

# Multistability of Self-Attention Dynamics in Transformers

Claudio Altafini\*

November 17, 2025

## Abstract

In machine learning, a self-attention dynamics is a continuous-time multiagent-like model of the attention mechanisms of transformers. In this paper we show that such dynamics is related to a multiagent version of the Oja flow, a dynamical system that computes the principal eigenvector of a matrix corresponding for transformers to the value matrix. We classify the equilibria of the “single-head” self-attention system into four classes: consensus, bipartite consensus, clustering and polygonal equilibria. Multiple asymptotically stable equilibria from the first three classes often coexist in the self-attention dynamics. Interestingly, equilibria from the first two classes are always aligned with the eigenvectors of the value matrix, often but not exclusively with the principal eigenvector.

## 1 Introduction

Less than a decade since their introduction [27], transformer architectures have become the *de facto* standard algorithm for many problems in machine learning and are widely adopted in various fields, such as natural language processing, computer vision and speech processing [13, 28, 12]. At the core of a transformer is a so-called self-attention mechanism, a set of operations performed on vectorial representations of the “tokens” i.e., elementary units of the objects under analysis (words for large language models (LLM), images patches in computer vision, etc.). These operations involve three matrices called query ( $Q$ ), key ( $K$ ) and value ( $V$ ) matrices, two of which,  $Q$  and  $K$ , are involved in an inner product, which, exponentiated and normalized by a partition function, yields a softmax function depending on the tokens. Such softmax function is the celebrated attention mechanism, and expresses how much attention a token  $i$  is giving to another token  $j$ , relative to the ensemble of all tokens. The attention coefficients provide the weight in the weighted sum of the product of the tokens by the value matrix  $V$ , hence forming a “self-attention” mechanism. This is the core of a transformer layer, which receives as

---

\*C. Altafini is with the Division of Automatic Control, Dept. of Electrical Engineering, Linköping University, SE-58183, Linköping, Sweden. email: [claudio.altafini@liu.se](mailto:claudio.altafini@liu.se). Work supported in part by the Swedish Research Council (grant n. 2024-04772) and by the ELLIIT framework program at Linköping University.

input the tokens and gives as output “transformed” tokens. To avoid a collapsing or exploding token norm due to these operations, the output is then normalized. Other operations (which we do not consider here) are typically present, like for instance multiple such mechanisms act simultaneously to form a “multi-head” attention, or the output just described is passed through a feedforward neural network. Overall this mechanism constitutes a layer of the transformer: multiple identical layers are concatenated to form what is normally referred to as a transformer.

As is often the case in machine learning, the mathematical understanding of a new approach (“why it works”) lags behind its practical use, and transformer are no exception [26]. However, given the incredible importance in everyday life that transformer-based applications like LLM are acquiring, investigating and understanding their behavior from a rational perspective appears an important and compelling issue.

One possible approach to investigating the behavior of transformers was provided recently in a series of papers [6, 21, 14, 7, 8, 11, 29]. The basic idea of these papers is to treat the repeated application of identical layers typical of a transformer as the unfolding in time of a dynamical system whose states are the tokens being modified by each transformer layer. Rather than dealing with a discrete-time dynamical system (as the setting would immediately suggest), [7, 8] opt for passing to a continuous-time description, which is more amenable to mathematical analysis and easier to characterize. The resulting ODE model corresponds to a transformer with an infinite number of layers, which is clearly an idealization (in practical implementations, a transformer may have tens or hundreds of layers). An interesting perspective that is suggested in [7, 8] is that such ODE can be seen as a multiagent dynamical system, in which each agent (called a “particle” in [7, 8]) is a token, and its update law depends on all the other tokens/agents. The resulting dynamics is nonlinear due to the attention mechanism, and evolves on a unit sphere because of the normalization operation.

Multiagent systems on spheres have been studied extensively in the control community [3, 16, 25, 33, 34], in particular for what concerns collective phenomena like consensus (all agents converge to the same point in the unit sphere) and more complex, yet related, behaviors like bipartite consensus (when some agents converge to a common point and some other to the antipodal point [3, 33]) which appear naturally because of the compact nature of the ambient manifold. These collective behaviors are highly relevant for transformers: it has in fact been observed repeatedly that transformers indeed tend to be subject to rank-collapse phenomena [4, 18] (also sometimes referred to as token-uniformity or over-smoothing [17, 23, 24, 32, 5]) which appear essentially when the tokens become equal or cluster into groups of equal tokens. Indeed in [7, 8] consensus is one of the main behaviors shown to occur for this continuous-time model of self-attention dynamics. A similar result is reported in [1] (paper which is closest to ours in terms of mathematical approach).

The scope of this paper is to make a thorough analysis of the asymptotic behavior of the continuous-time self-attention dynamics model of [7, 8] using tools from dynamical systems and control. In order to do so we establish a connection with another well-known model on the sphere, which, following [19, 20, 9, 30, 31], we call the Oja flow, but which is also related to the continuous-time Rayleigh quotient flow [9] and to the continuous-time power method, see eq. (3) of [15]. This is a much simpler dynamical system whose main feature is that it converges to the principal eigenvector of a matrix which in our setting

corresponds to the value matrix  $V$ . In fact, Oja flows are at the basis of algorithms that are used to compute the eigenvectors of a matrix, and have long been used for this scope as an alternative to power methods e.g. in principal component analysis [10, 22]. The Oja flow is insightful but far too simple to use for the self-attention dynamics. However a multiagent version of Oja flow, which we develop in the paper, is much more similar, and in fact it corresponds to the self-attention dynamics without the attention coefficients. Similarly to the Oja flow, the multiagent Oja flow discovers the principal eigenvector of the value matrix  $V$ , i.e., all agents converge to a consensus equilibrium which is aligned with the principal eigenvector of  $V$ . In addition, the consensus and bipartite consensus points aligned with the other eigenvectors of  $V$  are also equilibria, but always unstable. It can be shown explicitly that the multiagent Oja flow generically converges to consensus at the principal eigenvector of  $V$ .

