dynamics on the OA manifold.** We prove that, for a class of initializations associated with the Ott–Antonsen (OA) manifold, the linear self-attention model (LSA) exhibits clustering behavior for a broad class of matrices A and V , going substantially beyond the assumptions considered in the existing literature. We further show that, in other parameter regimes, the model can display a rich range of dynamical behaviors, including oscillations and bifurcations. These oscillations and bifurcations are characterized by uncovering a hidden Hamiltonian structure in the dynamics. The results are summarized in Table 1. To the best of our knowledge, these are the first theoretical characterizations linking such diverse dynamical behaviors in transformer models to explicit parameter choices.

	Class of matrices	Clustering	Convergence	Assumptions
Case 1 (Thm. 3.1)	A arbitrary, $V = I$	✓	✓	$\text{tr}(A) > 0$ or $\det(A) \leq 0$
Case 2 (Thm. 3.3)	$A = I$, V symmetric	✓	✓	$\lambda_{\max}(V) \geq 0$
Case 3 (Thm. 3.6)	A diagonal, V with two degrees of freedom	✓	✗	A and V satisfy (3.13)
Case 4 (Thm. 3.9)	$A = I$, $V = \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & -v_{11} \end{pmatrix}$	✗	✗	$v_{11} < v_{12}$

Table 1: Summary of clustering and convergence regimes across four case studies.

- (iii) **Stability beyond the OA manifold.** We establish a structural stability result showing that, for

certain choices of A and V , dynamics initialized near the OA manifold exhibit the same long-time behavior as those initialized exactly on the manifold. The corresponding results are presented in Theorems 4.7 and 4.8. Our argument goes beyond a standard Grönwall estimate by exploiting the structural properties of the dynamics induced by the linear self-attention model, which allows us to obtain the stability over an infinite time horizon. The techniques developed here may also be of independent interest for the study of Kuramoto-type collective dynamics.

- (iv) **Numerical evidence.** We provide numerical experiments in higher dimensions d showing that the qualitative behavior proved for linear self-attention persists beyond the 2-dimensional setting. We also observe analogous behavior for the softmax self-attention; see Section 5.

1.2 Related Works

Mathematical analysis of the transformer inference dynamics. One fundamental question in the mathematical study of transformer architectures is to characterize the behavior of token representations as they are propagated through the layers of the transformer model [48, 78, 92]. We refer to this evolution of token representations as the transformer inference dynamics.

The approach we adopt is to model the inference dynamics of encoder-only transformers with self-attention modules as interacting multi-particle systems, an idea pioneered in [41, 69]. This viewpoint offers a continuous-time interpretation of the transformer architecture and, in the long-context regime [82, 88, 89, 110], naturally leads to mean-field perspectives, a strategy that we adopted also in this work. Earlier studies in this direction have focused on self-attention-only transformers, due to the self-attention modules constituting the central novelty of the transformer architecture. They have connected the self-attention dynamics to transport-type PDEs [92] and have developed a mathematical framework for analyzing transformer architectures through the lens of Wasserstein gradient flows [21, 48, 92] and optimization perspectives [5].

Building on this dynamical-system perspective, one of the main behaviors that has been characterized is *token concentration* in the self-attention dynamics: after many self-attention layers, all token representations concentrate at a lower-dimensional region of the embedding space. In case of collapse to a single fixed point, this is referred to as token uniformity, over-smoothing [28, 29, 53, 108], or rank collapse [31, 38, 44, 51]. By analyzing the energy landscape of the Wasserstein gradient flow induced by the *softmax* self-attention layers, earlier works [32, 48, 87] have shown that, under relatively restrictive assumptions on the matrices A and V , such as $A = V = I$, the self-attention dynamics converge to a Dirac distribution. Subsequent works have relaxed these assumptions, either by refining the analytical techniques within the Wasserstein gradient flow framework [27] or by adopting a different geometric approach from control theory [1, 2, 9] and optimization [5], leading to more general statements about concentration phenomena in mean-field transformers. These types of cluster formation result have also been extended to hardmax-attention [6, 7], causal self-attention [1, 2, 57], multi-head attention [1, 2, 72], transformers with fully-connected MLP layers [4, 8, 10, 43, 61], and noisy transformer dynamics [4, 10, 11, 42, 43, 61, 74–76].

Beside token representation collapse to a single point, several other theoretical properties of the transformer dynamics have also been established, including metastability [19, 20, 47], connections to optimization algorithms [5, 7], and scaling limits of hyperparameters [18, 26, 51]. We refer the reader to [90] for a more systematic survey.

Our results in Theorems 3.1 and 3.3 also fall within the general direction of characterizing token clustering. However, we diverge from existing lines of work by developing an *entirely new analytical framework* by connecting the *linear* self-attention dynamics to Kuramoto-type models with pure second-harmonic coupling; see the next paragraphs for a review. This framework allows us to characterize cluster formation under more general assumptions on the parameter matrices A and V . More importantly, it enables us to identify parameter regimes in which the associated self-attention dynamics exhibit behaviors beyond cluster formation, including oscillation and bifurcation, as demonstrated by Theorems 3.6 and 3.9. This broader treatment comes at the cost of restricting our current analysis to the linear self-attention model in dimension $d = 2$. In Section 6 we discuss some potential directions to extend our analysis to higher dimensional settings.

Kuramoto model and transformer dynamics. The Kuramoto model [63, 64] is a classical mathematical model for studying synchronization in large populations of coupled oscillators, with broad applications in

the natural sciences [3, 16, 60], engineering [39, 45, 106], and more recently, machine learning [14, 73, 77, 97]. In its simplest form the model takes the form

$$\dot{\theta}_k = \omega_k + \frac{\kappa}{n} \sum_{j=1}^n \sin(\theta_j - \theta_k) \quad (1.5)$$

for n oscillators θ_k with natural frequencies ω_k and a global coupling constant κ . One important generalization of the original model is the Kuramoto–Daido model [33, 34], in which the sine coupling, corresponding to the first Fourier harmonic, is replaced by a general periodic coupling function that may contain higher-order or multiple Fourier modes. A recent connection has been established between transformer dynamics and Kuramoto–Daido dynamics [48], where in dimension $d = 2$, the model (1.1) with matrices $Q^\top K = A = V = I$ and a general attention function h reduce to the Kuramoto–Daido model. Subsequent works identified sufficient conditions on h for synchronization [32, 87], and analyzed bifurcation and phase-transition phenomena in noisy Kuramoto–Daido models through the associated McKean–Vlasov equations [11, 74]. However, in the context of transformer dynamics (1.1), these analyses are restricted to the special matrix choice $A = V = I$.

One important mathematical advance in the analysis of Kuramoto-type models comes from identifying low-dimensional structures in Kuramoto–Daido dynamics when the coupling function is a *pure* Fourier harmonic. This direction was pioneered by Watanabe and Strogatz [105], who developed a theory to explain phase coherence in coupled arrays of Josephson junctions. Their technique applies to the more general Kuramoto–Daido models, but it took some time for the underlying mechanism of the reduction to be discovered as a Möbius group action [71]. Independently, Ott and Antonsen [79, 80] showed that the mean-field Kuramoto dynamics further reduce to a single ODE for the order parameter when the oscillator positions are initially chosen from a certain family. It turns out that these reductions are closely related: the Ott–Antonsen ansatz corresponds to a special initialization of oscillator positions within the Watanabe–Strogatz framework [71]. These dimension-reduction techniques have also been extended to Kuramoto–Daido models whose coupling function are pure higher-order Fourier harmonics [52, 96].

In this paper, we make a new observation connecting the Kuramoto–Daido model [33] and the transformer dynamics (1.1). Specifically, Theorem 2.1 shows that the transformer dynamics (1.1) with the attention function $h(y) = y$ can be reformulated as a Kuramoto–Daido model with *pure second-harmonic coupling* for *arbitrary* matrices A and V . This observation allows us to apply the Watanabe–Strogatz transformation and Ott–Antonsen ansatz to reduce the mean-field linear self-attention dynamics to a single complex-valued governing ODE. We then use this reduced dynamics to study how different classes of matrices A and V determine the long-time behavior of the linear self-attention dynamics, going substantially beyond the previously studied parameter regimes. We also note that Ott and Antonsen demonstrated the OA manifold is globally attracting under mild assumptions on the oscillator frequencies [80]. Unfortunately, these assumptions do not apply to our model, hence we conduct a separate stability analysis in Section 4.

1.3 Outline

The rest of the paper is organized as follows. In Section 2, we revisit the interacting particle system induced by the linear self-attention model (LSA) in dimension $d = 2$ and show that it can be reformulated as a Kuramoto-type model with pure second-order harmonics. We further show that this system admits a low-dimensional structure through the Watanabe–Strogatz (WS) transformation and the Ott–Antonsen (OA) ansatz, leading to reduced and trackable dynamics. In Section 3, we present case studies for different choices of the matrices A and V , and analyze the long-time behavior of the reduced dynamics under the OA ansatz in each case. In Section 4, we establish structural stability results beyond the OA ansatz for some choices of A and V , thereby describing the asymptotic behavior of the dynamics without requiring special initializations. In Section 5, we present a variety of numerical experiments where we explore the validity of our theoretical findings from Section 3 for the 2-dimensional linear self-attention model (LSA) in higher dimensions. We observe that the behaviors identified in the case $d = 2$ also arise in higher dimensions. We conclude the paper in Section 6. The main notation used throughout the manuscript is summarized in Table 2.

2 Linear Self-Attention Model on the Circle and its Low-Dimensional Structure

In this section, we show that the linear self-attention model (LSA) admits an intrinsic low-dimensional structure by rewriting it, in dimension $d = 2$, as a Kuramoto-type model with pure second-order harmonics. In Section 2.1, we establish this connection and discuss the evolution of the mean-field's second order parameter $\mathcal{R}_2$. Thereafter, in Section 2.2, we apply the Watanabe–Strogatz transformation [105] to show that the original system of n coupled equations is completely governed by *two* coupled ODEs. For a special class of initializations, corresponding to the Ott–Antonsen (OA) ansatz [79], this WS-reduced system further simplifies, leading to a single closed complex-valued ODE capturing the behavior of the full linear self-attention dynamics (LSA).

2.1 Linear Self-Attention Model on the Circle

We consider the linear self-attention model (LSA) in the 2-dimensional setting $d = 2$, i.e., on the unit circle $\mathbb{S}^1$. In this case, each token $x_k(t) \in \mathbb{S}^1$ is completely determined by an angle $\theta_k(t) \in \mathbb{T}^1 = \mathbb{R}/2\pi\mathbb{Z}$ since $x_k(t) = \cos(\theta_k(t))e_1 + \sin(\theta_k(t))e_2$, where $e_1 = (1, 0)^\top$ and $e_2 = (0, 1)^\top$ denote the standard basis vectors in $\mathbb{R}^2$. In what follows, recall that we have set $\beta = 1$ without loss of generality.

For each $k \in \{1, \dots, n\}$, the induced dynamics for the angle θ_k is given by

$$\dot{\theta}_k(t) = -\frac{1}{n \sin(\theta_k(t))} \left(\sum_{j=1}^n \langle x_k(t), Ax_j(t) \rangle \langle Vx_k(t), e_1 - \langle x_k(t), e_1 \rangle x_k(t) \rangle \right). \quad (2.1)$$

This expression follows by differentiating the identity $\cos(\theta_k(t)) = \langle x_k(t), e_1 \rangle$ w.r.t. time t , which yields $\langle \dot{x}_k(t), e_1 \rangle = \frac{d}{dt} \cos(\theta_k(t)) = -\sin(\theta_k(t))\dot{\theta}_k(t)$, and then substituting the evolution equation (LSA) together with the definition of the projection $P_{x_k(t)}^\perp$.

To expand (2.1), we write the matrices A and V as

$$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad V := \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}. \quad (2.2)$$

With this notation, we show that the dynamics governed by (2.1) can be written as a Kuramoto-type model with pure second-order harmonics, as stated in the following proposition. The proof is deferred to Appendix B.2.

Theorem 2.1. *Let $\{\theta_k(t)\}_{k=1}^n \in \mathbb{T}^1$ denote the angular variables associated with the LSA dynamics (LSA) on $\mathbb{S}^1$. Define the second order parameter $\mathcal{R}_2^n$ as*

$$\mathcal{R}_2^n(t) := \frac{1}{n} \sum_{j=1}^n e^{i2\theta_j(t)}. \quad (2.3)$$

Then, for each $k \in \{1, \dots, n\}$, the angle θ_k satisfies the closed ODE

$$\dot{\theta}_k(t) = b(\theta_k(t), \mathcal{R}_2^n(t)), \quad (2.4)$$

where the vector field $b : \mathbb{T}^1 \times \mathbb{C} \mapsto \mathbb{R}$ is given by

$$b(\theta, r) := 2\operatorname{Re}(B(r)e^{i2\theta}) + C(r). \quad (2.5)$$

Here $B : \mathbb{C} \mapsto \mathbb{C}$ and $C : \mathbb{C} \mapsto \mathbb{R}$ are the coefficient functions defined in (A.4).

The resulting system (2.4), with vector field b of the form (2.5), is a Kuramoto-type model with *pure* second-order harmonics [96] (see Remark 2.3 below for a more detailed comparison with the standard Kuramoto model). A key feature of this class of models is its intrinsic low-dimensional structure. In particular, the n -particle dynamics (2.4) can be completely described by just *two* macroscopic variables satisfying a

closed low-dimensional system of coupled ODEs. This leads to a substantial reduction in complexity: instead of analyzing n coupled ODEs, or a PDE in the mean-field limit, one can study a system of only two coupled ODEs. We discuss this reduction in Section 2.2.

Before doing so, let us formulate the mean-field limit of the particle dynamics (2.4), which facilitates the analysis conducted in the subsequent sections. Specifically, by letting the number of particles $n \rightarrow \infty$, we formally consider a process $\Theta = (\Theta(t))_{t \geq 0}$ evolving according to

$$\dot{\Theta}(t) = b(\Theta(t), \mathcal{R}_2(t)), \quad \Theta(0) \sim f(0, \bullet), \quad (2.6)$$

where the second mean-field order parameter $\mathcal{R}_2$ is defined, accordingly, as

$$\mathcal{R}_2(t) := \int_0^{2\pi} e^{i2\theta} f(t, d\theta). \quad (2.7)$$

Here, f denotes the law of Θ , i.e., $f := \text{Law}(\Theta)$, and satisfies the continuity equation

$$\partial_t f + \partial_\theta (b(\bullet, \mathcal{R}_2(t))f) = 0 \quad (2.8)$$

with the vector field b as defined in (2.5). Throughout the paper, we use either f_t or $f(t, \bullet)$ to denote the distribution of $\Theta(t)$.

Remark 2.2. The dependence of the functions B and C on the matrices A and V is in general very complicated (see (A.4) for a full definition). In our analysis we have used computer algebra software¹ to aid in writing the ODE systems explicitly. Even in the simplest case $A = V = I$, it is not obvious that the system reduces to the standard Kuramoto-type model with pure second harmonic coupling. We choose to view the complicated dependence on A and V as a positive—our case analysis demonstrates the richness of dynamics induced by the linear self-attention layers.

To expose the low-dimensional structure of both the finite particle system (2.4) and its mean-field counterpart (2.6), and to track the behavior of the order parameter $\mathcal{R}_2$, it is convenient to work with the *double angles* $\xi_k := 2\theta_k$. Specifically, we introduce the change of variable $\xi_k = 2\theta_k \in \mathbb{T}^1$ and $\Xi := 2\Theta \in \mathbb{T}^1$, with both variables understood modulo 2π . The dynamics of the transformed particle system $\{\xi_k\}_{k=1}^n$ and its mean-field limit can then be derived directly from the corresponding equations (2.4) and (2.6) for $\{\theta_k\}_{k=1}^n$ and Θ , respectively. In particular, the variables $\{\xi_k\}_{k=1}^n$ satisfy

$$\dot{\xi}_k(t) = \tilde{b}(\xi_k(t), \mathcal{R}_2^n(t)), \quad (2.9)$$

where the vector field $\tilde{b}$ is induced by b in (2.5) through

$$\tilde{b}(\xi, r) := 2b(\xi/2, r). \quad (2.10)$$

Moreover, the second order parameter $\mathcal{R}_2^n$ can be expressed purely in terms of the double angles $\{\xi_k\}_{k=1}^n$ since

$$\mathcal{R}_2^n(t) = \frac{1}{n} \sum_{j=1}^n e^{i2\theta_j(t)} = \frac{1}{n} \sum_{j=1}^n e^{i\xi_j(t)}. \quad (2.11)$$

Analogously, at the mean-field we obtain a process Ξ evolving according to

$$\dot{\Xi}(t) = \tilde{b}(\Xi(t), \mathcal{R}_2(t)), \quad \Xi(0) \sim g(0, \bullet). \quad (2.12)$$

Here, g denotes the law of Ξ , i.e., $g := \text{Law}(\Xi)$, which satisfies the continuity equation

$$\partial_t g + \partial_\xi (\tilde{b}(\bullet, \mathcal{R}_2(t))g) = 0 \quad (2.13)$$

with the vector field $\tilde{b}$ as defined in (2.10). The second mean-field order parameter $\mathcal{R}_2$ defined in (2.7) admits the equivalent representation

$$\mathcal{R}_2(t) = \int_0^{2\pi} e^{i2\theta} f(t, d\theta) = \int_0^{2\pi} e^{i\xi} g(t, d\xi), \quad (2.14)$$

¹The associated code can be found in https://github.com/tmaranzatto/attention_dynamics.

where we recall that $f = \text{Law}(\Theta)$ is the law of the original angle variable Θ . The relationship between the distributions f and g is given by

$$g(t, \xi) = \frac{1}{2} (f(t, \xi/2) + f(t, \xi/2 + \pi)), \quad (2.15)$$

which is a consequence of the change of variables formula on $[0, 2\pi)$ due to $\Xi = 2\Theta \in \mathbb{T}^1$.

In the next section, we apply the Watanabe–Strogatz transformation [105] and the Ott–Antonsen ansatz [79] to the double-angle dynamics (2.9) (and to its mean-field limit (2.12)) to reveal its underlying low-dimensional structure. This will enable a more tractable mathematical analysis.

Remark 2.3. The classical Kuramoto model (1.5) can be expressed as $\dot{\theta}_k(t) = \omega_k + \text{Im}(H(t)e^{i\theta_k(t)})$, where H is typically a function of the first order parameter $\mathcal{R}_1^n(t) := \frac{1}{n} \sum_{j=1}^n e^{i\theta_j(t)}$. Our model differs in a few ways. As stated above, our angles are coupled via the second-order harmonics $e^{i2\theta}$. Furthermore, our field extracts the real part of the second harmonic instead of the imaginary part. The term $C(\mathcal{R}_2^n)$ can be interpreted as a time-varying external forcing which acts on all particles and is controlled by the second order parameter $\mathcal{R}_2^n$. Lastly, our model assumes that all intrinsic frequencies are zero, i.e., $\omega_k = 0$ for all $k = 1, \dots, n$. Despite these differences the techniques used to analyze the standard Kuramoto system still apply to our model.

2.2 Watanabe–Strogatz Transformation and Ott–Antonsen Ansatz

In this section, we investigate the low-dimensional structure of the finite-particle double-angle dynamics for $\{\xi_k\}_{k=1}^n$ in (2.9) and its mean-field counterpart Ξ in (2.12) using the Watanabe–Strogatz (WS) transformation [71, 105] and the Ott–Antonsen (OA) ansatz [79]. In particular, the WS transformation shows that the n -particle system (2.9) can be fully described by *two* macroscopic variables, denoted by α and η , which satisfy a coupled system of ODEs. In the mean-field regime, under a special choice of initialization for the mean-field variable Ξ , the dynamics of (2.12) reduces further to a *single* complex variable α , which, in this case, coincides with the order parameter $\mathcal{R}_2$, that is, $\alpha = \mathcal{R}_2$. This is the Ott–Antonsen ansatz.

2.2.1 Watanabe–Strogatz Transformation

We begin by introducing the Watanabe–Strogatz (WS) transformation, which reveals the intrinsic low-dimensional structure of the systems (2.9) and (2.12). The corresponding result is stated in the following proposition, whose proof is deferred to Appendix B.3.

For fixed $\alpha \in \mathbb{C}$ and $\eta \in \mathbb{R}$, we define the Möbius transformation $M_{\alpha, \eta} : \mathbb{C} \mapsto \mathbb{C}$ by

$$M_{\alpha, \eta}(u) := \frac{\alpha + e^{i\eta}u}{1 + \bar{\alpha}e^{i\eta}u}. \quad (2.16)$$

Proposition 2.4 (Watanabe–Strogatz Transformation [71]). *Let $\{\xi_k\}_{k=1}^n$ be the solution to (2.9). Then there exist time-independent angles $\{\varphi_k\}_{k=1}^n$ such that, for each $k \in \{1, \dots, n\}$, it holds*

$$e^{i\xi_k(t)} = M_{\alpha^n(t), \eta^n(t)}(e^{i\varphi_k}), \quad t \geq 0. \quad (2.17)$$

Here, $\alpha^n : [0, \infty) \mapsto \mathbb{C}$ and $\eta^n : [0, \infty) \mapsto \mathbb{R}$ satisfy the system of ODEs

$$\dot{\alpha}^n(t) = F(\alpha^n(t), \mathcal{R}_2^n(t)), \quad (2.18a)$$

$$\dot{\eta}^n(t) = G(\alpha^n(t), \mathcal{R}_2^n(t)), \quad (2.18b)$$

where the functions F and G are defined for $\alpha, r \in \mathbb{C}$ and $\eta \in \mathbb{R}$ by

$$F(\alpha, r) := 2i(B(r)\alpha^2 + C(r)\alpha + \bar{B}(r)), \quad G(\alpha, r) := 2(2\text{Re}(B(r)\alpha) + C(r)), \quad (2.19)$$

with B and C as in (A.4), and $\mathcal{R}_2^n$ is the second order parameter defined in (2.11).

The previous proposition says that the evolution of the n -particles system (2.9) is completely characterized by the two variables α and η , together with n time-independent angles $\{\varphi_k\}_{k=1}^n$. This reduces the description

of the dynamics from n coupled differential equations in (2.9) to the two equations in (2.18). For later reference, we refer to α^n and η^n , as well as α and η introduced shortly, as the *WS variables*; to $\{\varphi_k\}_{k=1}^n$ as the *constants of motion*; and to system (2.18) (thus also (2.21)) as the *WS system*.

In the mean-field limit, an analogous representation holds. The double angle Ξ , which evolves according to the mean-field dynamics (2.12), can be written in terms of a time-varying Möbius transformation on the complex unit circle, i.e.,

$$e^{i\Xi(t)} = M_{\alpha(t),\eta(t)}(e^{i\varphi}), \quad (2.20)$$

where φ is a time-independent random variable (in analogy with the finite particle system, we will continue to refer to it as a constant of motion) and where the Möbius transformation $M_{\alpha,\eta}$ is as in (2.16). The WS variables α and η satisfy the coupled ODE system

$$\dot{\alpha}(t) = F(\alpha(t), \mathcal{R}_2(t)), \quad (2.21a)$$

$$\dot{\eta}(t) = G(\alpha(t), \mathcal{R}_2(t)), \quad (2.21b)$$

where F and G are as in (2.19), and where $\mathcal{R}_2$ is the mean-field's second order parameter defined in (2.14).

Möbius transformation. Let us now discuss the Möbius transformation (2.16). Below we reuse the same variables as in Section 2.2.1, however for the current discussion these should be seen as arbitrary constants. It is straightforward to verify that, when $|\alpha| \neq 1$, $M_{\alpha,\eta}$ defines a bijection (and a diffeomorphism) from the unit circle $\mathbb{S}^1$ in the complex plane $\mathbb{C}$ onto itself. Its inverse is given by

$$M_{\alpha,\eta}^{-1}(v) := e^{-i\eta} \frac{v - \alpha}{1 - \bar{\alpha}v}, \quad (2.22)$$

which can also be written as $M_{\tilde{\alpha},\tilde{\eta}}$ for $\tilde{\eta} := -\eta$ and $\tilde{\alpha} := -e^{-i\eta}\alpha$. If, on the other hand, $|\alpha| = 1$, the transformation $M_{\alpha,\eta}$ is constant and equal to α . This can be seen from the fact that, when $|\alpha| = 1$, one can write $\alpha = e^{i\Phi}$ (for $\Phi \in \mathbb{R}$) and the Möbius map takes the form

$$M_{\alpha,\eta}(u) = \frac{\alpha + ue^{i\eta}}{1 + \bar{\alpha}ue^{i\eta}} = e^{i\Phi} \frac{1 + e^{-i\Phi}ue^{i\eta}}{1 + e^{-i\Phi}\bar{\alpha}ue^{i\eta}} = \alpha, \quad (2.23)$$

provided $ue^{i\eta} \neq -\alpha$.

From the above discussion we conclude that, regardless of the values of α and η , each Möbius transformation $M_{\alpha,\eta}$ induces a map

$$T_{\alpha,\eta} : [0, 2\pi) \mapsto [0, 2\pi), \quad T_{\alpha,\eta}(\varphi) = \text{Arg}\left(\frac{\alpha + e^{i(\eta+\varphi)}}{1 + \bar{\alpha}e^{i(\eta+\varphi)}}\right) \quad (2.24)$$

as illustrated in the diagram in Figure 2. Therefore, any probability distribution ν on the constant-of-motion variable φ induces a corresponding family of distributions g on the physical variable Ξ via pushforwards by the family of maps $\{T_{\alpha,\eta}\}_{\alpha \in \mathbb{C} \setminus \mathbb{S}^1, \eta \in \mathbb{R}}$. More precisely, each distribution in this family takes the form

$$g = (T_{\alpha,\eta})_{\#}\nu, \quad (2.25)$$

where $(T_{\alpha,\eta})_{\#}\nu$ denotes the pushforward of the measure ν under the map $T_{\alpha,\eta}$.

$$\begin{array}{ccc} \varphi \in [0, 2\pi) & \xrightarrow{p} & u = e^{i\varphi} \in \mathbb{S}^1 \\ \downarrow T_{\alpha,\eta} & & \downarrow M_{\alpha,\eta} \\ \Xi \in [0, 2\pi) & \xleftarrow{p^{-1}} & e^{i\Xi} = M_{\alpha,\eta}(u) \in \mathbb{S}^1. \end{array}$$

Figure 2: Diagram illustrating the Möbius transformation.

In light of (2.20), (2.14) and (2.16), we can thus express the second order parameter $\mathcal{R}_2(t)$ as

$$\mathcal{R}_2(t) = \int_0^{2\pi} e^{i\xi} g(t, d\xi) = \int_0^{2\pi} e^{i\xi} (T_{\alpha(t), \eta(t)})_{\#} \nu(d\xi) = \int_0^{2\pi} M_{\alpha(t), \eta(t)}(e^{i\varphi}) \nu(d\varphi). \quad (2.26)$$

We now discuss the interpretation of the WS variables $\alpha(t)$ and $\eta(t)$.

- **Variable $\alpha(t)$.** The variable $\alpha(t)$ characterizes the level of synchronization of the particles. In particular, $|\alpha(t)| = 1$ implies $|\mathcal{R}_2(t)| = 1$, which corresponds to complete synchronization (as discussed in previous sections). Indeed, when $|\alpha(t)| = 1$, (2.26) and (2.23) imply that $\mathcal{R}_2(t) = \alpha(t)$ and in particular $|\mathcal{R}_2(t)| = |\alpha(t)| = 1$. Conversely, it is straightforward to see that if the measure ν is assumed to have full support, then $|\mathcal{R}_2(t)| = 1$ only if $|\alpha(t)| = 1$.

As discussed in Figure 1, we know that $|\mathcal{R}_2(t)| = 1$ if and only if the corresponding particle distribution f is supported on a single point or two antipodal points. For this reason, we will be interested in tracking the long-term behavior of the variable $|\alpha|$ and determining whether $|\alpha(t)| \rightarrow 1$ as $t \rightarrow \infty$.

- **Variable $\eta(t)$.** The variable $\eta(t)$ can be interpreted as a rotating frame acting on all particles. Specifically, if we set $\alpha(t) \equiv 0$ for simplicity, then the Möbius transformation (2.16) reduces to

$$e^{i\xi_k(t)} = M_{\alpha(t), \eta(t)}(e^{i\varphi_k}) = \frac{\alpha(t) + e^{i\eta(t)} e^{i\varphi_k}}{1 + \overline{\alpha(t)} e^{i\eta(t)} e^{i\varphi_k}} = e^{i(\varphi_k + \eta(t))} \quad (2.27)$$

Hence, in this case, $\eta(t)$ simply shifts all phases by the same amount, corresponding to a global rotation. In general, one can view $M_{\alpha(t), \eta(t)}$ as the composition of a rotation (counterclockwise by $\eta(t)$) and the transformation $M_{\alpha(t), 0}$.

Despite the simpler form of the WS system (2.21), it remains difficult to analyze it in general since the variables $\alpha(t)$ and $\eta(t)$ are implicitly coupled through the second order parameter $\mathcal{R}_2(t)$, whose dependence on $\alpha(t)$ and $\eta(t)$ can be quite intricate, as can be appreciated from equation (2.26). However, as we will show in the next section, for a special choice of the distribution ν , or, equivalently, for a corresponding special initialization of the double angle $\Xi(0) = 2\Theta(0)$, the WS system (2.21) further decouples. In this case, $\mathcal{R}_2(t)$ coincides with the variable $\alpha(t)$ and becomes independent of the rotation variable $\eta(t)$. This is precisely the setting of the Ott–Antonsen ansatz [79].

Remark 2.5. The intrinsic coupling to $\mathcal{R}_2^n$ in the WS system (2.18) means there is no computational reduction compared to numerically evaluating the untransformed system. In the simplest setting, using Euler’s method to compute the next state requires $\mathcal{O}(n)$ operations for both the untransformed and WS systems. This is straightforward to see for the original system; as for the WS system note $\mathcal{R}_2$ is expressed in terms of the n constants of motion φ_k as seen in (2.26). Savings may be found in the mean-field limit, where the computation of $\mathcal{R}_2$ in the WS system involves integrating over the time-invariant measure ν , as opposed to the time-varying measure f . If the integral (2.26) can be efficiently approximated (or indeed analytically solved as in the next section), then there are potentially huge computational savings — the WS transform reduces an infinite dimensional PDE into a finite dimensional ODE.

2.2.2 Ott–Antonsen Ansatz

Recall from the diagram in Figure 2 that any probability distribution ν on the constant of motion φ induces a corresponding family of distributions g on the physical variable Ξ via pushforwards by the maps $T_{\alpha, \eta}$. In general, the resulting distributions $g = T_{\alpha, \eta\#} \nu$ may be quite complicated. However, if ν is the uniform distribution on $[0, 2\pi)$, and $|\alpha| < 1$, then the distribution of Ξ is a *wrapped Cauchy distribution* [70], associated with densities in the form of the Poisson kernel

$$\text{WC}(\xi; \alpha) := \frac{1}{2\pi} \frac{1 - |\alpha|^2}{|e^{i\xi} - \alpha|^2}, \quad (2.28)$$

parameterized by the single complex number $\alpha \in \mathbb{C}$ with $|\alpha| < 1$; see Figure 3 for an illustration.

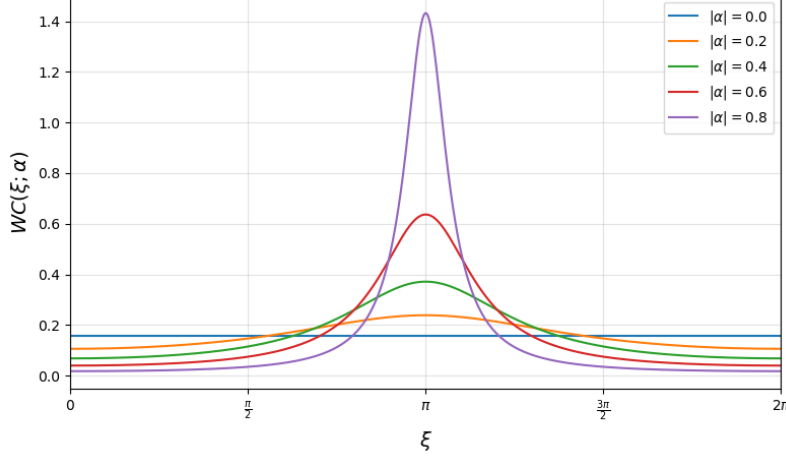


Figure 3: The family of wrapped Cauchy distributions generated by fixing the argument $\text{Arg}(\alpha) = \pi$ and varying the magnitude $|\alpha|$.

In particular, $g = (T_{\alpha,\eta})_{\#} \text{Unif}([0, 2\pi)) = \text{WC}(\cdot; \alpha)$ when $|\alpha| < 1$ (see Lemma 2.6 below). Note that the dependence on η of the induced measures disappears because the uniform distribution is invariant under rotations. Also, note that $\text{WC}(\cdot; \alpha)$ for $\alpha = 0$ is equal to $\text{Unif}([0, 2\pi))$ and that $\text{WC}(\cdot; \alpha)$ for $|\alpha| = 1$ can be identified with a Dirac delta at $\text{Arg}(\alpha)$. We refer to the induced family of distributions parameterized by $\alpha \in \mathbb{C}$ with $|\alpha| \leq 1$ as the *Ott–Antonsen (OA) manifold* [80].

From formula (2.26) it follows that the second order parameter $\mathcal{R}_2$ depends only on α and is thus independent of the rotating-frame parameter η . As a consequence, the dynamics of α (recall (2.21a)) are independent of η . This special case is exactly the setting of the Ott–Antonsen ansatz [79, 80]. Based on [71], we summarize the above discussion in the following lemma. A detailed proof is deferred to Appendix B.3.

Lemma 2.6. *Let $\Xi \in [0, 2\pi)$ and $\varphi \in [0, 2\pi)$ be two random variables satisfying the relation $\Xi = T_{\alpha,\eta}(\varphi)$ for some $\alpha \in \mathbb{C}$ with $|\alpha| < 1$, where the map $T_{\alpha,\eta}$ is as in (2.24). Then the distribution $g(\cdot)$ of Ξ is a wrapped Cauchy distribution of the form*

$$g(\xi) = \text{WC}(\xi; \alpha) \quad (2.29)$$

if and only if $\varphi \sim \nu = \text{Unif}([0, 2\pi))$.

Let us now discuss several direct consequences of Lemma 2.6, which highlight some favorable features of the setting of the Ott–Antonsen ansatz.

- **The wrapped Cauchy family is invariant under the dynamics (2.12).** Because of the representation (2.20), the constant of motion φ is completely determined by the initial phase $\Xi(0)$ together with the initial values $\alpha(0)$ and $\eta(0)$. More precisely, using (2.20) together with the inverse Möbius transform (2.22), we obtain

$$e^{i\varphi} = M_{\alpha(0), \eta(0)}^{-1} \left(e^{i\Xi(0)} \right).$$

Consequently, for any fixed $\eta(0)$, $\varphi \sim \nu = \text{Unif}([0, 2\pi))$ if and only if the initial data $\Xi(0)$ is distributed according to the wrapped Cauchy distribution $\text{WC}(\xi; \alpha(0))$. The value of $\eta(0)$ only rotates the constant of motion φ and therefore does not affect this distributional equivalence.

As a direct consequence, the mean-field dynamics (2.6) leaves the wrapped Cauchy family invariant: if the initial distribution belongs to this family, then it remains in the same family for all later times.

- **Closed dynamics for $\mathcal{R}_2$ and characterization of distribution g .** If the distribution g has the form $\text{WC}(\xi; \alpha)$, then it admits the Fourier representation

$$g(t, \xi) = \frac{1}{2\pi} \left(1 + 2 \sum_{m=1}^{\infty} \text{Re}(\bar{\alpha}^m e^{im\xi}) \right).$$

Consequently, the order parameter $\mathcal{R}_2$ defined in (2.14) satisfies

$$\mathcal{R}_2(t) = \alpha(t). \quad (2.30)$$

In particular, $\mathcal{R}_2$ depends only on α , and thus the WS system (2.21) decouples, as had already been mentioned earlier. Moreover, the second order parameter $\mathcal{R}_2$ satisfies the closed evolution equation

$$\dot{\mathcal{R}}_2(t) = F(\mathcal{R}_2(t), \mathcal{R}_2(t)) = 2i \left(B(\mathcal{R}_2(t))(\mathcal{R}_2(t))^2 + C(\mathcal{R}_2(t))\mathcal{R}_2(t) + \overline{B}(\mathcal{R}_2(t)) \right). \quad (2.31)$$

We further note that, in this case, the distribution g is fully determined by the order parameter $\mathcal{R}_2$. Hence, tracking the behavior of $\mathcal{R}_2$ is sufficient to characterize the full distribution of the double-angle variable Ξ .

We refer to the dynamics (2.31) as the *OA system*. Compared with the WS system (2.21), the OA system (2.31) is generally more amenable to analysis. Accordingly, in the subsequent Section 3, we investigate the long-time behavior of the OA system for different choices of the matrices A and V . The resulting analysis allows us to infer the synchronization properties of the mean-field particle system (2.12) through the behavior of $\mathcal{R}_2$.

It is important to note, however, that the finite-particle system (2.9) cannot satisfy the OA ansatz exactly, since it is necessarily associated with an atomic initial distributions $g(0) = (T_{\alpha(0), \eta(0)})_{\#} \nu$, and hence with an atomic distribution ν .

Remark 2.7. Let us comment on the relation between the distribution f of the original angle variable Θ and the distribution g of the double-angle variable $\Xi = 2\Theta$ in the Ott–Antonsen setting.

Using (2.15) and the definition of the wrapped Cauchy distribution in (2.28), we observe that if Θ is distributed according to a wrapped Cauchy distribution with parameter α , namely, $f = \text{WC}(\bullet; \alpha)$, then the double-angle variable Ξ is distributed according to a wrapped Cauchy distribution, now with parameter α^2 , that is, $g = \text{WC}(\bullet; \alpha^2)$.

The converse, however, does not hold in general: if g is a wrapped Cauchy distribution, f need not be a wrapped Cauchy. For example, let g be the uniform distribution $\text{Unif}([0, 2\pi])$, which corresponds to the wrapped Cauchy distribution (2.28) with $\alpha = 0$. Define $f(\bullet) = \frac{1}{2\pi}(1 + \cos(\bullet))$. Then f and g satisfy the relation (2.15), but f is not a wrapped Cauchy distribution. More generally, the relation (2.15) does not uniquely determine f . Indeed, whereas g belonging to the wrapped Cauchy family necessarily has at most one local maximum, one may perturb f by a sufficiently small π -antiperiodic function, provided the total mass and non-negativity are preserved, thereby obtaining admissible densities with arbitrary many local maxima.

Consequently, if the initial distribution of $\Theta(0)$ is a wrapped Cauchy distribution, then the induced initial distribution of $\Xi(0)$ is also a wrapped Cauchy distribution, and hence the dynamics of Ξ starts on the OA manifold. This manifold, however, is invariant only for the double angle dynamics (2.12), and not for the original dynamics (2.6) of Θ . In other words, under the setting of the Ott–Antonsen ansatz, only the distribution g remains in the wrapped Cauchy family, whereas the distribution f may not necessarily stay in this family over time. This can be illustrated by the fact that distributions in the Cauchy family are unimodal (see Figure 3), while the distribution f may concentrate around antipodal points.

3 Long-Time Behavior Within the Ott–Antonsen Ansatz: Case Studies

In this section, we investigate the long-time behavior of the dynamics (2.31) of the second mean-field order parameter $\mathcal{R}_2$ under the OA ansatz, with a particular emphasis on its dependence on the parameter matrices A and V of the transformer dynamics. We will show that, even for relatively simple choices of these parameter matrices, the system (2.31) exhibits remarkably rich behavior. These different dynamical regimes of $\mathcal{R}_2$ reveal the richness of the mean-field linear transformer dynamics (2.6).