The self-attention dynamics is obtained inserting an attention matrix in the multiagent Oja flow, and corresponds to replacing a constant average coupling among the agents (equal for all agents) with a weighted average coupling, which is varying from agent to agent and over time. Even restricting to time-invariant  $Q$ ,  $K$  and  $V$ , and to symmetric  $V$ , the asymptotic behavior changes significantly w.r.t. the multiagent Oja flow: while consensus at the principal eigenvector still remains a locally asymptotically stable equilibrium point, other locally asymptotically stable equilibria normally emerge, rendering the typical self-attention dynamics multistable. Most of the new attractors correspond to bipartite consensus equilibria, aligned with the principal eigenvector or with some other eigenvector of  $V$ , even though sometimes other locally stable equilibria, which we call clustering equilibria, may emerge. The name derives from the (numerical) observation that, just like consensus and bipartite consensus, even these extra equilibria are typically in the form of clusters of tokens, i.e., multiple tokens end-up in the same point on the sphere. All these equilibria correspond to low-rank attention matrices: rank-1 (uniform) for consensus, rank-2 for a bipartite consensus, and typically low rank also for the clustering equilibria. The complete classification of equilibria of the self-attention dynamics includes also the so-called polygonal equilibria [3], which are however always unstable. Bipartite consensus and clustering equilibria are not mentioned in papers like [1, 2], which focus only on consensus. While the stability properties of the consensus and bipartite consensus equilibria can be studied analytically through the Lyapunov indirect method, as we do in this paper, for the more complex clustering equilibria an analytical treatment seems still out of reach.

While convergence towards a consensus-type of equilibrium is known both experimentally and theoretically, the observation that convergence typically occurs towards one of the eigenvectors of the value matrix does not seem to appear in the literature<sup>1</sup>. Often the consensus or bipartite consensus aligns with the principal eigenvector of  $V$ , as in a multiagent Oja flow, but alignment with other eigenvectors of  $V$  can also occur, in particular with the one corresponding to the least (i.e., most negative) eigenvalue. This asymptotic behavior can be considered a form of nonlinear Perron-Frobenius property, and it might provide insight into the interpretation of a transformer, potentially verifiable on real data with pretrained transformer matrices.

---

<sup>1</sup>With the exception of [1], where asymptotic stability to a consensus aligned with the principal eigenvector of  $V$  is shown, but only for an autoregressive model, i.e., a model with a triangular attention matrix.

**Notation** Boldface letters denote vectors, lower case greek and roman letter scalars, and upper case letters matrices. The eigenvalues of a matrix  $A$  are denoted  $\mu_i[A]$ , except for the value matrix  $V$ , whose eigenvalues are denoted  $\lambda_i$ . The inner product is indicated  $\langle \cdot, \cdot \rangle$ , while  $A \succeq 0$  means positive semidefinite (psd). The expression  $\mathbf{x} \parallel \mathbf{y}$  means  $\mathbf{y} = \gamma \mathbf{x}$  for some scalar  $\gamma$ , while  $\mathbf{x} \nparallel \mathbf{y}$  means no such  $\gamma$  exists. Finally,  $\mathbb{1}$  is the vector of all 1.

## 2 Model formulation

Consider  $n$  tokens represented as unit length vectors  $\mathbf{x}_i \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ ,  $i = 1, \dots, n$ . Denote  $Q, K \in \mathbb{R}^{m \times d}$  the query and key matrices (for simplicity and without loss of generality we hereafter assume  $m = d$ ) and  $V \in \mathbb{R}^{d \times d}$  the value matrix.

Following [8], an ODE model for a single-head self-attention mechanism on  $n$  tokens can be formulated as follows

$$\begin{aligned}\dot{\mathbf{x}}_i &= (I - \mathbf{x}_i \mathbf{x}_i^T) \sum_{j=1}^n \frac{e^{\langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle}}{\sum_{\ell=1}^n e^{\langle Q\mathbf{x}_i, K\mathbf{x}_\ell \rangle}} V \mathbf{x}_j \\ &= (I - \mathbf{x}_i \mathbf{x}_i^T) V \sum_{j=1}^n A_{ij}(\mathbf{x}) \mathbf{x}_j, \quad i = 1, \dots, n\end{aligned}\tag{1}$$

where for the terms of (1) we have the following interpretation:

- $P_i = (I - \mathbf{x}_i \mathbf{x}_i^T)$  is the projection onto  $T_{\mathbf{x}_i} \mathbb{S}^{d-1}$ , the tangent space of  $\mathbb{S}^{d-1}$  at  $\mathbf{x}_i$ . This guarantees that  $\|\mathbf{x}_i(t)\| = 1$  for all  $t$ , i.e., that the flow of (1) evolves on the unit sphere  $\mathbb{S}^{d-1}$ . In fact, for any  $\mathbf{y} \in \mathbb{S}^{d-1}$ ,  $P_i \mathbf{y} = (I - \mathbf{x}_i \mathbf{x}_i^T) \mathbf{y} = \mathbf{y} - \langle \mathbf{x}_i, \mathbf{y} \rangle \mathbf{x}_i$  is always normal to  $\mathbf{x}_i$ , and  $\mathbf{x}_i^T \dot{\mathbf{x}}_i = 0$ . The projection models the layer normalization present on each layer of the transformer.
- $A_{ij}(\mathbf{x}) = \frac{e^{\langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle}}{\sum_{\ell=1}^n e^{\langle Q\mathbf{x}_i, K\mathbf{x}_\ell \rangle}}$  is the attention that the token  $\mathbf{x}_i$  gives to the token  $\mathbf{x}_j$ , computed through a softmax function ( $\mathbf{x}$  is the stack of  $\mathbf{x}_1, \dots, \mathbf{x}_n$  vectors).  $A_{ij}(\mathbf{x})$  is a nonnegative scalar. The attention matrix is then  $A(\mathbf{x}) = [A_{ij}(\mathbf{x})]$  and it is a row stochastic matrix, i.e.,  $A(\mathbf{x}) \mathbb{1} = \mathbb{1}$ .
- In the model (1) time corresponds to the layer index, hence a self-attention model in continuous-time can be interpreted as a “continuum of layers”. The asymptotic value of the ODE,  $\mathbf{x}_i(\infty)$ , corresponds to the output of a transformer with an infinite number of layers.

The model (1) represents each token  $\mathbf{x}_i$  as an “agent” (it is called a “particle” in [8]) evolving on the sphere  $\mathbb{S}^{d-1}$ . The total state space is given by  $\mathbf{x} = [\mathbf{x}_1^T \dots \mathbf{x}_n^T]^T \in (\mathbb{S}^{d-1})^n$ . Since the flow of (1) lies on the compact manifold  $(\mathbb{S}^{d-1})^n$ , the vector field is Lipschitz, so forward existence, uniqueness and boundedness for all  $t$  follow automatically.

The model (1) corresponds to an example of collective dynamics on the sphere similar to those investigated in e.g. [3, 16, 33, 34]: the evolution occurs on a product of unit spheres and it is driven by the interaction with the other agents. The difference with these other models of collective dynamics on the sphere is that in (1) the attention matrix

$A(\mathbf{x})$  provides the “interaction graph”. It is typically fully connected and time-varying, since it depends on  $\mathbf{x}$ .

We are interested in studying the dynamical behavior of (1), and in particular in investigating its equilibria and their stability properties. To do so, we exploit the fact that the model (1) has some similarities with the so-called Oja flow, reviewed in next Section, and especially with a multiagent version of Oja flow, investigated in Section 4.

### 3 Oja flow

In its simplest formulation, the Oja flow [19, 20, 9] is the following dynamical system

$$\dot{\mathbf{x}} = (I - \mathbf{x}\mathbf{x}^T)V\mathbf{x}, \quad \mathbf{x} \in \mathbb{S}^{d-1}. \quad (2)$$

Assume that  $V$  is symmetric of eigenvalues  $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_d$ , with  $\lambda_1$  simple, and  $\mathbf{v}_1, \dots, \mathbf{v}_d$  the associated eigenvectors, normalized s.t.  $\|\mathbf{v}_k\| = 1$ . The dynamical behavior of (2) is summarized in the next lemma. Consider the Rayleigh quotient  $R(\mathbf{x}) = \frac{\mathbf{x}^T V \mathbf{x}}{\|\mathbf{x}\|^2}$ , which on the unit sphere reduces to the quadratic form  $R(\mathbf{x}) = \mathbf{x}^T V \mathbf{x}$ .  $R(\mathbf{x})$  can be used to construct a Lyapunov function for (2).

**Lemma 1** *All eigenvectors  $\mathbf{v}_k$  of  $V$  (more precisely, the values  $\pm\mathbf{v}_k$ ,  $\|\mathbf{v}_k\| = 1$ ) are equilibria of (2). The function  $W(\mathbf{x}) = \frac{1}{2}(\lambda_1 - R(\mathbf{x}))$  is a Lyapunov function for (2) and guarantees that (2) converges to the principal eigenvector  $\pm\mathbf{v}_1$  of  $V$  for almost all initial conditions  $\mathbf{x}(0) \in \mathbb{S}^{d-1}$ , while all other  $\pm\mathbf{v}_k$ ,  $k = 2, \dots, d$ , are unstable.*

**Proof.** From  $\lambda_d = \mathbf{x}^T \lambda_d \mathbf{x} \leq R(\mathbf{x}) \leq \mathbf{x}^T \lambda_1 \mathbf{x} = \lambda_1$ ,  $R(\mathbf{x})$  is upper bounded by  $\lambda_1$  on  $\mathbb{S}^{d-1}$ , hence  $W(\mathbf{x}) \geq 0$  and  $W(\mathbf{x}) = 0$  only when  $\mathbf{x} = \pm\mathbf{v}_1$ , since  $\lambda_1 > \lambda_k$   $k = 2, \dots, d$ . Differentiating, we have

$$\begin{aligned} \dot{W}(\mathbf{x}) &= -\mathbf{x}^T V^2 \mathbf{x} + (\mathbf{x}^T V \mathbf{x})^2 \\ &= -\|(I - \mathbf{x}\mathbf{x}^T)V\mathbf{x}\|^2 \leq 0 \end{aligned} \quad (3)$$

with  $\dot{W}(\mathbf{x}) = 0$  iff  $\mathbf{x} = \pm\mathbf{v}_k$  where  $\mathbf{v}_k$  is an eigenvector of  $V$ . Also, from (3),  $\dot{W}(\mathbf{x}) = 0$  iff  $(I - \mathbf{x}\mathbf{x}^T)V\mathbf{x} = 0$  i.e.,  $V\mathbf{x}$  is collinear with  $\mathbf{x}$ , which guarantees that the eigenvectors  $\mathbf{v}_k$  of  $V$  (more precisely, on the sphere, the values  $\pm\mathbf{v}_k$  with  $\|\mathbf{v}_k\| = 1$ ) are the equilibria of (2). From LaSalle invariance principle, the only trajectories in the level surfaces of  $W(\mathbf{x})$  are the eigenvectors  $\mathbf{v}_k$  of  $V$ , which guarantees that (3) converges to  $\pm\mathbf{v}_k$  for some  $k = 1, \dots, d$ .

To show convergence to the principal eigenvector  $\mathbf{v}_1$  of  $V$  let us consider the linearization of (2) at  $\mathbf{v}_k$ . Let  $\mathbf{x} = \mathbf{v}_k + \mathbf{u}$  with  $\mathbf{u}$  a small increment s.t.  $\mathbf{u}^T \mathbf{v}_k = 0$  (so that the linearization indeed lies in  $T_{\mathbf{v}_k} \mathbb{S}^{d-1}$ , the tangent plane to the unit sphere at  $\mathbf{v}_k$ ). Computing the linearization, we get

$$\dot{\mathbf{u}} = (V - \lambda_k I)\mathbf{u}. \quad (4)$$

Expressing  $\mathbf{u}$  in the eigenbasis of  $V$ ,  $\mathbf{u} = \sum_{j=1}^d \eta_j \mathbf{v}_j$ , then for  $j \neq k$ , we can project (4) along  $\mathbf{v}_j$  getting the scalar ODE  $\dot{\eta}_j = (\lambda_j - \lambda_k)\eta_j$ . If  $k \neq 1$ , then for one of the projections it must be  $j = 1$ . Since  $\lambda_1$  is a strictly dominant eigenvalue, it is always  $\lambda_1 - \lambda_k > 0$ , i.e.,

each linearization of  $\mathbf{v}_k \neq \mathbf{v}_1$  is unstable. Only when  $\mathbf{v}_k = \mathbf{v}_1$  it is  $\lambda_j - \lambda_1 < 0$  for all  $j$ , meaning that the linearization at  $\mathbf{v}_1$  is locally asymptotically stable. Hence (2) converges (almost always) to  $\pm\mathbf{v}_1$  i.e. to the principal eigenvector of the matrix  $V$ . ■

Notice that, while convergence of  $\pm\mathbf{v}_1$  is generic, also the other eigenvalues  $\mathbf{v}_k$ ,  $k = 2, \dots, d$ , have stable submanifolds of various sizes, always strictly smaller than the ambient manifold  $\mathbb{S}^{d-1}$  (and hence of measure 0).