For the reader's convenience, let us recall the dynamics (2.31), which is given by

$$\dot{\mathcal{R}}_2(t) = F(\mathcal{R}_2(t), \mathcal{R}_2(t)) = 2i \left(B(\mathcal{R}_2(t))(\mathcal{R}_2(t))^2 + C(\mathcal{R}_2(t))\mathcal{R}_2(t) + \overline{B}(\mathcal{R}_2(t)) \right).$$

When the context is clear, we will simply write $F(\mathcal{R}_2(t))$ in place of $F(\mathcal{R}_2(t), \mathcal{R}_2(t))$. We also introduce the polar representation of the order parameter,

$$\mathcal{R}_2(t) = \rho(t)e^{i\phi(t)},$$

where $\rho : [0, \infty) \mapsto [0, 1]$ and $\phi : [0, \infty) \mapsto \mathbb{T}^1$. Substituting this representation into (2.31), we obtain the coupled ODE system:

$$\dot{\rho}(t) = \operatorname{Re} \left(e^{-i\phi(t)} F(\rho(t)e^{i\phi(t)}) \right), \quad \dot{\phi}(t) = \frac{1}{\rho(t)} \operatorname{Im} \left(e^{-i\phi(t)} F(\rho(t)e^{i\phi(t)}) \right). \quad (3.1)$$

In what follows, we present a series of case studies for four different choices of the matrices A and V . By analyzing either the complex ODE (2.31) directly or the polar-coordinate system (3.1) in each case, we show that $\mathcal{R}_2(t)$ can display a rich variety of long-time behaviors. These, in turn, correspond to distinct long-time behaviors of the mean-field linear transformer dynamics (2.6). We highlight three types of behavior for which we are able to draw definitive conclusions about the dynamics (2.6), and we further explain their interpretations in the context of linear transformers, where the particles are viewed as token embeddings evolving through linear self-attention layers.

- **(Type 1).** Both $\rho = |\mathcal{R}_2| \rightarrow 1$ and $\phi = \operatorname{Arg}(\mathcal{R}_2) \rightarrow \phi_\infty$ converge. Correspondingly, the distribution f of the particle Θ converges to a distribution supported on two fixed antipodal points on $\mathbb{T}^1$; see Case 1 and Case 2.

In the linear transformer interpretation, this means that, after sufficiently many layers, the token embeddings cluster into two *fixed* feature representations.

- **(Type 2).** While $\rho = |\mathcal{R}_2| \rightarrow 1$ converges, $\phi = \operatorname{Arg}(\mathcal{R}_2)$ does not. In this scenario, the distribution f of the particle Θ converges to a distribution supported on two antipodal points on $\mathbb{T}^1$. However, these support points themselves continue to move and do not converge; see Case 3.

In the linear transformer interpretation, this means that, after sufficiently many layers, the token embeddings cluster into two feature representations, but these representations continue to change as the number of layers increases.

- **(Type 3).** Neither $\rho = |\mathcal{R}_2|$ or $\phi = \operatorname{Arg}(\mathcal{R}_2)$ converges; instead, both continue to oscillate. Correspondingly, the distribution f does not converge to a stationary distribution; see Case 4.

In the linear transformer interpretation, this means that the token embeddings do not converge: even after arbitrarily many layers, the output continues to change as additional layers are applied.

Moreover, as a by-product of our analysis, we also identify bifurcations in the dynamics (2.31), highlighting the sensitivity of the long-time behavior to small changes in the matrices A and V ; see Case 4.

Before proceeding to the case studies, we introduce some notation. We denote the unit disk and its boundary by

$$D := \{z \in \mathbb{C} : |z| \leq 1\}, \quad \partial D := \{z \in \mathbb{C} : |z| = 1\}. \quad (3.2)$$

For the entries of the matrices A and V , we use the shorthand

$$\begin{aligned} a_+^d &:= a_{11} + a_{22}, & a_-^d &:= a_{11} - a_{22}, & a_+^o &:= a_{12} + a_{21}, & a_-^o &:= a_{12} - a_{21}, \\ v_+^d &:= v_{11} + v_{22}, & v_-^d &:= v_{11} - v_{22}, & v_+^o &:= v_{12} + v_{21}, & v_-^o &:= v_{12} - v_{21}. \end{aligned}$$

The proofs of theorems in this section are deferred to Appendix C.

3.1 Case 1: $V = I$, A Arbitrary

Let A be a real-valued matrix such that $A \neq 0$ and V be the identity matrix,

$$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad V := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (3.3)$$

In this case, the dynamics (2.31) of the order parameter $\mathcal{R}_2$ and its corresponding polar-coordinate system (3.1) simplify to

$$\dot{\mathcal{R}}_2 = \frac{1}{4} (1 - |\mathcal{R}_2|^2) ((a_+^d + ia_-^o) \mathcal{R}_2 + (a_-^d + ia_+^o)), \quad (3.4)$$

and

$$\dot{\rho} = \frac{1}{4} (1 - \rho^2) (a_+^d \rho + a_-^d \cos \phi + a_+^o \sin \phi), \quad \dot{\phi} = \frac{1}{4} (1 - \rho^2) \left(a_-^o - \frac{a_-^d \sin \phi - a_+^o \cos \phi}{\rho} \right). \quad (3.5)$$

The long-time behavior of the dynamical system (3.5) is summarized as follows.

Theorem 3.1. *Consider the dynamical system (3.5).*

1. *If the matrix A satisfies the condition:*

$$\text{tr}(A) > 0 \quad \text{or} \quad \det(A) \leq 0, \quad (3.6)$$

then for almost every initial condition $(\rho(0), \phi(0)) \in (0, 1] \times \mathbb{T}^1$, the solution of (3.5) converges to $(1, \phi_\infty)$ for some $\phi_\infty \in \mathbb{T}^1$ as $t \rightarrow \infty$.

2. *If the symmetric part of A is negative definite, i.e.*

$$\frac{A + A^\top}{2} \prec 0, \quad (3.7)$$

then for every initial condition $(\rho(0), \phi(0)) \in (0, 1) \times \mathbb{T}^1$, the solution of (3.5) converges to $(\rho_{\text{eq}}, \phi_{\text{eq}})$ as $t \rightarrow \infty$ with $\rho_{\text{eq}} \in (0, 1)$ and $\phi_{\text{eq}} \in \mathbb{T}^1$; see the explicit form of $(\rho_{\text{eq}}, \phi_{\text{eq}})$ in (C.14).

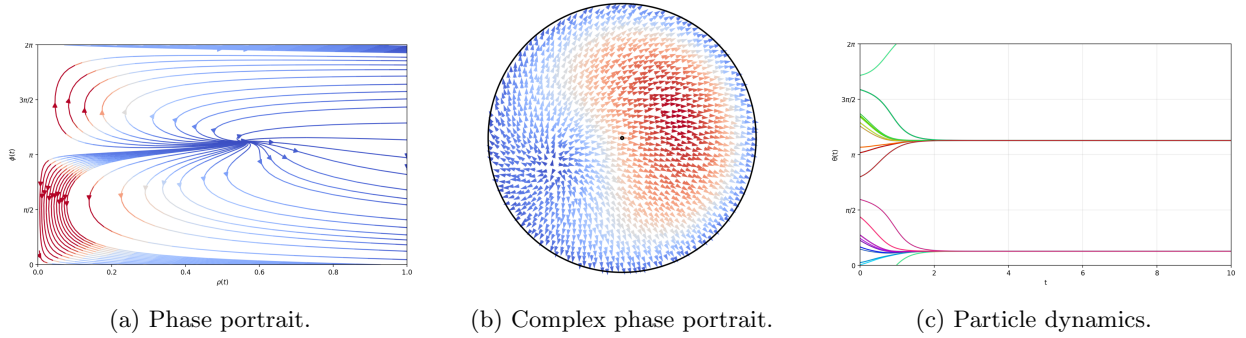


Figure 4: Cartesian and complex phase portraits of (ρ, ϕ) -dynamics (3.5) and the trajectories of particles evolving under the transformer dynamics (2.4) for $V = I$, $A = \begin{pmatrix} 4 & -1 \\ 1 & 1 \end{pmatrix}$, which satisfies condition (3.6). In the phase portraits, red flow lines or arrows indicate a fast velocity of (ρ, ϕ) , and blue indicates a slow velocity. The coloring in the cartesian and complex portraits do not align, due to the fast velocity near degenerate point $\rho = 0$.

Theorem 3.1 states that, when $V = I$, if the matrix A satisfies (3.6), then the order parameter $\mathcal{R}_2(t)$ converges to a fixed point on ∂D . Through the relation between $\mathcal{R}_2(t)$ and the distribution f of the mean-field angle Θ in (2.7), this further implies that f converges to a stationary distribution supported on two antipodal points.

Figure 4 illustrates an example in this regime. The left panel shows the phase portrait of the (ρ, ϕ) -dynamics in (3.5), where the vector field points toward the boundary $\rho = 1$ throughout the phase space. The right panel shows the trajectories of 20 angle particles $\{\theta_k\}$ evolving under the finite-particle transformer dynamics (2.4), and these particles converge to two antipodal points.

On the other hand, when the symmetric part of A is negative definite, the order parameter $\mathcal{R}_2$ converges asymptotically to an interior point of the unit disk D . In this case, since the distribution g of the double

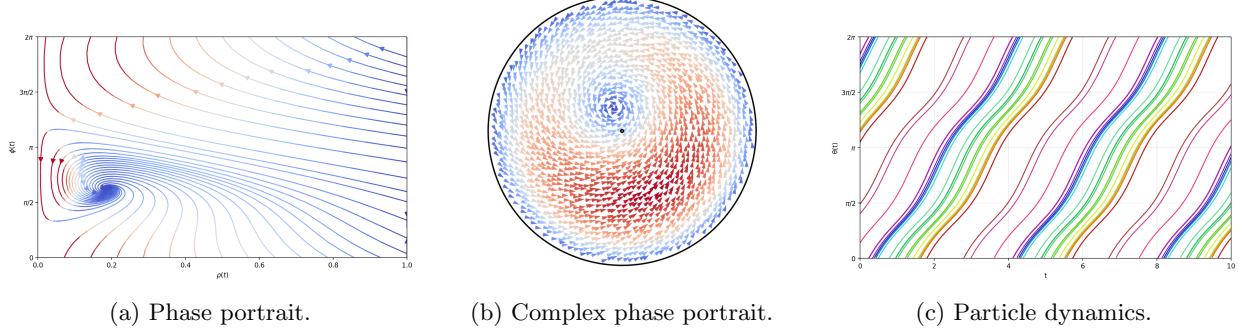


Figure 5: Cartesian and complex phase portraits of (ρ, ϕ) -dynamics (3.5) and the trajectories of particle evolving under the transformer dynamics (2.4) for $V = I$, $A = \begin{pmatrix} -1 & 4 \\ -3 & -2 \end{pmatrix}$, which satisfies condition (3.7). Again, red indicates fast flow and blue indicates slow flow in the phase diagrams.

angle $\Xi = 2\Theta$ remains a wrapped Cauchy distribution with parameter $\mathcal{R}_2(t)$, it follows that g converges to a stationary distribution in the wrapped Cauchy family. However, as explained in Remark 2.7, the distribution f of the original angle Θ does not necessarily remain in the wrapped Cauchy family, nor does it necessarily converge. Figure 5 illustrates an example in this regime. In the left panel, the phase portrait shows that the vector field of the (ρ, ϕ) -dynamics points toward the interior equilibrium, again consistent with the theoretical prediction. Meanwhile, the right panel shows that the corresponding particle dynamics does not converge, in agreement with the theoretical insight that the distribution f does not necessarily converge.

Remark 3.2. The union of the regimes described by conditions (3.6) and (3.7) does not cover the entire space $\mathbb{R}^{2 \times 2}$. In the remaining regimes, the long-time behavior of the order parameter $\mathcal{R}_2$ depends on the initial data. In particular, $\mathcal{R}_2(t)$ may converge to a fixed point on the boundary ∂D , converge to the interior equilibrium in D , or approach a periodic orbit encircling the interior equilibrium. We refer to the proof of Theorem 3.1 in Appendix C for a detailed discussion.

3.2 Case 2: $A = I$, V Symmetric

Let A be the identity matrix and V be a real-valued symmetric matrix,

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{pmatrix}. \quad (3.8)$$

In this case, the dynamics (2.31) of the order parameter $\mathcal{R}_2$ and its corresponding polar-coordinate system (3.1) simplify to

$$\dot{\mathcal{R}}_2 = \frac{1}{4} (v_+^d (1 - |\mathcal{R}_2|^2) \mathcal{R}_2 + 2(-v_+^o + iv_-^d) (\mathcal{R}_2)^2 + (v_+^o + iv_-^d) (1 + |\mathcal{R}_2|^2)), \quad (3.9)$$

and

$$\dot{\rho} = \frac{1}{4} (1 - \rho^2) (v_+^d \rho + v_-^d \cos \phi + v_+^o \sin \phi), \quad \dot{\phi} = -\frac{3\rho^2 + 1}{4\rho} (v_-^d \sin \phi - v_+^o \cos \phi). \quad (3.10)$$

The long-time behavior of the dynamical system (3.10) is summarized as follows.

Theorem 3.3. *Consider the dynamical system (3.10) and assume that V is symmetric. Let $\lambda_{\max}(V)$ denote the largest eigenvalue of V .*

1. *If $\lambda_{\max}(V) \geq 0$, then for almost every initial condition $(\rho(0), \phi(0)) \in (0, 1] \times \mathbb{T}^1$, the solution of (3.10) converges to a boundary equilibrium $(1, \phi_V)$ as $t \rightarrow \infty$. Here, ϕ_V is defined in (C.19).*
2. *If $\lambda_{\max}(V) < 0$, then for almost every initial condition $(\rho(0), \phi(0)) \in (0, 1) \times \mathbb{T}^1$, the solution of (3.10) converges to the interior equilibrium $(\rho_{\text{eq}}, \phi_V)$ as $t \rightarrow \infty$. Here, ρ_{eq} is given in (C.28).*

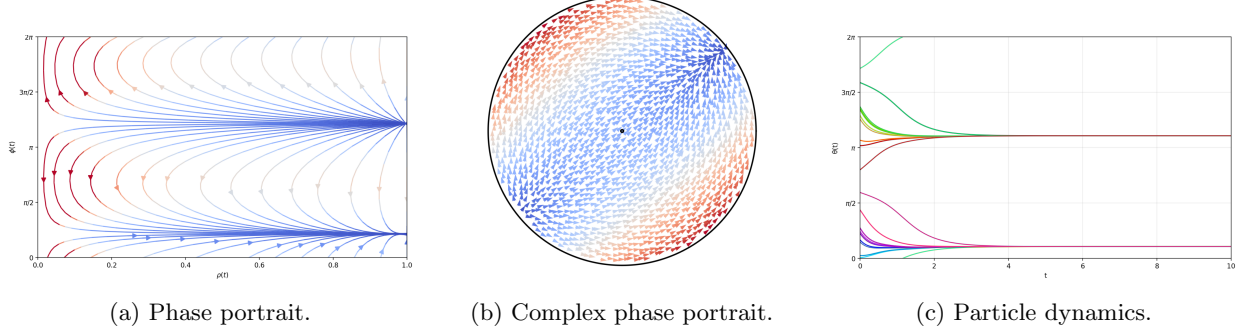


Figure 6: Phase portrait of (ρ, ϕ) -dynamics (3.10) and the trajectories of particle evolving under the transformer dynamics (2.4) for $A = I$, $V = \begin{pmatrix} 1 & 2 \\ 2 & -4 \end{pmatrix}$, which satisfies $\lambda_{\max}(V) \geq 0$.

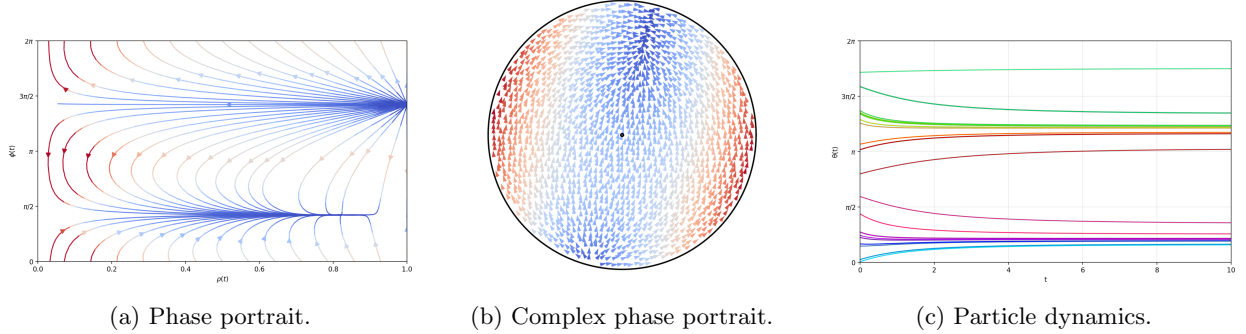


Figure 7: Phase portrait of (ρ, ϕ) -dynamics (3.10) and the trajectories of particle evolving under the transformer dynamics (2.4) for $A = I$, $V = \begin{pmatrix} -2 & 2 \\ 2 & -3 \end{pmatrix}$, which satisfies $\lambda_{\max}(V) < 0$.

Theorem 3.3 states that, when $A = I$ and the matrix V is symmetric, the long-term behavior of dynamics (3.10) is entirely determined by the sign of the largest eigenvalue of V . More precisely, if $\lambda_{\max}(V) \geq 0$, then the order parameter $\mathcal{R}_2(t)$ converges to a fixed point on the boundary ∂D . This in turn implies that the distribution f converges to a stationary distribution supported on two antipodal points. Figure 6 illustrate this regime: the phase portrait shows that the vector field points toward a boundary fixed point on ∂D , and the particle dynamics converges to two antipodal points.

By contrast, if $\lambda_{\max}(V) < 0$, then the order parameter $\mathcal{R}_2(t)$ converges to an interior equilibrium in D . Figure 7 illustrates an example in this regime. The right panel shows that the particles converge, but not to a fully synchronized configuration.

Remark 3.4 (Closest results available in the literature). The closest results in the literature concerning the role of the matrix V in the cluster formation of the transformer dynamics appear in [1, 2]. In those works, the authors analyze the *causal-attention* model and show that, when $A = I$ and V is symmetric with $\lambda_{\max}(V) > 0$, the associated dynamics collapse to a single point. This condition coincides with the one identified in our Theorem 3.3. We emphasize, however, that our setting differs from theirs in two important respects. First, we study the *self-attention* model, whose dynamical properties differ from those of the *causal-attention* model. Second, we focus on the linear attention kernel $h(y) = y$, whereas [1, 2] consider the softmax kernel $h(y) = \exp(y)$.

3.3 Case 3: Non-commuting A and V

Let A and V be given by

$$A := \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ -\frac{a_{22}}{a_{11}}v_{12} & \frac{a_{11}}{a_{22}}v_{11} \end{pmatrix}, \quad (3.11)$$

with $a_{11}, a_{22} \neq 0$.

Remark 3.5. Suppose A and V are as in (3.11). A direct computation reveals that if $a_{22} \neq a_{11}$ and $v_{12} \neq 0$, then the matrices A, V do not commute.

In this case, the polar-coordinate system (3.1) simplifies to²

$$\dot{\phi} = K(\phi) + (\rho - 1)Q(\rho, \phi), \quad (3.12)$$

where the functions H, K and Q depend on the matrices A and V ; see (C.29) in the Appendix for their explicit expressions.

In this example, our goal is to identify a parameter regime for which $\rho(t) \rightarrow 1$ while $\phi(t)$ does not converge. This behavior is guaranteed under the sufficient conditions

$$H(\phi) > 0, \quad |K(\phi)| > 0, \quad |Q(\rho, \phi)| < \infty,$$

for all $(\rho, \phi) \in [0, 1] \times \mathbb{T}^1$. The corresponding result is stated in the following theorem.

Theorem 3.6. Consider the dynamical system (3.12). If the matrices A and V defined in (3.11) satisfy:

$$\frac{v_{11}}{a_{22}} > 0, \quad \frac{1}{8} \frac{a_{22}}{a_{11}} \frac{(a_{11} - a_{22})^4}{a_{11}^2(a_{11}^2 + a_{22}^2)} < \frac{v_{11}^2}{v_{12}^2} < 4 \frac{a_{22}}{a_{11}} \frac{a_{22}^2}{(a_{11} - a_{22})^2}, \quad (3.13)$$

then for every initial condition $(\rho(0), \phi(0)) \in (0, 1] \times \mathbb{T}^1$, the solution of (3.12) satisfies that $\rho(t) \rightarrow 1$ as $t \rightarrow \infty$ while $\phi(t)$ never converges.

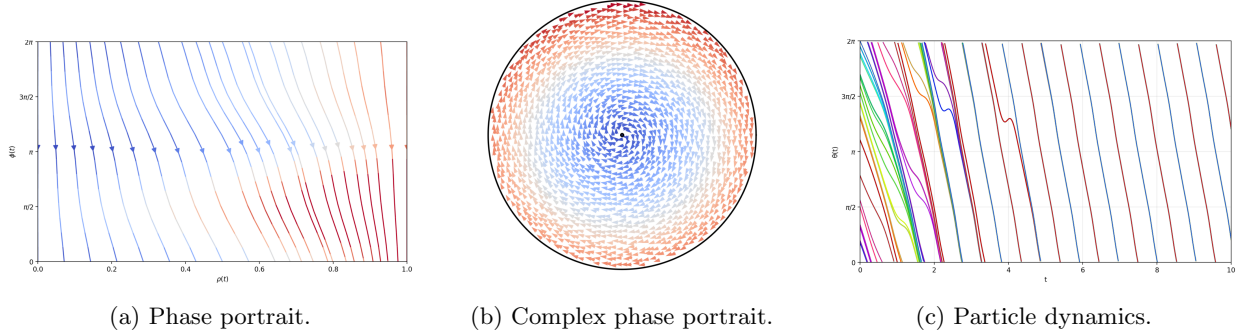


Figure 8: Phase portrait of (ρ, ϕ) -dynamics (3.12) and the trajectories of particle evolving under the transformer dynamics (2.4) for $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $V = \begin{pmatrix} 2 & 3 \\ -6 & 1 \end{pmatrix}$, which satisfies condition (3.13).

Theorem 3.6 states that, when the matrices A and V are given by (3.11)³, and their entries satisfy condition (3.13), the magnitude ρ of the order parameter $\mathcal{R}_2$ converges to 1, while its argument ϕ fails to converge. Through the relation between $\mathcal{R}_2$ and the distribution f of the mean-field particle Θ in (2.7), this implies that, for large times, the support of f is concentrated on two antipodal points. However, these two support points continue to rotate, and hence do not converge to fixed locations. Figure 8 illustrates a numerical example in this regime. In particular, the right panel shows the particle trajectories clustering around the trajectories of two antipodal points that continue to revolve around each other.

To the best of our knowledge, this is the first theoretical characterization of the long-time behavior of the transformer dynamics in a regime where both A and V are non-identity and non-commuting. The asymptotic behavior identified here is also new: the token embeddings asymptotically form two clusters, but the cluster locations continue to rotate and hence do not converge.

²We don't present the corresponding dynamics of $\mathcal{R}_2$ in this case, since the resulting expression is rather complicated and offers limited insight.

³The construction of the matrices A and V is explained in the proof of Theorem 3.6.

Remark 3.7 (Connection to the practical behavior of LLMs). We note that the clustering-without-convergence long-time behavior identified in Theorem 3.6 may not be merely an artifact of the mathematical model studied here. Indeed, the numerical results of [13] show that in Looped Transformers [46, 50, 111], where a trained LLM is recurrently applied, the output of the recurrent block follows a consistent cyclic trajectory. This observation is qualitatively reminiscent of the clustering-without-convergence phenomenon identified in our analysis, and suggests interesting directions of future investigation in practical settings.

Remark 3.8. The condition (3.13) is somewhat complicated and not immediately transparent. We therefore present here a simple family of matrices that falls into this parameter regime. In particular, consider matrices A and V of the form,

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix} \quad (3.14)$$

with $v_{11} > 0$ and $v_{12} \neq 0$. In this particular regime the matrices A and V commute, but we can more easily explain the mechanism behind the validity of Theorem 3.6. Indeed, in this regime the coupled ODE system (3.1) simplifies to

$$\dot{\rho} = \frac{v_{11}}{2}(1 - \rho^2)\rho, \quad \dot{\phi} = \frac{v_{12}}{2}(\rho^2 + 3). \quad (3.15)$$

Since $v_{11} > 0$, $\rho \in (0, 1]$, by Lemma C.8, we know that $\rho(t) \rightarrow 1$ for all initial condition $\rho(0) \in (0, 1]$. And since $|\dot{\phi}(t)| > 0$ for $t \geq 0$, we infer that $|\phi|$ does not converge in $\mathbb{T}$ as $t \rightarrow \infty$.

3.4 Case 4: Hidden Hamiltonian Dynamics

Let A and V be of the form,

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & -v_{11} \end{pmatrix} \quad (3.16)$$

with $v_{11}, v_{12} > 0$.

In this case, the dynamics (2.31) of the order parameter $\mathcal{R}_2$ and its corresponding polar-coordinate system (3.1) simplify to

$$\dot{\mathcal{R}}_2 = \frac{v_{11}}{2} (1 + |\mathcal{R}_2|^2 - 2(\mathcal{R}_2)^2) - \frac{iv_{12}}{2} (|\mathcal{R}_2|^2 + 3) \mathcal{R}_2, \quad (3.17)$$

and

$$\dot{\rho} = \frac{v_{11}}{2}(1 - \rho^2)\cos(\phi), \quad \dot{\phi} = -\frac{1}{2} \left(v_{12}(\rho^2 + 3) + v_{11} \left(3\rho + \frac{1}{\rho} \right) \sin \phi \right). \quad (3.18)$$

Interestingly, the system (3.18) turns out to be a Hamiltonian flow on the open cylinder $(0, 1) \times \mathbb{T}^1$ after an appropriate time rescaling. More precisely, define the Hamiltonian

$$H(\rho, \phi) = -\frac{v_{11}\rho}{(1 - \rho^2)^2} \sin(\phi) - \frac{v_{12}(\rho^2 + 1)}{2(1 - \rho^2)^2}.$$

Then, under the time change

$$\frac{d\tau}{dt} = \frac{(1 - \rho^2)^3}{2\rho}, \quad (3.19)$$

the dynamics (3.18) on the domain $(0, 1) \times \mathbb{T}^1$ can be rewritten in Hamiltonian form as

$$\frac{d\rho}{d\tau} = -\partial_\phi H, \quad \frac{d\phi}{d\tau} = \partial_\rho H. \quad (3.20)$$

Moreover, this system exhibits a saddle-node bifurcation [98], indicating that even a small perturbation of V may drastically alter the long-time behavior of the system. The long-time clustering behavior of (3.18) is described in the following proposition.

Theorem 3.9. *Consider the dynamical system (3.18), and assume that $v_{11}, v_{12} > 0$.*

1. If $v_{11} < v_{12}$, then for almost every initial condition $(\rho(0), \phi(0)) \in (0, 1) \times \mathbb{T}^1$ the associated solution is bounded away from the boundary $\rho \in \{0, 1\}$. Furthermore, the trajectory is periodic, and hence forms a closed orbit in $(0, 1) \times \mathbb{T}^1$.
2. If $v_{11} \geq v_{12}$, then for almost every initial condition $(\rho(0), \phi(0)) \in (0, 1] \times \mathbb{T}^1$ the associated solution converges to a boundary equilibrium $(1, \phi_\infty)$ as $t \rightarrow \infty$, where $\phi_\infty := 2\pi - \arcsin(v_{12}/v_{11})$.

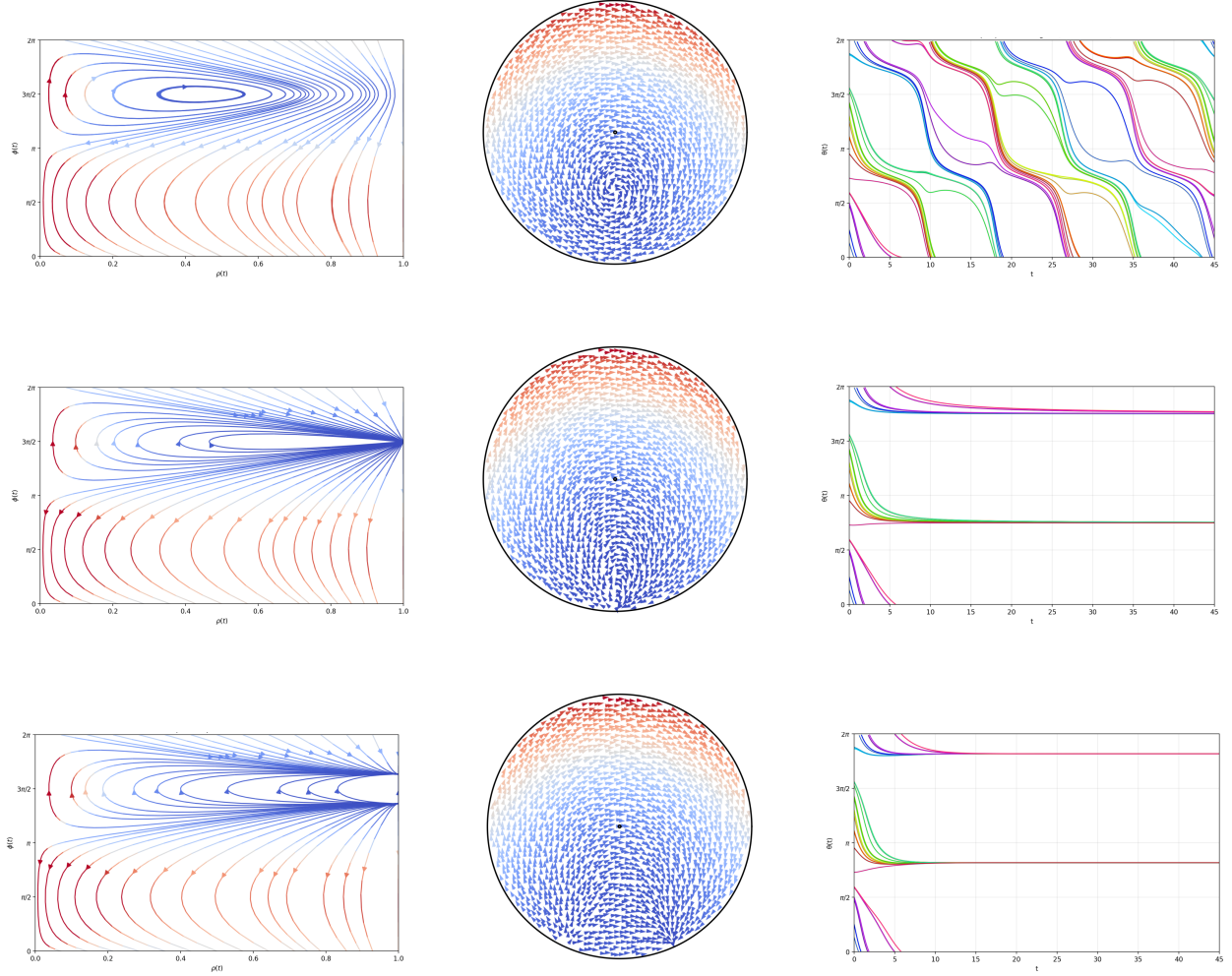


Figure 9: Figures showing the phase portraits and bifurcation phenomena in the dynamics (2.4) with $A = I$ and $V = \begin{pmatrix} v_{11} & 1 \\ -1 & -v_{11} \end{pmatrix}$. **Top:** $v_{11} = 0.9$. **Middle:** $v_{11} = 1.0$. **Bottom:** $v_{11} = 1.1$. The first column shows the cartesian phase portraits, the middle column shows the complex portraits, and the final column the particle trajectories. Note that $T = 45$ in the trajectories in order to highlight cyclic and slow convergence behavior. Moving from top to bottom, the panels illustrate the transition from a stable limit cycle, to a semistable fixed point, and then to a stable-unstable pair of equilibria. Note the slow convergence in the middle panel at the bifurcation point.

Theorem 3.9 states that, when $A = I$ and the matrix V is given by (3.16), the system (3.18) undergoes a saddle-node bifurcation and exhibits qualitatively distinct long-term behavior across different parameter regimes. In particular, when $v_{11} < v_{12}$, almost any solution (ρ, ϕ) to (3.18) is periodic and traces out a closed orbit. Since the distribution g of the double-angle variable $\Xi = 2\Theta$ is always wrapped Cauchy and is completely determined by the order parameter $\mathcal{R}_2$, it follows that g is likewise periodic and therefore does not converge to a stationary limit. Moreover, through the relationship between $\mathcal{R}_2$ and the distribution f

of the original particle Θ , this also implies that f continues to oscillate in time, possibly aperiodically. By contrast, when $v_{11} \geq v_{12}$, $\mathcal{R}_2(t)$ converges to a boundary equilibrium, which in turn implies that f converges to a stationary distribution supported on two antipodal points. This bifurcation behavior is illustrated in Figure 9.

To the best of our knowledge, this is the first theoretical characterization of bifurcation phenomena in the transformer dynamics. This result, together with Theorem 3.6, goes beyond the clustering and synchronization behaviors that have been the primary focus of the existing literature, and highlights the richness of dynamical behaviors that can already arise in linear transformers in dimension $d = 2$. It also suggests that, for certain choices of the matrices A and V , the transformer output may be sensitive even to small perturbations of these parameters. This observation may be particularly relevant in the context of transformer model quantization [55, 109].

Remark 3.10. We note that in the case $v_{11} \geq v_{12}$ we prove convergence to a fixed point despite the fact that the system can be described in Hamiltonian form (3.20). This is not a contradiction, as the Hamiltonian description is only valid on the *open* cylinder $(0, 1) \times \mathbb{T}^1$, whereas the full dynamics are well defined at $\rho = 1$. It turns out that the energy degenerates in a convenient way in the limit $\rho \rightarrow 1$, where all level sets agree on the same phase ϕ_∞ .

Remark 3.11. The proof in Appendix C.1 provides a more detailed description of the cyclic behavior in the regime $v_{11} < v_{12}$. In particular, we characterize the regions where dynamics flow *around* the cylinder (so called ‘rotations’) versus the regions where the dynamics flow in a closed region on the cylinder (so called ‘librations’). These can be seen in the top left panel of Figure 9, where rotations are observed as flow lines completing a cycle inside the plot, and librations are flow lines leaving the bottom of the plot and wrapping around to the top. These two behaviors are separated by a special Hamiltonian level set $\{H = -v_{12}/2\}$. Level sets with values above this separatrix correspond to rotations, while those below correspond to librations; see Figure 21a for an illustration.

Remark 3.12. In typical dynamical systems, slow dynamics occur in parameter regimes around a bifurcation. We have empirically observed such a bottleneck in our system. In particular, we have observed that when $v_{11} \approx v_{12}$ and $v_{11} < v_{12}$ the system spends most of its time close to $(1, \phi_\infty)$ (this is the point where all flow lines coalesce in the middle left figure). We view this phenomenon as an undesirable form of *metastability*, which was studied in a different context for the self-attention dynamics [48]. Generalizing and making this observation rigorous is an interesting question that we leave for future work.

4 Long-Time Behavior Beyond the Ott–Antonsen Ansatz: Stability Analysis

In Section 3, we analyzed the long-time behavior of order parameter $\mathcal{R}_2$ under the Ott–Antonsen ansatz, thereby characterizing the corresponding behavior of the particle dynamics (2.6) for different choices of the matrices A and V . However, this analysis relies on the assumption that the double-angle particle $\Xi = 2\Theta$ is initialized within the wrapped Cauchy family (2.28). In this section, we go beyond the OA ansatz by studying the long-time behavior of $\mathcal{R}_2$ in the WS system (2.21).

More specifically, recall that in the general WS system (2.21), the order parameter $\mathcal{R}_2$ depends on both WS variables α and η through (2.26), which makes its long-time behavior difficult to analyze directly. On the other hand, as discussed in Section 2.2.1, whenever $|\alpha| \rightarrow 1$ we also have $\mathcal{R}_2 = \alpha$ and thus $|\mathcal{R}_2| \rightarrow 1$, corresponding to the full synchronization of the particle dynamics. Crucially, this conclusion holds *independently* of the OA ansatz.

Motivated by this observation, in Section 4.1 we analyze the structural stability of the α -dynamics (2.21a). We then apply this result in Section 4.2 to characterize the long-time behavior of α and identify the choices of the matrices A and V for which $|\alpha| \rightarrow 1$. In turn, this yields the convergence $|\mathcal{R}_2| \rightarrow 1$.

Before proceeding, we recall several equations from Section 2 that will be used later. We consider the WS system in (2.21), which we rewrite here for convenience:

$$\dot{\alpha}(t) = F(\alpha(t), \mathcal{R}_2(t)), \quad (4.1a)$$

$$\dot{\eta}(t) = G(\alpha(t), \mathcal{R}_2(t)), \quad (4.1b)$$

where the functions F and G are defined in (2.19). We also introduce the polar-coordinate form $\alpha = \gamma e^{i\Phi}$ with $\gamma : [0, \infty) \mapsto [0, 1]$ and $\Phi : [0, \infty) \mapsto \mathbb{T}^1$. Recall that applying the Möbius transform (2.16) to the double-angle variable Ξ defined in (2.12) gives

$$e^{i\Xi(t;\varphi)} = M_{\alpha(t), \eta(t)}(e^{i\varphi}) = \frac{\alpha(t) + e^{i(\varphi + \eta(t))}}{1 + \bar{\alpha}(t)e^{i(\varphi + \eta(t))}}. \quad (4.2)$$

The constant of motion φ is time-independent, and hence its associated distribution ν is preserved over time. Moreover, we have the relation $\Xi = T_{\alpha, \eta}(\varphi)$, where the map $T_{\alpha, \eta}$ is defined in (2.24). We also recall from (2.26) that the order parameter can be expressed as the pushforward average of the distribution ν induced by the WS variables α and η :

$$\mathcal{R}_2(t) = \int_0^{2\pi} e^{i\xi} (T_{\alpha(t), \eta(t)})_{\#} \nu(d\xi) = \int_0^{2\pi} M_{\alpha(t), \eta(t)}(e^{i\varphi}) \nu(d\varphi). \quad (4.3)$$

Equations (4.1)–(4.3) provide a closed WS representation of the mean-field dynamics (2.6) in terms of (α, η) and the fixed distribution ν of the constant of motion φ .

4.1 Stability of the WS Dynamics Beyond the OA-manifold

Recall from Section 2.2.2 that the α -dynamics (4.1a) reduces to the OA dynamics (2.31) when the distribution of the double-angle variable Ξ is initialized within the wrapped Cauchy family (2.28). This condition on Ξ is equivalent to the distribution ν being uniform (see Lemma 2.6). In this section, we perform a stability analysis of the α -dynamics (4.1a) under perturbations away from the OA regime by quantifying its deviation from the OA dynamics (2.31) when ν is not the uniform measure.

We begin by outlining the main idea of the argument. Our analysis is based on a comparison between the α -dynamics (4.1a) with a OA-closure counterpart

$$\dot{\alpha}_{\text{OA}}(t) = F(\alpha_{\text{OA}}(t), \alpha_{\text{OA}}(t)). \quad (4.4)$$

To this end, we rewrite (4.1a) in the form of a perturbation of the OA dynamics (4.4):

$$\dot{\alpha}(t) = F(\alpha(t), \mathcal{R}_2(t)) = F(\alpha(t), \alpha(t)) + \Delta(t), \quad (4.5)$$

where the perturbation term $\Delta : [0, \infty) \mapsto \mathbb{C}$ is defined by

$$\Delta(t) := F(\alpha(t), \mathcal{R}_2(t)) - F(\alpha(t), \alpha(t)). \quad (4.6)$$

From (2.19), the function F appearing in (4.1a) is smooth and globally Lipschitz on $D \times D$, where D is the closed complex unit disc. We denote the corresponding Lipschitz constant by L_F . Assuming that (4.4) and (4.5) share the same initial condition $\alpha(0) = \alpha_{\text{OA}}(0)$, then a standard Grönwall argument comparing the two equations gives, for every $T > 0$,

$$\sup_{t \in [0, T]} |\alpha(t) - \alpha_{\text{OA}}(t)| \leq \int_0^T e^{L_F(T-s)} |\Delta(s)| ds.$$

Moreover, again by the Lipschitz continuity of F , we have

$$|\Delta(t)| = |F(\alpha(t), \mathcal{R}_2(t)) - F(\alpha(t), \alpha(t))| \leq L_F |\mathcal{R}_2(t) - \alpha(t)|. \quad (4.7)$$

Therefore, the problem is reduced to estimating the discrepancy between the order parameter $\mathcal{R}_2$ and the WS parameter α within the WS system (4.1).