## 4 Multiagent Oja flow

In this section we propose an extension of the Oja flow in the style of multiagent systems. It corresponds essentially to the self-attention system (1) without the attention matrix. Consider  $n$  vectors  $\mathbf{x}_i \in \mathbb{S}^{d-1}$  obeying the coupled ODEs:

$$\dot{\mathbf{x}}_i = \frac{1}{n}(I - \mathbf{x}_i \mathbf{x}_i^T) \sum_{j=1}^n V \mathbf{x}_j, \quad i = 1, \dots, n. \quad (5)$$

The key difference w.r.t. the Oja flow (2) is that in the right hand side of the ODEs the action of a single agent is replaced by the mean of the  $n$  agents  $\mathbf{m} = \mathbf{m}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$ .

We make the following assumption which holds throughout the rest of the paper.

**Assumption 1** *The value matrix  $V \in \mathbb{R}^{d \times d}$  is symmetric, of eigenvalues  $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_d$  with  $\lambda_1 > 0$  simple and positive.*

Let  $\mathbf{v}_1, \dots, \mathbf{v}_d$  be the associated eigenvectors, normalized s.t.  $\|\mathbf{v}_k\| = 1$ .

### 4.1 Equilibria

Let us begin by expressing the notion of consensus and bipartite consensus for multiagent systems on the sphere that will be used in this paper. The points  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{S}^{d-1}$  are said to be in a *consensus state* if  $\mathbf{x}_i = \mathbf{x}_j = \mathbf{v}_k \forall i, j = 1, \dots, n$  and for some  $k = 1, \dots, d$ . They are said to be in a *bipartite consensus state* if  $\mathbf{x}_i = \pm\mathbf{x}_j = \pm\mathbf{v}_k \forall i, j = 1, \dots, n$  and for some  $k = 1, \dots, d$ .

**Remark 1** W.r.t. the literature, [1, 33, 34], we expressly require a consensus or bipartite consensus point to be aligned with one of the eigenvectors of  $V$ . This choice will be useful when we treat the self-attention dynamics in Section 5.

Let  $\mathbf{y} = V\mathbf{m}(\mathbf{x}) = \frac{1}{n}V \sum_{j=1}^n \mathbf{x}_j \in \mathbb{R}^d$  be the total influence of all agents on agent  $i$  (which is the same for all agents).

**Lemma 2** *The system (5) has the following classes of equilibria:*

1. *consensus:*  $\mathbf{x}_i = \mathbf{v}_k \forall i$ , and  $k = 1, \dots, d$ ;
2. *bipartite consensus:*  $\mathbf{x}_i = \pm\mathbf{v}_k \forall i$  and  $k = 1, \dots, d$ ;
3. *polygonal equilibria:*  $\{\mathbf{x}_i \in \mathbb{S}^{d-1} \text{ s.t. } V \sum_{j=1}^n \mathbf{x}_j = 0\}$ .

**Proof.** The proof follows the reasoning of [3]. Using  $\mathbf{y}$ , (5) becomes

$$\dot{\mathbf{x}}_i = (I - \mathbf{x}_i \mathbf{x}_i^T) \mathbf{y} = \mathbf{y} - \mathbf{x}_i \langle \mathbf{x}_i, \mathbf{y} \rangle, \quad (6)$$

and an equilibrium is a point  $\mathbf{x} \in (\mathbb{S}^{d-1})^n$  in which  $\mathbf{y}$  is collinear with  $\mathbf{x}_i \forall i = 1, \dots, n$ , or in which  $\mathbf{y}$  vanishes. This can happen in 3 cases:

1.  $\mathbf{y} = \gamma_i \mathbf{x}_i$ , for some scalar  $\gamma_i > 0$  (consensus);
2.  $\mathbf{y} = -\gamma_i \mathbf{x}_i$ , for some scalar  $\gamma_i > 0$  (bipartite consensus);
3.  $\mathbf{y} = 0$  (polygonal equilibria).

In fact, in the first two cases, (6) becomes  $\dot{\mathbf{x}}_i = \pm \gamma_i (\mathbf{x}_i - \mathbf{x}_i \mathbf{x}_i^T \mathbf{x}_i) = 0$ , since  $\mathbf{x}_i^T \mathbf{x}_i = 1$ . The third case follows trivially from (6). To show that consensus must be an eigenvector of  $V$ , observe that at this equilibrium point the total influence  $\mathbf{y}$  can be written as  $\mathbf{y} = V \mathbf{x}_i$ , since  $\mathbf{x}_i = \mathbf{x}_j$ . From the expression above, it is also  $\mathbf{y} = \gamma_i \mathbf{x}_i$ . Putting together these two expressions of  $\mathbf{y}$ :  $V \mathbf{x}_i = \gamma_i \mathbf{x}_i$ , i.e.,  $\mathbf{x}_i$  is an eigenvector of  $V$  and  $\gamma_i$  one of its eigenvalues. The argument for bipartite consensus is identical. ■

For a given  $\mathbf{v}_k$  there are  $2^n$  possible consensus or bipartite consensus points. These are always paired by a global symmetry w.r.t. the origin, i.e., if  $\mathbf{x}_i = \mathbf{v}_k \forall i$  is a consensus point, its antipodal point  $\mathbf{x}_i = -\mathbf{v}_k \forall i$  is also a consensus point, and similarly for the bipartite consensus equilibria.

The name polygonal equilibria [3] is due to the observation that  $\mathbf{z}_j = V \mathbf{x}_j$  must sum to 0, hence they must correspond to the vertices of a spherical polygon, see also [16]. Notice that while the collinear equilibria (consensus and bipartite consensus) are isolated points in  $(\mathbb{S}^{d-1})^n$ , polygonal equilibria form instead a set. The set is of zero measure in  $(\mathbb{S}^{d-1})^n$ , as it is determined by algebraic constraints. When  $V$  is invertible, the polygonal equilibria are the manifold  $\{\sum_{j=1}^n \mathbf{x}_j = 0\} \cap (\mathbb{S}^{d-1})^n$ .

## 4.2 Stability analysis

The following theorem summarizes the stability and convergence properties of the multiagent Oja system (5).