In the sequel, we first establish upper bounds for $|\mathcal{R}_2(t) - \alpha(t)|$ in Proposition 4.1, and then use these estimates to derive a stability bound for $\sup_{t \in [0, T]} |\alpha(t) - \alpha_{\text{OA}}(t)|$ in Proposition 4.3.

Before stating the main results, we recall the bounded-Lipschitz distance between measures $\mu, \tilde{\mu}$ over $[0, 2\pi)$, which is given by

$$d_{\text{BL}}(\mu, \tilde{\mu}) = \sup \left\{ \left| \int_0^{2\pi} g d\mu - \int_0^{2\pi} g d\tilde{\mu} \right| : \|g\|_\infty \leq 1, \text{Lip}(g) \leq 1 \right\}. \quad (4.8)$$

We also clarify a potential source of confusion concerning the notation. In the WS system, the order parameter $\mathcal{R}_2(t)$ does not, in general, evolve according to the OA dynamics (2.31) studied in Section 3, which is only valid under the OA ansatz. To avoid ambiguity, we denote by α_{OA} the variable satisfying the OA dynamics, as we wrote in (4.4), which coincides with (2.31), while reserving the notation $\mathcal{R}_2$ for the actual order parameter.

Proposition 4.1. *Let Unif denote the uniform probability measure on $[0, 2\pi)$, and let ν be a probability measure on $[0, 2\pi)$. Let $\alpha(t) = \gamma(t)e^{i\Phi(t)}$ be a solution of the dynamics (4.1a), and let $\mathcal{R}_2(t)$ denote the corresponding order parameter defined by (4.3) with respect to ν . Then, for every $t \geq 0$, the difference between $\mathcal{R}_2(t)$ and $\alpha(t)$ admits the following two upper bounds:*

$$|\mathcal{R}_2(t) - \alpha(t)| \leq C_\nu(\gamma(t))(1 - \gamma(t)), \quad (4.9)$$

where

$$C_\nu(\gamma) := \sup_{\chi \in [0, 2\pi)} \int_0^{2\pi} \frac{2}{|1 + \gamma e^{i(\varphi + \chi)}|} \nu(d\varphi), \quad (4.10)$$

and

$$|\mathcal{R}_2(t) - \alpha(t)| \leq \left(1 + \frac{2}{1 - \gamma(t)}\right) d_{\text{BL}}(\nu, \text{Unif}). \quad (4.11)$$

Before turning to the proof, we briefly discuss the interpretation of Proposition 4.1. The proposition shows that the discrepancy between the order parameter $\mathcal{R}_2$ and the WS parameter α within the WS system (4.1) can be small in two different ways, provided we can control $C_\nu(\gamma)$ and $(1 - \gamma)$. On the one hand, inequality (4.9) shows that the discrepancy is controlled by the factor $1 - \gamma$, and is therefore small when the modulus γ of α is close to 1, assuming that the prefactor $C_\nu(\gamma)$ remains controlled. On the other hand, inequality (4.11) shows that the discrepancy is small when the probability measure ν is close to the uniform measure Unif in bounded-Lipschitz distance, as long as γ remains uniformly bounded away from 1.

Proof of Proposition 4.1. We show that the difference $|\mathcal{R}_2(t) - \alpha(t)|$ admits two different upper bounds. For notational simplicity, we suppress the time dependence throughout the proof.

We first verify inequality (4.9). Using the form of $\mathcal{R}_2$ in (4.3), we compute

$$\begin{aligned} |\mathcal{R}_2 - \alpha| &= \left| \int_0^{2\pi} \frac{\alpha + e^{i(\varphi + \eta)}}{1 + \bar{\alpha} e^{i(\varphi + \eta)}} d\nu(\varphi) - \alpha \right| = \left| \int_0^{2\pi} \frac{e^{i(\varphi + \eta)}}{1 + \bar{\alpha} e^{i(\varphi + \eta)}} d\nu(\varphi) \right| (1 - |\alpha|^2) \\ &\leq \int_0^{2\pi} \frac{1}{|1 + \gamma e^{i(\varphi + \eta - \Phi)}|} d\nu(\varphi) (1 - \gamma^2) \leq C_\nu(\gamma)(1 - \gamma). \end{aligned} \quad (4.12)$$

Here, the first inequality applies the integral triangle inequality and uses the polar-representation $\alpha = \gamma e^{i\Phi}$, and the second inequality makes use of $(1 + \gamma) \leq 2$ and the definition of $C_\nu(\gamma)$.

Next we prove inequality (4.11). Recall the Möbius transform from (2.16). By Lemma D.1, we have

$$\alpha = \int_0^{2\pi} M_{\alpha, \eta}(e^{i\varphi}) d\text{Unif}(\varphi).$$

Combining this identity with (4.3), we find that

$$\mathcal{R}_2 - \alpha = \int_0^{2\pi} M_{\alpha, \eta}(e^{i\varphi}) d(\nu - \text{Unif})(\varphi).$$

By the definition of d_{BL} in (4.8) and Lemma D.1, we have

$$|\mathcal{R}_2 - \alpha| \leq (\|M_{\alpha, \eta}\|_\infty + \text{Lip}(M_{\alpha, \eta})) d_{\text{BL}}(\nu, \text{Unif}) \leq \left(1 + \frac{2}{1 - \gamma}\right) d_{\text{BL}}(\nu, \text{Unif}).$$

□

Using the polar-coordinate representations $\alpha = \gamma e^{i\Phi}$ and $\mathcal{R}_2 = \rho e^{i\phi}$, together with elementary inequalities, we obtain the following corollary as a direct consequence of Proposition 4.1. The associated proof is deferred to Appendix D.

Corollary 4.2. *Let $\alpha(t) = \gamma(t)e^{i\Phi(t)}$ be a solution of the dynamics (4.1a), and let $\mathcal{R}_2(t) = \rho(t)e^{i\phi(t)}$ be defined by (4.3). Then the following inequalities hold*

$$|\rho(t) - \gamma(t)| \leq C_\nu(\gamma)(1 - \gamma(t)), \quad |e^{i\Phi(t)} - e^{i\phi(t)}| \leq 2C_\nu(\gamma) \frac{1 - \gamma(t)}{\gamma(t)}, \quad |\Phi(t) - \phi(t)| \leq 2\pi C_\nu(\gamma) \frac{1 - \gamma(t)}{\gamma(t)}, \quad (4.13)$$

where $C_\nu(\gamma)$ is defined in (4.10).

Using the discrepancy estimates between $\mathcal{R}_2(t)$ and $\alpha(t)$ derived in Proposition 4.1, together with the discussion at the beginning of this section, we are now in a position to establish the stability of the α -dynamics (4.1a) with respect to its OA counterpart (4.4).

Proposition 4.3 (Stability of the α -dynamics). *Let ν be a probability measure on $[0, 2\pi)$ and Unif denote the uniform probability measure on $[0, 2\pi)$. Let (α, η) be a solution of the dynamics (4.1) with initial data $(\alpha(0), \eta(0))$, and α_{OA} be a solution of the dynamics (4.4) satisfying $\alpha_{\text{OA}}(0) = \alpha(0)$. Then, for arbitrary $\eta(0) \in \mathbb{T}^1$ and every $T > 0$,*

$$\sup_{t \in [0, T]} |\alpha(t) - \alpha_{\text{OA}}(t)| \leq L_F e^{L_F T} \int_0^T \min \left\{ C_\nu(\gamma(t))(1 - \gamma(t)), \left(1 + \frac{2}{1 - \gamma(t)} \right) d_{\text{BL}}(\nu, \text{Unif}) \right\} dt.$$

Proposition 4.3 follows a standard Grönwall argument applied to the comparison between the two ODEs (4.4) and (4.5), together with the two upper bound estimates on $|\mathcal{R}_2 - \alpha|$ established in Proposition 4.1. The resulting estimate is therefore in the same spirit as Proposition 4.1. In particular, it states that, on any finite time interval $[0, T]$, the deviation of the α -dynamics from its OA counterpart α_{OA} is controlled in two ways: first, by the factor $1 - \gamma(t)$, which measures the degree of synchronization of the particle dynamics; and second, by the proximity of the measure ν to the uniform distribution Unif, both up to a multiplicative constant. This estimate will serve as one of the key ingredients in identifying different choices of the matrices A and V such that $|\alpha| \rightarrow 1$. We develop these applications in Section 4.2.

Although the proof of Proposition 4.3 is straightforward, we include it for completeness.

Proof of Proposition 4.3. Let $e(t) = \alpha(t) - \alpha_{\text{OA}}(t)$. Subtracting (4.4) and (4.5), and using the Lipschitz continuity of F , we obtain

$$|\dot{e}(t)| \leq |F(\alpha(t), \alpha(t)) - F(\alpha_{\text{OA}}(t), \alpha_{\text{OA}}(t))| + |\Delta(t)| \leq L_F |e(t)| + |\Delta(t)|.$$

Since $\alpha(0) = \alpha_{\text{OA}}(0)$, we have $e(0) = 0$. Integrating the above inequality over $[0, t]$ with $0 \leq t \leq T$, we therefore obtain

$$|e(t)| \leq L_F \int_0^t |e(s)| ds + \int_0^t |\Delta(s)| ds.$$

Applying Grönwall inequality yields

$$|e(t)| \leq \int_0^t e^{L_F(t-s)} |\Delta(s)| ds \leq e^{L_F T} \int_0^T |\Delta(s)| ds.$$

Using again the Lipschitz continuity of F and the Proposition 4.1, we obtain that

$$\begin{aligned} |\Delta(t)| &= |F(\alpha(t), \mathcal{R}_2(t)) - F(\alpha(t), \alpha(t))| \\ &\leq L_F |\mathcal{R}_2(t) - \alpha(t)| \leq L_F \min \left\{ C_\nu(\gamma(t))(1 - \gamma(t)), \left(1 + \frac{2}{1 - \gamma(t)} \right) d_{\text{BL}}(\nu, \text{Unif}) \right\}. \end{aligned}$$

We conclude the proof by combining the above estimates and taking the supremum over $t \in [0, T]$. $\square$

4.2 Long-Time Behavior

In this section, we revisit Cases 1 and 2 studied in Section 3 and focus on the matrices A and V for which the associated OA dynamics (4.4) converges to full synchronization, namely, $|\alpha_{\text{OA}}| \rightarrow 1$. Building on the stability of the α -dynamics (4.5) in the WS system established in Section 4.1, we prove that the α -dynamics likewise satisfies $|\alpha(t)| \rightarrow 1$ as $t \rightarrow \infty$. This requires slightly stronger assumptions on A and V , together with the assumption that the distribution ν satisfies an L^p regularity condition and is sufficiently close to the uniform measure Unif . The corresponding main results for Cases 1 and 2 are presented in Theorem 4.7 and Theorem 4.8, respectively.

Before turning to the case studies, we formally state the L^p regularity assumption on ν and briefly explain its role.

Assumption 4.4 (Bounded L^p density). *For some $p > 1$ and $M > 0$, the probability measure ν has a density, still denoted by ν , satisfying*

$$\|\nu\|_{L^p([0, 2\pi))} \leq M.$$

Remark 4.5 (Relation to the initial condition $f(0)$). The regularity assumption $\nu \in L^p([0, 2\pi))$ is equivalent to the condition $f(0) \in L^p([0, 2\pi))$, where $f(0)$ is the initial distribution of the physical variable Θ in the mean-field linear self-attention model (2.6).

In the rest of this remark, all quantities are evaluated at $t = 0$. Recall from (2.25) that the distribution g of the double-angle variable $\Xi = 2\Theta$ is given by $g = (T_{\alpha, \eta})_{\#} \nu$, where $T_{\alpha, \eta}$ is defined in (2.24). As discussed in Section 2.2.1, whenever $|\alpha| \neq 1$, the map $T_{\alpha, \eta}$ is a diffeomorphism. It follows that $\nu \in L^p([0, 2\pi))$ if and only if $g \in L^p([0, 2\pi))$. By the relationship between f and g given in (2.15), it is equivalent to require $f(0) \in L^p([0, 2\pi))$.

Remark 4.6 (The role of Assumption 4.4). Assumption 4.4 is motivated by the upper bound (4.9) on the discrepancy between the order parameter $\mathcal{R}_2$ and the WS variable α established in Proposition 4.1. In particular, to prove convergence of the α -dynamics to full synchronization in Theorems 4.7 and 4.8, one needs to ensure that $|\mathcal{R}_2 - \alpha| \rightarrow 0$ as the modulus $\gamma = |\alpha| \rightarrow 1$. In view of (4.9), it is enough to impose the following abstract regularity condition on ν :

$$C_\nu(\gamma)(1 - \gamma) \rightarrow 0 \quad \text{as } \gamma \rightarrow 1, \quad (4.14)$$

where $C_\nu(\gamma)$ is defined in (4.10). Assumption 4.4 provides a sufficient condition that ensures that (4.14) holds; see Lemma D.3.

We note that $\nu \in L^p([0, 2\pi))$ does not cover the atomic distributions. Identifying more general assumptions on ν that also allow for atomic components requires a further analysis of the WS system (4.1). We leave this question for future work.

We are now ready to revisit the case studies and identify the parameter regimes of the matrices A and V for which the α -dynamics converges to full synchronization. Our analysis is carried out using the polar-coordinate representation $\alpha = \gamma e^{i\Phi}$. In particular, the α -dynamics in (4.5) can be rewritten in terms of (γ, Φ) as

$$\dot{\gamma} = \text{Re}(e^{-i\Phi} F(\alpha, \alpha)) + \Delta_\gamma, \quad (4.15a)$$

$$\dot{\Phi} = \frac{1}{\gamma} \text{Im}(e^{-i\Phi} F(\alpha, \alpha)) + \Delta_\Phi \quad (4.15b)$$

where

$$\Delta_\gamma(t) := \text{Re}(e^{-i\Phi} \Delta(t)), \quad \Delta_\Phi(t) := \frac{1}{\gamma} \text{Im}(e^{-i\Phi} \Delta(t)) \quad (4.16)$$

with $\Delta(t)$ defined in (4.6).

Case 1 Revisited.

We consider the case where A is a real-valued matrix and V is the identity matrix, as defined in (3.3). In this setting, the polar-coordinate system (4.15) simplifies to

$$\begin{aligned}\dot{\gamma} &= \frac{1}{4} (1 - \gamma^2) (a_+^d \gamma + a_-^d \cos(\Phi) + a_+^o \sin(\Phi)) + \Delta_\gamma, \\ \dot{\Phi} &= \frac{1}{4} (1 - \gamma^2) \left(a_-^o - \frac{a_-^d \sin(\Phi) - a_+^o \cos(\Phi)}{\gamma} \right) + \Delta_\Phi,\end{aligned}\tag{4.17}$$

where the explicit forms of Δ_γ and Δ_Φ are provided in (D.23).

The long-term behavior of (4.17) is stated in the following theorem.

Theorem 4.7. *Fix any $\iota > 0$, and consider the dynamical system (4.17) with initial data*

$$(\gamma(0), \Phi(0)) \in \{(\gamma, \Phi) \in (0, 1] \times \mathbb{T}^1 : |(\gamma, \Phi) - (\rho_{\text{eq}}, \phi_{\text{eq}})| \geq \iota\},\tag{4.18}$$

where $(\rho_{\text{eq}}, \phi_{\text{eq}})$ is defined in (C.14). Assume that the matrix A satisfies

$$\frac{A + A^\top}{2} \succ 0.\tag{4.19}$$

Moreover, assume that the probability measure ν satisfies Assumption 4.4 and

$$d_{\text{BL}}(\nu, \text{Unif}) < \varepsilon(\iota, p, M, A),\tag{4.20}$$

where $\varepsilon(\iota, p, M, A) > 0$ is a sufficiently small constant depending only on ι , the constants p and M in Assumption 4.4, and the matrix A . Then the solution of (4.17) converges to $(1, \Phi_\infty)$ as $t \rightarrow \infty$ for some $\Phi_\infty \in \mathbb{T}^1$.

Theorem 4.7 shows that the WS variable α converges to a fixed point on the boundary ∂D , provided that $V = I$, the symmetric part of A is positive definite, and the distribution $\nu \in L^p([0, 2\pi))$ is sufficiently close to the uniform measure Unif on $[0, 2\pi)$. As discussed in Section 2.2.1, one has $\mathcal{R}_2 = \alpha$ whenever $|\alpha| = 1$. Consequently, the order parameter converges to the same boundary fixed point as $t \rightarrow \infty$. This implies that the distribution f of the physical variable Θ , induced by the linear self-attention model (2.6), concentrates at two a pair of fixed antipodal points.

We also briefly comment on the assumption on the initial condition in (4.18). This assumption merely excludes the case in which the initial datum $(\gamma(0), \Phi(0))$ coincides with $(\rho_{\text{eq}}, \phi_{\text{eq}})$, the interior equilibrium of the OA dynamics (4.4) in this parameter regime; see Figure 4a for an illustration. By Theorem 3.1, the condition (4.18) ensures that the corresponding solution of the OA dynamics (4.4) with this initialization converges to full synchronization, namely, $|\alpha_{\text{OA}}| \rightarrow 1$. This convergence is a key ingredient in establishing the corresponding convergence of the WS variable α .

Case 2 Revisited.

We revisit now the case where A is the identity matrix and V is a symmetric real-valued matrix, as in (3.8). In this setting, the polar-coordinate system (4.15) simplifies to

$$\begin{aligned}\dot{\gamma} &= \frac{1}{4} (1 - \gamma^2) (v_+^d \gamma + v_-^d \cos(\Phi) + v_+^o \sin(\Phi)) + \Delta_\gamma \\ \dot{\Phi} &= -\frac{1}{4} \frac{3\gamma^2 + 1}{\gamma} (v_-^d \sin(\Phi) - v_+^o \cos(\Phi)) + \Delta_\Phi,\end{aligned}\tag{4.21}$$

where the explicit forms of Δ_γ and Δ_Φ are provided in (D.25).

The long-term behavior of (4.21) is stated in the following theorem.

Theorem 4.8. *Fix any $\iota > 0$, and consider the dynamical system (4.21) with initial data*

$$(\gamma(0), \Phi(0)) \in \{(\gamma, \Phi) \in (0, 1] \times \mathbb{T}^1 : |\Phi - (\Phi_V + \pi)| \geq \iota\},\tag{4.22}$$

where Φ_V is defined in (D.4). Assume that the matrix V satisfies

$$\lambda_{\max}(V) > 0.\tag{4.23}$$

Moreover, assume that the probability measure ν satisfies Assumption 4.4 and

$$d_{\text{BL}}(\nu, \text{Unif}) < \varepsilon(\iota, p, M, V) \quad (4.24)$$

for a sufficiently small constant $\varepsilon(\iota, p, M, V) > 0$ depending only on ι , the constants p and M in Assumption 4.4, and on the matrix V . Then the solution of (4.21) converges to $(1, \Phi_\infty)$ as $t \rightarrow \infty$ for some $\Phi_\infty \in \mathbb{T}^1$.

Theorem 4.8 establishes a similar qualitative result as Theorem 3.3: the WS variable α converges globally to a boundary fixed point on ∂D , provided that $A = I$ and the largest eigenvalue of the symmetric matrix V is strictly positive. It follows that the order parameter $\mathcal{R}_2(t)$ also converges to a boundary fixed point, and the associated distribution f converges to a stationary distribution supported on two antipodal points.

As with condition (4.18) in Theorem 4.7, the assumption on the initial condition in stated in (4.22) excludes an exceptional equilibrium of the OA dynamics. More precisely, this assumption rules out that $\Phi(0) = \Phi_V + \pi$, which is an equilibrium of the angular dynamics in the OA system (4.4) in this parameter regime. By Theorem 3.3, condition (4.22) ensures that the corresponding solution of the OA dynamics converges to full synchronization. This convergence again serves as a key ingredient in establishing the corresponding convergence of the WS variable α .

The overall proof strategies for Theorems 4.7 and 4.8 are similar. We summarize the main ideas below and defer the full details to Appendix D.

Proof sketch for Theorems 4.7 and 4.8. In Section 3, we identified parameter regimes for the matrices A and V under which the OA dynamics (4.4) converges to a full synchronization state; that is, $|\alpha_{\text{OA}}(t)| \rightarrow 1$ for almost every initial condition. We now explain why, in these regimes, the associated α -dynamics (4.5) also converge to full synchronization states, provided that the distribution ν is close enough to the uniform measure over $[0, 2\pi)$. The argument consists of two key steps, which show the convergence of the modulus $\gamma(t) \rightarrow 1$.

- **Step 1.** We first show that, there exists a threshold value $\gamma_{\text{th}} \in (0, 1]$ so that, if $\gamma(t_0) \in [\gamma_{\text{th}}, 1]$, then $\gamma(t)$ converges to 1 exponentially fast for $t \geq t_0$. This follows from the fact that, whenever $\gamma(t) \in [\gamma_{\text{th}}, 1]$, the first term in (4.15a) is positive and strictly dominate the perturbation term Δ_γ .
- **Step 2.** It remains to show that, for almost every initial condition $\gamma(0) \notin [\gamma_{\text{th}}, 1]$, the dynamics (4.15a) enters this regime in finite time. This follows from the the stability of the α -dynamics (4.5) with respect to the perturbation of the OA dynamics (4.4), as established in Proposition 4.3, together with the asymptotic convergence of the OA dynamics to the fully synchronized state. Indeed, since $|\alpha_{\text{OA}}(t)| \rightarrow 1$, then there exists a time t_0 so that $|\alpha_{\text{OA}}(t_0)| > \gamma_{\text{th}}$. Proposition 4.3 then implies that, if ν is sufficiently close to the uniform measure Unif, then

$$\sup_{t \in [0, t_0]} |\alpha(t) - \alpha_{\text{OA}}(t)|$$

is sufficiently small, and hence $\gamma(t_0) = |\alpha(t_0)| \geq \gamma_{\text{th}}$.

Lastly, by using the convergence of the modulus $\gamma(t)$, together with the properties of the Φ -dynamics, we further show that the angular variable $\Phi(t)$ converges to a fixed point in $\mathbb{T}^1$. $\square$

We note that our proof does not rely solely on a standard Grönwall-type estimate established in Proposition 4.3, which by itself is insufficient to conclude the asymptotic convergence of the α -dynamics (4.5) without stronger assumptions such as $\nu \rightarrow \text{Unif}$ in d_{BL} . Indeed, we additionally need to exploit the structural properties of the dynamics induced by the linear self-attention model for the particular choices of the matrices A and V , as reflected in Step 1 of the proof sketch above. This additional structure allows us to extend the Grönwall-type argument to the infinite time horizon.

We close this section by comparing Theorems 4.7 and 4.8 with their OA counterparts, Theorems 3.1 and 3.3. The conditions on the matrices A and V in Theorems 4.7 and 4.8, which ensure the desired convergence of the α -dynamics (4.5) in the WS system, are slightly stronger than the corresponding conditions used to establish the convergence of the OA dynamics (4.4) in Theorems 3.1 and 3.3. This is natural since the general α -dynamics (4.5) involves an additional perturbation term Δ that is not present in the OA dynamics. As the

proofs of Theorems 4.7 and 4.8 will show, establishing convergence of the α -dynamics requires controlling Δ by the first term in (4.5), and this leads to the stronger conditions on A and V that we impose.

Nevertheless, Theorems 4.7 and 4.8 characterize the long-term behaviors of the general WS system (4.1) for certain choices A and V . In turn, they extend the analysis of the mean-field dynamics (2.6) induced by the linear self-attention model beyond the special class of wrapped Cauchy initializations. Specifically, they allow for initial distributions that are close to, but not necessarily exactly on, the OA manifold.

5 Experiments

In this section, we present numerical simulations of the continuous-time self-attention dynamics

$$\dot{x}_k(t) = P_{x_k(t)}^\perp \left(\frac{1}{n} \sum_{j=1}^n h(\beta \langle Ax_k(t), x_j(t) \rangle) V x_j(t) \right), \quad k = 1, \dots, n, \quad (5.1)$$

on the unit sphere $\mathbb{S}^{d-1}$, considering both linear self-attention (LSA), corresponding to $h(y) = y$, and unnormalized softmax attention (USA), corresponding to $h(y) = \exp(y)$. The latter model has been analyzed for some specific choices of the parameter matrices A and V in [1, 2, 9, 27, 47, 48].

We conduct simulations in higher dimensions $d \gg 2$ under parameter regimes analogous to those analytically identified in Section 3. The experimental results suggest that the theoretical characterizations of the long-time behavior of the 2-dimensional linear self-attention model, presented in Sections 3.1–3.4, persist in higher-dimensional settings. In addition, we observe that, within the identified parameter regimes, the USA model exhibits similar qualitative behaviors such as clustering and cyclicity, while also displaying some differences from the LSA model.

Before presenting our experimental results, let us first specify the diagnostics used to quantify clustering and cyclic behaviors, as well as the general experimental setup.

Clustering and cyclicity diagnostics. We track three quantities to measure clustering behavior and one quantity to measure cyclic behavior.

For clustering, we use the mean-squared cosine similarity

$$\widehat{\mathcal{R}}_2(t) := \frac{1}{\binom{n}{2}} \sum_{j < k} \langle x_j(t), x_k(t) \rangle^2 \in [0, 1] \quad (5.2)$$

as the primary diagnostic. This quantity is equal to 1 whenever all tokens collapse to a pair of antipodal points $\{x^*, -x^*\}$, allowing the two clusters to contain different numbers of tokens; this includes the case in which all tokens collapse to a single point x^* . We note that the clustering diagnostic $\widehat{\mathcal{R}}_2(t)$ is directly related to the second order parameter $\mathcal{R}_2^n(t)$ defined in (2.3). More precisely, in dimension $d = 2$, it holds

$$|\mathcal{R}_2^n(t)|^2 = \widehat{\mathcal{R}}_2(t) + \frac{n-2}{n}(\widehat{\mathcal{R}}_2(t) - 1).$$

In particular, $|\mathcal{R}_2^n(t)| = 1$ if and only if $\widehat{\mathcal{R}}_2(t) = 1$.

We also track the mean cosine similarity

$$\widehat{\mathcal{R}}_1(t) := \frac{1}{\binom{n}{2}} \sum_{j < k} \langle x_j(t), x_k(t) \rangle \in [0, 1], \quad (5.3)$$

which equals 1 if and only if all tokens collapse to a single point x^* . This diagnostic is particularly useful for distinguishing consensus formation at a single point from antipodal clustering. In particular, while $\widehat{\mathcal{R}}_2(t)$ is insensitive to the distinction between x^* and $-x^*$, the quantity $\widehat{\mathcal{R}}_1(t)$ detects whether the tokens have collapsed to one cluster or split into two antipodal clusters. Clearly, $\widehat{\mathcal{R}}_1(t) = 1$ implies $\widehat{\mathcal{R}}_2(t) = 1$.

In addition to tracking these two scalar clustering diagnostics, we also plot the Gram matrix $G(t)$, whose entries are given by $G_{ij}(t) = \langle x_i(t), x_j(t) \rangle$, at $t = 0$ and $t = T$. For Gram matrix plots, we sort the tokens

according to their locations on $\mathbb{S}^{d-1}$ at time $t = T$. The formation of two antipodal clusters is indicated by a 2×2 block structure, with positive diagonal blocks and negative off-diagonal blocks, and where the block sizes reflect the corresponding cluster sizes.

Finally, to quantify cyclic behavior, we track the coordinate-wise mean of the token representation vectors. For $l = 1, \dots, d$, writing $x_j = (x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(d)})^\top$, we define

$$m_l(t) := \frac{1}{n} \sum_{j=1}^n x_j^{(l)}(t). \quad (5.4)$$

If the clustering diagnostics indicate consensus while the trajectory $m(t) = (m_1(t), \dots, m_d(t))$ fails to converge, this suggests that the system has formed clusters that continue to rotate on $\mathbb{S}^{d-1}$.

Remark 5.1. To avoid presenting an overwhelming number of figures, we omit certain diagnostic plots when they do not provide substantial additional information or insight. In particular, we will not present the Gram matrix for the USA model in regimes where the dynamics cluster to a single point, since in such cases the Gram matrix provides little additional information beyond the scalar diagnostics $\widehat{\mathcal{R}}_1$ and $\widehat{\mathcal{R}}_2$. Moreover, in cases where the dynamics both cluster and converge, we present only the clustering diagnostics and omit the coordinate-wise mean plot, describing the convergence behavior in words instead.

Experimental setup. In all experiments, we take the embedding dimension to be $d = 100$ and use $n = 200$ tokens. The choice $n > d$ is motivated not only by the mean-field nature of our analysis, but also by the fact that large-token regimes are common in modern large language models [17, 66, 99, 100]. We have observed similar phenomena in other embedding dimensions, but focus on this parameter setting for conciseness. The initial tokens $\{x_k(0)\}_{k=1}^n \subset \mathbb{S}^{d-1}$ are drawn independently from the uniform distribution on the unit sphere.

5.1 Case 1: $V = I$, A Arbitrary

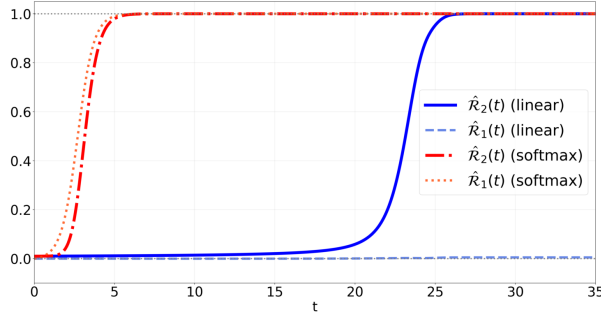
Our theoretical findings from Theorem 3.1 characterize parameter regimes of the matrix A associated with clustering and non-clustering behaviors for the 2-dimensional LSA model. In particular, if $\text{tr}(A) > 0$ or $\det(A) \leq 0$, then the LSA dynamics converge to a pair of antipodal points. On the other hand, if the symmetric part of A satisfies $(A + A^\top)/2 \prec 0$, then the dynamics do not exhibit clustering.

The corresponding numerical results are presented in Figures 10–12. We observe that the qualitative behavior predicted by Theorem 3.1 for the 2-dimensional LSA model persists in higher dimensions. Specifically, when A satisfies (3.6), the LSA dynamics converge to a pair of antipodal points, as indicated by the clustering diagnostic $\widehat{\mathcal{R}}_2(t)$ converging to 1 while $\widehat{\mathcal{R}}_1(t)$ remains close to 0; see Figures 10 and 11. In contrast, when $(A + A^\top)/2 \prec 0$, the LSA dynamics do not synchronize, as reflected by $\widehat{\mathcal{R}}_1(t) = \widehat{\mathcal{R}}_2(t) = 0$ in Figure 12. For the USA model, the dynamics cluster to a single point in all three parameter regimes considered here, as indicated by the convergence of both $\widehat{\mathcal{R}}_1(t)$ and $\widehat{\mathcal{R}}_2(t)$ to 1 in Figures 10–12. Thus, while the USA dynamics exhibits clustering in regimes where LSA forms antipodal clusters, it also synchronizes in the regime where LSA does not cluster. This highlights a qualitative difference between the two models: the USA dynamics tends to produce single-cluster synchronization, which is consistent with the existing theoretical findings in [21, 48], whereas the LSA dynamics can either form antipodal clusters or fail to cluster, depending on the spectral structure of the matrix A .

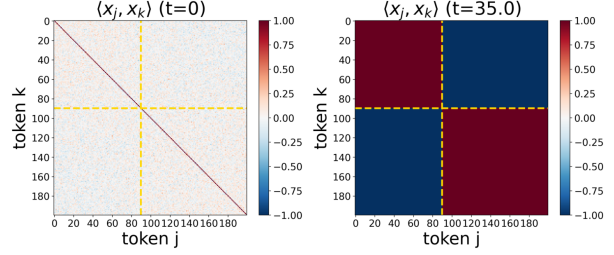
We also observe quantitative differences in convergence speed across models and parameter regimes. Compared with the regime $\text{tr}(A) > 0$ shown in Figure 10, the convergence of the LSA model appears to be slower when $\text{tr}(A) \leq 0$ and $\det(A) \leq 0$ as shown in Figure 11. In contrast, the USA model exhibits more uniform convergence speeds across these regimes and generally converges faster than the LSA model.

5.2 Case 2: $A = I$, V Symmetric

Our theoretical findings from Theorem 3.3 characterize the role of the dominant eigenvalue $\lambda_{\max}(V)$ in the 2-dimensional LSA model. In particular, if $\lambda_{\max}(V) \geq 0$, then the LSA dynamics converge to a pair of antipodal points. If $\lambda_{\max}(V) < 0$, then the dynamics converge without forming clusters.

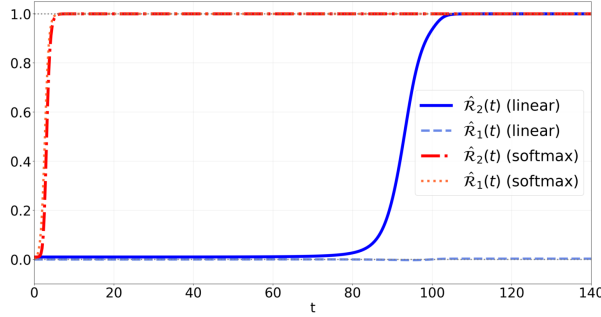


(a) Evolution of the clustering diagnostics $\widehat{R}_1(t)$ and $\widehat{R}_2(t)$ for the LSA (blue) and USA (red) models.

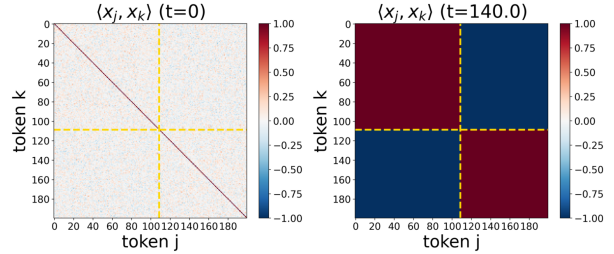


(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 35$.

Figure 10: **Case 1, parameter regime:** $\text{tr}(A) > 0$ and $V = I$. For both the LSA and USA models, the mean-squared cosine similarity $\widehat{R}_2(t)$ converges to 1. However, the mean cosine similarity $\widehat{R}_1(t)$ converges to 1 for the USA model, shown by the red dotted line, but to 0 for the LSA model, shown by the blue dashed line. This indicates that the USA dynamics clusters at a single point, whereas the LSA dynamics converges to two antipodal clusters. The Gram matrices G for the LSA model, shown on the right, suggest that the tokens approximately evenly split between the two clusters.



(a) Evolution of the clustering diagnostics $\widehat{R}_1(t)$ and $\widehat{R}_2(t)$ for the LSA (blue) and USA (red) models.



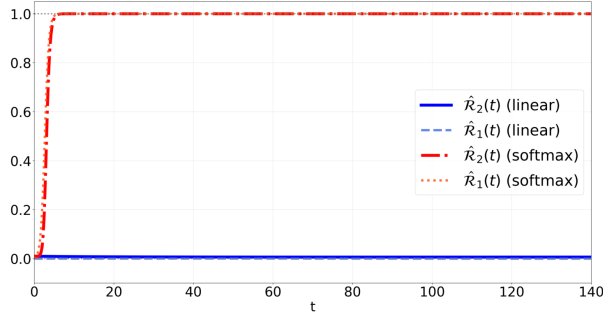
(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 130$.

Figure 11: **Case 1, parameter regime:** $\text{tr}(A) \leq 0$, $\det(A) \leq 0$ and $V = I$. The same qualitative behavior as Figure 10 is observed: the USA dynamics clusters at a single point, while the LSA dynamics converges to two approximately balanced antipodal clusters.

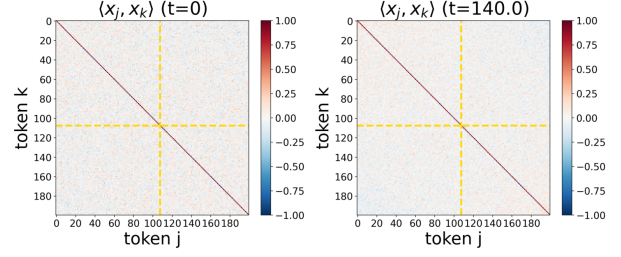
The corresponding numerical results are presented in Figures 13–15. For the LSA model, the high-dimensional simulations agree with the 2-dimensional theory: the sign of $\lambda_{\max}(V)$ determines whether clustering occurs. When $\lambda_{\max}(V) \geq 0$, the tokens converge to two antipodal points, as shown in Figures 13 and 14. When $\lambda_{\max}(V) < 0$, the tokens converge without cluster formation, as indicated by $\widehat{R}_1(t) = \widehat{R}_2(t) = 0$ together with the convergence of the coordinate-wise empirical mean of the tokens $m_i(t)$; see Figure 15.

For the USA model, the behavior also depends on $\lambda_{\max}(V)$, but the resulting cluster configuration can differ from that of the LSA model. When $\lambda_{\max}(V) > 0$, the USA dynamics converge to a single point, in contrast to the antipodal clustering of the LSA dynamics. At the boundary regime $\lambda_{\max}(V) = 0$, Figure 14 shows that the USA model can also exhibit antipodal clustering similarly to the LSA model. Lastly, when $\lambda_{\max}(V) < 0$, both models converge without forming clusters.

For both the USA and LSA models, the boundary regime $\lambda_{\max}(V) = 0$ also displays a notable quantitative feature: clustering occurs much slower than in the positive-eigenvalue regime. In particular, synchronization is rapid when $\lambda_{\max}(V) > 0$, whereas the boundary case requires orders of magnitude longer to approach the clustering state.

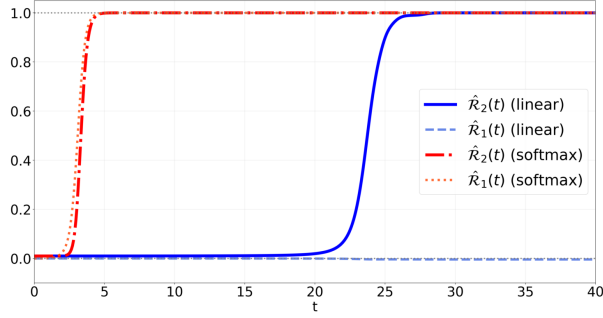


(a) Evolution of the clustering diagnostics $\hat{R}_1(t)$ and $\hat{R}_2(t)$ for the LSA (blue) and USA (red) models.

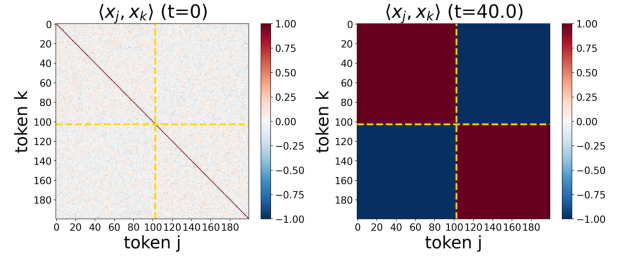


(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 120$.

Figure 12: **Case 1, parameter regime:** $(A + A^\top)/2 \prec 0$ and $V = I$. Both clustering diagnostics $\hat{R}_1(t)$ and $\hat{R}_2(t)$ converge to almost 0 for the LSA model and converge to 1 for the USA model. This implies that the LSA model does not synchronize whereas the USA model clusters at a single point.



(a) Evolution of the clustering diagnostics $\hat{R}_1(t)$ and $\hat{R}_2(t)$ for the LSA (blue) and USA (red) models.



(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 40$.

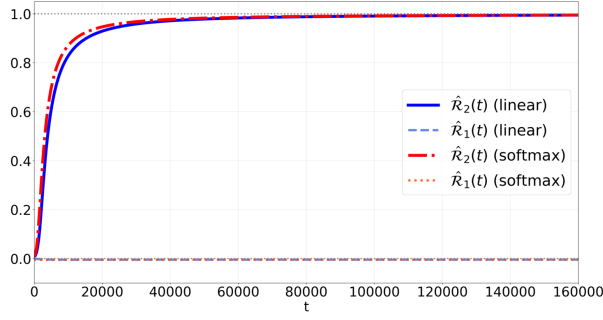
Figure 13: **Case 2, parameter regime:** $A = I$ and $\lambda_{\max}(V) > 0$. For both the LSA and USA models, the mean-squared cosine similarity $\hat{R}_2(t)$ converges to 1. However, the mean cosine similarity $\hat{R}_1(t)$ converges to 1 for the USA model, shown by the red dotted line, but to 0 for the LSA model, shown by the blue dashed line. This indicates that the USA dynamics clusters at a single point, whereas the LSA dynamics converges to two antipodal clusters. The Gram matrices G for the LSA model, shown on the right, suggest that the tokens approximately evenly split between the two clusters.

5.3 Case 3: Non-commuting A and V

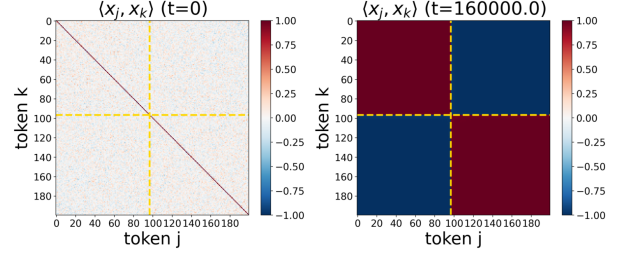
Our theoretical findings from Theorem 3.6 identify parameter regimes for the matrices A and V under which the associated 2-dimensional LSA dynamics form antipodal clusters without converging. As a representative example from Remark 3.8, we take $A = I$ and $V = \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix}$ with $v_{11} > 0$ and $v_{12} \neq 0$. In this regime, the LSA dynamics forms two antipodal clusters, but the cluster locations continue to oscillate periodically.

To extend this construction to higher dimensions, we set $A = I$ and V to be a block-diagonal matrix whose blocks are 2×2 blocks of the form $\begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix}$.

The corresponding numerical results are presented in Figure 16. For the LSA model, the observed behavior of the higher-dimensional dynamics with $d > 2$ agrees with the 2-dimensional theory from Theorem 3.6. Specifically, the tokens form two antipodal clusters, as shown by $\hat{R}_2(t)$ converging to 1 while $\hat{R}_1(t)$ remains close to 0 in Figure 16a. However, the token trajectories do not converge to fixed locations; instead, they oscillate periodically, as indicated by the oscillatory behavior of the per-dimensional means $m_l(t)$ in Figure 16c.

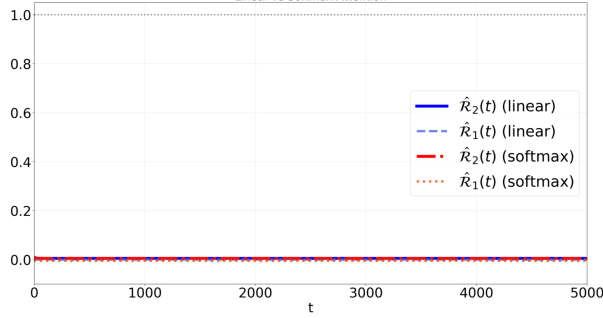


(a) Evolution of the clustering diagnostics $\hat{\mathcal{R}}_1(t)$ and $\hat{\mathcal{R}}_2(t)$ for the LSA (blue) and USA (red) models.

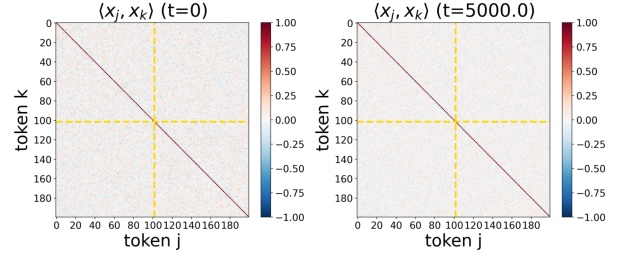


(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 150,000$.

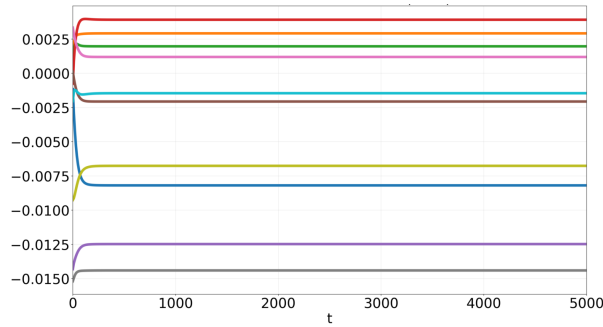
Figure 14: **Case 2, parameter regime:** $A = I$ and $\lambda_{\max}(V) = 0$. For both the LSA and USA models, the mean-squared cosine similarity $\hat{\mathcal{R}}_2(t)$ converges to 1, while the mean cosine similarity $\hat{\mathcal{R}}_1(t)$ remains close to 0. This indicates that both dynamics converge to antipodal clustering configurations. Compared with the regime $\lambda_{\max}(V) > 0$, the convergence in this boundary regime is significantly slower for both models.



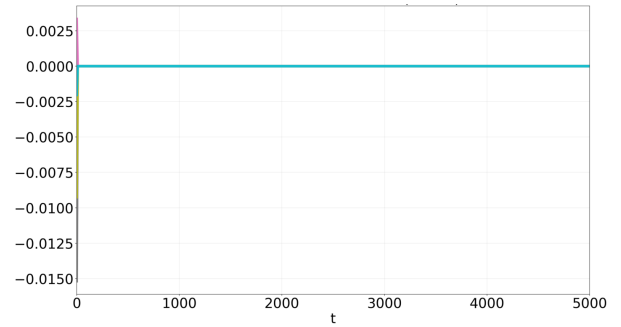
(a) Evolution of the clustering diagnostics $\hat{\mathcal{R}}_1(t)$ and $\hat{\mathcal{R}}_2(t)$ for the LSA (blue) and USA (red) models.



(b) The Gram matrices $G(t)$ for LSA model at $t = 0$ and $t = 5000$.



(c) Coordinate-wise mean $m_l(t)$ for the LSA model.

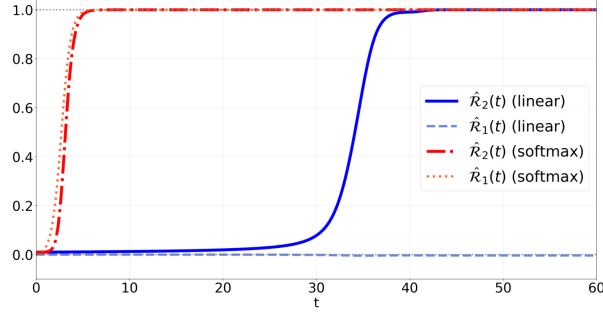


(d) Coordinate-wise mean $m_l(t)$ for the USA model.

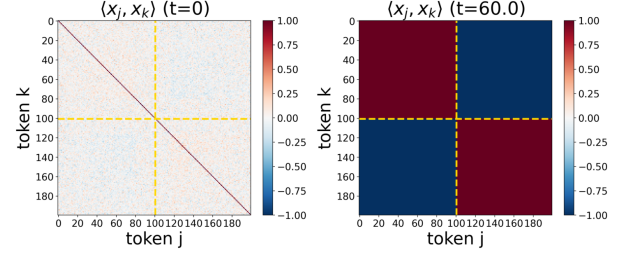
Figure 15: **Case 2, parameter regime:** $A = I$ and $\lambda_{\max}(V) < 0$. For both the LSA and USA models, the clustering diagnostics remain at $\hat{\mathcal{R}}_1(t) = \hat{\mathcal{R}}_2(t) = 0$, while the per-dimensional means $m_l(t)$ converge across dimensions. These observations indicate that the dynamics converge, but without cluster formation.

For the USA model, the dynamics exhibit a similar combination of clustering and non-convergence. The main difference lies in the cluster configuration: the USA dynamics collapses to a single cluster, whereas the

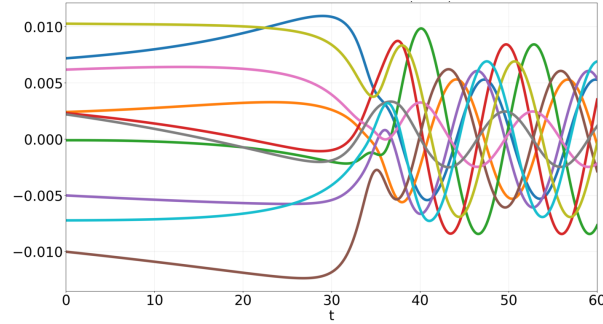
LSA dynamics form two antipodal clusters. Nevertheless, in both models, the oscillatory behavior of the per-dimensional means indicates that the token dynamics remain time-periodic rather than converging to a stationary configuration.



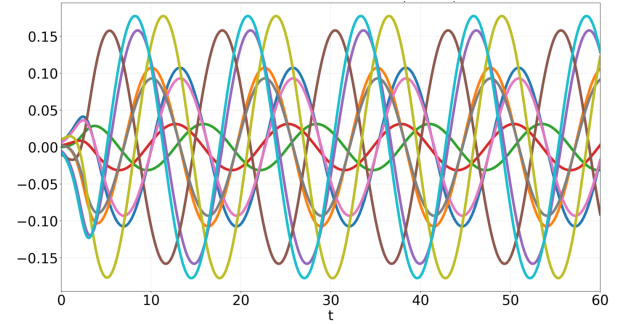
(a) Evolution of the clustering diagnostics $\hat{\mathcal{R}}_1(t)$ and $\hat{\mathcal{R}}_2(t)$ for the LSA (blue) and USA (red) models.



(b) The Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 60$.



(c) Coordinate-wise mean $m_l(t)$ for the LSA model.



(d) Coordinate-wise mean $m_l(t)$ for the USA model.

Figure 16: **Case 3, parameter regime:** $A = I$ and $V = \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix}$ with $v_{11} > 0$ and $v_{12} \neq 0$. For both the LSA and USA models, the mean-squared cosine similarity $\hat{\mathcal{R}}_2(t)$ converges to 1. However, the mean cosine similarity $\hat{\mathcal{R}}_1(t)$ converges to 1 for the USA model, shown by the red dotted line, but to 0 for the LSA model, shown by the blue dashed line; see Figure 16a. Thus, the USA dynamics clusters at a single point, whereas the LSA dynamics forms two antipodal clusters. The periodic oscillations of the coordinate-wise means $m_l(t)$, shown in Figures 16c and 16d, indicate that the token dynamics remain oscillatory.

5.4 Case 4: Hidden Hamiltonian Dynamics

Our theoretical findings from Theorem 3.9 identify the parameter regimes under which the associated 2-dimensional LSA model exhibits a “hidden” Hamiltonian structure after suitable transformations. Specifically, we take $A = I$ and $V = \begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix}$, which depends on two parameters. When $v_{11} < v_{12}$, the LSA dynamics exhibits persistent periodic oscillations. In contrast, when $v_{11} \geq v_{12}$, the LSA dynamics converge to a pair of antipodal points.

To extend this construction to higher dimensions, we take V to be a block-diagonal matrix whose blocks are 2×2 blocks of the form $\begin{pmatrix} v_{11} & v_{12} \\ -v_{12} & v_{11} \end{pmatrix}$.

The corresponding numerical results are presented in Figures 17–19. For the LSA model, the long-time behavior is governed by the relation between v_{11} and v_{12} , in agreement with the 2-dimensional theory from Theorem 3.9. When $v_{11} \geq v_{12}$, the tokens cluster into a pair of antipodal points, as indicated by $\hat{\mathcal{R}}_2(t)$ converging to 1 while $\hat{\mathcal{R}}_1(t)$ remains close to 0 in Figures 17 and 18. When $v_{11} < v_{12}$, the token trajectories are periodic and do not form clusters. In this regime, the clustering diagnostics remain at $\hat{\mathcal{R}}_1(t) = \hat{\mathcal{R}}_2(t) = 0$, as shown in Figure 19a, while the per-dimensional means $m_l(t)$ exhibit sustained periodic oscillations, as

shown in Figure 19c.

For the USA model, the same parameter threshold separates clustering from oscillatory non-clustering behavior. When $v_{11} \geq v_{12}$, the USA model converges to a single point, as indicated by $\hat{\mathcal{R}}_1(t) = \hat{\mathcal{R}}_2(t) = 1$ in Figures 17 and 18. This differs from the antipodal clustering observed in the LSA model. When $v_{11} < v_{12}$, the USA dynamics also fails to synchronize and instead exhibits persistent periodic oscillations, as shown by the oscillatory behavior of $m_l(t)$ in Figure 19d.

Finally, we observe a quantitative slowdown near the transition. In the regime $v_{11} > v_{12}$, as shown in Figure 17, convergence is relatively rapid for both models. However, at the bifurcation point $v_{11} = v_{12}$, shown in Figure 18, convergence is substantially slower and takes orders of magnitude longer to stabilize. Consistent with the previous cases, the USA model generally converges faster than the LSA model in the regimes where clustering occurs.

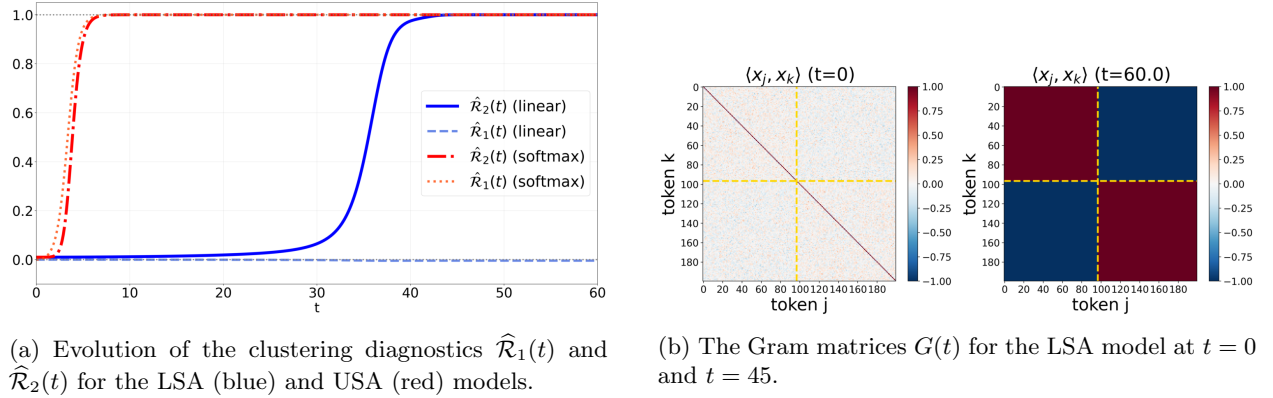


Figure 17: **Case 4, parameter regime:** $v_{11} > v_{12}$. For both the LSA and USA models, the mean-squared cosine similarity $\hat{\mathcal{R}}_2(t)$ converges to 1. However, the mean cosine similarity $\hat{\mathcal{R}}_1(t)$ converges to 1 for the USA model, shown by the red dotted line, but to 0 for the LSA model, shown by the blue dashed line. This indicates that the USA dynamics clusters at a single point, whereas the LSA dynamics converges to two antipodal clusters. The Gram matrices G for the LSA model, shown on the right, suggest that the tokens are not entirely evenly split between the two clusters.

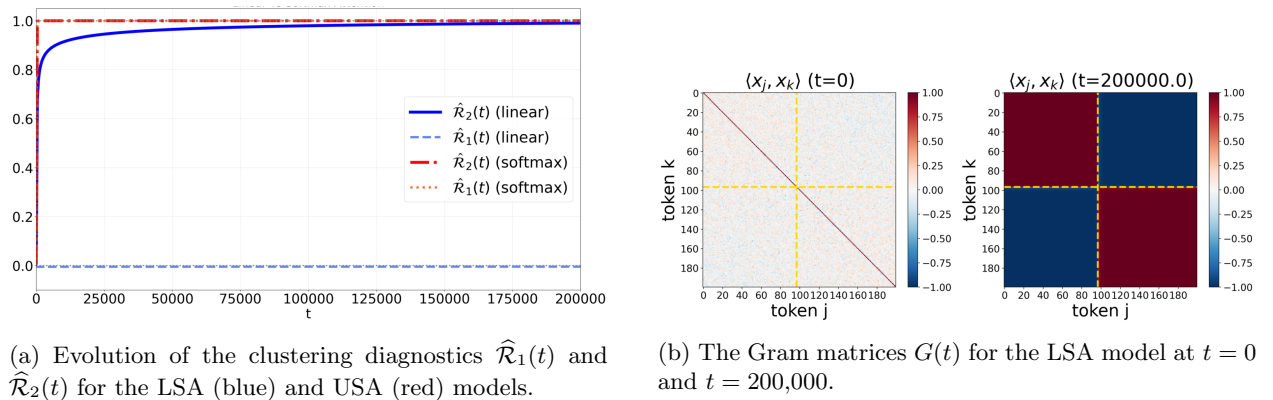
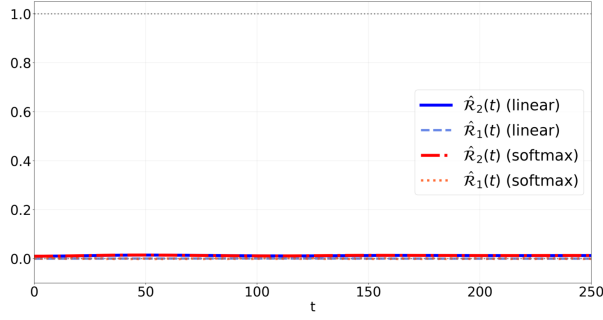
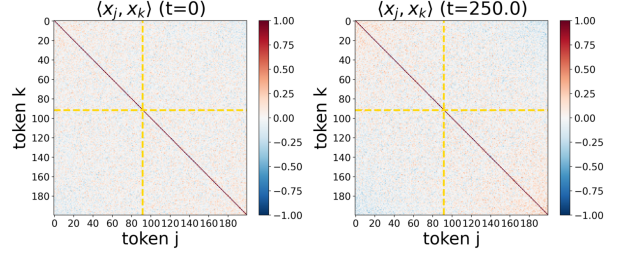


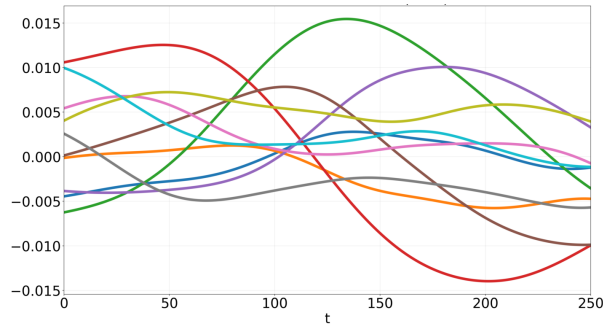
Figure 18: **Case 4, parameter regime:** $v_{11} = v_{12}$. The qualitative behavior is the same as in the regime $v_{11} > v_{12}$: for both the LSA and USA models, $\hat{\mathcal{R}}_2(t)$ converges to 1, while $\hat{\mathcal{R}}_1(t)$ converges to 1 for the USA model and to 0 for the LSA model. This indicates that the USA dynamics clusters at a single point, whereas the LSA dynamics converges to two antipodal clusters. However, convergence is notably slower in this regime, taking approximately $t = 150,000$ steps.



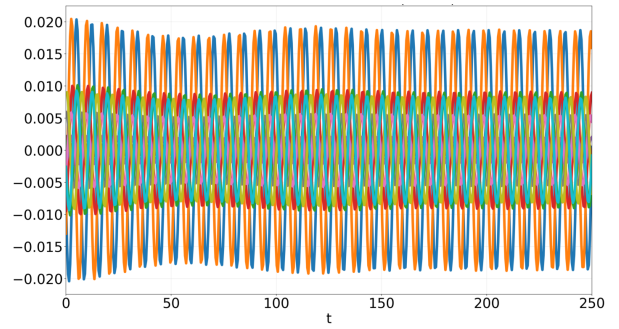
(a) Evolution of the clustering diagnostics $\hat{\mathcal{R}}_1(t)$ and $\hat{\mathcal{R}}_2(t)$ for the LSA (blue) and USA (red) models.



(b) Gram matrices $G(t)$ for the LSA model at $t = 0$ and $t = 250$.



(c) Coordinate-wise mean $m_l(t)$ for the LSA model.



(d) Coordinate-wise mean $m_l(t)$ for the USA model.

Figure 19: **Case 4, parameter regime:** $v_{11} < v_{12}$. For both the LSA and USA models, the clustering diagnostics remain at $\hat{\mathcal{R}}_1(t) = \hat{\mathcal{R}}_2(t) = 0$, while the per-dimensional means $m_l(t)$ oscillate periodically. This indicates persistent oscillatory dynamics without cluster formation.

6 Conclusion

In this paper, we develop an interacting-particle-system-based approach to analyze the inference-time behavior of deep linear encoder-only transformers. Our main results demonstrate that such dynamics exhibit a rich variety of dynamical behaviors beyond token synchronization and clustering. To reveal these behaviors, we establish a connection between two-dimensional linear self-attention dynamics and generalized Kuramoto models with pure second-harmonic coupling. This connection allows us to apply dimension-reduction techniques, including the Watanabe–Strogatz transformation and the Ott–Antonsen ansatz, to derive analytically tractable reduced dynamics. The resulting framework identifies the mechanisms underlying several distinct long-time behaviors, including clustering with convergence, clustering without convergence, oscillations, and bifurcations. In particular, our theory characterizes how specific choices of the key, query, and value matrices can lead to fundamentally different long-time behaviors, while our stability analysis shows that, under some parameter regimes, these behaviors persist beyond initializations exactly on the OA manifold. Our numerical experiments confirm these theoretical observations, and further suggest that the mechanisms identified in dimension $d = 2$ continue to appear in higher-dimensional linear self-attention dynamics and may also be relevant to softmax self-attention.

Open questions. Let us conclude by discussing several open questions and research directions that may broaden the scope of the presented analytical framework, deepen the connection between transformer dynamics and other models of collective dynamics, as well as contribute to a more comprehensive understanding of transformer architectures.

- A first natural question is whether the analytical framework developed here for dimension $d = 2$

can be extended to higher-dimensional settings. Our analysis in dimension $d = 2$ builds on the fact that the linear self-attention model (LSA) possesses a low-dimensional structure, revealed through the WS transformation and the OA ansatz, that is substantially more amenable to analysis. It is therefore natural to ask whether also the higher-dimensional LSA model admits an analogous reduced description governing the evolution of the particle dynamics.

Higher-dimensional generalizations of the classical Kuramoto model (1.5) and the corresponding extensions of the WS transformation and the OA ansatz have been studied in [25, 65, 67, 68, 81]. We leave the investigation of analogous reduced structures for higher-dimensional LSA models to future work.

- A second important question is whether our analytical framework can be extended to the transformer dynamics (1.1) in case of more general attention kernels h . As shown in Theorem 2.1, the linear kernel $h(y) = y$ gives rise to a Kuramoto model with *pure* second-harmonic coupling, placing the dynamics within a class for which powerful dimension-reduction techniques are available.

By contrast, the nonlinear attention kernel $h(y) = \exp(y)$, corresponding to softmax self-attention [103], produces dynamics with *mixed-order* harmonics coupling. In this setting, the WS transformation and the OA ansatz are no longer directly applicable [52]. It would therefore be interesting to develop alternative dimension-reduction techniques or identify approximate low-dimensional structures for such models [24, 101, 102].

A related direction is to investigate regimes in which one harmonic dominates the remaining coupling terms. Under suitably prepared initializations, one may ask whether the resulting mixed-harmonic dynamics remain close to those of the corresponding pure-harmonic model over a finite time interval. A rigorous approximation result of this type could help explain the metastable clustering observed in the softmax self-attention model [19, 20, 47]. In particular, the dominant harmonic may initially govern the dynamics and induce a transient multi-cluster state resembling that of the corresponding pure-harmonic model, before the cumulative influence of the remaining harmonics becomes significant.

- A third direction is to extend the present framework to linear transformer architectures incorporating additional model components. For instance, the fully connected layer may contribute an active-rotator term, leading to an active-rotator model that generalizes the classical Kuramoto dynamics [49, 95], or may be viewed as an external control input [22]. Similarly, causal attention dynamics [57, 58] is related to Kuramoto-type model with a directed and nonuniform interaction structure [56, 91], while time-dependent parameter matrices lead naturally to Kuramoto models with time-varying coefficients [84, 85]. Investigating how these additional structures interact with the long-time behaviors and parameter regimes identified in the present work would constitute an important step toward understanding more realistic transformer architectures.
- A fourth direction is to clarify the possible connection between transformer dynamics and Vicsek-type models of collective alignment [35, 36, 104]. In two dimensions, continuous-time Vicsek-type orientation dynamics can be viewed as a spatially coupled form of Kuramoto alignment, in which the interaction network is determined by the evolving particle configuration. The state-dependent interaction structure of the transformer dynamics (1.1) is therefore reminiscent of such models, although general attention weights and value transformations need not represent standard attractive alignment. It would be interesting to identify choices of attention kernels and parameter regimes for which transformer dynamics admit a precise interpretation as a Vicsek-type model, and to investigate whether analytical tools developed for collective alignment systems can be adapted to this setting.

Acknowledgements

SL and NGT were supported by NSF-DMS grant 2236447. They also gratefully acknowledge support from the IFDS at UW-Madison through NSF TRIPODS grant 2023239. JP was supported by the Polish National Science Centre’s Grant No. 2025/58/E/ST1/00482 (Sonata Bis).

For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

- [1] Á. R. Abella, J. P. Silvestre, and P. Tabuada. The asymptotic behavior of attention in transformers. *arXiv preprint arXiv:2412.02682*, 2024.
- [2] Á. R. Abella, J. P. Silvestre, and P. Tabuada. Consensus is all you get: The role of attention in transformers. In *Forty-second International Conference on Machine Learning*, 2025.
- [3] J. A. Acebrón, L. L. Bonilla, C. J. Pérez Vicente, F. Ritort, and R. Spigler. The kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137–185, 2005.
- [4] A. Agazzi, G. Bruno, E. M. García, S. Saviozzi, and M. Romito. Stochastic scaling limits and synchronization by noise in deep transformer models. *arXiv preprint arXiv:2604.26898*, 2026.
- [5] A. Alcalde, L. Bungert, K. Riedl, and T. Roith. Quantifying concentration phenomena of mean-field transformers in the low-temperature regime. *arXiv preprint arXiv:2605.10931*, 2026.
- [6] A. Alcalde, G. Fantuzzi, and E. Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *SIAM Journal on Mathematics of Data Science*, 7(3):1367–1393, 2025.
- [7] A. Alcalde, B. Geshkovski, and D. Ruiz-Balet. Attention’s forward pass and frank-wolfe. *arXiv preprint arXiv:2508.09628*, 2025.
- [8] A. Alcalde, Z. Ji, and E. Zuazua. Reachability and asymptotics of gaussian transformer dynamics. *arXiv preprint arXiv:2606.07600*, 2026.
- [9] C. Altafini. Multistability of self-attention dynamics in transformers. *IEEE Transactions on Automatic Control*, 2026.
- [10] A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Perceptrons and localization of attention’s mean-field landscape. *arXiv preprint arXiv:2601.21366*, 2026.
- [11] K. Balasubramanian, S. Banerjee, and P. Rigollet. On the structure of stationary solutions to mckean-vlasov equations with applications to noisy transformers. *arXiv preprint arXiv:2510.20094*, 2025.
- [12] I. Barbălat. Systèmes d’équations différentielles d’oscillations non linéaires. *Rev. Math. Pures Appl.*, 4:267–270, 1959. see p. 269.
- [13] H. Blayney, Á. Arroyo, J. Obando-Ceron, P. S. Castro, A. Courville, M. M. Bronstein, and X. Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026.
- [14] C. Böhm, C. Plant, J. Shao, and Q. Yang. Clustering by synchronization. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–592, 2010.
- [15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [16] M. Breakspear, S. Heitmann, and A. Daffertshofer. Generative models of cortical oscillations: Neurobiological implications of the kuramoto model. *Frontiers in Human Neuroscience*, Volume 4 - 2010, 2010.
- [17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

- [18] G. Bruno, S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Scaling limits of long-context transformers. *arXiv preprint arXiv:2605.08505*, 2026.
- [19] G. Bruno, F. Pasqualotto, and A. Agazzi. Emergence of meta-stable clustering in mean-field transformer models. *arXiv preprint arXiv:2410.23228*, 2024.
- [20] G. Bruno, F. Pasqualotto, and A. Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. *arXiv preprint arXiv:2509.25040*, 2025.
- [21] M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. *Philos. Trans. Roy. Soc. A*, 383(2298):Paper No. 20240233, 48, 2025.
- [22] M. Caponigro, A. C. Lai, and B. Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete and Continuous Dynamical Systems*, 35(9):4241–4268, 2015.
- [23] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.
- [24] R. Cestnik and A. Pikovsky. Hierarchy of exact low-dimensional reductions for populations of coupled oscillators. *Physical Review Letters*, 128(5):054101, 2022.
- [25] S. Chandra, M. Girvan, and E. Ott. Complexity reduction ansatz for systems of interacting orientable agents: Beyond the kuramoto model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(5), 2019.
- [26] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Critical attention scaling in long-context transformers. *arXiv preprint arXiv:2510.05554*, 2025.
- [27] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.
- [28] T. Chen, Z. Zhang, Y. Cheng, A. Awadallah, and Z. 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.
- [29] Z. Chen, Z. Lin, S. Chen, Y. Polyanskiy, and P. Rigollet. Residual connections provably mitigate oversmoothing in graph neural networks. *arXiv preprint arXiv:2501.00762*, 2025.
- [30] K. M. Choromanski, V. Likhoshervstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.
- [31] A. Cowsik, T. Nebabu, X. Qi, and S. Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Physical Review E*, 112(5):055301, 2025.
- [32] C. Criscitiello, Q. Rebjock, A. D. McRae, and N. Boumal. Synchronization on Circles and Spheres with Nonlinear Interactions. *SIAM J. Appl. Dyn. Syst.*, 25(2):1207–1237, 2026.
- [33] H. Daido. Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. *Progress of theoretical physics*, 88(6):1213–1218, 1992.
- [34] H. Daido. Generic scaling at the onset of macroscopic mutual entrainment in limit-cycle oscillators with uniform all-to-all coupling. *Phys. Rev. Lett.*, 73:760–763, Aug 1994.
- [35] P. Degond, A. Frouvelle, and J.-G. Liu. Macroscopic limits and phase transition in a system of self-propelled particles. *Journal of nonlinear science*, 23(3):427–456, 2013.
- [36] P. Degond and S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1193–1215, 2008.

- [37] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [38] Y. Dong, J.-B. Cordonnier, and A. 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.
- [39] F. Dörfler and F. Bullo. Synchronization in complex networks of phase oscillators: A survey. *Automatica*, 50(6):1539–1564, 2014.
- [40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [41] S. Dutta, T. Gautam, S. Chakrabarti, and T. Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 5531–5544, 2021.
- [42] M. Engel and A. Shalova. Random quadratic form on a sphere: Synchronization by common noise. *arXiv preprint arXiv:2603.06187*, 2026.
- [43] L. Fedorov, M. E. Sander, R. Elie, P. Marion, and M. Laurière. Clustering in deep stochastic transformers. *arXiv preprint arXiv:2601.21942*, 2026.
- [44] R. Feng, K. Zheng, Y. Huang, D. Zhao, M. Jordan, and Z.-J. Zha. Rank diminishing in deep neural networks. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [45] G. Filatrella, A. H. Nielsen, and N. F. Pedersen. Analysis of a power grid using a kuramoto-like model. *The European Physical Journal B*, 61(4):485–491, 2008.
- [46] J. Geiping, S. M. McLeish, N. Jain, J. Kirchenbauer, S. Singh, B. R. Bartoldson, B. Kailkhura, A. Bhatele, and T. Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.
- [47] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet. Dynamic metastability in the self-attention model, 2024.
- [48] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc. (N.S.)*, 62(3):427–479, 2025.
- [49] G. Giacomini, K. Pakdaman, X. Pellegrin, and C. Poquet. Transitions in active rotator systems: invariant hyperbolic manifold approach. *SIAM Journal on Mathematical Analysis*, 44(6):4165–4194, 2012.
- [50] A. Giannou, S. Rajput, J.-y. Sohn, K. Lee, J. D. Lee, and D. Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pages 11398–11442. PMLR, 2023.
- [51] A. Giorlandino and S. Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *The Fourteenth International Conference on Learning Representations*, 2026.

- [52] C. C. Gong and A. Pikovsky. Low-dimensional dynamics for higher-order harmonic, globally coupled phase-oscillator ensembles. *Physical Review E*, 100(6), Dec. 2019.
- [53] X. Guo, Y. Wang, T. Du, and Y. Wang. Contranorm: A contrastive learning perspective on over-smoothing and beyond. In *The Eleventh International Conference on Learning Representations*, 2023.
- [54] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [55] C. Hooper, S. Kim, H. Mohammadzadeh, M. W. Mahoney, Y. S. Shao, K. Keutzer, and A. Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.
- [56] P. Ji, T. K. Peron, F. A. Rodrigues, and J. Kurths. Low-dimensional behavior of kuramoto model with inertia in complex networks. *Scientific reports*, 4(1):4783, 2014.
- [57] N. Karagodin, Y. Polyanskiy, and P. Rigollet. Clustering in causal attention masking, 2024.
- [58] N. Karagodin, Y. Polyanskiy, and P. Rigollet. Clustering in causal attention masking. *Advances in neural information processing systems*, 37:115652–115681, 2024.
- [59] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.
- [60] I. Z. Kiss, Y. Zhai, and J. L. Hudson. Emerging coherence in a population of chemical oscillators. *Science*, 296(5573):1676–1678, 2002.
- [61] H. Koubbi, B. Geshkovski, and P. Rigollet. Homogenized transformers. *arXiv preprint arXiv:2604.01978*, 2026.
- [62] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [63] Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In H. Araki, editor, *International Symposium on Mathematical Problems in Theoretical Physics*, pages 420–422, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.
- [64] Y. Kuramoto. Chemical turbulence. In *Chemical oscillations, waves, and turbulence*, pages 111–140. Springer, 1984.
- [65] M. Lipton, R. Mirollo, and S. H. Strogatz. The kuramoto model on a sphere: Explaining its low-dimensional dynamics with group theory and hyperbolic geometry. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(9):093113, 09 2021.
- [66] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [67] M. A. Lohe. Higher-dimensional generalizations of the watanabe–strogatz transform for vector models of synchronization. *Journal of Physics A: Mathematical and Theoretical*, 51(22):225101, may 2018.
- [68] M. A. Lohe. Exact reduction of synchronized systems in higher-dimensional spaces. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(2), 2025.
- [69] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T. Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *CoRR*, abs/1906.02762, 2019.
- [70] K. V. Mardia and P. E. Jupp. *Directional statistics*. John Wiley & Sons, 2009.

- [71] S. A. Marvel, R. E. Mirollo, and S. H. Strogatz. Identical phase oscillators with global sinusoidal coupling evolve by möbius group action. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(4), 2009.
- [72] A. Massucco, L. Del Grande, M. Carioni, C. Brune, and C.-B. Schönlieb. Multi-headed transformer architectures as time-dependent wasserstein gradient flows. *arXiv preprint arXiv:2605.18870*, 2026.
- [73] T. Miyato, S. Löwe, A. Geiger, and M. Welling. Artificial kuramoto oscillatory neurons. In *International Conference on Learning Representations*, volume 2025, pages 44278–44322, 2025.
- [74] K. Mun and M. Rosenzweig. Phase transitions and linear stability for the mean-field kuramoto-daido model. *arXiv preprint arXiv:2602.14954*, 2026.
- [75] K. Mun and M. Rosenzweig. Phase transitions for the noisy transformer model in arbitrary dimension. *arXiv preprint arXiv:2606.05140*, 2026.
- [76] K. Mun and M. Rosenzweig. Phase transitions in doi-onsager, noisy transformer, and other multimodal models. *arXiv preprint arXiv:2604.16288*, 2026.
- [77] T. Nguyen, H. Honda, T. Sano, V. Nguyen, S. Nakamura, and T. M. 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.
- [78] L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.
- [79] E. Ott and T. M. Antonsen. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(3), 2008.
- [80] E. Ott and T. M. Antonsen. Long time evolution of phase oscillator systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(2), May 2009.
- [81] H. Park. The watanabe-strogatz transform and constant of motion functionals for kinetic vector models. *Journal of Differential Equations*, 323:113–151, 2022.
- [82] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pages 31932–31951, 2024.
- [83] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. Smith, and L. Kong. Random feature attention. In *International Conference on Learning Representations*, 2021.
- [84] S. Petkoski and A. Stefanovska. Kuramoto model with time-varying parameters. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 86(4):046212, 2012.
- [85] B. Pietras and A. Daffertshofer. Ott-antonsen attractiveness for parameter-dependent oscillatory systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(10), 2016.
- [86] A. Pikovsky and M. Rosenblum. Dynamics of globally coupled oscillators: Progress and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9), 2015.
- [87] Y. Polyanskiy, P. Rigollet, and A. Yao. Synchronization of mean-field models on the circle. *arXiv preprint arXiv:2507.22857*, 2025.
- [88] K. C. Puvvada, F. Ladhak, S. Akle Serano, C.-P. Hsieh, S. Acharya, S. Majumdar, F. Jia, S. Kri-man, S. Sun, D. Rekish, and B. Ginsburg. SWAN: An efficient and scalable approach for long-context language modeling. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2424–2438, Suzhou, China, Nov. 2025. Association for Computational Linguistics.

- [89] J. Qu, D. Holzmüller, G. Varoquaux, and M. L. Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.
- [90] P. Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.
- [91] F. A. Rodrigues, T. K. D. Peron, P. Ji, and J. Kurths. The kuramoto model in complex networks. *Physics Reports*, 610:1–98, 2016.
- [92] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.
- [93] I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.
- [94] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.
- [95] S. Shinomoto and Y. Kuramoto. Phase transitions in active rotator systems. *Progress of Theoretical Physics*, 75(5):1105–1110, 1986.
- [96] P. S. Skardal, E. Ott, and J. G. Restrepo. Cluster synchrony in systems of coupled phase oscillators with higher-order coupling. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(3):036208, 2011.
- [97] Y. Song, A. Keller, S. Brodjian, T. Miyato, Y. Yue, P. Perona, and M. Welling. Kuramoto orientation diffusion models. *Advances in Neural Information Processing Systems*, 38:4047–4076, 2026.
- [98] S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.
- [99] G. Team. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.
- [100] L. Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- [101] K. Tokunaga. Low-dimensional reduction theory for populations of globally coupled phase oscillators with multiharmonic coupling: A method based on opuc theory. *arXiv preprint arXiv:2604.14611*, 2026.
- [102] R. Tönjes and A. Pikovsky. Low-dimensional description for ensembles of identical phase oscillators subject to cauchy noise. *Physical Review E*, 102(5):052315, 2020.
- [103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [104] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.
- [105] S. Watanabe and S. H. Strogatz. Integrability of a globally coupled oscillator array. *Phys. Rev. Lett.*, 70:2391–2394, Apr 1993.
- [106] K. Wiesenfeld, P. Colet, and S. H. Strogatz. Frequency locking in josephson arrays: Connection with the kuramoto model. *Physical Review E*, 57(2):1563, 1998.
- [107] S. Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*. Springer, 2003.
- [108] X. Wu, A. Ajorlou, Z. Wu, and A. Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36:35084–35106, 2023.