**Theorem 1** *For the system (5), under Assumption 1, the consensus equilibrium at the principal eigenvector  $\mathbf{v}_1$  of  $V$  is asymptotically stable, while the other consensus equilibria  $\mathbf{v}_k$ ,  $k = 2, \dots, d$ , the bipartite consensus equilibria and the polygonal equilibria are all unstable. The trajectories of (5) converge to  $\mathbf{v}_1$  for almost all initial conditions  $\mathbf{x}_i(0) \in (\mathbb{S}^{d-1})^n$ .*

To prove this theorem we need a series of preliminary lemmas. We start by computing the Jacobian linearization of (5). Denote  $f_i(\mathbf{x}) = \frac{1}{n}(I - \mathbf{x}_i \mathbf{x}_i^T) \sum_{j=1}^n V \mathbf{x}_j$  and  $f(\mathbf{x}) = [f_1(\mathbf{x})^T \dots f_n(\mathbf{x})^T]^T$  the  $(nd)$ -dimensional vector field associated to the stacked state vector  $\mathbf{x} = [\mathbf{x}_1^T \dots \mathbf{x}_n^T]^T \in (\mathbb{S}^{d-1})^n$ .

**Lemma 3** *The Jacobian of (5),  $F(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[ \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} \right]$  has the following components:*

- *diagonal terms:*

$$\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_i} = \frac{1}{n} \left( (I - \mathbf{x}_i \mathbf{x}_i^T) V - \sum_{j=1}^n (\mathbf{x}_i^T V \mathbf{x}_j I + \mathbf{x}_i \mathbf{x}_j^T V) \right),$$

- *off-diagonal terms:*  $\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} = \frac{1}{n} (I - \mathbf{x}_i \mathbf{x}_i^T) V, h \neq i.$

**Proof.** For the diagonal terms direct calculations give

$$\begin{aligned} \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_i} &= \frac{1}{n} \left( \frac{\partial}{\partial \mathbf{x}_i} V \left( \sum_{j=1}^n \mathbf{x}_j \right) - \frac{\partial}{\partial \mathbf{x}_i} V \left( \mathbf{x}_i \mathbf{x}_i^T V \sum_{j=1}^n \mathbf{x}_j \right) \right) \\ &= \frac{1}{n} \left( V - \mathbf{x}_i^T V \sum_{j=1}^n \mathbf{x}_j I - \mathbf{x}_i \left( \mathbf{x}_i^T V + \sum_{j=1}^n \mathbf{x}_j^T V \right) \right). \end{aligned}$$

For the off-diagonal terms, instead we have

$$\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} = \frac{1}{n} (I - \mathbf{x}_i \mathbf{x}_i^T) V \frac{\partial}{\partial \mathbf{x}_h} \sum_{j=1}^n \mathbf{x}_j.$$

■

The linearization can be used to determine the local stability character of the equilibria of (5).

**Lemma 4** *The consensus point  $\mathbf{x}_i = \mathbf{v}_1 \forall i$  is locally asymptotically stable for (5), while the remaining consensus equilibria  $\mathbf{x}_i = \mathbf{v}_k \forall i$ , with  $k = 2, \dots, d$ , are all unstable. At a consensus equilibrium  $\mathbf{v}_k$ , the eigenvalues of  $F(\mathbf{v}_k)$  are*

1.  $-2\lambda_k$  of multiplicity  $n$ ,
2.  $\lambda_h - \lambda_k$ , for  $h = 1, \dots, k-1, k+1, \dots, d$ ,
3.  $-\lambda_k$  of multiplicity  $nd - n - d + 1$ .

*The bipartite consensus and polygonal equilibria are all unstable.*

**Proof.** To prove this lemma, we follow the same procedure of Lemma 10 of [33]. Remarkably, our system (5) and the (different) system in [33] share the same eigenvectors even though they have different eigenvalues. Notice first that, when computed in an eigenvector  $\mathbf{v}_k$ , the Jacobian of (5) can be compactly expressed using tensor products as

$$\begin{aligned} F(\mathbf{v}_k) &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \Big|_{\mathbf{x}_i=\mathbf{v}_k} \\ &= \frac{1}{n} \mathbf{1} \mathbf{1}^T \otimes (I - \mathbf{v}_k \mathbf{v}_k^T) V - I \otimes (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V). \end{aligned} \tag{7}$$

The first term represents a factor present in all entries of  $F$ , while the second one is present only on the diagonal. For  $F(\mathbf{v}_k)$  there are 3 classes of eigenvectors:

1. The first class is given by  $\mathbf{p}^\ell = [\underbrace{0 \dots 0}_{\ell-1} \mathbf{v}_k^T \underbrace{0 \dots 0}_{n-\ell}]^T$ . There are  $n$  such eigenvectors, and they are obviously all orthogonal to each other. Since

$$(I - \mathbf{v}_k \mathbf{v}_k^T) V \mathbf{v}_k = \lambda_k (\mathbf{v}_k - \mathbf{v}_k \mathbf{v}_k^T \mathbf{v}_k) = 0,$$

and

$$\begin{aligned} (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V) \mathbf{v}_k &= \lambda_k (\mathbf{v}_k^T \mathbf{v}_k \mathbf{v}_k + \mathbf{v}_k \mathbf{v}_k^T \mathbf{v}_k) \\ &= 2\lambda_k \mathbf{v}_k, \end{aligned}$$

it is

$$F(\mathbf{v}_k) \mathbf{p}^\ell = -2\lambda_k \mathbf{p}^\ell, \quad \ell = 1, \dots, n,$$

i.e.,  $\mathbf{p}^\ell$  is an eigenvector of  $F(\mathbf{v}_k)$  of eigenvalue  $-2\lambda_k$ .

2. The second class of eigenvectors is given by  $\mathbf{q}^h = [\mathbf{v}_h^T \dots \mathbf{v}_h^T]^T$ , where  $\mathbf{v}_h$  is an eigenvector of  $V$  associated to  $\lambda_h$  with  $h \neq k$ , so that  $\mathbf{v}_k^T \mathbf{v}_h = 0$ . There are  $d-1$  such  $\mathbf{q}^h$  vectors. Computing

$$\begin{aligned} (I - \mathbf{v}_k \mathbf{v}_k^T) V \mathbf{v}_h &= \lambda_h (\mathbf{v}_h - \mathbf{v}_k \mathbf{v}_k^T \mathbf{v}_h) = \lambda_h \mathbf{v}_h \\ (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V) \mathbf{v}_h &= \lambda_k \mathbf{v}_k^T \mathbf{v}_k \mathbf{v}_h + \lambda_h \mathbf{v}_k \mathbf{v}_k^T \mathbf{v}_h \\ &= \lambda_k \mathbf{v}_h, \end{aligned}$$

hence

$$F(\mathbf{v}_k) \mathbf{q}^h = (\lambda_h - \lambda_k) \mathbf{q}^h, \quad h = 1, \dots, k-1, k+1, \dots, d.$$