- [109] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR, 2023.
- [110] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [111] L. Yang, K. Lee, R. D. Nowak, and D. Papailiopoulos. Looped transformers are better at learning learning algorithms. In *The Twelfth International Conference on Learning Representations*, 2024.

Appendix Contents

A	Collected Notation and Quantities	45
B	Proof Details for Section 2	46
B.1	Proof of Lemma B.1	46
B.2	Proof of Theorem 2.1	47
B.3	Proof Details for Section 2.2	49
C	Proof Details for Section 3	50
C.1	Proofs of Main Results	50
C.2	Auxiliary Lemmas	68
D	Proof Details for Section 4	70
D.1	Proof of Main Results	70
D.2	Auxiliary Lemmas	76

A Collected Notation and Quantities

In Table 2, we collect some notation used throughout the manuscript.

Table 2: Notation.

Symbol	Meaning	Symbol	Meaning
$\mathbb{S}^1$	Unit circle in $\mathbb{R}^2$.	$\mathbb{T}^1$	One-dimensional torus $\mathbb{R}/2\pi\mathbb{Z}$.
$x_k(t)$	Representation of the k -th token.	A, V	Parameter matrices.
$P_x^\perp(y)$	Projection of y onto the tangent space at x .	$\theta_k(t)$	Angular variable associated with $x_k(t)$.
$\Theta(t)$	Mean-field angular variable.	$f(t, \theta)$	Law of $\Theta(t)$.
$\xi_k(t), \Xi(t)$	Double-angle variables $2\theta_k(t)$ and $2\Theta(t)$.	$g(t, \xi)$	Law of $\Xi(t)$.
$\mathcal{R}_2^n(t), \mathcal{R}_2(t)$	Finite-particle and mean-field second-order parameters.	$\rho(t), \phi(t)$	Writing $\mathcal{R}_2(t) = \rho(t)e^{i\phi(t)}$.
$\alpha(t), \eta(t)$	Watanabe–Strogatz variables.	$\gamma(t), \Phi(t)$	Writing $\alpha(t) = \gamma(t)e^{i\Phi(t)}$.
α_{OA}	Ott–Antonsen closure counterpart of α (4.4).	$M_{\alpha, \eta}(\omega)$	Möbius transform (2.16).
$T_{\alpha, \eta}$	Map induced by $M_{\alpha, \eta}$ (2.24).	φ_k, φ	Finite-particle and mean-field constants of motion.
ν	Law of φ .	$\text{WC}(\xi; \alpha)$	Wrapped Cauchy distribution with parameter α .

We now collect some notation used in Section 2. Recall the parameter matrices A and V in dimension $d = 2$,

$$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad V := \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}. \quad (\text{A.1})$$

We define the following real-valued constants depending on the matrices A and V :

$$\begin{aligned}
w_0 &:= v_{11}a_{21} + v_{12}a_{22} - v_{21}a_{11} - v_{22}a_{12}, & w_1 &:= v_{11}a_{11} + v_{12}a_{12} - v_{21}a_{21} - v_{22}a_{22}, \\
w_2 &:= -v_{11}a_{21} - v_{12}a_{22} - v_{21}a_{11} - v_{22}a_{12}, & w_3 &:= v_{11}a_{21} - v_{12}a_{22} - v_{21}a_{11} + v_{22}a_{12}, \\
w_4 &:= v_{11}a_{22} + v_{12}a_{21} - v_{21}a_{12} - v_{22}a_{11}, & w_5 &:= v_{11}a_{11} - v_{12}a_{12} - v_{21}a_{21} + v_{22}a_{22}, \\
w_6 &:= v_{11}a_{12} + v_{12}a_{11} - v_{21}a_{22} - v_{22}a_{21}, & w_7 &:= v_{11}a_{21} - v_{12}a_{22} + v_{21}a_{11} - v_{22}a_{12}, \\
w_8 &:= v_{11}a_{22} + v_{12}a_{21} + v_{21}a_{12} + v_{22}a_{11}.
\end{aligned} \quad (\text{A.2})$$

This enables us to compactly write the following complex-valued constants,

$$\begin{aligned} b_1 &:= \frac{1}{16}(iw_5 + w_6 + w_7 - iw_8), & b_2 &:= \frac{1}{16}(iw_5 - w_6 + w_7 + iw_8), & b_3 &:= \frac{1}{8}(iw_1 - w_2), \\ b_4 &:= \frac{1}{8}(-w_3 + iw_4), & b_5 &:= -\frac{1}{4}w_0. \end{aligned} \quad (\text{A.3})$$

The functions B and C , as introduced in Theorem 2.1, are defined as,

$$B(r) := b_1 r + b_2 \bar{r} + b_3 \quad \text{and} \quad C(r) := b_4 r + \bar{b}_4 \bar{r} + b_5. \quad (\text{A.4})$$

Finally, the definitions of F and G are derived from B and C as in (2.19)

$$F(\alpha, r) := 2i (B(r)\alpha^2 + C(r)\alpha + \bar{B}(r)), \quad G(\eta, r) := 2 (2\text{Re}(B(r)\eta) + C(r)), \quad (\text{A.5})$$

B Proof Details for Section 2

In this appendix, we provide the technical proofs for the results presented in Section 2.

B.1 Proof of Lemma B.1

Lemma B.1. *Consider the model*

$$\dot{x}_k(t) = P_{x_k(t)}^\perp \left(\frac{1}{n} \sum_{j=1}^n h(\beta \langle x_k(t), Ax_j(t) \rangle) V x_j(t) \right), \quad k = 1, \dots, n. \quad (\text{B.1})$$

Suppose that the attention kernel h admits an anti-derivative H , and that the matrices A and V satisfy

$$A^\top = A, \quad V = cA,$$

for some constant $c \in \mathbb{R}$. Then (B.1) can be written as a Wasserstein gradient flow associated with the energy functional

$$E[\mu] := -\frac{c}{2\beta} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} H(\beta \langle x, Ay \rangle) d\mu(x) d\mu(y).$$

Proof. Let

$$W(x, y) := H(\beta \langle x, Ay \rangle).$$

Since $A^\top = A$, one has $W(x, y) = W(y, x)$, which fits the classical framework of gradient flows of symmetric interaction energies. Hence, for

$$E[\mu] = -\frac{c}{2\beta} \iint W(x, y) d\mu(x) d\mu(y),$$

the standard first-variation formula for symmetric interaction energies gives

$$\frac{\delta E}{\delta \mu}(x) = -\frac{c}{\beta} \int W(x, y) d\mu(y).$$

Therefore, since $H' = h$,

$$\nabla_{\mathbb{S}^{d-1}} \frac{\delta E}{\delta \mu}(x) = -c P_x^\perp \int h(\beta \langle x, Ay \rangle) Ay d\mu(y).$$

Consequently, the characteristic velocity of the Wasserstein gradient flow is

$$v(x) = -\nabla_{\mathbb{S}^{d-1}} \frac{\delta E}{\delta \mu}(x) = c P_x^\perp \int h(\beta \langle x, Ay \rangle) Ay d\mu(y) = P_x^\perp \int h(\beta \langle x, Ay \rangle) V y d\mu(y).$$

Taking $\mu_t^n = \frac{1}{n} \sum_{j=1}^n \delta_{x_j(t)}$ yields exactly (B.1). $\square$

B.2 Proof of Theorem 2.1

Proof of Theorem 2.1. For the matrices A and V as defined in (2.2), and for each $k \in \{1, \dots, n\}$ we can first rewrite the dynamics (2.1) satisfied by the angle $\theta_k(t)$ as

$$\begin{aligned} \dot{\theta}_k = & -\frac{1}{n} \sum_{j=1}^n (a_{11} \cos \theta_k \cos \theta_j + a_{12} \cos \theta_k \sin \theta_j + a_{21} \sin \theta_k \cos \theta_j + a_{22} \sin \theta_k \sin \theta_j) \\ & \cdot (v_{11} \sin \theta_k \cos \theta_j + v_{12} \sin \theta_k \sin \theta_j - v_{21} \cos \theta_k \cos \theta_j - v_{22} \cos \theta_k \sin \theta_j). \end{aligned} \quad (\text{B.2})$$

To obtain Equation (B.2), recall that $x_k = (\cos \theta_k, \sin \theta_k)^\top$ and analogously for x_j . Hence,

$$\langle x_k, Ax_j \rangle = a_{11} \cos \theta_k \cos \theta_j + a_{12} \cos \theta_k \sin \theta_j + a_{21} \sin \theta_k \cos \theta_j + a_{22} \sin \theta_k \sin \theta_j.$$

Since V is real-valued, we can compute

$$\frac{1}{\sin \theta_k} \langle Vx_j, e_1 - \langle x_k, e_1 \rangle x_k \rangle = v_{11} \sin \theta_k \cos \theta_j + v_{12} \sin \theta_k \sin \theta_j - v_{21} \cos \theta_k \cos \theta_j - v_{22} \cos \theta_k \sin \theta_j,$$

which concludes the computations for Equation (B.2).

Next we reorganize the different summands in (B.2), which requires us to use the quantities defined in (A.2). After expanding the product on the right-hand side of (B.2) by using the trigonometric identities $\sin \theta \cos \theta = \frac{1}{2} \sin 2\theta$, $(\sin \theta)^2 = \frac{1}{2}(1 - \cos 2\theta)$, and $(\cos \theta)^2 = \frac{1}{2}(1 + \cos 2\theta)$, we arrive for (B.2) at

$$\begin{aligned} \dot{\theta}_k = & -\frac{1}{n} \sum_{j=1}^n \left(\left(v_{11} a_{11} \sin \theta_k \cos \theta_k (\cos \theta_j)^2 + v_{11} a_{12} \sin \theta_k \cos \theta_k \cos \theta_j \sin \theta_j \right. \right. \\ & + v_{11} a_{21} (\sin \theta_k)^2 (\cos \theta_j)^2 + v_{11} a_{22} (\sin \theta_k)^2 \cos \theta_j \sin \theta_j \Big) \\ & + \left(v_{12} a_{11} \cos \theta_k \sin \theta_k \sin \theta_j \cos \theta_j + v_{12} a_{12} \cos \theta_k \sin \theta_k (\sin \theta_j)^2 \right. \\ & + v_{12} a_{21} (\sin \theta_k)^2 \sin \theta_j \cos \theta_j + v_{12} a_{22} (\sin \theta_k)^2 (\sin \theta_j)^2 \Big) \\ & - \left(v_{21} a_{11} (\cos \theta_k)^2 (\cos \theta_j)^2 + v_{21} a_{12} (\cos \theta_k)^2 \cos \theta_j \sin \theta_j \right. \\ & + v_{21} a_{21} \sin \theta_k \cos \theta_k (\cos \theta_j)^2 + v_{21} a_{22} \sin \theta_k \cos \theta_k \cos \theta_j \sin \theta_j \Big) \\ & - \left(v_{22} a_{11} (\cos \theta_k)^2 \sin \theta_j \cos \theta_j + v_{22} a_{12} (\cos \theta_k)^2 (\sin \theta_j)^2 \right. \\ & + v_{22} a_{21} \sin \theta_k \cos \theta_k \sin \theta_j \cos \theta_j + v_{22} a_{22} \sin \theta_k \cos \theta_k (\sin \theta_j)^2 \Big) \Big) \\ = & -\frac{1}{4} (v_{11} a_{21} + v_{12} a_{22} - v_{21} a_{11} - v_{22} a_{12}) \\ & -\frac{1}{4} \left((v_{11} a_{11} + v_{12} a_{12} - v_{21} a_{21} - v_{22} a_{22}) \sin 2\theta_k + (-v_{11} a_{21} - v_{12} a_{22} - v_{21} a_{11} - v_{22} a_{12}) \cos 2\theta_k \right) \\ & -\frac{1}{4n} \sum_{j=1}^n \left((v_{11} a_{21} - v_{12} a_{22} - v_{21} a_{11} + v_{22} a_{12}) \cos 2\theta_j + (v_{11} a_{22} + v_{12} a_{21} - v_{21} a_{12} - v_{22} a_{11}) \sin 2\theta_j \right) \\ & -\frac{1}{4n} \sum_{j=1}^n \left((v_{11} a_{11} - v_{12} a_{12} - v_{21} a_{21} + v_{22} a_{22}) \sin 2\theta_k \cos 2\theta_j \right. \\ & + ((v_{11} a_{12} + v_{12} a_{11} - v_{21} a_{22} - v_{22} a_{21}) \sin 2\theta_k \sin 2\theta_j) \\ & - ((v_{11} a_{21} - v_{12} a_{22} + v_{21} a_{11} - v_{22} a_{12}) \cos 2\theta_k \cos 2\theta_j) \\ & - ((v_{11} a_{22} + v_{12} a_{21} + v_{21} a_{12} + v_{22} a_{11}) \cos 2\theta_k \sin 2\theta_j) \Big) \\ = & \left(-\frac{1}{4} (w_0 + w_1 \sin 2\theta_k + w_2 \cos 2\theta_k) - \frac{1}{4n} \sum_{j=1}^n (w_3 \cos 2\theta_j + w_4 \sin 2\theta_j) \right) \end{aligned} \quad (\text{B.3})$$

$$-\frac{1}{4n} \sum_{j=1}^n (w_5 \sin 2\theta_k \cos 2\theta_j + w_6 \sin 2\theta_k \sin 2\theta_j - w_7 \cos 2\theta_k \cos 2\theta_j - w_8 \cos 2\theta_k \sin 2\theta_j) := T_1 + T_2.$$

Here, $\{w_i\}_{i=0}^8$ are defined in (A.2). Thanks to Euler's identity, we observe that

$$\operatorname{Re}(\mathcal{R}_2^n) = \frac{1}{n} \sum_{j=1}^n \cos 2\theta_j \quad \text{and} \quad \operatorname{Im}(\mathcal{R}_2^n) = \frac{1}{n} \sum_{j=1}^n \sin 2\theta_j.$$

We now simplify each of the terms T_1 and T_2 .

Term T_1 . For term T_1 , we have

$$\begin{aligned} T_1 &= -\frac{1}{4} (w_0 + w_1 \sin 2\theta_k + w_2 \cos 2\theta_k) - \frac{1}{4} \left(w_3 \left(\frac{1}{n} \sum_{j=1}^n \cos 2\theta_j \right) + w_4 \left(\frac{1}{n} \sum_{j=1}^n \sin 2\theta_j \right) \right) \\ &= -\frac{1}{4} (w_0 + w_1 \operatorname{Im}(e^{2i\theta_k}) + w_2 \operatorname{Re}(e^{2i\theta_k})) - \frac{1}{4} (w_3 \operatorname{Re}(\mathcal{R}_2^n) + w_4 \operatorname{Im}(\mathcal{R}_2^n)). \end{aligned}$$

Recalling that $\operatorname{Re}(z) = \frac{1}{2}(z + \bar{z})$ and $\operatorname{Im}(z) = \frac{1}{2i}(z - \bar{z})$, we obtain

$$\begin{aligned} T_1 &= -\frac{1}{4} \left(w_0 + w_1 \frac{e^{2i\theta_k} - e^{-2i\theta_k}}{2i} + w_2 \frac{e^{2i\theta_k} + e^{-2i\theta_k}}{2} \right) - \frac{1}{4} \left(w_3 \frac{\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}}{2} + w_4 \frac{\mathcal{R}_2^n - \overline{\mathcal{R}_2^n}}{2i} \right) \\ &= b_3 e^{2i\theta_k} + \bar{b}_3 e^{-2i\theta_k} + C(\mathcal{R}_2^n), \end{aligned}$$

with b_3 and $C(r)$ as in (A.3) and (A.4).

Term T_2 . For term T_2 , we have

$$\begin{aligned} T_2 &= -\frac{1}{4n} \sum_{j=1}^n (w_5 \sin 2\theta_k \cos 2\theta_j + w_6 \sin 2\theta_k \sin 2\theta_j - w_7 \cos 2\theta_k \cos 2\theta_j - w_8 \cos 2\theta_k \sin 2\theta_j) \\ &= -\frac{1}{4} (w_5 \operatorname{Im}(e^{2i\theta_k}) \operatorname{Re}(\mathcal{R}_2^n) + w_6 \operatorname{Im}(e^{2i\theta_k}) \operatorname{Im}(\mathcal{R}_2^n) - w_7 \operatorname{Re}(e^{2i\theta_k}) \operatorname{Re}(\mathcal{R}_2^n) - w_8 \operatorname{Re}(e^{2i\theta_k}) \operatorname{Im}(\mathcal{R}_2^n)). \end{aligned}$$

By applying the real and complex identities, we obtain

$$\begin{aligned} T_2 &= -\frac{1}{4} \left(w_5 \frac{e^{2i\theta_k} - e^{-2i\theta_k}}{2i} \frac{\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}}{2} + w_6 \frac{e^{2i\theta_k} - e^{-2i\theta_k}}{2i} \frac{\mathcal{R}_2^n - \overline{\mathcal{R}_2^n}}{2i} \right. \\ &\quad \left. - w_7 \frac{e^{2i\theta_k} + e^{-2i\theta_k}}{2} \frac{\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}}{2} - w_8 \frac{e^{2i\theta_k} + e^{-2i\theta_k}}{2} \frac{\mathcal{R}_2^n - \overline{\mathcal{R}_2^n}}{2i} \right) \\ &= \frac{1}{16} e^{2i\theta_k} (iw_5 (\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}) + w_6 (\mathcal{R}_2^n - \overline{\mathcal{R}_2^n}) + w_7 (\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}) - iw_8 (\mathcal{R}_2^n - \overline{\mathcal{R}_2^n})) \\ &\quad + \frac{1}{16} e^{-2i\theta_k} (iw_5 (\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}) - w_6 (\mathcal{R}_2^n - \overline{\mathcal{R}_2^n}) - w_7 (\mathcal{R}_2^n + \overline{\mathcal{R}_2^n}) + iw_8 (\mathcal{R}_2^n - \overline{\mathcal{R}_2^n})) \\ &= B_2(\mathcal{R}_2^n) e^{2i\theta_k} + \overline{B_2}(\mathcal{R}_2^n) e^{-2i\theta_k} \end{aligned}$$

with

$$B_2(r) := b_1 r + b_2 \bar{r}.$$

Conclusion. Combining now the expressions for T_1 and T_2 and inserting them back into (B.3) yields for (B.2) the equivalent formulation

$$\dot{\theta}_k = B(\mathcal{R}_2^n) e^{2i\theta_k} + \overline{B}(\mathcal{R}_2^n) e^{-2i\theta_k} + C(\mathcal{R}_2^n),$$

where $B(r)$ and $C(r)$ are defined in (A.4). □

B.3 Proof Details for Section 2.2

Proof of Proposition 2.4. The proof follows computations similar to those presented in [52,86]. For notational simplicity, let us abbreviate $B(\mathcal{R}_2^n)$ and $C(\mathcal{R}_2^n)$ as B and C , respectively.

Denoting $z_k := e^{i\xi_k}$, we compute with chain rule and using the dynamics (2.9) that

$$\frac{d}{dt}z_k = i\dot{\xi}_k z_k = 2i(Bz_k^2 + Cz_k + \bar{B}) \quad (\text{B.4})$$

for all $k \in \{1, \dots, n\}$. Observe that each z_k solves a Riccati equation (with the same coefficients for all k) and satisfies the constraint $|z_k(t)| = 1$ for all $t \geq 0$. Hence, z_k can be represented as

$$e^{i\xi_k} = z_k = \frac{\alpha^n + e^{i\vartheta_k}}{1 + \bar{\alpha}^n e^{i\vartheta_k}} \quad (\text{B.5})$$

with $\alpha^n : [0, \infty) \mapsto \mathbb{C}$ satisfying $|\alpha^n| < 1$ and $\vartheta_k \in \mathbb{T}$ for $k \in \{1, \dots, n\}$. For notational simplicity, we shorthand α^n by α , and denote $u_k := e^{i\vartheta_k}$.

Expanding the left-hand side of (B.4), by plugging in the transformation (B.5), we get

$$\begin{aligned} \frac{d}{dt}z_k &= \frac{d}{dt} \left(\frac{\alpha + u_k}{1 + \bar{\alpha}u_k} \right) = \frac{\dot{\alpha} + i\dot{\vartheta}_k u_k}{1 + \bar{\alpha}u_k} - \frac{(\alpha + u_k) \left(\dot{\bar{\alpha}}u_k + i\dot{\vartheta}_k u_k \bar{\alpha} \right)}{(1 + \bar{\alpha}u_k)^2} \\ &= \frac{1}{(1 + \bar{\alpha}u_k)^2} \left(\dot{\alpha} + \left(\dot{\alpha}\bar{\alpha} - \alpha\dot{\bar{\alpha}} + i\dot{\vartheta}_k(1 - |\alpha|^2) \right) u_k - \dot{\bar{\alpha}}u_k^2 \right). \end{aligned}$$

Likewise, the right-hand side of (B.4) becomes

$$\begin{aligned} 2i(Bz_k^2 + Cz_k + \bar{B}) &= 2i \left(B \left(\frac{\alpha + u_k}{1 + \bar{\alpha}u_k} \right)^2 + C \frac{\alpha + u_k}{1 + \bar{\alpha}u_k} + \bar{B} \right) \\ &= \frac{2i}{(1 + \bar{\alpha}u_k)^2} \left((B\alpha^2 + C\alpha + \bar{B}) + (4\text{Re}(B\alpha) + C(1 + |\alpha|^2)) u_k + (B + C\bar{\alpha} + \bar{B}\bar{\alpha}^2) u_k^2 \right). \end{aligned}$$

By combining the constant terms as well as the coefficients of u_k, u_k^2 , we obtain the condition

$$K_1 + K_2 u_k - \bar{K}_1 u_k^2 = 0 \quad (\text{B.6})$$

for all $k \in \{1, \dots, n\}$, where the coefficients are defined as

$$K_1 := \dot{\alpha} - 2i(B\alpha^2 + C\alpha + \bar{B}), \quad K_2 := \dot{\alpha}\bar{\alpha} - \alpha\dot{\bar{\alpha}} + i\dot{\vartheta}_k(1 - |\alpha|^2) - 2i(4\text{Re}(B\alpha) + C(1 + |\alpha|^2)).$$

We now choose α and ϑ_k so that the identity (B.6) holds for all u_k , for which it suffices to set the coefficients to zero, i.e., $K_1 = K_2 = 0$.

From $K_1 = 0$, we obtain the dynamics that α needs to satisfy. Specifically,

$$\dot{\alpha} = 2i(B\alpha^2 + C\alpha + \bar{B}). \quad (\text{B.7})$$

From $K_2 = 0$, we deduce using (B.7) that

$$\dot{\vartheta}_k = 2(2\text{Re}(B\alpha) + C) \quad (\text{B.8})$$

for all $k \in \{1, \dots, n\}$. Noticing that the right-hand side of (B.8) is independent from the particle index k implies that all angles $\{\vartheta_k\}_{k=1}^n$ rotate at the same speed. Therefore, we can introduce a new time-dependent but particle-independent parameter η which has the same rotational speed as all $\{\vartheta_k\}$'s, i.e., $\dot{\vartheta}_k = \dot{\eta}$, and define

$$\eta(t) = \vartheta_k(t) - \varphi_k$$

for all $k \in \{1, \dots, n\}$, where $\{\varphi_k\}_{k=1}^n$ are the constants of motion.

We thus arrive at the family of Möbius transformations as defined in (2.16), mapping constants φ_k to phases ξ_k according to

$$M_{\alpha,\eta} : \varphi_k \mapsto \xi_k, \quad e^{i\xi_k} = M_{\alpha,\eta}(e^{i\varphi_k}) = \frac{\alpha + e^{i(\varphi_k+\eta)}}{1 + \bar{\alpha}e^{i(\varphi_k+\eta)}},$$

with the WS variables α and η following the coupled ODE system given by

$$\dot{\alpha} = 2i(B\alpha^2 + C\alpha + \bar{B}), \quad \dot{\eta} = 2(2\operatorname{Re}(B\alpha) + C).$$

Moreover, when the parameter α has modulus less than one, the Möbius transformation is a one-to-one mapping and thus admits a corresponding inverse transformation, which can be explicitly written as in (2.22). $\square$

Proof of Lemma 2.6. We fix $t \geq 0$ and suppress the time dependence for notational simplicity. Let $\alpha \in \mathbb{C}$ be such that $|\alpha| < 1$. Since $T_{\alpha,\eta}$ is an one-to-one map with $|\alpha| < 1$, it suffices to prove that the pushforward of the uniform distribution under $T_{\alpha,\eta}$ is the wrapped Cauchy distribution $\operatorname{WC}(\bullet; \alpha)$.

Suppose $\nu = \operatorname{Unif}([0, 2\pi))$, and let $g(\xi)$ denote the density of the distribution $T_{\alpha,\eta\sharp}(\operatorname{Unif}([0, 2\pi)))$. Then, using the definitions of pushforward and the of the map $T_{\alpha,\eta}$, we obtain, for all $m \in \mathbb{Z}$,

$$\begin{aligned} \int_0^{2\pi} g(\xi) e^{im\xi} d\xi &= \frac{1}{2\pi} \int_0^{2\pi} e^{imT_{\alpha,\eta}(\varphi)} d\varphi = \frac{1}{2\pi} \int_0^{2\pi} (e^{iT_{\alpha,\eta}(\varphi)})^m d\varphi = \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\alpha + e^{i(\varphi+\eta)}}{1 + \bar{\alpha}e^{i(\varphi+\eta)}} \right)^m d\varphi \\ &= \frac{-i}{2\pi} \oint_{\mathbb{S}^1} \left(\frac{\alpha + z}{1 + \bar{\alpha}z} \right)^m \frac{1}{z} dz = \alpha^m, \end{aligned}$$

where we wrote the second to last integral as a line integral in the complex plane (with the change of variables $z = e^{i(\varphi+\eta)}$), and used the residue theorem to obtain the last equality. The above allows us to write g in Fourier basis as

$$g(\xi) = \frac{1}{2\pi} \sum_{m=0}^{\infty} (\alpha^m e^{-im\xi} + \bar{\alpha}^m e^{im\xi}) = \frac{1}{2\pi} \left(1 + 2 \sum_{m=1}^{\infty} \operatorname{Re}(\bar{\alpha}^m e^{im\xi}) \right) = \frac{1}{2\pi} \operatorname{Re} \left(\frac{1 + \bar{\alpha}e^{i\xi}}{1 - \bar{\alpha}e^{i\xi}} \right) = \frac{1}{2\pi} \frac{1 - |\alpha|^2}{|e^{i\xi} - \alpha|^2},$$

which is precisely the formula for the density of $\operatorname{WC}(\bullet; \alpha)$ stated in (2.28). Note that to get the third equality we used the formula for the geometric series. $\square$

C Proof Details for Section 3

In this appendix, we provide the technical details for the proofs of Section 3.

C.1 Proofs of Main Results

Case 1.

Proof of Theorem 3.1. Recall from (3.4) that the dynamics for the order parameter $\mathcal{R}_2$ are given by

$$\dot{\mathcal{R}}_2 = \frac{1}{4} (1 - |\mathcal{R}_2|^2) ((a_+^d + ia_-^o) \mathcal{R}_2 + (a_-^d + ia_+^o)). \quad (\text{C.1})$$

For notational convenience, we set

$$z(t) := \mathcal{R}_2(t), \quad a := a_+^d, \quad b := a_-^o, \quad c := a_-^d + ia_+^o, \quad R := |c|, \quad \Phi_A := \operatorname{Arg}(c). \quad (\text{C.2})$$

We set $\Phi_A = 0$ if $R = 0$. We also denote the closed unit disk and its boundary by

$$D := \{w : |w| \leq 1\}, \quad \partial D := \{w : |w| = 1\}.$$

We first consider the general case where $a \neq 0$ or $b \neq 0$. The equation of z can be written as

$$\dot{z} = \frac{1}{4} (a + ib) (1 - |z|^2) (z - z_{\text{eq}}), \quad z(0) = z_0, \quad (\text{C.3})$$

where

$$z_{\text{eq}} := -\frac{c}{a + ib}. \quad (\text{C.4})$$

Two important quantities and associated observations. It is straightforward to verify that the equilibrium set of the dynamics (C.3) is $\partial D \cup \{z_{\text{eq}}\}$. To understand the asymptotic behavior of $z(t)$, we examine how the trajectory $z(t)$ moves relative to the interior equilibrium z_{eq} , and how it may interact with the boundary equilibrium set ∂D .

To this end, we introduce two quantities associated with the displacement $z(t) - z_{\text{eq}}$. The first is the energy function

$$\mathcal{E}(t) = |z - z_{\text{eq}}|^2, \quad (\text{C.5})$$

which measures the squared distance between $z(t)$ and z_{eq} . A direct computation gives

$$\dot{\mathcal{E}} = \frac{1}{2}a(1 - |z|^2)\mathcal{E}. \quad (\text{C.6})$$

Consequently, integrating above equation yields

$$\mathcal{E}(t) = \mathcal{E}(0) \exp \left\{ \frac{a}{2} \int_0^t (1 - |z(s)|^2) ds \right\}. \quad (\text{C.7})$$

An important observation is that the sign of a determines the monotonicity of the energy function $\mathcal{E}(t)$, and hence describes how the trajectory $z(t)$ moves relative to z_{eq} . More precisely,

- when $a > 0$, $\dot{\mathcal{E}} > 0$, and thus the trajectory $z(t)$ is pushed away from z_{eq} ;
- when $a = 0$, $\dot{\mathcal{E}} = 0$, and thus the distance between $z(t)$ and z_{eq} is preserved;
- when $a < 0$, $\dot{\mathcal{E}} < 0$, and thus the trajectory $z(t)$ is drawn toward z_{eq} .

Thus, in each of these three regimes, the long-term behavior of $z(t)$ is determined by how the trajectory moves relative to z_{eq} and whether it reaches the boundary ∂D . Roughly speaking, $z(t)$ converges to a boundary equilibrium point if the trajectory is pushed away from z_{eq} , or if, while cycling around or moving toward z_{eq} , it reaches the boundary ∂D . By contrast, if the trajectory remains entirely inside the unit disk D , then $z(t)$ either keeps cycling around z_{eq} or converges to it; see Figure 20 as an illustrative example. A more detailed and rigorous characterization is provided in Claims 1, 2, and 3 below.

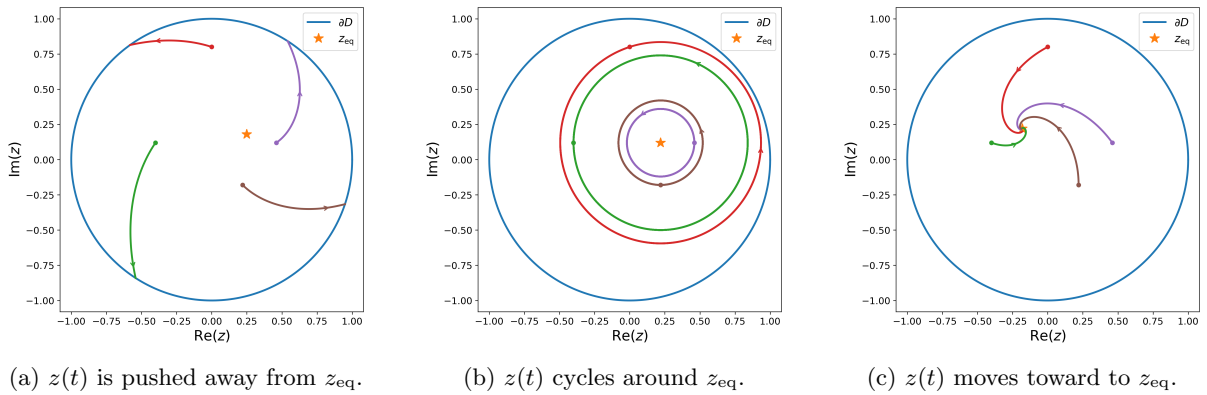


Figure 20: The visualization of the trajectory $z(t)$ relative to the interior equilibrium z_{eq} and the boundary equilibrium set ∂D . The solid dot is the initialization point, while the arrow indicates the moving direction of $z(t)$.

The second important quantity is the angular motion of the displacement $z(t) - z_{\text{eq}}$, that is,

$$\zeta(t) := \text{Arg}(z(t) - z_{\text{eq}}). \quad (\text{C.8})$$

This quantity describes how the trajectory $z(t)$ rotates around z_{eq} , and will be useful for determining whether the angular component of the trajectory converges.

A direct computation shows that the angle ζ satisfies

$$\dot{\zeta} = \frac{b}{4}(1 - |z|^2). \quad (\text{C.9})$$

Integrating in time yields

$$\zeta(t) = \zeta(0) + \frac{b}{4} \int_0^t (1 - |z(s)|^2) ds. \quad (\text{C.10})$$

Thus, the convergence of ζ is determined by the following condition:

$$\lim_{t \rightarrow \infty} \int_0^t (1 - |z(s)|^2) ds = C < \infty. \quad (\text{C.11})$$

We now formalize the above observations by analyzing the three parameter regimes: $a > 0$, $a = 0$ & $b \neq 0$, and $a < 0$.

Claim 1 (regime $a > 0$): when $a > 0$, for any initial condition $z_0 \neq z_{\text{eq}}$ and $z_0 \in D$, the solution of (C.3) converges to a point $z_* \in \partial D$.

Proof of Claim 1. Since ∂D is an equilibrium set and $z_0 \in D$, we have $z(t) \in D$ and an estimate

$$0 \leq \mathcal{E}(t) \leq 2(|z|^2 + |z_{\text{eq}}|^2) \leq 2 + \frac{2R}{a^2 + b^2} < \infty, \quad \text{for all } t \geq 0.$$

On the other hand, from (C.7), $a > 0$ and $\mathcal{E}(t) < \infty$, we know that the condition (C.11) must hold. Together with (C.10), this implies the convergence of the angle ζ to some $\zeta_\infty \in \mathbb{T}^1$. By (C.8), we thus know that $\text{Arg}(z)$ also converges to a fixed value in $\mathbb{T}^1$.

To conclude that $|z| \rightarrow 1$, we note that from the dynamics of z in (C.3), we know that $|z(t)|$ is Lipchitz, and thus the integrand $1 - |z(t)|^2$ is Lipchitz. Together with (C.11) and Barb lat's lemma [12], this implies that $1 - |z(t)|^2 \rightarrow 0$.

Combining the convergence of both modulus and argument of z , we conclude the proof. $\square$

Here, we record an additional quantitative convergence estimate in the sub-regime $a > R$, which will be used in the proof of Theorem 4.7. We defer the proof until after the proof of Theorem 3.1.

Proposition C.1. *Let $z(t)$ be the solution to (C.3) with initial data $z(0) \neq z_{\text{eq}}$, and let $z_* \in \partial D$ denote the limiting point, so that $\lim_{t \rightarrow \infty} z(t) = z_*$. For any prescribed $\epsilon > 0$, define*

$$T_\epsilon := \inf \{t \geq 0 : |z(t) - z_*| \leq \epsilon\}.$$

Suppose that $a > R$. Then there exist constants $C_1, C_2 > 0$, depending only on the matrix A , such that

$$T_\epsilon \leq C_1 \log \frac{1}{|z_0 - z_{\text{eq}}|} + C_2 \log \frac{1}{\epsilon}.$$

Here, z_{eq} is defined in (C.4).

Claim 2 (regime $a = 0$, $b \neq 0$): suppose $a = 0$, then

$$z(t) = z_{\text{eq}} + |z_0 - z_{\text{eq}}| e^{i\zeta(t)} \quad \text{for all } t \geq 0,$$

Furthermore, if

- $|b| > R$ and $|z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D)$, then $z(t)$ keeps circling within $D \setminus \partial D$ for all $t \geq 0$.
- $0 < |b| \leq R$ or $|z_0 - z_{\text{eq}}| \geq \text{dist}(z_{\text{eq}}, \partial D)$, then $\lim_{t \rightarrow \infty} z(t) = z_* \in \partial D$.

Proof of Claim 2. Setting $a = 0$ in (C.6), we obtain $\dot{\mathcal{E}} = 0$ and hence the dynamics (C.3) of z remain on a circle $C_{z_{\text{eq}}}$ centered around z_{eq} , that is,

$$z(t) \in C_{z_{\text{eq}}} := \{z : |z - z_{\text{eq}}| = |z_0 - z_{\text{eq}}|\}, \quad \text{for all } t \geq 0. \quad (\text{C.12})$$

Together with the polar representation (C.8), we have

$$z(t) - z_{\text{eq}} = |z(t) - z_{\text{eq}}|e^{i\zeta(t)} = |z_0 - z_{\text{eq}}|e^{i\zeta(t)} \implies z(t) = z_{\text{eq}} + |z_0 - z_{\text{eq}}|e^{i\zeta(t)}.$$

Next, we discuss the long-time behavior of z under different parameter regimes of A and initial condition of z_0 .

• **Parameter regime:** $|b| > R$.

Recall from (C.4) that $z_{\text{eq}} = -\frac{c}{a+ib}$. Together with $a = 0$ and $|b| > R$, we know that $z_{\text{eq}} \in D \setminus \partial D$.

In this parameter regime, we further consider two conditions on the initial data z_0 .

$$- |z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D).$$

Since $z_{\text{eq}} \in D$, we have

$$|z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D) = 1 - |z_{\text{eq}}|.$$

Therefore, we know that there exists $\delta \in (0, 1)$ such that

$$|z(t)| - |z_{\text{eq}}| \leq |z(t) - z_{\text{eq}}| \leq 1 - \delta - |z_{\text{eq}}|,$$

and thus

$$|z(t)| < 1 - \delta, \quad \text{for all } t \geq 0.$$

Therefore, from the dynamics (C.9) and $\delta \in (0, 1)$, we have for $b \neq 0$

$$|\dot{\zeta}| = \frac{|b|}{4}(1 - |z|^2) \geq \frac{|b|}{4}(2\delta - \delta^2) > 0,$$

Thus, the angular velocity is bounded from away from zero for all time and in turn implies that $\zeta(t)$ does not converge. Together with the fact that $z(t)$ always lies on the circle $C_{z_{\text{eq}}}$ as defined in (C.12), and the initial condition $|z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D)$, we know that the circle $C_{z_{\text{eq}}}$ is strictly contained in the unit disc D . Therefore the trajectory of z rotates around z_{eq} within $D \setminus \partial D$ perpetually.

$$- |z_0 - z_{\text{eq}}| \geq \text{dist}(z_{\text{eq}}, \partial D).$$

In this case, the circle $C_{z_{\text{eq}}}$ intersects with the boundary of the unit disc ∂D . Recall the dynamics of angle ζ in (C.9) and $b \neq 0$, thus we know that ζ is strictly monotone when $z \in D \setminus \partial D$. Together with $z(0) \in D \setminus \partial D$, this implies that the trajectory of $z(t)$ must converge to one of the points in $C_{z_{\text{eq}}} \cap \partial D$. Equivalently, we have $\lim_{t \rightarrow \infty} z(t) = z_* \in \partial D$.

• **Parameter regime:** $0 < |b| \leq R$.

Recall from (C.4) that $z_{\text{eq}} = -\frac{c}{a+ib}$. Together with $a = 0$ and $0 < |b| \leq R$, we know that $z_{\text{eq}} \in (D \setminus \partial D)^c$. In this case, the circle $C_{z_{\text{eq}}}$ intersects with the boundary of the unit disc ∂D . Recalling the dynamics of angle ζ in (C.9) and $b \neq 0$, thus we know that ζ is strictly monotone when $z \in D \setminus \partial D$. Together with $z(0) \in D \setminus \partial D$, this implies that the trajectory of $z(t)$ must converge to one of the points in $C_{z_{\text{eq}}} \cap \partial D$. Equivalently, we have $\lim_{t \rightarrow \infty} z(t) = z_* \in \partial D$.

□

Claim 3 (regime $a < 0$): For $a < 0$ and $z_0 \in D \setminus \partial D$,

- if $a^2 + b^2 \leq R^2$, then $\lim_{t \rightarrow \infty} z(t) = z_* \in \partial D$.
- if $a < -R$ or $(a^2 + b^2 > R^2 \ \& \ |z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D))$, then $\lim_{t \rightarrow \infty} z(t) = z_{\text{eq}}$.

Proof of Claim 3. When $a < 0$, recall from (C.6) that $\dot{\mathcal{E}} \leq 0$, and thus $\mathcal{E}$ is a Lyapunov function. Since D is a positive invariant compact set of the dynamics (C.3) and the zero derivative set of $\mathcal{E}$ coincides with the equilibrium set of (C.3):

$$\{z \in D : \dot{z} = 0\} = \{z \in D : \dot{\mathcal{E}} = 0\} = \partial D \cup \{z_{\text{eq}}\},$$

then the LaSalle invariance Principle implies that for all $z_0 \in D$, $z(t)$ converges to $\partial D \cup \{z_{\text{eq}}\}$ as $t \rightarrow \infty$. Next, we discuss three different parameter regimes, respectively.

- **Parameter regime:** $a^2 + b^2 \leq R^2$.

In this case, recall the definition of z_{eq} from (C.4), we thus know that $z_{\text{eq}} = -\frac{c}{a+ib} \in (D \setminus \partial D)^c$. This implies that for all $z_0 \in D$, $z(t)$ only converges to ∂D , that is, $|z(t)| \rightarrow 1$ as $t \rightarrow \infty$.

Next, we show that $\text{Arg}(z)$ also converges to a fixed point in $\mathbb{T}^1$. We discuss it in two sub-cases.

1. $z \rightarrow z_{\text{eq}}$. In this case, we by default obtain that $\text{Arg}(z) \rightarrow \text{Arg}(z_{\text{eq}}) \in \mathbb{T}^1$ as $t \rightarrow \infty$.
2. $z \not\rightarrow z_{\text{eq}}$. In this case, from (C.7), we know that the energy function $\mathcal{E}(t) > 0$ for all $t \geq 0$. Together with $a < 0$, this implies the condition (C.11) holds. Then again from the representation (C.10) of $\zeta(t)$, we know that $\zeta(t) \rightarrow \zeta_\infty \in \mathbb{T}^1$. From the relationship of $z(t)$ and $\zeta(t)$ in (C.8), we thus conclude that $\text{Arg}(z)$ converges in $\mathbb{T}^1$ as $t \rightarrow \infty$.

Based on the above discussions, we can conclude that when $a < 0$ and $a^2 + b^2 \leq R^2$, the dynamics of $z(t)$ satisfies $\lim_{t \rightarrow \infty} z(t) = z_* \in \partial D$.

- **Parameter regime:** $a < -R$.

From the LaSalle invariance principle, we have already shown that $z(t) \rightarrow \partial D \cup \{z_{\text{eq}}\}$. We now show that $z(t)$ cannot converge to ∂D , that is, $|z(t)| \not\rightarrow 1$, and therefore conclude that $z(t) \rightarrow z_{\text{eq}}$. To this end, it is convenient to work in polar coordinates.

Plugging the polar representation of the order parameter $z(t) = \mathcal{R}_2(t) = \rho(t)e^{i\phi(t)}$ in (C.3), together with $\Phi_A = \arg(c)$, we obtain the dynamics of ρ as

$$\dot{\rho} = \frac{1}{4}(1 - \rho^2)(a\rho + R\cos(\phi - \Phi_A)).$$

Now fix any $\rho \in (\rho_{\text{th}}, 1)$, where $\rho_{\text{th}} := \frac{R}{|a|} \in (0, 1)$. Since $a < 0$, we have

$$a\rho + R\cos(\phi - \Phi_A) \leq a\rho + R < a\rho_{\text{th}} + R = 0.$$

Moreover, for $\rho \in (\rho_{\text{th}}, 1)$, we have $1 - \rho^2 > 0$. Hence $\dot{\rho} < 0$ whenever $\rho \in (\rho_{\text{th}}, 1)$. This shows that the trajectory of $\rho(t)$ cannot approach $\rho = 1$ within the interval $(0, 1)$. Equivalently, for every initial condition $z_0 \notin \partial D$, we have $z(t) \not\rightarrow \partial D$. Combining with the conclusion from the LaSalle invariance principle, we therefore obtain $\lim_{t \rightarrow \infty} z(t) = z_{\text{eq}}$.

- **Parameter regime:** $a^2 + b^2 > R^2$ & $|z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D)$.

Since $a^2 + b^2 > R^2$, it follows from (C.4) that $z_{\text{eq}} = -\frac{c}{a+ib} \in D \setminus \partial D$. Together with the initial condition on z_0 : $|z_0 - z_{\text{eq}}| < \text{dist}(z_{\text{eq}}, \partial D)$, this implies that there exists a $\delta > 0$ such that

$$|z_0 - z_{\text{eq}}| \leq \text{dist}(z_{\text{eq}}, \partial D) - \delta = 1 - |z_{\text{eq}}| - \delta.$$

On the other hand, by $a < 0$ and (C.6), the energy function $\mathcal{E}(t) = |z(t) - z_{\text{eq}}|^2$ is non-increasing. Therefore, for all $t \geq 0$,

$$|z(t) - z_{\text{eq}}| \leq |z_0 - z_{\text{eq}}| \leq 1 - |z_{\text{eq}}| - \delta.$$

Using the inverse triangle inequality we obtain

$$|z(t)| \leq |z_{\text{eq}}| + |z(t) - z_{\text{eq}}| \leq 1 - \delta < 1, \quad \text{for all } t \geq 0.$$

Hence $z(t)$ cannot converge to the boundary ∂D . Combining this with the convergence conclusion from the LaSalle invariance principle, we deduce that $\lim_{t \rightarrow \infty} z(t) = z_{\text{eq}}$.

□

Lastly, we consider a boundary case $a = b = 0$, where dynamics (C.1) of z take the form

$$\dot{z} = \frac{c}{4}(1 - |z|^2), \quad z(0) = z_0. \quad (\text{C.13})$$

Note that $R = |c| > 0$ unless A is the zero matrix, which corresponds to a trivial situation.

Claim 4 (regime $a = 0$ and $b = 0$): For any initial condition $z_0 \in D$, the solution of (C.13) converges to a boundary point $z_* \in \partial D$.

Proof of Claim 4. If $z_0 \in \partial D$, then $|z_0| = 1$, and hence $\dot{z} = 0$. Thus the claim is immediate. It remains to consider $z_0 \in D \setminus \partial D$.

Rotating the complex plane by $e^{-i\Phi_A}$, set $\tilde{z} := e^{-i\Phi_A} z$. Then

$$\dot{\tilde{z}} = \frac{R}{4}(1 - |\tilde{z}|^2).$$

Write $\tilde{z}(t) = x(t) + iy(t)$, and then the dynamics of x and y are given by

$$\dot{x} = \frac{R}{4}(1 - x^2 - y^2), \quad \dot{y} = 0.$$

with initial condition $x(0) = x_0 := \text{Re}(\tilde{z}_0)$ and $y(0) = y_0 := \text{Im}(\tilde{z}_0)$. Thus $y(t) \equiv y_0$, and x satisfies

$$\dot{x} = \frac{R}{4}(\alpha^2 - x^2), \quad \alpha := \sqrt{1 - y_0^2}.$$

Since $\tilde{z}_0 \in D \setminus \partial D$, we have $x_0 \in (-\alpha, \alpha)$. Moreover, when $x(t) \in (-\alpha, \alpha)$, the condition $R > 0$ implies that $\dot{x} > 0$. Thus $x(t)$ is strictly increasing until it reaches the equilibrium α . Therefore,

$$\lim_{t \rightarrow \infty} z(t) = \lim_{t \rightarrow \infty} e^{i\Phi_A} \tilde{z}(t) = \lim_{t \rightarrow \infty} e^{i\Phi_A} (x(t) + iy_0) = e^{i\Phi_A} (\alpha + iy_0) =: z_*,$$

and by construction $|z_*| = 1$, so $z_* \in \partial D$. □

Conclusion for the order parameter $\mathcal{R}_2$. Recall that $\mathcal{R}_2(t) = z(t)$. We first summarize the long-term behavior of $\mathcal{R}_2(t)$ under different parameter regimes of the matrix A , by simply combining the four claims established above. We include only those regimes where the asymptotic behavior holds for all initial conditions, and omit cases that require additional assumptions on the initial data.

- If the matrix A satisfies $a > 0$ or $a \leq 0$ & $a^2 + b^2 \leq R^2$, then for every initial condition $\mathcal{R}_2(0) \neq z_{\text{eq}}$, the solution of (C.3) satisfies $\lim_{t \rightarrow \infty} \mathcal{R}_2(t) = z_* \in \partial D$.
- If the matrix A satisfies $a < -R$, then for every initial condition $\mathcal{R}_2(0) \in D \setminus \partial D$, the solution of (C.3) satisfies $\lim_{t \rightarrow \infty} \mathcal{R}_2(t) = z_{\text{eq}}$, where $z_{\text{eq}} \in D \setminus \partial D$ is defined in (C.4).

Using the notation $a = a_+^d = \text{tr}(A)$, $b = a_-^o$, $R = |c| = \sqrt{(a_-^d)^2 + (a_+^o)^2}$, together with basic algebraic identities, the above conditions on A can be equivalently expressed as the following final results:

- If the matrix A satisfies:

$$\text{tr}(A) > 0 \quad \text{or} \quad \det(A) \leq 0,$$

then the order parameter $\mathcal{R}_2(t) = \rho(t)e^{i\phi(t)}$ satisfies $\rho(t) \rightarrow 1$ and $\phi(t) \rightarrow \phi_\infty$ for some $\phi_\infty \in \mathbb{T}^1$ as $t \rightarrow \infty$.

- If the matrix A satisfies:

$$\frac{A + A^\top}{2} \prec 0,$$

then the order parameter $\mathcal{R}_2(t) = \rho(t)e^{i\phi(t)}$ satisfies $\rho(t) \rightarrow \rho_{\text{eq}}$ and $\phi(t) \rightarrow \phi_{\text{eq}}$ as $t \rightarrow \infty$, where

$$\rho_{\text{eq}} := \frac{\sqrt{(a_-^d)^2 + (a_+^o)^2}}{\sqrt{(a_+^d)^2 + (a_-^o)^2}} \in (0, 1), \quad \phi_{\text{eq}} := \pi + \text{Arg}(a_-^d + ia_+^o) - \text{Arg}(a_+^d + ia_-^o) \pmod{2\pi}. \quad (\text{C.14})$$

□

Proof of Proposition C.1. For notation simplicity, we use the notation introduced in the proof of Theorem 3.1; see the shorthand notation in (C.2), the dynamics for $z(t)$ in (C.3), and the definition of z_{eq} in (C.4). By Theorem 3.1, when $a > R > 0$, for any initial condition $z(0) \neq z_{\text{eq}}$, the solution $z(t)$ converges to a boundary equilibrium $z_* \in \partial D$.

Denote $q(t) := 1 - |z(t)|^2$. Using (C.3), a straightforward calculation gives

$$\dot{q}(t) = -\kappa(t)q(t), \quad \text{where } \kappa(t) := \frac{1}{2} (a|z(t)|^2 + \text{Re}(\bar{z}(t)c)) . \quad (\text{C.15})$$

Furthermore, by the elementary inequality, we obtain $\kappa(t) \geq \frac{1}{2}|z(t)|(a|z(t)| - R)$. Thus, if we choose any number $r \in (R/a, 1)$, then whenever $|z(t)| \geq r$, we have

$$\kappa(t) \geq \mu_r := \frac{1}{2}r(ar - R) > 0 . \quad (\text{C.16})$$

Define

$$T_r := \inf \{t \geq 0 : |z(t)| \geq r\} . \quad (\text{C.17})$$

Then for $t \geq T_r$, applying the inequality (C.16) in (C.15) and then integrating the differential inequality over time gives

$$q(t) \leq q(T_r) \exp(-\mu_r(t - T_r)) \leq (1 - r^2) \exp(-\mu_r(t - T_r)) .$$

On the other hand, from (C.3), we have

$$|\dot{z}(t)| \leq \frac{1}{4} (1 - |z(t)|^2) (\sqrt{a^2 + b^2} + R) = Kq(t), \quad \text{where } K := \frac{\sqrt{a^2 + b^2} + R}{4} .$$

Then we can compute that, for $t \geq T_r$,

$$|z(t) - z_*| \leq \int_t^\infty |\dot{z}(s)| ds \leq K \int_t^\infty q(s) ds \leq \frac{K}{\mu_r} (1 - r^2) \exp(-\mu_r(t - T_r)) .$$

Letting the right-hand side of above inequality less or equal to ϵ gives

$$t \geq T_r + \frac{1}{\mu_r} \log \frac{K(1 - r^2)}{\mu_r \epsilon} =: T_\epsilon .$$

Thus, it remains to estimate T_r as defined in (C.17).

Denote $w(t) := z(t) - z_{\text{eq}}$. Using (C.3), a direct computation gives

$$\frac{d}{dt}|w(t)| = \frac{a}{4}(1 - |z(t)|^2)|w(t)| .$$

For $t < T_r$, we have $|z(t)| < r$, and therefore $\frac{d}{dt}|w(t)| \geq \frac{a}{4}(1 - r^2)|w(t)|$. Integrating it over time gives

$$|w(t)| \geq |w(0)| \exp\left(\frac{a}{4}(1 - r^2)t\right), \quad t < T_r . \quad (\text{C.18})$$

On the other hand, since $|z(t)| < r$, we have $|w(t)| = |z(t) - z_{\text{eq}}| \leq |z(t)| + |z_{\text{eq}}| < r + |z_{\text{eq}}|$. Together with (C.18), this implies that

$$T_r \leq \frac{4}{a(1 - r^2)} \log \left(\frac{r + |z_{\text{eq}}|}{|z(0) - z_{\text{eq}}|} \right) .$$

Therefore, we obtain the final estimate on T_ϵ :

$$\begin{aligned} T_\epsilon &\leq \frac{4}{a(1 - r^2)} \log \left(\frac{r + |z_{\text{eq}}|}{|z(0) - z_{\text{eq}}|} \right) + \frac{1}{\mu_r} \log \frac{K(1 - r^2)}{\mu_r \epsilon} \quad \text{for any } r \in (R/a, 1), \\ &\leq C_1 \log \frac{1}{|z(0) - z_{\text{eq}}|} + C_2 \log \frac{1}{\epsilon}, \end{aligned}$$

where the constants $C_1, C_2 > 0$ only depend on the matrix A . □

Case 2.

Proof of Theorem 3.3. We begin with a change of variables. Denote $R := \sqrt{(v_-^d)^2 + (v_+^o)^2}$. In the following, we assume $R > 0$; the case $R = 0$ will be treated separately. Define the angle ϕ_V by

$$\cos \phi_V = \frac{v_-^d}{R}, \quad \sin \phi_V = \frac{v_+^o}{R}, \quad (\text{C.19})$$

Then, by the trigonometric identities, we have

$$v_-^d \cos \phi + v_+^o \sin \phi = R \cos(\phi - \phi_V), \quad v_-^d \sin \phi - v_+^o \cos \phi = R \sin(\phi - \phi_V).$$

For notational convenience, we set

$$\psi(t) := \phi(t) - \phi_V, \quad v := v_+^d, \quad \kappa(t) := \frac{1}{4} R \frac{3\rho(t)^2 + 1}{\rho(t)}. \quad (\text{C.20})$$

With these definitions, the original system (3.10) becomes

$$\dot{\rho} = \frac{1}{4}(1 - \rho^2)(v\rho + R \cos \psi), \quad \dot{\psi} = -\kappa \sin \psi. \quad (\text{C.21})$$

The equilibria of (C.21) are easily identified:

$$\begin{aligned} \{(\rho, \psi) = (1, 0)\}, & \quad \{(\rho, \psi) = (1, \pi)\}, \\ \{(\rho, \psi) = (\rho_{\text{eq}}, 0)\} \text{ (if } v + R < 0\text{)}, & \quad \{(\rho, \psi) = (-\rho_{\text{eq}}, \pi)\} \text{ (if } v - R > 0\text{)}, \end{aligned}$$

where we denote

$$\rho_{\text{eq}} := -\frac{R}{v}. \quad (\text{C.22})$$

We now analyze the long-time behavior. We first exclude a special initialization $\psi(0) = \pi$. This case will be discussed at the end.

We first show that $\rho(t) \in (0, 1]$ for all $t \geq 0$, which in turn ensures that $\kappa(t)$ stays positive and does not diverge. The upper bound $\rho(t) \leq 1$ follows from the invariance of the boundary $\rho = 1$ to the ρ -dynamics (C.21) and the assumption $\rho(0) \leq 1$. To prove $\rho(t) > 0$, we momentarily convert to Cartesian coordinates $x = \rho \cos \psi$ and $y = \rho \sin \psi$. A direct computation gives

$$\dot{x} = \frac{1}{4} [v(1 - x^2 - y^2)x + R(1 - x^2 + 3y^2)], \quad \dot{y} = \frac{1}{4} [v(1 - x^2 - y^2) - 4Rx]y.$$

The Cartesian system is smooth and is globally well-posed on the invariant unit disk $x^2 + y^2 \leq 1$. Integrating the y -equation gives

$$y(t) = y(0) \exp \left\{ \frac{1}{4} \int_0^t [v(1 - \rho(s)^2) - 4Rx(s)] ds \right\}.$$

If $y(0) \neq 0$, then the apriori estimates $\rho(t) \leq 1$ and $|x(t)| \leq 1$ ensure that the exponent in the y -formulation is finite on every finite time interval, which implies $y(t) \neq 0$ for every finite $t \geq 0$. If $y(0) = 0$, then $y(t) = 0$ for all $t \geq 0$, and the x -equation reduces to

$$\dot{x} = \frac{1}{4}(1 - x^2)(vx + R), \quad x(0) = x_0.$$

Since $\rho(0) > 0$ and $\psi(0) \neq \pi$, we have $x(0) > 0$. Moreover, note that $\dot{x}|_{x=0} = \frac{R}{4} > 0$, we thus know that $x(t)$ cannot reach zero.

Combining the above discussions, we show that $\rho(t) = \sqrt{x^2(t) + y^2(t)} \in (0, 1]$ for all $t \geq 0$, and hence $\kappa(t)$ is well-defined and positive. More precisely, $\kappa(t) = \frac{R}{4}(3\rho(t) + 1/\rho(t)) \geq \frac{R}{4}$.

Next, we analysis the ψ -dynamics. We first observe that $\psi = 0$ is an invariance state of the ψ -dynamics in (C.21). Hence, if $\psi(0) = 0$, then $\psi(t) = 0$ for all $t \geq 0$. Thus, in the remaining discussion, we restrict our attention to the case $\psi(0) \neq 0, \pi$.

Introduce the half-angle variable

$$u(t) := \tan(\psi(t)/2).$$

Using the identities

$$\sin \psi = \frac{2u}{1+u^2}, \quad \dot{u} = \frac{1}{2}(1+u^2)\dot{\psi},$$

we find that $u(t)$ solves the linear equation:

$$\dot{u} = -\kappa(t)u.$$

Therefore,

$$u(t) = u(0) \exp \left\{ - \int_0^t \kappa(s) ds \right\}. \quad (\text{C.23})$$

Since $\kappa(t) \geq \frac{R}{4}$, together with $|\psi(t)| = 2|\arctan(u(t))| \leq 2|u(t)|$ gives

$$|\psi(t)| \leq 2|u(0)| \exp \left\{ - \int_0^t \kappa(s) ds \right\} \leq 2|u(0)| \exp \left(-\frac{R}{4}t \right). \quad (\text{C.24})$$

Since $\psi(0) \neq 0, \pi$, which implies $|u(0)| < \infty$, we therefore conclude from the above inequality that $\psi(t)$ converges to 0 exponentially fast as $t \rightarrow \infty$.

Next, we turn to the ρ -dynamics. Define

$$G(\rho) := \frac{1}{4}(1-\rho^2)(v\rho + R), \quad E(t) := \frac{R}{4}(1-\rho^2)(1-\cos \psi). \quad (\text{C.25})$$

Then the ρ -equation in (C.21) can be written as

$$\dot{\rho} = G(\rho) - E(t), \quad \rho(0) = \rho_0. \quad (\text{C.26})$$

Since $\rho(t) \leq 1$, $\cos \psi(t) \leq 1$, and $R > 0$, we know $E(t)$ is non-negative. Moreover, by the exponential convergence of $\psi(t)$ to 0, established in (C.24), $E(t) \rightarrow 0$ exponentially fast as $t \rightarrow \infty$. Thus the long-time behavior of $\rho(t)$ is governed, up to an exponentially decaying perturbation, by the vector field G . We make this observation mathematically precise in the following lemma.

Lemma C.2. *Let $a < b$. Suppose that $\varrho : [0, \infty) \rightarrow [a, b]$ solves*

$$\dot{\varrho}(t) = g(\varrho(t)) + e(t),$$

where $g : [a, b] \rightarrow \mathbb{R}$ is Lipschitz, and $e : [0, \infty) \mapsto \mathbb{R}$ is continuous and satisfies

$$\lim_{t \rightarrow \infty} e(t) = 0.$$

Assume that there exists $\varrho_ \in [a, b]$ such that*

$$g(\varrho_*) = 0,$$

and

$$(\varrho - \varrho_*)g(\varrho) < 0 \quad \text{for every } \varrho \in [a, b] \setminus \{\varrho_*\}.$$

Then

$$\lim_{t \rightarrow \infty} \varrho(t) = \varrho_*.$$

Proof. Fix $\varepsilon > 0$ so that the set

$$K_\varepsilon := \{\varrho \in [a, b] : |\varrho - \varrho_*| \geq \varepsilon\}$$

is nonempty. By the continuity of g and the strict sign assumption, we have

$$m_\varepsilon := \min_{\varrho \in K_\varepsilon} |g(\varrho)| > 0.$$

Since $e(t) \rightarrow 0$, there exists $T_\varepsilon > 0$ such that

$$|e(t)| \leq \frac{m_\varepsilon}{2} \quad \text{for all } t \geq T_\varepsilon.$$

Therefore, for $t \geq T_\varepsilon$, if

$$\varrho(t) \leq \varrho_* - \varepsilon,$$

then

$$\dot{\varrho}(t) = g(\varrho(t)) + e(t) \geq m_\varepsilon - |e(t)| \geq \frac{m_\varepsilon}{2} > 0.$$

Similarly, if

$$\varrho(t) \geq \varrho_* + \varepsilon,$$

then

$$\dot{\varrho}(t) = g(\varrho(t)) + e(t) \leq -m_\varepsilon + |e(t)| \leq -\frac{m_\varepsilon}{2} < 0.$$

Define

$$I_\varepsilon := [a, b] \cap [\varrho_* - \varepsilon, \varrho_* + \varepsilon].$$

We claim that the solution enters I_ε in finite time after T_ε .

Indeed, if the solution stays below $\varrho_* - \varepsilon$, it would increase at speed at least $m_\varepsilon/2$, while if the solution stays above $\varrho_* + \varepsilon$, it would decrease at speed at least $m_\varepsilon/2$. Moreover, once the solution has entered I_ε after time T_ε , it cannot leave: at the left endpoint $\varrho_* - \varepsilon$, whenever this point belongs to $[a, b]$, the derivative is strictly positive, and at the right endpoint $\varrho_* + \varepsilon$, whenever this point belongs to $[a, b]$, the derivative is strictly negative. Hence there exists $\tilde{T}_\varepsilon \geq T_\varepsilon$ such that

$$|\varrho(t) - \varrho_*| \leq \varepsilon \quad \text{for all } t \geq \tilde{T}_\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that

$$\lim_{t \rightarrow \infty} \varrho(t) = \varrho_*.$$

□

With the above lemma, we are ready to discuss the asymptotic behavior of ρ -dynamics (C.26) in two parameter regimes.

Claim 1. If $v + R \geq 0$, the solution $\rho(t)$ of (C.26) satisfies $\lim_{t \rightarrow \infty} \rho(t) = 1$.

Proof of Claim 1. We focus on the non-trivial case $\rho(0) < 1$. For any $\rho \in [0, 1)$, our assumption implies that $v\rho + R > 0$, and $G(\rho) > 0$, as G defined in (C.25). Then the ρ -dynamics in (C.26) satisfies the assumptions in Lemma C.2 with choosing $a = 0$, $b = 1$, and $\varrho_* = 1$, and thus we conclude the convergence $\lim_{t \rightarrow \infty} \rho(t) = 1$. □

Here, we record an additional quantitative estimate in the sub-regime $v + R > 0$, which will be used in the proof of Theorem 4.8. We defer the proof until after the proof of Theorem 3.3.

Proposition C.3. Assume that $R > 0$ and $v + R > 0$. Let $(\rho(t), \psi(t))$ be the solution to (C.21) with initial data $\rho(0) \in (0, 1]$ and $\psi(0) \in \mathbb{T}^1 \setminus \{\pi\}$. Then there exist constant $0 < \epsilon_* \leq \frac{1}{2\pi}$ and $C > 0$, depending only on V , such that for every $0 < \epsilon \leq \epsilon_*$, if

$$T_\epsilon := \inf \{t \geq 0 : \rho(t) \geq 1 - \epsilon, |\psi(t)| \leq \epsilon\}, \quad (\text{C.27})$$

then

$$T_\epsilon \leq C \log \frac{1}{\epsilon |\psi(0) - \pi|}.$$

Here $\psi(0)$ is represented in $[0, 2\pi)$.

Claim 2. If $v + R < 0$ and $\rho(0) \neq 1$, the solution $\rho(t)$ of (C.26) satisfies $\lim_{t \rightarrow \infty} \rho(t) = \rho_{\text{eq}}$, with ρ_{eq} defined in (C.22).

Proof of Claim 2. Since $v + R < 0$, we have

$$v\rho + R \begin{cases} > 0, & \rho < \rho_{\text{eq}}, \\ = 0, & \rho = \rho_{\text{eq}}, \\ < 0, & \rho > \rho_{\text{eq}}, \end{cases} \quad \text{and therefore } G(\rho) \begin{cases} > 0, & \rho < \rho_{\text{eq}}, \\ = 0, & \rho = \rho_{\text{eq}}, \\ < 0, & \rho > \rho_{\text{eq}}, \end{cases}.$$

Then the ρ -dynamics in (C.26) satisfies the assumptions in Lemma C.2 with choosing $a = 0$, $b = 1$, and $\varrho_* = \rho_{\text{eq}}$, and thus we conclude the convergence $\lim_{t \rightarrow \infty} \rho(t) = \rho_{\text{eq}}$. $\square$

Finally, we consider the boundary case $R = 0$. In this regime, we restrict our attention to $v \neq 0$ since $v = 0$ and $R = 0$ corresponds to the trivial case where the matrix V is the zero matrix.

When $R = 0$, the dynamics (C.21) simplifies to

$$\dot{\rho} = \frac{1}{4}(1 - \rho^2)v\rho, \quad \dot{\psi} = 0.$$

The angular motion ψ is thus a constant over time. For the ρ -equation, we can easily identify the two fixed points: $\rho = 0$ and $\rho = 1$. The long-time behavior is also straightforward. If $v > 0$, Lemma C.2, applied with $\varrho(t) = \rho(t)$, $\varrho_* = 1$, and $e(t) = 0$, gives that for any initial condition $\rho(0) \in (0, 1]$, $\lim_{t \rightarrow \infty} \rho(t) = 1$. If $v < 0$, we set $\varrho(t) = -\rho(t)$, $\varrho_* = 0$, and it shows that for any initial condition $\rho(0) \in [0, 1)$, $\lim_{t \rightarrow \infty} \rho(t) = 0$.

A note on the measure-zero initialization set $\{(\rho(0), \psi(0)) : \psi(0) = \pi\}$. Since $\psi = \pi$ is a fixed point of the ψ -dynamics in (C.21), we have $\psi(t) = \psi(0)$ for all $t \geq 0$. In this case, the ρ -equation reduces to

$$\dot{\rho} = \frac{1}{4}(1 - \rho^2)(v\rho - R).$$

We note that, in both parameter regimes $v + R \geq 0$ and $v + R < 0$, the limiting value of $\rho(t)$ may depend on the initial value $\rho(0)$. Since this exceptional set has measure zero and its initial-condition-dependent behavior is not the main focus of this paper, we omit the detailed discussion.

Conclusion for the order parameter $\mathcal{R}_2$. Combining the preceding arguments with the relation $\psi(t) = \phi(t) - \phi_V$, we obtain the following convergence results.

- If $v + R \geq 0$, then for any initial condition satisfying $\psi(0) \neq \pi$, the solution (ρ, ϕ) of (3.10) converges to $(1, \phi_V)$.
- If $v + R < 0$, then for any initial condition satisfying $\rho(0) \neq 1$ and $\psi(0) \neq \pi$, the solution (ρ, ϕ) of (3.10) converges to $(\rho_{\text{eq}}, \phi_V)$, where ρ_{eq} is defined in (C.22).

It remains only to translate these conditions into the parameter regimes stated in Theorem 3.3. This follows from the definitions $v = v_+^{\text{d}}$, $R = \sqrt{(v_-^{\text{d}})^2 + (v_+^{\text{o}})^2}$, together with elementary algebraic identities. For the reader's convenience, we also record the expression

$$\rho_{\text{eq}} := -\frac{\sqrt{(v_-^{\text{d}})^2 + (v_+^{\text{o}})^2}}{v_+^{\text{d}}}. \quad (\text{C.28})$$

$\square$

Proof of Proposition C.3. For notation simplicity, we use the notation introduced in the proof of Theorem 3.3; see the shorthand notation in (C.20), and the dynamics for $(\rho(t), \psi(t))$ in (C.21).

From the ψ -dynamics, we know that if $\psi(0) = 0$, then $\psi(t) = 0$ for all $t \geq 0$, satisfying $|\psi(t)| \leq \epsilon$. Next, we consider the initial data $\psi(0) \neq \{0, \pi\}$. By the inequalities (C.24) and $|\tan \psi| \leq \frac{1}{|\psi - \pi/2|}$ for $\psi \in [0, \pi]$, we have

$$|\psi(t)| \leq 2|\tan(\psi(0)/2)| \exp\left(-\frac{R}{4}t\right) \leq \frac{4}{|\psi(0) - \pi|} \exp\left(-\frac{R}{4}t\right).$$

By letting the right-hand side of the above inequality less or equal to ϵ , we get

$$t \geq \frac{4}{R} \log \frac{4}{\epsilon |\psi(0) - \pi|} =: T_{\psi, \epsilon}.$$

Next, we turn to the ρ -dynamics in (C.21). Define

$$\chi := \min_{\rho \in [0, 1]} (v\rho + R) = \min \{R, v + R\}.$$

Since $R > 0$ and $v + R > 0$, we have $\chi > 0$. Choose $0 < \epsilon_* \leq 1/(2\pi)$, depending only on the matrix V , such that $R(1 - \cos \epsilon_*) \leq \frac{\chi}{2}$. Then for every $0 < \epsilon \leq \epsilon_*$, whenever $|\psi| \leq \epsilon$, we have

$$v\rho + R \cos \psi = (v\rho + R) - R(1 - \cos \psi) \geq \frac{\chi}{2}.$$

Consequently, after time $T_{\psi, \epsilon}$, the ρ -dynamics satisfies

$$\dot{\rho} = \frac{1}{4}(1 - \rho^2)(v\rho + R \cos \psi) \geq \frac{\chi}{8}(1 - \rho^2) \geq \frac{\chi}{8}(1 - \rho).$$

Equivalently, $q(t) := 1 - \rho(t)$ satisfies $\dot{q}(t) \leq -\frac{\chi}{8}q(t)$ for $t \geq T_{\psi, \epsilon}$. Integrating it over time gives

$$q(t) \leq q(T_{\psi, \epsilon}) \exp\left(-\frac{\chi}{8}(t - T_{\psi, \epsilon})\right) \leq \exp\left(-\frac{\chi}{8}(t - T_{\psi, \epsilon})\right).$$

Hence $\rho(t) \geq 1 - \epsilon$ if

$$t \geq T_{\psi, \epsilon} + \frac{8}{\chi} \log \frac{1}{\epsilon}.$$

Since $|\psi(t)|$ continues to decrease after $T_{\psi, \epsilon}$, we also have $|\psi(t)| \leq \epsilon$. Therefore,

$$T_\epsilon = T_{\psi, \epsilon} + \frac{8}{\chi} \log \frac{1}{\epsilon} \leq \frac{4}{R} \log \frac{4}{\epsilon |\psi(0) - \pi|} + \frac{8}{\chi} \log \frac{1}{\epsilon} \leq C \log \frac{1}{\epsilon |\psi(0) - \pi|},$$

where constant $C > 0$ only depends on the matrix V . □

Case 3.

Proof of Theorem 3.6. We start from the (ρ, ϕ) -dynamics (3.1) associated with arbitrary matrices A and V defined in (2.2). By Lemma C.7, this system admits the equivalent representation

$$\dot{\rho} = 2(1 - \rho^2)(S_1(\phi)\rho + S_2(\phi)), \quad \dot{\phi} = 2\left(S_3(\phi)(\rho^2 + 1) + S_4(\phi)\frac{\rho^2 + 1}{\rho} + 2S_5(\phi)\rho + \operatorname{Re}(b_5)\right), \quad (\text{C.29})$$

where the constant b_5 and the functions S_j are defined in (A.3) and (C.54), respectively.

Our goal is to identify a family of matrices A and V , together with explicit parameter regimes, such that the corresponding solution of (C.29) satisfies, as $t \rightarrow \infty$,

$$\rho(t) \rightarrow 1 \quad \text{while} \quad \phi(t) \text{ does not converge}.$$

ρ -dynamics. We first analyze the ρ -dynamics in (C.29). Since $\rho \in (0, 1)$, we obtain the following sufficient condition for $\rho(t) \rightarrow 1$:

$$S_1(\phi) > 0, \quad \text{and} \quad S_2(\phi) \geq 0, \quad \text{for all } \phi \in \mathbb{T}^1. \quad (\text{C.30})$$

Indeed, under this condition,

$$\dot{\rho} = 2(1 - \rho^2)(S_1(\phi)\rho + S_2(\phi)) \geq 0$$

with equality only holding at $\rho = 1$. Therefore, by Lemma C.8, $\rho(t) \rightarrow 1$ as $t \rightarrow \infty$ for every $\rho(0) \in (0, 1]$.

We next translate the condition (C.30) into explicit constraints on the matrices A and V . From (C.54), we recall

$$S_1(\phi) = \text{Im}(b_1 e^{i2\phi} + b_2), \quad S_2(\phi) = \text{Im}(b_3 e^{i\phi}),$$

where b_1, b_2 , and b_3 are defined in (A.3). Hence, the conditions $S_1(\phi) > 0$ and $S_2(\phi) = 0$ for all $\phi \in [0, 2\pi)$ is equivalent to

$$\text{Im}(b_2) > |b_1| \quad \text{and} \quad b_3 = 0.$$

Using the definitions of b_1, b_2 , and b_3 , these conditions can be rewritten in terms of the coefficients w_i introduced in (A.2) as follows:

- (i). $w_1 = w_2 = 0$, and
- (ii). $w_5 + w_8 > \sqrt{(w_6 + w_7)^2 + (w_5 - w_8)^2}$.

Since translating these conditions into explicit constraints on general matrices A and V is algebraically involved, in the following we restrict attention to the important special case where A is diagonal:

$$A := \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}, \quad a_{11}, a_{22} \neq 0.$$

We now characterize the family of matrices V satisfying the above two conditions in this setting.

Under this assumption on A , condition (i) becomes

$$w_1 = a_{11}v_{11} - a_{22}v_{22}, \quad w_2 = -(a_{22}v_{12} + a_{11}v_{21}).$$

Hence, $w_1 = w_2 = 0$ is equivalent to

$$v_{22} = \frac{a_{11}}{a_{22}}v_{11}, \quad v_{21} = -\frac{a_{22}}{a_{11}}v_{12}.$$

Accordingly, the corresponding admissible family of matrices V can be written as

$$V := \begin{pmatrix} v_{11} & v_{12} \\ -\frac{a_{22}}{a_{11}}v_{12} & \frac{a_{11}}{a_{22}}v_{11} \end{pmatrix}, \quad v_{11}, v_{12} \in \mathbb{R}.$$

We next consider condition (ii). For the matrices A and V defined above, one can verify that

$$w_5 + w_8 > \sqrt{(w_6 + w_7)^2 + (w_5 - w_8)^2} \quad \Longleftrightarrow \quad \frac{v_{11}}{a_{22}} > 0, \quad 8v_{11}^2(a_{11}^2 + a_{22}^2)\frac{a_{11}}{a_{22}} > (a_{11} - a_{22})^4 \frac{v_{12}^2}{a_{11}^2}.$$

Therefore, for the family of matrices

$$A := \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}, \quad V := \begin{pmatrix} v_{11} & v_{12} \\ -\frac{a_{22}}{a_{11}}v_{12} & \frac{a_{11}}{a_{22}}v_{11} \end{pmatrix}, \quad (\text{C.31})$$

with the entries $a_{11}, a_{22} \neq 0, v_{11}, v_{12} \in \mathbb{R}$ satisfying

$$\frac{v_{11}}{a_{22}} > 0, \quad 8\frac{a_{11}}{a_{22}}(a_{11}^2 + a_{22}^2)v_{11}^2 > \frac{(a_{11} - a_{22})^4}{a_{11}^2}v_{12}^2, \quad (\text{C.32})$$

the ρ -equation satisfies $\rho(t) \rightarrow 1$ as $t \rightarrow \infty$.

ϕ -dynamics. We now turn to the ϕ -dynamics in (C.33). For the matrices A and V above, the system simplifies to

$$\dot{\rho} = \rho(1 - \rho^2)H(\phi), \quad \dot{\phi} = K(\phi) + (\rho - 1)Q(\rho, \phi). \quad (\text{C.33})$$

Here, the functions H , K , and Q are given by

$$H(\phi) := \frac{v_{11}}{a_{22}}(a_+^d)^2 + (a_-^d)^2 \left(-\frac{v_{11}}{a_{22}} \cos(2\phi) + \frac{v_{12}}{a_{11}} \sin(2\phi) \right), \quad (\text{C.34})$$

$$K(\phi) := \frac{1}{2} (a_+^d + a_-^d \cos \phi) P(\phi), \quad (\text{C.35})$$

where

$$P(\phi) := \left(\frac{v_{11}}{a_{22}} a_-^d \sin \phi + \frac{v_{12}}{a_{11}} a_-^d \cos \phi - \frac{v_{12}}{a_{11}} a_+^d \right), \quad (\text{C.36})$$

and

$$Q(\rho, \phi) := \frac{1}{2} \frac{v_{11}}{a_{22}} a_+^d a_-^d \sin \phi + \frac{1}{8} (\rho + 1) \left(-\frac{v_{12}}{a_{11}} (a_+^d)^2 + \frac{v_{12}}{a_{11}} (a_-^d)^2 \cos(2\phi) + \frac{v_{11}}{a_{22}} (a_-^d)^2 \sin(2\phi) \right). \quad (\text{C.37})$$

To show that $\phi(t)$ does not converge, it suffices to prove that there exists $T \geq 0$ such that

$$|\dot{\phi}(t)| > 0, \quad \forall t \geq T.$$

We now identify an additional sufficient condition which, together with the previous condition on A and V , guarantees this property.