3. The remaining  $nd - n - d + 1$  eigenvectors are assembled by considering vectors  $\mathbf{r} = [(\mathbf{z}^1)^T \ (\mathbf{z}^2)^T \ \dots \ (\mathbf{z}^n)^T]^T$  s.t.  $\mathbf{v}_k^T \mathbf{z}^i = 0$  and  $\mathbf{v}_h^T \mathbf{z}^i = 0$  for all  $i = 1, \dots, n$  and all  $h = 1, \dots, k-1, k+1, \dots, d$ . Since the number of such constraints is  $d-1+n$ , there exist  $nd - n - d + 1$  such vectors  $\mathbf{r}$ . Notice that the  $\mathbf{z}^i$  can always be chosen so that  $\sum_{i=1}^n \mathbf{z}^i = 0$ . Computing

$$\begin{aligned} (I - \mathbf{v}_k \mathbf{v}_k^T) V \mathbf{z}^i &= V \mathbf{z}^i - \lambda_k \mathbf{v}_k \mathbf{v}_k^T \mathbf{z}^i = V \mathbf{z}^i \\ (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V) \mathbf{z}^i &= \lambda_k \mathbf{v}_k^T \mathbf{v}_k \mathbf{z}^i + \lambda_k \mathbf{v}_k \mathbf{v}_k^T \mathbf{z}_i \\ &= \lambda_k \mathbf{z}^i, \end{aligned}$$

hence

$$F(\mathbf{v}_k) \mathbf{r} = \begin{bmatrix} V \sum_i \mathbf{z}^i - \lambda_k \mathbf{z}^1 \\ \vdots \\ V \sum_i \mathbf{z}^i - \lambda_k \mathbf{z}^n \end{bmatrix} = -\lambda_k \mathbf{r}.$$

Therefore the eigenvalue  $-\lambda_k$  has multiplicity  $nd - n - d + 1$  for  $F(\mathbf{v}_k)$ .

When  $k > 1$ , then at least one of the eigenvalues of the second class is  $\lambda_1 - \lambda_k > 0$ , hence the equilibrium  $\mathbf{v}_k$  is unstable. When  $k = 1$ , as by assumption  $\lambda_1 > 0$  and  $\lambda_1 > \lambda_k$  for all  $k = 2, \dots, d$ , all eigenvalues of  $F(\mathbf{v}_1)$  are negative, meaning that  $\mathbf{v}_1$  is locally asymptotically stable for (5).

Consider now a bipartite consensus point  $\mathbf{x}_i = \pm \mathbf{v}_k$ . Assume that  $n_1$  agents are equal to  $\mathbf{v}_k$  and  $n_2 = n - n_1$  agent to  $-\mathbf{v}_k$ . Denote  $\mathcal{V}_1$  and  $\mathcal{V}_2$  the associated sets of indices. In the Jacobian matrix  $F(\mathbf{x})$  computed at such bipartite consensus we have now the following cases:

- if  $i \in \mathcal{V}_1$  and  $h = i$ :

$$[F(\mathbf{v}_k)]_{ih} = \frac{1}{n}(I - \mathbf{v}_k \mathbf{v}_k^T) - \frac{n_1 - n_2}{n}(\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V)$$

- if  $i \in \mathcal{V}_2$  and  $h = i$ :

$$[F(\mathbf{v}_k)]_{ih} = \frac{1}{n}(I - \mathbf{v}_k \mathbf{v}_k^T) + \frac{n_1 - n_2}{n}(\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V)$$

- if  $h \neq i$

$$[F(\mathbf{v}_k)]_{ih} = \frac{1}{n}(I - \mathbf{v}_k \mathbf{v}_k^T)$$

i.e., some of the diagonal terms switch sign. In compact form, assuming that the first  $n_1$  indices are in  $\mathcal{V}_1$  and the remaining  $n_2$  in  $\mathcal{V}_2$ ,

$$\begin{aligned} F(\mathbf{v}_k) &= \frac{1}{n} \mathbf{1} \mathbf{1}^T \otimes (I - \mathbf{v}_k \mathbf{v}_k^T) V - \frac{n_1 - n_2}{n} \\ &\cdot \text{diag}[\underbrace{1 \dots 1}_{n_1 \text{ times}} \underbrace{-1 \dots -1}_{n_2 \text{ times}}] \otimes (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V). \end{aligned}$$

Computing eigenvalues in the first of the three classes mentioned above, we have

$$F(\mathbf{v}_k) \mathbf{p}^\ell = \begin{cases} -\frac{2(n_1 - n_2)}{n} \lambda_k \mathbf{p}^\ell & \text{if } k \in \mathcal{V}_1 \\ \frac{2(n_1 - n_2)}{n} \lambda_k \mathbf{p}^\ell & \text{if } k \in \mathcal{V}_2. \end{cases}$$

Regardless of the sign of  $\lambda_k$  and of the cardinality of the  $\mathcal{V}_1/\mathcal{V}_2$  partition, the bipartite consensus point is always unstable, since both  $\pm \frac{2(n_1 - n_2)}{n} \lambda_k$  are eigenvalues.

Consider now a polygonal equilibrium point  $\mathbf{x} = \mathbf{s} = [\mathbf{s}_1^T \dots \mathbf{s}_n^T]^T \in (\mathbb{S}^{d-1})^n$  where  $\mathbf{s}$  is s.t.  $\mathbf{y} = \frac{1}{n} V \sum_{j=1}^n \mathbf{s}_j = 0$ . Let us compute the linearization of (5) at  $\mathbf{s}$ , obtained perturbing  $\mathbf{s}$  with a perturbation  $\mathbf{x} = \mathbf{s} + \mathbf{u}$  belonging to the tangent space  $T_{\mathbf{s}_i} \mathbb{S}^{d-1}$  of each agent: if  $\mathbf{u} = [\mathbf{u}_1 \dots \mathbf{u}_n]$ , it is  $\mathbf{s}_i^T \mathbf{u}_i = 0$ ,  $i = 1, \dots, n$  and

$$\begin{aligned} \dot{\mathbf{u}}_i &= \frac{1}{n} (I - (\mathbf{s}_i + \mathbf{u}_i)(\mathbf{s}_i + \mathbf{u}_i)^T) V \sum_{j=1}^n (\mathbf{s}_j + \mathbf{u}_j) \\ &\approx \frac{1}{n} (I - \mathbf{s}_i \mathbf{s}_i^T) V \underbrace{\sum_{j=1}^n \mathbf{s}_j}_{=0} + (I - \mathbf{s}_i \mathbf{s}_i^T) V \sum_{j=1}^n \mathbf{u}_j \\ &\quad - (\mathbf{u}_i \mathbf{s}_i^T + \mathbf{s}_i \mathbf{u}_i^T) V \underbrace{\sum_{j=1}^n \mathbf{s}_j}_{=0} + h.o.t. \end{aligned}$$