Although the expressions for K and Q are somewhat involved, the key observation is simple. Under the previous condition (C.32), we already know that $\rho(t) \rightarrow 1$ as $t \rightarrow \infty$. Moreover, since $\rho \in [0, 1]$ and $a_{11}, a_{22} \neq 0$, there exists a constant $C < \infty$ such that

$$|Q(\rho, \phi)| \leq C, \quad \forall (\rho, \phi) \in [0, 1] \times \mathbb{T}^1.$$

Therefore, for every $\varepsilon > 0$, there exists $T_\varepsilon < \infty$ such that

$$|(\rho - 1)Q(\rho, \phi)| \leq \varepsilon, \quad \forall t \geq T_\varepsilon.$$

Thus, in order to ensure $|\dot{\phi}| > 0$ for all sufficiently large t , it remains to impose a condition guaranteeing that $|K(\phi)| > 0$ for all $\phi \in \mathbb{T}^1$.

Note that the previous condition (C.32) implies that a_{11} and a_{22} have the same sign. Hence,

$$\begin{aligned} |a_+^d + a_-^d \cos \phi| &= |(a_{11} + a_{22}) + (a_{11} - a_{22}) \cos \phi| = \left| 2 \left(a_{11} \cos^2 \frac{\phi}{2} + a_{22} \sin^2 \frac{\phi}{2} \right) \right| \\ &\geq 2 \min \{|a_{11}|, |a_{22}|\} =: c_0 > 0, \quad \forall \phi \in \mathbb{T}^1. \end{aligned}$$

Therefore, $|K(\phi)| > 0$ for all $\phi \in \mathbb{T}^1$ if and only if $|P(\phi)| > 0$ for all $\phi \in \mathbb{T}^1$. From the definition of P , a sufficient condition for $|P(\phi)| > 0$ for all $\phi \in \mathbb{T}^1$ is

$$\left| \frac{v_{12}}{a_{11}} a_+^d \right| > |a_-^d| \sqrt{\left(\frac{v_{11}}{a_{22}} \right)^2 + \left(\frac{v_{12}}{a_{11}} \right)^2}. \quad (\text{C.38})$$

Indeed, under this condition,

$$|P(\phi)| = \left| \frac{v_{11}}{a_{22}} a_-^d \sin \phi + \frac{v_{12}}{a_{11}} a_-^d \cos \phi - \frac{v_{12}}{a_{11}} a_+^d \right| \geq \left| \frac{v_{12}}{a_{11}} a_+^d \right| - |a_-^d| \sqrt{\left(\frac{v_{11}}{a_{22}} \right)^2 + \left(\frac{v_{12}}{a_{11}} \right)^2} =: c_1 > 0.$$

It follows that

$$|K(\phi)| \geq \frac{1}{2} c_0 c_1 > 0, \quad \forall \phi \in \mathbb{T}^1.$$

Combining the sufficient conditions above, we conclude that $\rho(t) \rightarrow 1$ as $t \rightarrow \infty$, while $\phi(t)$ cannot converge. This completes the proof. $\square$

Case 4.

Proof of Theorem 3.9. For convenience we relabel $v_{11} = a, v_{12} = b$.

Step 1: Characterization of the fixed points. By letting $\dot{\rho} = \dot{\phi} = 0$, we obtain the following fixed points of the dynamics (3.18):

- If $a \geq b$, there are two fixed points:

$$E_1 := \{(\rho, \phi) = (1, \pi + \arcsin(b/a))\}, \quad E_2 := \{(\rho, \phi) = (1, 2\pi - \arcsin(b/a))\}. \quad (\text{C.39})$$

When $a = b$, these two fixed points coincide, namely $E_1 = E_2 = \{(\rho, \phi) = (1, 3\pi/2)\}$.

- If $a < b$, there is a unique fixed point:

$$E_3 := \left\{(\rho, \phi) = \left(\tanh\left(\frac{1}{3} \operatorname{arctanh}\left(\frac{a}{b}\right)\right), \frac{3\pi}{2}\right)\right\}. \quad (\text{C.40})$$

We now provide the detailed derivation. First, consider the boundary case $\rho = 1$. In this case, we automatically have $\dot{\rho} = 0$, while $\dot{\phi} = 0$ if and only if $b + a \sin(\phi) = 0$. Since $a, b > 0$, this equation has no solutions when $a < b$, has the unique solution at $\phi = \frac{3\pi}{2}$ when $a = b$, and has two solutions when $a > b$. In the latter case, the two fixed points are E_1 and E_2 as defined in (C.39).

Next, consider the interior case $0 < \rho < 1$. From $\dot{\rho} = 0$, we obtain $\phi \in \{\frac{\pi}{2}, \frac{3\pi}{2}\}$. Imposing $\dot{\phi} = 0$ then gives the following constraints on ρ :

$$\begin{cases} f_1(\rho) := -b\rho^3 - 3a\rho^2 - 3b\rho - a = 0, & \text{if } \phi = \frac{\pi}{2}, \\ f_2(\rho) := -b\rho^3 + 3a\rho^2 - 3b\rho + a = 0, & \text{if } \phi = \frac{3\pi}{2}. \end{cases}$$

The equation $f_1(\rho) = 0$ has no solution in $[0, 1]$ for any $a, b > 0$, since $f_1(0) = -a$ and $f'_1(\rho) < 0$ for $\rho \in [0, 1]$.

It remains to analyze $f_2(\rho) = 0$. Rewriting this equation gives

$$\frac{a}{b} = \frac{\rho^3 + 3\rho}{3\rho^2 + 1} =: q(\rho).$$

The function q is monotone increasing on $[0, 1]$, since $q'(\rho) = \frac{3(1-\rho^2)^2}{(3\rho^2+1)^2} \geq 0$, and it satisfies $q(0) = 0$ and $q(1) = 1$. Therefore, if $a > b$, then $a/b > 1$ and there is no solution in $[0, 1]$. If $a = b$, the unique solution is $\rho = 1$. If $a < b$, then $a/b \in (0, 1)$, and hence there is a unique solution $\rho \in (0, 1)$. Using the identity

$$\tanh(3u) = \frac{\tanh^3 u + 3 \tanh u}{3 \tanh^2 u + 1},$$

this solution is given explicitly as in (C.40).

Next, we perform the linear stability analysis on these fixed points. The Jacobian matrix (3.18) is

$$\mathcal{J} = \begin{bmatrix} -2a\rho \cos(\phi) & -a(1-\rho^2) \sin(\phi) \\ -2b\rho - 3a \sin(\phi) + \frac{a}{\rho^2} \sin(\phi) & -3a\rho \cos(\phi) - \frac{a}{\rho} \cos(\phi) \end{bmatrix}. \quad (\text{C.41})$$

When $a > b$, evaluating $\mathcal{J}$ at the two boundary fixed points shows that both E_1 and E_2 are hyperbolic. By the Hartman–Grobman Theorem, E_1 is unstable, while E_2 is asymptotically stable. When $a = b$, the unique equilibrium is $E_3 = (1, 3\pi/2)$. In this case, evaluating the Jacobian gives $\mathcal{J} = 0$, so E_3 is a degenerate equilibrium. We postpone the characterization of its behavior to the energy analysis below. Finally, when $a < b$, the two eigenvalues of $\mathcal{J}$ evaluated on E_3 are pure imaginary, and thus E_3 is a linear center. In the next step, we perform an energy analysis and show that, in this case, E_3 is in fact a stable nonlinear center.

Step 2: Energy Analysis.

$$H(\rho, \phi) = \frac{-a\rho \sin(\phi) - \frac{b}{2}(\rho^2 + 1)}{(1 - \rho^2)^2}. \quad (\text{C.42})$$

We first show that H is a first integral for (3.18).

Lemma C.4. *If (ρ, ϕ) follow the dynamics (3.18) and $0 < \rho < 1$, then $\dot{H}(\rho, \phi) = 0$.*

Proof. We can first compute that

$$\nabla H(\rho, \phi) = \begin{pmatrix} \frac{\partial H}{\partial \rho} \\ \frac{\partial H}{\partial \phi} \end{pmatrix} = \begin{pmatrix} -\frac{a(1+3\rho^2)\sin\phi + b\rho(\rho^2+3)}{(1-\rho^2)^3} \\ -\frac{a\rho\cos\phi}{(1-\rho^2)^2} \end{pmatrix}. \quad (\text{C.43})$$

Denote $\mu(\rho) = \frac{2\rho}{(1-\rho^2)^3}$ and $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, then we have

$$\begin{pmatrix} \dot{\rho} \\ \dot{\phi} \end{pmatrix} = \frac{1}{\mu(\rho)} J \nabla H. \quad (\text{C.44})$$

Therefore,

$$\dot{H} = \nabla H \cdot \begin{pmatrix} \dot{\rho} \\ \dot{\phi} \end{pmatrix} = \frac{1}{\mu(\rho)} \nabla H \cdot J \nabla H = 0,$$

where the last equality follows from the skew-symmetry of J . This proves that H is conserved along trajectories in the interior region $0 < \rho < 1$. $\square$

We now analyze the level set of the Hamiltonian H , since every trajectory $(\rho(t), \phi(t)) \in (0, 1) \times \mathbb{T}^1$ evolves along a level set of H . For notation simplicity, we write $\{H(\rho, \phi) = c\}$, or simply $\{H = c\}$, to denote the level set of H at value $c \in \mathbb{R}$ within the open cylinder $(0, 1) \times \mathbb{T}^1$.

In the following, we first show in Lemma C.5 that for almost every $c \in \mathbb{R}$, the level set $\{H = c\}$ is a *connected* one-dimensional manifold whenever it is nonempty. Building upon this geometric property, we discuss in two regimes.

- In the parameter regime $a < b$, we show that $\{H = c\}$, whenever nonempty, is a *compact* subset of the open cylinder $(0, 1) \times \mathbb{T}^1$; see Figure 21a for a visual illustration.

Consequently, any trajectory $(\rho(t), \phi(t))$ of the dynamics (3.18) that evolves along such a level set forms a closed periodic orbit.

- In the parameter regime $a \geq b$, we show that $\{H = c\}$, whenever nonempty, has closure intersecting the boundary $\{\rho = 1\}$; see Figures 21b and 21c for a visual illustration. In particular, such a level set is *not compact* as a subset of the open cylinder $(0, 1) \times \mathbb{T}^1$.

Consequently, no trajectory $(\rho(t), \phi(t))$ of the dynamics (3.18) evolving along such a level set can form a nontrivial periodic orbit in the open cylinder $(0, 1) \times \mathbb{T}^1$. Moreover, Step 1 shows that the closed cylinder $[0, 1] \times \mathbb{T}^1$ is a compact positive invariant set for the dynamics, and it contains an unstable fixed point E_1 and an asymptotically stable fixed point E_2 . Then the Poincaré–Bendixson theorem [107] implies that, for almost every initial condition $(\rho(0), \phi(0)) \in [0, 1] \times \mathbb{T}^1$, the solution converges to E_2 .

We introduce some notation that will be used throughout the analysis and record a useful geometric observation. Define

$$W(\rho) = W(\rho; a, b, c) := -\frac{c(1-\rho^2)^2 + \frac{b}{2}(\rho^2+1)}{a\rho}. \quad (\text{C.45})$$

Then the level set of H at value $c \in \mathbb{R}$ can be written as

$$\{H(\rho, \phi) = c\} = \{(\rho, \phi) \in (0, 1) \times \mathbb{T}^1 : \sin\phi = W(\rho)\}. \quad (\text{C.46})$$

Define

$$I := \{\rho \in (0, 1) : |W(\rho)| \leq 1\}. \quad (\text{C.47})$$

Then the level set can be decomposed into two branches,

$$\{H(\rho, \phi) = c\} = \{(\rho, \arcsin(W(\rho))) : \rho \in I\} \cup \{(\rho, \pi - \arcsin(W(\rho))) : \rho \in I\} =: \Gamma_1 \cup \Gamma_2. \quad (\text{C.48})$$

There are two special level sets that we exclude from the analysis below. First, for any $a, b > 0$, there is a unique level set whose closure intersects the boundary $\{\rho = 0\}$, that is, $\{H = -b/2\}$; see the dashed orange

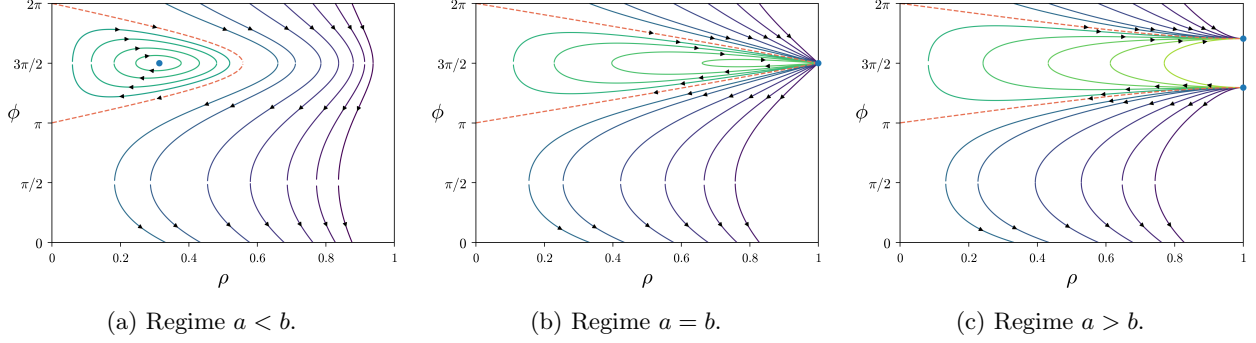


Figure 21: Level sets of the Hamiltonian $H(\rho, \phi)$ for different parameter regimes. The dashed orange curve indicates the special level set $\{H = -b/2\}$, while the colored solid curves represent other level sets. The blue dots denote the fixed points of the dynamics (3.18), introduced in (C.39) and (C.40). The arrows indicate the direction of the flow.

curve in Figure 21. Second, when $a < b$, Step 1 shows that the dynamics (3.18) has a unique fixed point E_3 , as defined in (C.40). Together with the relation (C.44), this implies that $\nabla H(\rho, \phi) = 0$ only at E_3 . We denote the corresponding critical value as c^* and the associated level set by $\{H = c^*\}$.

The long-time behavior of (3.18) initialized on either of these exceptional level sets depends sensitively on the precise initial position and is not the focus of the present analysis. We therefore omit a detailed study of trajectories lying on these two level sets.

Finally, we record a geometric property of the remaining level sets in the following lemma.

Lemma C.5. *For any $a, b > 0$, and $c \in \mathbb{R} \setminus \{-b/2, c^*\}$, the level set $\{H(\rho, \phi) = c\}$ is either empty or a connected one-dimensional manifold.*

Proof. Fix $c \in \mathbb{R} \setminus \{-b/2, c^*\}$. Recall the decomposition of the level set $\{H = c\}$ into the two branches Γ_1 and Γ_2 over the set I defined in (C.47) and (C.48). If $I = \emptyset$, then both Γ_1 and Γ_2 are empty, and therefore $\{H = c\}$ is empty. Next, we focus on $I \neq \emptyset$. In this case, we will show that both Γ_1 and Γ_2 are connected, and $\Gamma_1 \cap \Gamma_2 \neq \emptyset$, which implies that $\{H = c\}$ is connected. Moreover, since $\nabla H(\rho, \phi) \neq 0$ for every $(\rho, \phi) \in \{H = c\}$, therefore by the regular level set theorem, $\{H = c\}$ is a connected one-dimensional manifold.

We first justify that both Γ_1 and Γ_2 are connected. Since W defined in (C.45) is continuous, it is enough to show that I is connected. A direct calculation gives

$$W'(\rho) = \frac{(1 - \rho^2)(b + 2c(1 + 3\rho^2))}{2a\rho^2}.$$

Since $0 < \rho < 1$, the sign of $W'(\rho)$ is determined by the sign of

$$w(\rho) := b + 2c(1 + 3\rho^2). \quad (\text{C.49})$$

The function w is affine in ρ^2 , and therefore has at most one zero in $(0, 1)$. Hence W is either strictly monotone, or it has exactly one critical point. We now split the argument into two cases.

First, suppose $c \leq -\frac{b}{2}$. Then, for every $\rho \in (0, 1)$, $w(\rho) = b + 2c(1 + 3\rho^2) < 0$. Therefore, W is strictly decreasing on $(0, 1)$. Since I is the preimage of the interval $[-1, 1]$ under a strictly monotone continuous function and is nonempty, then I is an interval and therefore connected.

Next, suppose $c > -\frac{b}{2}$. If W is monotone, then again I is an interval. Otherwise, W has a unique critical point, denoted as $\rho_* \in (0, 1)$. At this point,

$$w(\rho_*) = b + 2c(1 + 3\rho_*^2) = 0.$$

Substituting this relation into the definition of W , we obtain

$$W(\rho_*) = -\frac{b\rho_*(3 + \rho_*^2)}{a(1 + 3\rho_*^2)} < 0.$$

Since ρ_* is the unique maximum point of W , then we know $W(\rho) \leq W(\rho_*) < 0$ for all $\rho \in (0, 1)$. Consequently,

$$I = \{\rho \in (0, 1) : |W(\rho)| \leq 1\} = \{\rho \in (0, 1) : W(\rho) \geq -1\}.$$

Together with W increasing first and then decreasing, this implies the superlevel set $\{W \geq -1\}$ is an interval. Therefore, in all cases, I is connected, and consequently both Γ_1 and Γ_2 are connected.

It remains to show that $\Gamma_1 \cap \Gamma_2 \neq \emptyset$. By the definitions of I in (C.47) and the branch decomposition in (C.48), it suffices to show that there exists $\kappa \in (0, 1)$ such that $|W(\kappa)| = 1$. We split the argument into two regimes.

First, consider the regime $a < b$. Let $\kappa_R := \sup I$. Since W extends continuously to $\rho = 1$ and $W(1) = -b/a < -1$, there exists a neighborhood of $\rho = 1$ on which $W < -1$. Together with $I \neq \emptyset$, this implies that $\kappa_R < 1$. Moreover, because $c \neq -b/2$, we have $|W(\rho)| \rightarrow \infty$ as $\rho \rightarrow 0^+$, so I is bounded away from $\rho = 0$. Hence $\kappa_R \in (0, 1)$. By the continuity of W and the definition of I , we get $\kappa_R \in I$ and $W(\kappa_R) = -1$. Therefore, the two branches coincide at the point $(\kappa_R, 3\pi/2)$, and hence $\Gamma_1 \cap \Gamma_2 \neq \emptyset$.

Next, consider the regime $a \geq b$. Let $\kappa_L := \inf I$. Since $c \neq -b/2$, we have

$$\lim_{\rho \rightarrow 0^+} W(\rho) = -\infty \quad \text{if } c > -b/2, \quad \lim_{\rho \rightarrow 0^+} W(\rho) = +\infty \quad \text{if } c < -b/2.$$

Thus I is bounded away from $\rho = 0$, and so $\kappa_L > 0$. Moreover, we have $W(1) = -b/a \geq -1$. Then by the continuity of W and the definition of I , we know that $\kappa_L \in I$ and $|W(\kappa_L)| = 1$. More precisely,

$$W(\kappa_L) = -1 \quad \text{if } c > -b/2, \quad W(\kappa_L) = 1 \quad \text{if } c < -b/2.$$

Consequently, the two branches coincide at $(\kappa_L, 3\pi/2)$ when $c > -b/2$, and at $(\kappa_L, \pi/2)$ when $c < -b/2$. This proves that $\Gamma_1 \cap \Gamma_2 \neq \emptyset$, completing the proof. $\square$

Now, we are ready to study the structure of the level sets of H based on the relation between the parameters a and b .

Regime $a < b$. Fix $c \in \mathbb{R} \setminus \{-b/2, c^*\}$. By Lemma C.5, the level set $\{H = c\}$ is a connected one-dimensional manifold whenever it is nonempty. Thus it remains to show that $\{H = c\}$ is compact as a subset of the open cylinder $(0, 1) \times \mathbb{T}^1$.

To this end, we first note that, by the definition of H in (C.42), for every $\phi \in \mathbb{T}^1$, we have $H(\rho, \phi) \rightarrow -\infty$ as $\rho \rightarrow 1^-$. Together with the continuity of H , this implies that, for any finite $c \in \mathbb{R}$, the level set $\{H = c\}$ stays bounded away from the boundary $\{\rho = 1\}$. Combining this with the fact that $\{H = -b/2\}$ is the only level set whose closure intersects with the boundary $\{\rho = 0\}$, we see that, for every $c \neq -b/2$, there exist $0 < \rho_-^c < \rho_+^c < 1$ such that $\{H = c\} \subset K_c$ with $K_c := [\rho_-^c, \rho_+^c] \times \mathbb{T}^1$. Since H is continuous, $\{H = c\}$ is closed in K_c . Because K_c is compact, it follows that $\{H = c\}$ is also compact. We thus conclude that, for every $c \in \mathbb{R} \setminus \{-b/2, c^*\}$, the level set $\{H = c\}$ is a compact, connected one-dimensional manifold whenever it is nonempty.

Remark C.6. We record a qualitative consequence of the above analysis in the regime $a < b$. Since distinct level sets are disjoint, and since no level set other than $\{H = -b/2\}$ can approach the boundary $\{\rho = 0\}$, the exceptional curve $\{H = -b/2\}$ separates the cylinder $(0, 1) \times \mathbb{T}^1$ into two regions. The region $\{H > -b/2\}$ consists of closed loops on the cylinder, corresponding to libration in the dynamics (3.18). By contrast, the region $\{H < -b/2\}$ consists of loops that wrap around the cylinder, corresponding to rotation. This is also illustrated in Figure 21a.

Regime $a \geq b$. Fix $c \in \mathbb{R} \setminus \{-b/2, c^*\}$. By Lemma C.5, the level set $\{H = c\}$, whenever nonempty, is a connected one-dimensional manifold. It remains to show that the closure of $\{H = c\}$ intersects the boundary $\{\rho = 1\}$. In fact, the interaction points are precisely the boundary fixed points E_1 and E_2 defined in (C.39); see Figures 21b and 21c.

Recall the definition of set I in (C.47) and the branch decomposition of $\{H = c\}$ in (C.48). Together with the continuity of function W (see (C.45)), it suffices to show that the closure of I reaches $\rho = 1$ whenever $I \neq \emptyset$.

If $a > b$, this is immediate. Indeed, W is continuous, and $W(1) = -b/a > -1$. Hence $|W(\rho)| \leq 1$ for all ρ sufficiently close to 1, and therefore the closure of I reaches $\rho = 1$.

It remains to consider the borderline case $a = b$. In this case, $W(1) = -1$. Moreover, a direct calculation gives

$$W(\rho) + 1 = -\frac{(1-\rho)^2}{a\rho} \left(\frac{a}{2} + c(1+\rho)^2 \right). \quad (\text{C.50})$$

Since $I \neq \emptyset$, there exists some $\rho \in (0, 1)$ such that $|W(\rho)| \leq 1$ and hence $W(\rho) + 1 \geq 0$. By (C.50), $W(\rho) + 1 \geq 0$ can occur only if $a/2 + c(1+\rho)^2 \leq 0$ for some $\rho \in (0, 1)$, which forces $c < -a/8$. Consequently, for all ρ sufficiently close to 1, we have $a/2 + c(1+\rho)^2 < 0$ and therefore $W(\rho) + 1 > 0$. Together with $W(1) = -1$ and the definition of I , this implies that the closure of I reaches $\rho = 1$, completing the proof. $\square$

C.2 Auxiliary Lemmas

Lemma C.7. *The polar-coordinate system (3.1) admits the equivalent representation*

$$\dot{\rho} = 2(1 - \rho^2) (S_1(\phi)\rho + S_2(\phi)), \quad \dot{\phi} = 2 \left(S_3(\phi)(\rho^2 + 1) + S_4(\phi)\frac{\rho^2 + 1}{\rho} + 2S_5(\phi)\rho + \text{Re}(b_5) \right),$$

where both functions S_j constant b_5 depend on the matrices A and V . The explicit form of S_j and b_5 are provided in (C.54) and (A.3), respectively.

Proof. For notational convenience we drop the time dependence on all variables.

Substituting $\mathcal{R}_2 = \rho e^{i\phi}$ and $\overline{\mathcal{R}_2} = \rho e^{-i\phi}$ into the ODE (2.31) satisfied by $\mathcal{R}_2$ gives

$$\frac{d}{dt}\mathcal{R}_2 = \frac{d}{dt}(\rho e^{i\phi}) = e^{i\phi} (\dot{\rho} + i\rho\dot{\phi}). \quad (\text{C.51})$$

By using the definitions of B and C from (A.4), we can expand the left-hand side of the above equation:

$$\begin{aligned} \text{LHS} &= \frac{d}{dt}\mathcal{R}_2 = 2i (B(\mathcal{R}_2)(\mathcal{R}_2)^2 + C(\mathcal{R}_2)\mathcal{R}_2 + \overline{B}(\mathcal{R}_2)) \\ &= 2i (b_1\mathcal{R}_2^3 + b_2\mathcal{R}_2^2\overline{\mathcal{R}_2} + b_3\mathcal{R}_2^2 + b_4\mathcal{R}_2^2 + \overline{b_4}\mathcal{R}_2\mathcal{R}_2 + b_5\mathcal{R}_2 + \overline{b_1}\mathcal{R}_2 + \overline{b_2}\mathcal{R}_2 + \overline{b_3}) \\ &= 2i (b_1\rho^3 e^{i3\phi} + b_2\rho^3 e^{i\phi} + (b_3 + b_4)\rho^2 e^{i2\phi} + \overline{b_4}\rho^2 + (\overline{b_2} + b_5)\rho e^{i\phi} + \overline{b_1}\rho e^{-i\phi} + \overline{b_3}) \\ &= e^{i\phi} (2 (ib_1 e^{i2\phi}\rho^3 + i\overline{b_1} e^{-i2\phi}\rho + i(b_3 + b_4)e^{i\phi}\rho^2 + i(\overline{b_3} + \overline{b_4}\rho^2)e^{-i\phi} + i(b_2\rho^3 + (\overline{b_2} + b_5)\rho))) = e^{i\phi} G(\rho, \phi), \end{aligned}$$

where we defined G implicitly in the last step.

Comparing the two expressions for the left-hand and right-hand sides of (C.51) results in the complex-valued identity

$$\dot{\rho} + i\rho\dot{\phi} = G(\rho, \phi). \quad (\text{C.52})$$

Recalling that ρ and ϕ are both real-valued, we separate the real and imaginary parts of the function G . In particular, letting

$$G(\rho, \phi) = G_R(\rho, \phi) + iG_I(\rho, \phi) \quad \text{with} \quad G_R := \text{Re}(G) \quad \text{and} \quad G_I := \text{Im}(G),$$

we can match the real and imaginary parts of (C.52) to obtain the real-valued coupled ODE system

$$\dot{\rho} = G_R(\rho, \phi), \quad \dot{\phi} = \frac{1}{\rho} G_I(\rho, \phi), \quad (\text{C.53})$$

and it remains to write down G_R as well as G_I explicitly.

To do so, we introduce some notation and set

$$\begin{aligned} G(\rho, \phi) &= 2 (ib_1 e^{i2\phi}\rho^3 + i\overline{b_1} e^{-i2\phi}\rho + i(b_3 + b_4)e^{i\phi}\rho^2 + i(\overline{b_3} + \overline{b_4}\rho^2)e^{-i\phi} + i(b_2\rho^3 + (\overline{b_2} + b_5)\rho)) \\ &= 2(A_1 + A_2 + A_3 + A_4 + A_5). \end{aligned}$$

We can then compute for each term A_j the real and imaginary part.

- Term A_1 :

$$\operatorname{Re}(A_1) = \operatorname{Re}(ib_1 e^{i2\phi} \rho^3) = -\operatorname{Im}(b_1 e^{i2\phi}) \rho^3, \quad \operatorname{Im}(A_1) = \operatorname{Im}(ib_1 e^{i2\phi} \rho^3) = \operatorname{Re}(b_1 e^{i2\phi}) \rho^3.$$

- Term A_2 :

$$\operatorname{Re}(A_2) = \operatorname{Re}(i\bar{b}_1 e^{-i2\phi} \rho) = \operatorname{Im}(b_1 e^{i2\phi}) \rho, \quad \operatorname{Im}(A_2) = \operatorname{Im}(i\bar{b}_1 e^{-i2\phi} \rho) = \operatorname{Re}(b_1 e^{i2\phi}) \rho.$$

- Term A_3 :

$$\begin{aligned} \operatorname{Re}(A_3) &= \operatorname{Re}(i(b_3 + b_4)e^{i\phi} \rho^2) = -\operatorname{Im}((b_3 + b_4)e^{i\phi}) \rho^2, \\ \operatorname{Im}(A_3) &= \operatorname{Im}(i(b_3 + b_4)e^{i\phi} \rho^2) = \operatorname{Re}((b_3 + b_4)e^{i\phi}) \rho^2. \end{aligned}$$

- Term A_4 :

$$\begin{aligned} \operatorname{Re}(A_4) &= \operatorname{Re}(i(\bar{b}_3 + \bar{b}_4 \rho^2)e^{-i\phi}) = \operatorname{Im}(b_3 e^{i\phi}) + \operatorname{Im}(b_4 e^{i\phi}) \rho^2, \\ \operatorname{Im}(A_4) &= \operatorname{Im}(i(\bar{b}_3 + \bar{b}_4 \rho^2)e^{-i\phi}) = \operatorname{Re}(b_3 e^{i\phi}) + \operatorname{Re}(b_4 e^{i\phi}) \rho^2. \end{aligned}$$

- Term A_5 :

$$\begin{aligned} \operatorname{Re}(A_5) &= \operatorname{Re}(i(b_2 \rho^3 + (\bar{b}_2 + b_5)\rho)) = -\operatorname{Im}(b_2) \rho^3 + \operatorname{Im}(b_2) \rho, \\ \operatorname{Im}(A_5) &= \operatorname{Im}(i(b_2 \rho^3 + (\bar{b}_2 + b_5)\rho)) = \operatorname{Re}(b_2) \rho^3 + (\operatorname{Re}(b_2) + \operatorname{Re}(b_5))\rho. \end{aligned}$$

For the last term, recall that b_5 is real-valued.

With those equalities we have

$$\begin{aligned} G_R(\rho, \phi) &= 2(\operatorname{Re}(A_1) + \operatorname{Re}(A_2) + \operatorname{Re}(A_3) + \operatorname{Re}(A_4) + \operatorname{Re}(A_5)) \\ &= 2(1 - \rho^2) (\operatorname{Im}(b_1 e^{i2\phi} + b_2) \rho + \operatorname{Im}(b_3 e^{i\phi})), \end{aligned}$$

and

$$\begin{aligned} G_I(\rho, \phi) &= 2(\operatorname{Im}(A_1) + \operatorname{Im}(A_2) + \operatorname{Im}(A_3) + \operatorname{Im}(A_4) + \operatorname{Im}(A_5)) \\ &= 2((\rho^2 + 1) (\operatorname{Re}(b_1 e^{i2\phi} + b_2) \rho + \operatorname{Re}(b_3 e^{i\phi})) + \rho (2\operatorname{Re}(b_4 e^{i\phi}) \rho + \operatorname{Re}(b_5))). \end{aligned}$$

By introducing the notation

$$\begin{aligned} S_1(\phi) &:= \operatorname{Im}(b_1 e^{i2\phi} + b_2), & S_2(\phi) &:= \operatorname{Im}(b_3 e^{i\phi}), \\ S_3(\phi) &:= \operatorname{Re}(b_1 e^{i2\phi} + b_2), & S_4(\phi) &:= \operatorname{Re}(b_3 e^{i\phi}), & S_5(\phi) &:= \operatorname{Re}(b_4 e^{i\phi}), \end{aligned} \tag{C.54}$$

we eventually obtain the coupled ODE system

$$\begin{aligned} \dot{\rho} &= 2(1 - \rho^2) (S_1(\phi) \rho + S_2(\phi)) \\ \dot{\phi} &= 2 \left(S_3(\phi) (\rho^2 + 1) + S_4(\phi) \frac{\rho^2 + 1}{\rho} + 2S_5(\phi) \rho + \operatorname{Re}(b_5) \right), \end{aligned}$$

which concludes the proof. $\square$

Lemma C.8. *Let $U : [0, \infty) \rightarrow [0, 1)$ be an absolutely continuous function satisfying*

$$\dot{U}(t) = (1 - U(t)^2) S(t) \tag{C.55}$$

for some locally integrable function $S : [0, \infty) \rightarrow \mathbb{R}$. Assume that

$$\lim_{t \rightarrow \infty} \int_0^t S(\tau) d\tau = \infty. \tag{C.56}$$

Then

$$\lim_{t \rightarrow \infty} U(t) = 1.$$

Proof. For $U(t) \in [0, 1)$ define

$$u(t) := \operatorname{artanh}(U(t)) = \frac{1}{2} \log \frac{1+U(t)}{1-U(t)}.$$

Since $\frac{d}{dU} \operatorname{artanh}(U) = (1-U^2)^{-1}$, the chain rule yields

$$\dot{u}(t) = \frac{\dot{U}(t)}{1-U(t)^2} = S(t) \quad \text{for a.e. } t \geq 0.$$

Integrating in time, we obtain

$$u(t) = u(0) + \int_0^t S(\tau) d\tau.$$

By assumption (C.56), the right-hand side diverges to ∞ as $t \rightarrow \infty$. Therefore $u(t) \rightarrow \infty$, and since $U(t) = \tanh(u(t))$, it follows that

$$U(t) \xrightarrow[t \rightarrow \infty]{} 1.$$

□

D Proof Details for Section 4

In this appendix, we provide the technical details for the proofs of Section 4.

D.1 Proof of Main Results

Proof of Corollary 4.2. We prove the three inequalities in (4.13) one by one using elementary inequalities together with the estimate established in Proposition 4.1, namely,

$$|\mathcal{R}_2 - \alpha| \leq C_\nu(\gamma)(1 - \gamma).$$

We recall the polar-representation $\alpha = \gamma e^{i\Phi}$ and $\mathcal{R}_2 = \rho e^{i\phi}$.

We begin with the first inequality. By the reverse triangle inequality,

$$|\rho - \gamma| = ||\mathcal{R}_2| - |\alpha|| \leq |\mathcal{R}_2 - \alpha| \leq C_\nu(\gamma)(1 - \gamma).$$

Next, for the second inequality, we write

$$\mathcal{R}_2 - \alpha = \rho e^{i\phi} - \gamma e^{i\Phi} = (\rho - \gamma)e^{i\phi} + \gamma(e^{i\phi} - e^{i\Phi}).$$

Rearranging terms and applying the triangle inequality gives

$$\gamma|e^{i\phi} - e^{i\Phi}| \leq |\mathcal{R}_2 - \alpha| + |\rho - \gamma| \leq 2C_\nu(\gamma)(1 - \gamma),$$

and therefore

$$|e^{i\phi} - e^{i\Phi}| \leq 2C_\nu(\gamma) \frac{(1 - \gamma)}{\gamma}.$$

Finally, we prove the third inequality. Choosing the representation of $\Phi - \phi$ modulo 2π so that $|\Phi - \phi| \leq \pi$, we have

$$|\Phi - \phi| = 2 \left| \frac{\Phi - \phi}{2} \right| \leq 2\pi \left| \sin \frac{\Phi - \phi}{2} \right| = \pi |e^{i\phi} - e^{i\Phi}| \leq 2\pi C_\nu(\gamma) \frac{1 - \gamma}{\gamma}.$$

This completes the proof. □

Proof of Theorem 4.7. We start the proof with a change of variables. Define a scalar

$$R := \sqrt{(a_-^d)^2 + (a_+^o)^2},$$

and an angle Φ_A satisfying

$$a_-^d = R \cos \Phi_A, \quad a_+^o = R \sin \Phi_A.$$

When $R = 0$, we set $\Phi_A = 0$. Then we can compute using trigonometric identities that

$$\begin{aligned} a_-^d \cos \Phi + a_+^o \sin \Phi &= R \cos \Phi_A \cos \Phi + R \sin \Phi_A \sin \Phi = R \cos(\Phi - \Phi_A), \\ a_-^d \sin \Phi - a_+^o \cos \Phi &= R \cos \Phi_A \sin \Phi - R \sin \Phi_A \cos \Phi = R \sin(\Phi - \Phi_A). \end{aligned}$$

Therefore, the original system (4.17) can be written as

$$\begin{aligned} \dot{\gamma} &= \frac{1}{4}(1 - \gamma^2) (a_+^d \gamma + R \cos(\Phi - \Phi_A)) + \Delta_\gamma, \\ \dot{\Phi} &= \frac{1}{4}(1 - \gamma^2) \left(a_-^o - \frac{R}{\gamma} \sin(\Phi - \Phi_A) \right) + \Delta_\Phi. \end{aligned} \quad (\text{D.1})$$

By Lemma D.2, the error terms Δ_γ and Δ_Φ satisfy

$$|\Delta_\gamma| \leq C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma} (1 - \gamma^2), \quad |\Delta_\Phi| \leq C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma},$$

where $C_{A,V}$ is a constant that depends on A and V .

Next, we focus on the convergence of the modulus γ . First, we show that there exists a threshold value $\gamma_{\text{th}} \in [0, 1]$ so that, if $\gamma(t_0) \in [\gamma_{\text{th}}, 1]$ for some t_0 , then $\gamma(t)$ converges to 1 exponentially fast for $t \geq t_0$. In particular, since the matrix A satisfies condition (4.19), together with the definitions of a_+^d and R , there exists a constant $\delta > 0$ so that

$$\lambda_{\min}(A + A^\top) = a_+^d - R \geq \delta.$$

This in turn implies that

$$a_+^d \gamma + R \cos(\Phi - \Phi_A) \geq (R + \delta)\gamma - R = \delta\gamma - R(1 - \gamma).$$

We infer that

$$\dot{\gamma} \geq \frac{1}{4}(1 - \gamma^2)(\delta\gamma - R(1 - \gamma) - C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma}) \geq \frac{1}{8}(1 - \gamma^2)\delta\gamma_{\text{th}}, \quad (\text{D.2})$$

for all $\gamma \in [\gamma_{\text{th}}, 1]$, provided that γ_{th} is sufficiently close to 1. This follows from Assumption 4.4 and Lemma D.3, which imply that $C_\nu(\gamma)(1 - \gamma) \rightarrow 0$ as $\gamma \rightarrow 1$. We note that γ_{th} depends on the constant M which controls the L^p -norm of ν .

From the above argument, we know that if there exists a time $0 < t_0 < \infty$ so that $\gamma(t_0) \in [\gamma_{\text{th}}, 1]$, then $\gamma(t)$ remains in this interval and increases monotonically toward the boundary value 1. In particular, inequality (D.2) continues to hold for all $t \geq t_0$. Moreover,

$$\frac{d}{dt}(1 - \gamma) \leq -\frac{\delta}{8}\gamma_{\text{th}}(1 - \gamma^2) \leq -c(1 - \gamma), \quad \text{where } c := \frac{\delta}{8}\gamma_{\text{th}}(1 + \gamma_{\text{th}}) > 0.$$

Then Grönwall's inequality gives the exponential convergence of $\gamma(t)$ to 1, namely, for $t \geq t_0$,

$$1 - \gamma(t) \leq (1 - \gamma_{\text{th}}) \exp(-c(t - t_0)). \quad (\text{D.3})$$

Therefore, it remains to show that for any given initial data $\gamma(0)$ satisfying our assumptions, there exists $t_0 \in (0, \infty)$ such that $\gamma(t_0) \geq \gamma_{\text{th}}$.

Let $\alpha_{\text{OA}} = \gamma_{\text{OA}} e^{i\Phi_{\text{OA}}}$ be the solution of the OA system (3.5), with initial data $\alpha(0)$ satisfying (4.18). By Theorem 3.1 and Proposition C.1, under the condition (4.19) on the matrix A there exists a time $t_0 > 0$, depending on ι in (4.18), such that $\gamma_{\text{OA}}(t_0) \geq \frac{1 + \gamma_{\text{th}}}{2}$. Then, we conclude from Lemma D.4 and Corollary 4.2 that there exists a constant $\varepsilon(\iota, p, M, A) > 0$, depending on ι , the matrix A , and p, M through γ_{th} , that is sufficiently small so that the closeness condition (4.20) implies

$$|\gamma(t_0) - \gamma_{\text{OA}}(t_0)| \leq \frac{1 - \gamma_{\text{th}}}{2}.$$

The reverse triangle inequality then implies

$$\gamma(t_0) \geq \gamma_{\text{OA}}(t_0) - \frac{1 - \gamma_{\text{th}}}{2} \geq \gamma_{\text{th}},$$

which is what we needed to conclude.

Finally, we establish the convergence of the angular variable Φ . From the Φ -equation in (D.1), Lemma D.2, and the fact that $\gamma(t) \in [\gamma_{\text{th}}, 1]$, for all $t \geq t_0$, we obtain

$$|\dot{\Phi}| = \left| \frac{1}{4}(1 - \gamma^2) \left(a_-^\circ - \frac{R}{\gamma} \sin(\Phi - \Phi_A) \right) + \Delta_\Phi \right| \leq C(1 - \gamma) + C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma} \leq C(1 + C_\nu(\gamma))(1 - \gamma).$$

Here, the constant $C > 0$ may change from one inequality to the next. By Assumption 4.4, Lemma D.3, and the exponential convergence of $\gamma(t)$ to 1 established in (D.3), the right-hand side of above inequality is integrable over $[t_0, \infty)$, and hence $|\dot{\Phi}| \in L^1([t_0, \infty))$. This implies that $\Phi(t)$ converges to some $\Phi_\infty \in \mathbb{T}^1$ as $t \rightarrow \infty$. $\square$

Proof of Theorem 4.8. We start the proof with a change of variables. Define a scalar

$$R := \sqrt{(v_-^d)^2 + (v_+^\circ)^2},$$

and an angle Φ_V satisfying

$$v_-^d = R \cos \Phi_V, \quad v_+^\circ = R \sin \Phi_V. \quad (\text{D.4})$$

When $R = 0$, we set $\Phi_V = 0$. Then we can compute using trigonometric identities that

$$\begin{aligned} v_-^d \cos \Phi + v_+^\circ \sin \Phi &= R \cos \Phi_V \cos \Phi + R \sin \Phi_V \sin \Phi = R \cos(\Phi - \Phi_V), \\ v_-^d \sin \Phi - v_+^\circ \cos \Phi &= R \cos \Phi_V \sin \Phi - R \sin \Phi_V \cos \Phi = R \sin(\Phi - \Phi_V). \end{aligned}$$

Therefore, the original system (4.21) can be written as

$$\begin{aligned} \dot{\gamma}(t) &= \frac{1}{4}(1 - \gamma(t)^2) (v_+^d \gamma(t) + R \cos(\Phi(t) - \Phi_V)) + \Delta_\gamma(t), \\ \dot{\Phi}(t) &= -\frac{1}{4} \frac{3\gamma(t)^2 + 1}{\gamma(t)} R \sin(\Phi(t) - \Phi_V) + \Delta_\Phi(t). \end{aligned}$$

Claim: There exist constants $\gamma_{\text{th}} \in [0, 1]$ and $\Phi_{\text{th}} > 0$ so that, if $(\gamma(t_0), \Phi(t_0))$ satisfies

$$\gamma(t_0) \in [\gamma_{\text{th}}, 1] \quad \text{and} \quad |\Phi(t_0) - \Phi_V| \leq \Phi_{\text{th}} \quad (\text{D.5})$$

for some $t_0 > 0$, then $(\gamma(t), \Phi(t)) \rightarrow (1, \Phi_V)$ as $t \rightarrow \infty$.

Assuming momentarily that the above claim has been proved, it would remain to show that for any initial condition $(\gamma(0), \Phi(0))$ satisfying our assumptions there exists a time $t_0 > 0$ such that $(\gamma(t_0), \Phi(t_0))$ satisfies condition (D.5). To verify the latter assertion, let $\alpha_{\text{OA}}(t) = \gamma_{\text{OA}}(t)e^{i\Phi_{\text{OA}}(t)}$ be the solution to the OA system (3.10), with initial data $\alpha_{\text{OA}}(0) = \alpha(0)$ satisfying (4.22). By Theorem 3.3 and Proposition C.3, under the condition (4.23) on the matrix V , there exists a time $t_0 > 0$, depending on ι in (4.22), such that $\gamma_{\text{OA}}(t_0) \geq \frac{1+\gamma_{\text{th}}}{2}$ and $|\Phi_{\text{OA}}(t_0) - \Phi_V| \leq \frac{\Phi_{\text{th}}}{2}$. Moreover, by Lemma D.4 and Corollary 4.2, there exists a constant $\varepsilon(\iota, p, M, V) > 0$, depending on ι , the matrix V , and p, M through γ_{th} , sufficiently small so that the closeness condition (4.24) implies

$$|\gamma(t_0) - \gamma_{\text{OA}}(t_0)| \leq \frac{1 - \gamma_{\text{th}}}{2} \quad \text{and} \quad |\Phi(t_0) - \Phi_{\text{OA}}(t_0)| \leq \frac{\Phi_{\text{th}}}{2}.$$

Therefore, by the triangle inequality, we have

$$\gamma(t_0) \geq \gamma_{\text{OA}} - \frac{1 - \gamma_{\text{th}}}{2} \geq \gamma_{\text{th}}, \quad \text{and} \quad |\Phi(t_0) - \Phi_V| \leq |\Phi(t_0) - \Phi_{\text{OA}}(t_0)| + |\Phi_{\text{OA}}(t_0) - \Phi_V| \leq \Phi_{\text{th}}.$$

Proof of the claim. We first show the convergence $\gamma(t) \rightarrow 1$, and then show that $\Phi(t) \rightarrow \Phi_V$.

Step 1: $\gamma(t) \rightarrow 1$ as $t \rightarrow \infty$.

Since the matrix V satisfies condition (4.23), there exists a constant $\delta > 0$ so that $\lambda_{\max}(V) \geq \delta$. Together with Lemma D.2, we can compute that

$$\begin{aligned}\dot{\gamma}(t) &= \frac{1}{4}(1 - \gamma(t)^2) (v_+^d \gamma(t) + R \cos(\Phi(t) - \Phi_V)) + \Delta_\gamma(t) \\ &\geq \frac{1}{4}(1 - \gamma(t)^2) \left(\frac{\delta}{2} \gamma(t) + R (\cos(\Phi(t) - \Phi_V) - \gamma(t)) \right) + \Delta_\gamma(t) \\ &\geq \frac{1}{8}(1 - \gamma(t)^2) \delta \left(1 - \frac{2}{k} \right) \gamma(t) - C_{A,V} C_\nu(\gamma(t)) \frac{1 - \gamma(t)}{\gamma(t)} (1 - \gamma(t)^2),\end{aligned}\tag{D.6}$$

provided that

$$R (\cos(\Phi(t) - \Phi_V) - \gamma(t)) \geq -\frac{\delta}{k} \gamma(t)\tag{D.7}$$

with any $k > 0$ to be specified later. Here, the first inequality uses the condition (4.23) on V , while the second inequality makes use of the condition (D.7) and Lemma D.2.

Furthermore, since ν satisfies Assumption 4.4, there exists a constant $\bar{\gamma}_{\text{th}} \in [0, 1]$, depending on constant M , so that

$$C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma} \leq \frac{\delta}{16} \left(1 - \frac{2}{k} \right) \gamma$$

for all $\gamma \in [\bar{\gamma}_{\text{th}}, 1]$, provided $k > 2$. The above inequality, together with (D.6), implies that, within the interval $[\bar{\gamma}_{\text{th}}, 1]$, we have

$$\dot{\gamma}(t) \geq \frac{\delta}{16} \left(1 - \frac{2}{k} \right) \gamma(t) (1 - \gamma(t)^2).\tag{D.8}$$

Then by Lemma C.8, we know that if $\gamma(t)$ enters the interval $[\bar{\gamma}_{\text{th}}, 1]$, then $\gamma(t) \rightarrow 1$ as $t \rightarrow \infty$, provided the condition (D.7) holds. We note that when $R = 0$, condition (D.7) is automatically satisfied whenever $k > 2$. Therefore, for the remaining proof in Step 1, it suffices to consider the case $R > 0$.

Let $k > 2$ be chosen such that $0 < 1 - \frac{\delta}{kR} < 1$. Next, we show that, for a time $t_0 > 0$, if $(\gamma(t_0), \Phi(t_0))$ satisfies the conditions:

$$\gamma(t_0) \in [\gamma_{\text{th}}, 1],\tag{D.9a}$$

$$|\Phi(t_0) - \Phi_V| \leq \Phi_{\text{th}} := \arccos \left(1 - \frac{\delta}{kR} \right),\tag{D.9b}$$

where $\gamma_{\text{th}} \geq \bar{\gamma}_{\text{th}}$ and γ_{th} satisfies

$$\left(\frac{4C_{A,V}}{R} \frac{S_p(\gamma_{\text{th}})}{\gamma_{\text{th}}(3\gamma_{\text{th}}^2 + 1)} \right)^2 \leq 1 - \left(1 - \frac{\delta}{kR} \right)^2,\tag{D.10}$$

then the condition (D.7) holds for all $t \geq t_0$. Here, the function $S_p : [0, 1] \mapsto \mathbb{R}$ is defined by

$$S_p(\gamma) := \begin{cases} C_p M (1 - \gamma)^{1-1/p}, & 1 < p < \infty, \\ C_\infty M (1 - \gamma) \log \frac{2}{1-\gamma}, & p = \infty, \end{cases}\tag{D.11}$$

where the constants C_p , for $1 < p \leq \infty$, are given in Lemma D.3, and M is introduced in Assumption 4.4.

We first introduce some notation for later use. Denote $\tilde{\Phi}(t) := \Phi(t) - \Phi_V$. Then the angular dynamics of $\tilde{\Phi}$ can be written as

$$\dot{\tilde{\Phi}}(t) = -\frac{1}{4} \frac{3\gamma(t)^2 + 1}{\gamma(t)} R \sin \tilde{\Phi}(t) + \Delta_\Phi(t).\tag{D.12}$$

By Lemma D.2, Lemma D.3, and the definition of function S_p in (D.11), the perturbation satisfies

$$|\Delta_\Phi(t)| \leq C_{A,V} C_\nu(\gamma(t)) \frac{1 - \gamma(t)}{\gamma(t)} \leq \frac{C_{A,V}}{\gamma(t)} S_p(\gamma(t)) = \frac{1}{4} \frac{3\gamma(t)^2 + 1}{\gamma(t)} R Z(\gamma(t)),\tag{D.13}$$

where we denote

$$Z(\gamma) := \frac{4C_{A,V}}{R} \frac{S_p(\gamma)}{3\gamma^2 + 1}. \quad (\text{D.14})$$

Lastly, we recall the constant

$$\Phi_{\text{th}} = \arccos\left(1 - \frac{\delta}{kR}\right) \in (0, \pi/2). \quad (\text{D.15})$$

Next, we prove that the interval $[-\Phi_{\text{th}}, \Phi_{\text{th}}]$ is forward invariant for $\tilde{\Phi}(t)$; see the definition of constant Φ_{th} in (D.15). That is, we prove that if $\tilde{\Phi}(t)$ satisfies the initial condition $|\tilde{\Phi}(t_0)| \leq \Phi_{\text{th}}$, then

$$|\tilde{\Phi}(t)| \leq \Phi_{\text{th}} \quad \text{for all } t \geq t_0.$$

Once this is proved, the condition (D.7) follows immediately. Indeed, if $|\tilde{\Phi}(t)| \leq \Phi_{\text{th}}$, together with $\gamma(t) \in [0, 1]$, we have

$$\cos \tilde{\Phi}(t) \geq \cos \Phi_{\text{th}} = 1 - \frac{\delta}{kR} \geq \left(1 - \frac{\delta}{kR}\right) \gamma(t).$$

Then, first subtracting $\gamma(t)$ and then multiplying R on both sides to the above inequality gives the condition (D.7).

Thus, it remains to prove the forward invariance of $[-\Phi_{\text{th}}, \Phi_{\text{th}}]$ with respect to the angular dynamics $\tilde{\Phi}$ in (D.12). Suppose, for contradiction, that $\tilde{\Phi}(t)$ exits the interval $[-\Phi_{\text{th}}, \Phi_{\text{th}}]$. Let $T > t_0$ be the first exit time. Then

$$|\tilde{\Phi}(t)| \leq \Phi_{\text{th}} \quad \text{for all } t \in [t_0, T],$$

and at the exit time T , one has either

$$\tilde{\Phi}(T) = \Phi_{\text{th}} \quad \text{or} \quad \tilde{\Phi}(T) = -\Phi_{\text{th}}.$$

For every $t \in [t_0, T]$, since $|\tilde{\Phi}(t)| \leq \Phi_{\text{th}}$, the previous argument implies the condition (D.7) holds. Together with the differential inequality (D.8) and $\gamma(t_0) \in [\gamma_{\text{th}}, 1]$, we know that $\gamma(t)$ is monotonic increasing on $[t_0, T]$. Moreover, by the definitions of S_p in (D.11) and $Z(\gamma)$ in (D.14), the function $Z(\gamma)$ is nonincreasing. As a consequence,

$$Z(\gamma(t)) \leq Z(\gamma(t_0)) \quad \text{for all } t \in [t_0, T].$$

Therefore by the initial condition (D.10), together with $Z(\gamma) \geq 0$ and $\sin \Phi_{\text{th}} \geq 0$, we have

$$Z(\gamma(t)) \leq Z(\gamma(t_0)) \leq \sin \Phi_{\text{th}} \quad \text{for all } t \in [t_0, T]. \quad (\text{D.16})$$

We now check the direction of the vector field for $\tilde{\Phi}$ -dynamics at the boundary point $\tilde{\Phi}(T) = \Phi_{\text{th}}$ and $\tilde{\Phi}(T) = -\Phi_{\text{th}}$.

First suppose that $\tilde{\Phi}(T) = \Phi_{\text{th}}$. Then by the angular dynamics (D.12) and the perturbation bound (D.13), we get

$$\dot{\tilde{\Phi}}(T) = -\frac{1}{4} \frac{3\gamma(T)^2 + 1}{\gamma(T)} R \sin \Phi_{\text{th}} + \Delta_{\Phi}(T) \leq -\frac{1}{4} \frac{3\gamma(T)^2 + 1}{\gamma(T)} R (\sin \Phi_{\text{th}} - Z(\gamma(T))) \leq 0,$$

where the last inequality uses (D.16). Thus, at the right endpoint $\tilde{\Phi} = \Phi_{\text{th}}$, the vector field points inward or is tangent. This implies that the solution $\tilde{\Phi}(t)$ cannot exit the interval through the right endpoints.

With similar argument, we can show that the solution $\tilde{\Phi}(t)$ cannot exit the interval $[-\Phi_{\text{th}}, \Phi_{\text{th}}]$ through the left endpoints either. Indeed, suppose that $\tilde{\Phi}(T) = -\Phi_{\text{th}}$, again using the angular dynamics (D.12) and the perturbation bound (D.13), we obtain

$$\dot{\tilde{\Phi}}(T) = \frac{1}{4} \frac{3\gamma(T)^2 + 1}{\gamma(T)} R \sin \Phi_{\text{th}} + \Delta_{\Phi}(T) \geq \frac{1}{4} \frac{3\gamma(T)^2 + 1}{\gamma(T)} R (\sin \Phi_{\text{th}} - Z(\gamma(T))) \geq 0.$$

This shows that the interval $[-\Phi_{\text{th}}, \Phi_{\text{th}}]$ is forward-invariance with respect to the $\tilde{\Phi}$ -dynamics on $[t_0, T]$, which contradicts the definition of T as the first exit time. Therefore, we conclude that $|\tilde{\Phi}(t)| \leq \Phi_{\text{th}}$ for all $t \geq t_0$ and then (D.7) holds for all $t \geq t_0$. Hence, $\gamma(t) \rightarrow 1$ follows immediately.

We further show that $\gamma(t)$ converges to 1 exponentially fast for $t \geq t_0$. Indeed, since $\gamma(t) \in [\gamma_{\text{th}}, 1]$ for all $t \geq t_0$, inequality (D.8) gives

$$1 - \dot{\gamma}(t) \leq -\frac{\delta}{16} \left(1 - \frac{2}{k}\right) \gamma(t) (1 - \gamma(t)^2) \leq -c(1 - \gamma(t)), \quad \text{where } c := \frac{\delta}{16} \left(1 - \frac{2}{k}\right) \gamma_{\text{th}}(1 + \gamma_{\text{th}}).$$

Then Grönwall's inequality gives

$$1 - \gamma(t) \leq (1 - \gamma_{\text{th}}) \exp(-c(t - t_0)), \quad t \geq t_0. \quad (\text{D.17})$$

Step 2: $\Phi(t) \rightarrow \Phi_V$ as $t \rightarrow \infty$.

We begin by recalling several formulas and conclusions established above. In particular, recall the angular dynamics (D.12) of $\tilde{\Phi}(t) = \Phi(t) - \Phi_V$:

$$\dot{\tilde{\Phi}}(t) = -c(t) \sin \tilde{\Phi}(t) + \Delta_{\Phi}(t),$$

where we define

$$c(t) := \frac{1}{4} \frac{3\gamma(t)^2 + 1}{\gamma(t)} R. \quad (\text{D.18})$$

From the proof in Step 1, if $(\gamma(t_0), \Phi(t_0))$ satisfies condition (D.9), then

$$\gamma(t) \in [\gamma_{\text{th}}, 1] \quad \text{and} \quad |\tilde{\Phi}(t)| \leq \Phi_{\text{th}} \in (0, \pi/2) \quad \text{for all } t \geq t_0.$$

Moreover, $\gamma(t)$ converges to 1 exponentially fast as $t \rightarrow \infty$. We now establish the convergence of the angular variable $\Phi(t)$ by considering the cases $R = 0$ and $R > 0$ separately.

Regime $R = 0$. In this case, the Φ -equation in (C.21), together with Lemma D.2, gives

$$|\dot{\Phi}| = |\Delta_{\Phi}| \leq C_{A,V} C_{\nu}(\gamma) \frac{1 - \gamma}{\gamma} \leq C C_{\nu}(\gamma) (1 - \gamma),$$

where $C > 0$ is a fixed constant. By Assumption 4.4, Lemma D.3, and the exponential convergence of $\gamma(t)$ to 1 established in (D.17), the right-hand side of above inequality is integrable over $[t_0, \infty)$, and hence $|\dot{\Phi}| \in L^1([t_0, \infty))$. This implies that $\Phi(t)$ converges to some $\Phi_{\infty} \in \mathbb{T}^1$ as $t \rightarrow \infty$.

Regime $R > 0$. In this regime, the coefficient $c(t)$, as defined in (D.18), is uniformly bounded below by a positive constant, namely,

$$c(t) \geq c_* := \frac{R}{4} \min_{\gamma \in [\gamma_{\text{th}}, 1]} \frac{3\gamma^2 + 1}{\gamma} > 0 \quad \text{for all } t \geq t_0.$$

Next, recall from (D.13), the perturbation $\Delta_{\Phi}(t)$ satisfies

$$|\Delta_{\Phi}(t)| \leq C_{A,V} C_{\nu}(\gamma(t)) \frac{1 - \gamma(t)}{\gamma(t)}.$$

By Assumption 4.4, Lemma D.3, and $\gamma(t) \rightarrow 1$, we obtain

$$\Delta_{\Phi}(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Now, we are ready to prove that $\tilde{\Phi}(t) = \Phi(t) - \Phi_V \rightarrow 0$. Consider the Lyapunov function

$$L(t) = \tilde{\Phi}(t)^2.$$

Using the angular dynamics of $\tilde{\Phi}$, we get

$$\dot{L}(t) = 2\tilde{\Phi}(t)\dot{\tilde{\Phi}}(t) = -2c(t)\tilde{\Phi}(t) \sin \tilde{\Phi}(t) + 2\tilde{\Phi}(t)\Delta_{\Phi}(t).$$

Since $|\tilde{\Phi}(t)| \leq \Phi_{\text{th}} < \pi/2$, we have the elementary inequality

$$\tilde{\Phi} \sin \tilde{\Phi} \geq c_{\text{th}} \tilde{\Phi}^2 \quad \text{for all } |\tilde{\Phi}| \leq \Phi_{\text{th}},$$

where

$$c_{\text{th}} := \frac{\sin \Phi_{\text{th}}}{\Phi_{\text{th}}} > 0.$$

Therefore, we have the following differential inequality

$$\begin{aligned} \dot{L}(t) &\leq -2c_* c_{\text{th}} L(t) + 2|\tilde{\Phi}(t)| |\Delta_{\Phi}(t)| \\ &\leq -CL(t) + \frac{1}{C} |\Delta_{\Phi}(t)|^2, \end{aligned}$$

where the last inequality applies Young's inequality and uses the notation $C := c_* c_{\text{th}}$. By Grönwall's inequality, for $t \geq t_0$,

$$L(t) \leq e^{-C(t-t_0)} L(t_0) + \frac{1}{C} \int_{t_0}^t e^{-C(t-s)} |\Delta_{\Phi}(s)|^2 ds.$$

On the right-hand side of above inequality, the first term clearly converges to 0 as $t \rightarrow \infty$. For the second term, since $|\Delta_{\Phi}(s)|^2 \rightarrow 0$ as $s \rightarrow \infty$, we also have

$$\lim_{t \rightarrow \infty} \int_{t_0}^t e^{-C(t-s)} |\Delta_{\Phi}(s)|^2 ds = 0.$$

Therefore, $L(t) = \tilde{\Phi}(t)^2 \rightarrow 0$, and thus $\tilde{\Phi}(t) = \Phi(t) - \Phi_V \rightarrow 0$. □

□

D.2 Auxiliary Lemmas

Lemma D.1 (Properties of the Möbius Transformation). *Let $M_{\alpha, \eta}$ be the Möbius transformation defined in (2.16) considered as a mapping from the unit circle into itself. It has the following properties.*

(i). $\alpha = \int_0^{2\pi} M_{\alpha, \eta}(e^{i\varphi}) d\text{Unif}(\varphi)$, where Unif denotes the uniform probability measure on $[0, 2\pi)$.

(ii). $\text{Lip}(M_{\alpha, \eta}) \leq \frac{2}{1-\gamma}$, where $\gamma = |\alpha|$.

Proof. (i). If $|\alpha| = 1$, then $M_{\alpha, \eta}(u) = \alpha$ for all $u \in \mathbb{S}^1$, so the identity is clear in this case. If, on the other hand, $|\alpha| < 1$, for every $u \in \mathbb{S}^1$ we may write

$$\frac{1}{1 + \bar{\alpha} e^{i\eta} u} = \sum_{k=0}^{\infty} (-\bar{\alpha} e^{i\eta} u)^k.$$

Then, we by the definition of $M_{\alpha, \eta}$ in (2.16), we can directly compute that

$$\begin{aligned} \int_0^{2\pi} M_{\alpha, \eta}(e^{i\varphi}) d\text{Unif}(\varphi) &= \int_0^{2\pi} \frac{\alpha + e^{i\eta} e^{i\varphi}}{1 + \bar{\alpha} e^{i\eta} e^{i\varphi}} d\text{Unif}(\varphi) \\ &= \int_0^{2\pi} (\alpha + e^{i\eta} e^{i\varphi}) \sum_{k=0}^{\infty} (-\bar{\alpha} e^{i\eta} e^{i\varphi})^k d\text{Unif}(\varphi) = \alpha. \end{aligned}$$

The last equality uses that all terms with nonzero Fourier frequency vanish after integration.

(ii). We start by writing $M_{\alpha, \eta}$ as

$$M_{\alpha, \eta}(e^{i\varphi}) = E_{\alpha}(e^{i(\varphi+\eta)}),$$

where, for $u \in \mathbb{S}^1$,

$$E_{\alpha}(u) := \frac{\alpha + u}{1 + \bar{\alpha} u}.$$

A direct computation gives

$$\frac{dE_\alpha}{du} = \frac{(1 + \bar{\alpha}u) - \bar{\alpha}(\alpha + u)}{(1 + \bar{\alpha}u)^2} = \frac{1 - |\alpha|^2}{(1 + \bar{\alpha}u)^2}.$$

Since $|u| = 1$ we have $|1 + \bar{\alpha}u| \geq 1 - |\alpha|$, and consequently

$$\left| \frac{dE_\alpha}{du} \right| \leq \frac{1 - |\alpha|^2}{(1 - |\alpha|)^2} \leq \frac{1 + |\alpha|}{1 - |\alpha|} \leq \frac{2}{1 - |\alpha|}.$$

Moreover,

$$\left| \frac{d}{d\varphi} e^{i(\varphi+\eta)} \right| = |ie^{i(\varphi+\eta)}| = 1$$

Therefore, by the chain rule,

$$|\partial_\varphi M_{\alpha,\eta}(\varphi)| = \left| \frac{dE_\alpha}{du} \right| \left| \frac{d}{d\varphi} e^{i(\varphi+\eta)} \right| \leq \frac{2}{1 - |\alpha|} = \frac{2}{1 - \gamma}.$$

This gives

$$\text{Lip}(M_{\alpha,\eta}) \leq \frac{2}{1 - \gamma}.$$

□

Lemma D.2. *Let $\alpha = \gamma e^{i\Phi}$ be a solution to (4.1a).*

- **Case 1.** *For the matrices A and V given in (3.3), the (γ, Φ) -dynamics is*

$$\begin{aligned} \dot{\gamma} &= \frac{1}{4} (1 - \gamma^2) (a_+^d \gamma + a_-^d \cos(\Phi) + a_+^o \sin(\Phi)) + \Delta_\gamma, \\ \dot{\Phi} &= \frac{1}{4} (1 - \gamma^2) \left(a_-^o - \frac{a_-^d \sin(\Phi) - a_+^o \cos(\Phi)}{\gamma} \right) + \Delta_\Phi. \end{aligned} \tag{D.19}$$

- **Case 2.** *For the matrices A and V given in (3.8), the (γ, Φ) -dynamics is*

$$\begin{aligned} \dot{\gamma} &= \frac{1}{4} (1 - \gamma^2) (v_+^d \gamma + v_-^d \cos(\Phi) + v_+^o \sin(\Phi)) + \Delta_\gamma \\ \dot{\Phi} &= -\frac{1}{4} \frac{3\gamma^2 + 1}{\gamma} (v_-^d \sin(\Phi) - v_+^o \cos(\Phi)) + \Delta_\Phi. \end{aligned} \tag{D.20}$$

Here, Δ_γ and Δ_Φ denote the corresponding error terms in each case. Moreover, within each case, there exists a constant $C_{A,V} > 0$ depending on the matrices A and V such that

$$|\Delta_\gamma| \leq (1 - \gamma^2) C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma}, \quad |\Delta_\Phi| \leq C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma},$$

where C_ν is defined in (4.10).

Proof. For notation simplicity, we suppress the time dependence throughout the proof.

Recall from (4.15) that the (γ, Φ) -dynamics is given by

$$\dot{\gamma} = \text{Re} (e^{-i\Phi} F(\alpha, \alpha)) + \Delta_\gamma, \quad \dot{\Phi} = \frac{1}{\gamma} \text{Im} (e^{-i\Phi} F(\alpha, \alpha)) + \Delta_\Phi.$$

Here, Δ_γ and Δ_Φ are defined in (4.16), and for any $\alpha, r \in \mathbb{C}$,

$$F(\alpha, r) = 2i (B(r)\alpha^2 + C(r)\alpha + \bar{B}(r)),$$

where the functions B and C are given in (A.4). We note that B and C depend on the matrices A and V , and hence so does the function F .

As an immediate consequence of (4.7) and Proposition 4.1, we obtain

$$|\Delta_\Phi(t)| = \left| \frac{1}{\gamma} \operatorname{Im} (e^{-i\Phi}) \Delta(t) \right| \leq \frac{1}{\gamma} |\Delta(t)| \leq L_F(A, V) C_\nu(\gamma) \frac{1 - \gamma(t)}{\gamma(t)}. \quad (\text{D.21})$$

Here, $L_F(A, V)$ denotes the Lipschitz constant of F , which depends on the matrices A and V .

Next, we derive the explicit form of the (γ, Φ) -dynamics and estimate $|\Delta_\gamma|$ separately for the two cases under consideration.

- **Case 1.** For the matrices A and V in (3.3), the functions B and C take the form

$$B(r) = Q_1 \bar{r} + Q_2, \quad C(r) = -2\operatorname{Re}(Q_2 r) + Q_3,$$

where

$$Q_1 := \frac{1}{8} (-a_-^o + i a_+^d), \quad Q_2 := \frac{1}{8} (a_+^o + i a_-^d), \quad Q_3 := \frac{1}{4} a_-^o.$$

Substituting these expressions into F , we obtain

$$F(\alpha, \alpha) = 2i(1 - |\alpha|^2) (-Q_1 \alpha + \bar{Q}_2),$$

and

$$F(\alpha, r) - F(\alpha, \alpha) = 2i \left((\bar{Q}_1 - \alpha Q_2) (\mathcal{R}_2 - \alpha) + \alpha (\alpha Q_1 - \bar{Q}_2) (\bar{\mathcal{R}}_2 - \bar{\alpha}) \right).$$

Using these identities together with (4.15), a direct computation yields the explicit (γ, Φ) -system

$$\begin{aligned} \dot{\gamma} &= 2(1 - \gamma^2) (\operatorname{Im}(Q_1) \gamma + \operatorname{Im}(Q_2 e^{i\Phi})) + \Delta_\gamma, \\ \dot{\Phi} &= 2(1 - \gamma^2) \left(\frac{Q_3}{2} + \frac{1}{\gamma} \operatorname{Re}(Q_2 e^{i\Phi}) \right) + \Delta_\Phi, \end{aligned} \quad (\text{D.22})$$

where

$$\begin{aligned} \Delta_\gamma &= 2(1 - \gamma^2) \operatorname{Im} \left(Q_1 \left(\rho e^{i(\Phi - \phi)} - \gamma \right) \right), \\ \Delta_\Phi &= \frac{2(1 + \gamma^2)}{\gamma} \operatorname{Re} \left(Q_1 \left(\rho e^{i(\Phi - \phi)} - \gamma \right) \right) + 4(\gamma \operatorname{Re}(Q_2 e^{i\Phi}) - \rho \operatorname{Re}(Q_2 e^{i\phi})). \end{aligned} \quad (\text{D.23})$$

Next, we estimate $|\Delta_\gamma|$. Using the triangle inequality, the bound $\rho, \gamma \leq 1$, and Proposition 4.3, we deduce

$$\begin{aligned} |\Delta_\gamma| &= \left| 2(1 - \gamma^2) \operatorname{Im} \left(Q_1 \left(\rho e^{i(\Phi - \phi)} - \gamma \right) \right) \right| \\ &\leq 2(1 - \gamma^2) |Q_1| \left| \rho e^{i(\Phi - \phi)} - \gamma \right| \\ &\leq 2(1 - \gamma^2) |Q_1| \left(|\rho - \gamma| + \rho |e^{i(\Phi - \phi)} - 1| \right) \\ &\leq 2(1 - \gamma^2) |Q_1| (|\rho - \gamma| + |\Phi - \phi|) \\ &\leq (1 - \gamma^2) C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma}. \end{aligned}$$

Finally, substituting the definitions of Q_1, Q_2 , and Q_3 into (D.22) gives the claimed form of the dynamics.

- **Case 2.** For the matrices A and V given in (3.8), the functions B and C take the form

$$B(r) = Q_1 \bar{r} + Q_2, \quad C(r) = 2\operatorname{Re}(Q_2 r)$$

with

$$Q_1 := \frac{i}{8} v_+^d, \quad Q_2 := \frac{1}{8} (v_+^o + i v_-^d).$$

Substituting these expressions into F , we obtain

$$F(\alpha, \alpha) = 2i (Q_1 |\alpha|^2 \alpha + 2Q_2 \alpha^2 + \overline{Q}_1 \alpha + (1 + |\alpha|^2) \overline{Q}_2),$$

and

$$F(\alpha, r) - F(\alpha, \alpha) = 2i ((\overline{Q}_1 + Q_2 \alpha) (r - \alpha) + \alpha (Q_1 \alpha + \overline{Q}_2) (\bar{r} - \bar{\alpha})).$$

Using these identities, together with (4.15), a direct computation yields the explicit (γ, Φ) -system

$$\begin{aligned} \dot{\gamma} &= 2(1 - \gamma^2) (\operatorname{Im}(Q_1) \gamma + \operatorname{Im}(Q_2 e^{i\Phi})) + \Delta_\gamma, \\ \dot{\Phi} &= \frac{2(3\gamma^2 + 1)}{\gamma} \operatorname{Re}(Q_2 e^{i\Phi}) + \Delta_\Phi, \end{aligned} \quad (\text{D.24})$$

where

$$\begin{aligned} \Delta_\gamma &:= 2(1 - \gamma^2) \operatorname{Im}(Q_1) (\rho \operatorname{Re}(e^{i(\phi - \Phi)}) - \gamma), \\ \Delta_\Phi &:= 2 \frac{1 + \gamma^2}{\gamma} \operatorname{Im}(Q_1) \rho \operatorname{Im}(e^{i(\phi - \Phi)}) + 4 (\rho \operatorname{Re}(Q_2 e^{i\phi}) - \gamma \operatorname{Re}(Q_2 e^{i\Phi})). \end{aligned} \quad (\text{D.25})$$

With a similar argument as for Case 1, we obtain

$$|\Delta_\gamma| \leq (1 - \gamma^2) C_{A,V} C_\nu(\gamma) \frac{1 - \gamma}{\gamma}.$$

Finally, substituting the definitions of Q_1, Q_2 , and Q_3 into (D.24) gives the claimed form of the dynamics.

□

Lemma D.3. *Let $1 \leq p \leq \infty$, and let ν be a probability density on $[0, 2\pi)$. Then there exists a constant $C_p < \infty$ such that, for every $\gamma \in [0, 1)$,*

$$C_\nu(\gamma)(1 - \gamma) \leq \begin{cases} C_p \|\nu\|_{L^p([0, 2\pi))} (1 - \gamma)^{1-1/p}, & 1 < p < \infty, \\ C_\infty \|\nu\|_{L^\infty([0, 2\pi))} (1 - \gamma) \log \frac{2}{1 - \gamma}, & p = \infty, \\ C_1, & p = 1. \end{cases}$$

In particular, when $1 < p \leq \infty$ and $\|\nu\|_{L^p([0, 2\pi))} < \infty$, we have $C_\nu(\gamma)(1 - \gamma) \rightarrow 0$ as $\gamma \rightarrow 1$.

Proof. Fix $\chi \in [0, 2\pi)$. By Hölder's inequality and the fact that $\|\cdot\|_{L^p([0, 2\pi))}$ is invariant by rotations we deduce

$$\int_0^{2\pi} \frac{1}{|1 + \gamma e^{i(\varphi + \chi)}|} \nu(d\varphi) \leq \|\nu\|_{L^p([0, 2\pi))} \left\| \frac{1}{|1 + \gamma e^{i\varphi}|} \right\|_{L^q([0, 2\pi))}, \quad (\text{D.26})$$

where q is p 's Hölder conjugate exponent. It thus suffices to estimate

$$K_q(\gamma) := \left\| \frac{1}{|1 + \gamma e^{i\varphi}|} \right\|_{L^q([0, 2\pi))}.$$

If $0 \leq \gamma \leq \frac{1}{2}$, then $|1 + \gamma e^{i\varphi}| \geq 1 - \gamma \geq \frac{1}{2}$, and hence

$$K_q(\gamma) \leq 2 \leq C_p (1 - \gamma)^{-\frac{1}{p}},$$

in case $p < \infty$, and

$$K_1(\gamma) \leq 2 \leq C_p \log \frac{2}{1 - \gamma},$$

if case $p = \infty$.

Assume now that $\frac{1}{2} \leq \gamma < 1$. Writing $\tilde{\varphi} = \varphi - \pi$, we obtain

$$|1 + \gamma e^{i\varphi}|^2 = (1 - \gamma)^2 + 2\gamma(1 - \cos \tilde{\varphi}) \geq (1 - \gamma)^2 + c_1 \tilde{\varphi}^2$$

for some $c_1 > 0$. Hence

$$K_q(\gamma)^q \leq C \int_{-\pi}^{\pi} \frac{d\tilde{\varphi}}{((1 - \gamma)^2 + \tilde{\varphi}^2)^{q/2}}.$$

If $1 < q < \infty$, a change of variables $\tilde{\varphi} = (1 - \gamma)u$ gives

$$K_q(\gamma)^q \leq C(1 - \gamma)^{1-q} \int_{\mathbb{R}} \frac{du}{(1 + u^2)^{q/2}},$$

and since $q > 1$, the integral is finite. Therefore

$$K_q(\gamma) \leq C_p(1 - \gamma)^{-1/p}.$$

If $q = 1$ (i.e. $p = \infty$), we obtain, instead,

$$K_1(\gamma) \leq C \int_{-\pi}^{\pi} \frac{d\theta}{\sqrt{(1 - \gamma)^2 + \theta^2}} \leq C \log \frac{2}{1 - \gamma}.$$

Finally, if $q = \infty$ (i.e. $p = 1$), we simply use

$$K_{\infty}(\gamma) = \sup_{\varphi} \frac{1}{|1 + \gamma e^{i\varphi}|} = \frac{1}{1 - \gamma}.$$

Inserting the above bounds in (D.26), and multiplying both sides of (D.26) by $1 - \gamma$, we obtain the desired upper bound on $C_{\nu}(\gamma)(1 - \gamma)$. $\square$

Lemma D.4. *Under the assumptions of Proposition 4.3, suppose, additionally, that ν satisfies Assumption 4.4. For any prescribed $\varsigma > 0$ and $T > 0$, there exists a constant $\varepsilon(\varsigma, T, p, M, L_F) > 0$, depending on ς, T , the constants p, M in Assumption 4.4, and the Lipschitz constant L_F in Proposition 4.3, such that if*

$$d_{\text{BL}}(\nu, \text{Unif}) \leq \varepsilon(\varsigma, T, p, M, L_F),$$

then

$$\sup_{t \in [0, T]} |\alpha(t) - \alpha_{\text{OA}}(t)| \leq \varsigma.$$

Proof. Since ν satisfies Assumption 4.4, by Lemma D.3 there exists a constant $\tilde{\gamma}_{\text{th}} > 0$, depending on T, ς, p, M , and L_F so that

$$TL_F e^{L_F T} C_{\nu}(\gamma)(1 - \gamma) \leq \frac{\varsigma}{2}, \quad \text{for } \forall \gamma \in [\tilde{\gamma}_{\text{th}}, 1].$$

Then, by Proposition 4.3, we have

$$\begin{aligned} & \sup_{t \in [0, T]} |\alpha(t) - \alpha_{\text{OA}}(t)| \\ & \leq L_F e^{L_F T} \int_0^T \min \left\{ C_{\nu}(\gamma(t))(1 - \gamma(t)), \left(1 + \frac{2}{1 - \gamma(t)}\right) d_{\text{BL}}(\nu, \text{Unif}) \right\} dt \\ & \leq L_F e^{L_F T} \int_0^T \left(C_{\nu}(\gamma(t))(1 - \gamma(t)) \mathbf{1}_{\{\gamma(t) \geq \tilde{\gamma}_{\text{th}}\}} + \left(1 + \frac{2}{1 - \gamma(t)}\right) d_{\text{BL}}(\nu, \text{Unif}) \mathbf{1}_{\{\gamma(t) \leq \tilde{\gamma}_{\text{th}}\}} \right) dt \\ & \leq \frac{\varsigma}{2} + TL_F e^{L_F T} \frac{3}{1 - \tilde{\gamma}_{\text{th}}} d_{\text{BL}}(\nu, \text{Unif}) \leq \varsigma, \end{aligned}$$

where in the final inequality we choose $\varepsilon > 0$, which depends on $T, L_F, \tilde{\gamma}_{\text{th}}$, and ς , sufficiently small so that the condition $d_{\text{BL}}(\nu, \text{Unif}) \leq \varepsilon$ ensures that the second term in the preceding inequality is bounded by $\varsigma/2$. $\square$