Recall that  $V$  has always at least one positive eigenvalue  $\lambda_1 > 0$ . We can always choose  $\mathbf{u}$  s.t. each  $\mathbf{u}_j$  has a nonzero component in the direction of the associated eigenvector  $\mathbf{v}_1$ :

expanding in the eigenbasis of  $V$ ,  $\mathbf{u}_j = \sum_{k=1}^d \eta_k^j \mathbf{v}_k$  with  $\eta_1^j \neq 0$ . If  $\mathbf{s}_i = \sum_{k=1}^d \zeta_k^i \mathbf{v}_k$ , and using orthogonality,

$$\begin{aligned}\dot{\mathbf{u}}_i &= \sum_k \dot{\eta}_k^i \mathbf{v}_k \\ &= \frac{1}{n} \left( \left( I - \sum_k \zeta_k^i \mathbf{v}_k \sum_\ell \zeta_\ell^i \mathbf{v}_\ell^T \right) V \sum_j \sum_k \eta_k^j \mathbf{v}_k \right) \\ &= \frac{1}{n} \left( \left( I - \sum_k \zeta_k^i \mathbf{v}_k \sum_\ell \zeta_\ell^i \mathbf{v}_\ell^T \right) \sum_j \sum_k \eta_k^j \lambda_k \mathbf{v}_k \right) \\ &= \frac{1}{n} \left( \sum_{j,k} \eta_k^j \lambda_k \mathbf{v}_k - \sum_k \zeta_k^i \mathbf{v}_k \sum_\ell \zeta_\ell^i \lambda_\ell \sum_j \eta_\ell^j \right)\end{aligned}$$

or, projecting along  $\mathbf{v}_1$ ,

$$\dot{\eta}_1^i = \frac{1}{n} \left( \sum_j \eta_1^j \lambda_1 - \zeta_1^i \sum_\ell \zeta_\ell^i \lambda_\ell \sum_j \eta_\ell^j \right).$$

To conclude the argument, it is enough to show that instability occurs along a specific  $\mathbf{u}$ . One such direction is  $\mathbf{u}$  that perturbs  $\mathbf{s}$  only along the first eigenvector  $\mathbf{v}_1$  for each  $i = 1, \dots, n$ , i.e., for all  $i$ ,  $\eta_1^i \neq 0$  and  $\eta_k^i = 0$  for  $k = 2, \dots, d$ . In this case in fact we get

$$\dot{\eta}_1^i = \frac{1}{n} (1 - (\zeta_1^i)^2) \lambda_1 \sum_j \eta_1^j,$$

or, in vector form (collecting only the  $\eta_1^i$  components,  $\boldsymbol{\eta}_1 = [\eta_1^1 \ \dots \ \eta_1^n]^T$ ),

$$\dot{\boldsymbol{\eta}}_1 = \frac{\lambda_1}{n} (I - \Psi_1^2) \mathbf{1} \mathbf{1}^T \boldsymbol{\eta}_1$$

where  $\Psi_1 = \text{diag}(\zeta_1^1, \dots, \zeta_1^n)$ . The matrix  $\mathbf{1} \mathbf{1}^T$  is a rank-1 matrix of eigenvalues 1 and 0, while  $|\zeta_1^i| < 1$  because  $\mathbf{s}$  is not an eigenvector of  $V$ , hence  $I - \Psi_1^2 \succ 0$ . Since  $\lambda_1 > 0$ , the matrix  $\frac{\lambda_1}{n} (I - \Psi_1^2) \mathbf{1} \mathbf{1}^T$  has at least one positive eigenvalue, hence  $\mathbf{s}$  is an unstable equilibrium for (5). ■

Finally, the following lemma shows that a Lyapunov stability argument can be set up for (5).

**Lemma 5** *The function  $W(\mathbf{x}) = \frac{1}{2} \left( \lambda_1 - \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j^T V \sum_{j=1}^n \mathbf{x}_j \right)$  is a Lyapunov function for (5) and guarantees that (5) converges to one of the equilibria determined in Lemma 2.*

**Proof.** Consider  $\mathbf{m} = \mathbf{m}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$  and  $\mathbf{y} = V\mathbf{m}$ . Differentiating  $W(\mathbf{x}) =$

$\frac{1}{2} (\lambda_1 - \frac{1}{n} \mathbf{m}^T V \mathbf{m})$  gives

$$\begin{aligned}
\dot{W}(\mathbf{x}) &= -\frac{1}{n} \mathbf{m}^T V \sum_{i=1}^n (I - \mathbf{x}_i \mathbf{x}_i^T) V \mathbf{m} \\
&= -\frac{n}{n} \mathbf{m}^T V V \mathbf{m} + \frac{1}{n} \mathbf{m}^T V \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T V \mathbf{m} \\
&= -\mathbf{y}^T \mathbf{y} + \frac{1}{n} \mathbf{y}^T \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mathbf{y} \\
&\leq -\|\mathbf{y}\|^2 + \frac{1}{n} \mu_{\max} \left[ \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right] \|\mathbf{y}\|^2 \\
&= \left( -1 + \frac{n}{n} \right) \|\mathbf{y}\|^2 = 0
\end{aligned}$$

where we have used that each psd matrix  $\mathbf{x}_i \mathbf{x}_i^T \preceq I$  since it projects onto the direction  $\mathbf{x}_i$ , and hence  $0 \preceq \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \preceq nI$ , from which  $0 \leq \mu_k [\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T] \leq n$  for all eigenvalues of  $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ .

Since the state space is compact, trajectories exist for all  $t$  and have limit points. Hence LaSalle's invariance principle applies. In particular, trajectories converge to the largest invariant set contained in

$$\begin{aligned}
\mathcal{L} &= \{\mathbf{x} \in (\mathbb{S}^{d-1})^n \text{ s. t. } \dot{W}(\mathbf{x}) = 0\} \\
&= \{\mathbf{y} \text{ s. t. } n\|\mathbf{y}\|^2 = \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{y})^2\}.
\end{aligned}$$

For the system (5),  $\mathbf{x} \in \mathcal{L}$  when  $\mathbf{y} = 0$  or, from the calculations above for  $\dot{W}$ , when equality holds in  $\mu_{\max} [\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T] \leq n$ , i.e., when all  $\mathbf{x}_i$  are collinear:  $\mathbf{x}_i = \pm \mathbf{v}$  for some  $\mathbf{v}$ . In fact, in this case,  $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T = n \mathbf{v} \mathbf{v}^T$ . If  $\mathbf{v} = \mathbf{v}_k$  for some  $k = 1, \dots, d$ , then we have an invariant point. If instead  $\mathbf{v} \neq \mathbf{v}_k$ ,  $k = 1, \dots, d$  (i.e., the consensus point is not an equilibrium), then it is  $\dot{\mathbf{x}}_i|_{\mathbf{x}_i=\mathbf{v}} = (I - \mathbf{v} \mathbf{v}^T) V \mathbf{v} = \dot{\mathbf{x}}_j|_{\mathbf{x}_j=\mathbf{v}} \neq 0$ , for all  $i, j$ , and the dynamics become  $n$  identical copies of the Oja flow (2). From Lemma 1, all these  $n$  copies converge to  $\pm \mathbf{v}_k$  for some  $k$  (almost always to  $\pm \mathbf{v}_1$ ). Summarizing, the largest invariant set in  $\mathcal{L}$  is given by the set of equilibria of (5), hence all trajectories of (5) converge to one of the equilibria in  $\mathcal{L}$  computed in Lemma 2. ■

**Proof of Theorem 1.** From Lemma 5 all trajectories converge to one of the equilibria computed in Lemma 2. Lemma 4 says that only the consensus point  $\mathbf{v}_1$  corresponding to the principal eigenvalue of  $V$  is locally asymptotically stable, while all other equilibria are unstable. Since all trajectories of (5) have a limit point, generically such limit point must be  $\mathbf{v}_1$ . ■

**Remark 2** Even though all equilibria except one (or one pair, if one counts also the antipodal point) are unstable, they typically have a basin of attraction associated to a stable submanifold. These stable submanifolds are however necessarily of measure 0 in  $(\mathbb{S}^{d-1})^n$ , and hence so must be the basins of attraction of the unstable equilibria.

## 5 Self-Attention dynamics

The self-attention system (1) differs from (5) in the fact that in the right hand side of the ODEs, the average of the states of the  $n$  agents,  $\frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$ , is replaced by a weighted average  $\mathbf{m} = \frac{1}{n} \sum_{j=1}^n A_{ij}(\mathbf{x})\mathbf{x}_j$ , with (state-dependent) weights given by the attention coefficients.

### 5.1 Equilibria

Apart from the three classes of equilibria already obtained for the multiagent Oja flow, in the self-attention dynamics we have an extra class, due to the fact that the total influence on agent  $i$  of all other agents  $\mathbf{y}_i = \mathbf{y}_i(\mathbf{x}) = V \sum_{j=1}^n A_{ij}(\mathbf{x})\mathbf{x}_j$  now differs from agent to agent.

**Lemma 6** *The system (1) has the following classes of equilibria*

1. *consensus:*  $\mathbf{x}_i = \mathbf{v}_k \forall i$ , and  $k = 1, \dots, d$ ;
2. *bipartite consensus:*  $\mathbf{x}_i = \pm \mathbf{v}_k \forall i$  and  $k = 1, \dots, d$ ;
3. *polygonal equilibria:*  $\{\mathbf{x}_i \in \mathbb{S}^{d-1} \text{ s.t. } V \sum_{j=1}^n A_{ij}(\mathbf{x})\mathbf{x}_j = 0\}$ ,  $i = 1, \dots, n$ ;
4. *clustering equilibria:*  $\{\mathbf{x}_i \in \mathbb{S}^{d-1} \text{ s.t. } \gamma_i \mathbf{x}_i = V \sum_{j=1}^n A_{ij}(\mathbf{x})\mathbf{x}_j\}$  for some scalars  $\gamma_i$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, d$ .

**Proof.** For the first three cases, the proof is identical to that of Lemma 2, provided that  $\mathbf{y}$  is replaced by  $\mathbf{y}_i = \sum_{j=1}^n A_{ij}(\mathbf{x})V\mathbf{x}_j$ . Concerning the clustering equilibria, these correspond to  $\mathbf{y}_i \in \ker(I - \mathbf{x}_i\mathbf{x}_i^T)$ , or, equivalently,  $\mathbf{y}_i$  aligned with  $\mathbf{x}_i$ ,  $i = 1, \dots, n$  but not necessarily aligned with an eigenvector  $\mathbf{v}_k$ , i.e.,  $\mathbf{y}_i = \gamma_i \mathbf{x}_i$  for some scalar  $\gamma_i$ , but possibly  $\mathbf{y}_i \not\parallel \mathbf{v}_k \forall k$ . ■

**Remark 3** Clustering equilibria own their name to the fact that typically multiple agents are found at the same value. In particular we say that  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are at an  $m$ -clustering equilibrium point if  $\mathbf{x}_i \in \{\mathbf{w}_1, \dots, \mathbf{w}_m\}$  with  $1 \leq m \leq n$ , i.e., if  $\mathbf{x}_1, \dots, \mathbf{x}_n$  cluster at the  $m$  vectors  $\mathbf{w}_1, \dots, \mathbf{w}_m$ . Notice that some  $\mathbf{w}_i$  can be eigenvectors of  $V$ .

**Remark 4** A clustering equilibrium can be a consensus, i.e.,  $\mathbf{x}_i = \mathbf{x}_j$  for all  $i, j$ , but with  $\mathbf{x}_i$  which is not aligned with any eigenvector of  $V$ , i.e.,  $\mathbf{x}_i \not\parallel \mathbf{v}_k \forall k$ . We refer to these as 1-clustering, while the “consensus” characterization is reserved for the case  $\mathbf{x}_i \parallel \mathbf{v}_k$ . A 2-clustering instead typically is s.t.  $\mathbf{x}_i \in \{\mathbf{w}_1, \mathbf{w}_2\} \forall i$ , with  $\mathbf{w}_1 \neq -\mathbf{w}_2$ .

**Proposition 1** *The system (1) has clustering equilibria iff  $\exists$  scalars  $\gamma_1, \dots, \gamma_n$  s.t. the matrix  $(I_n \otimes I_d - (\Gamma \otimes I_d)(A(\mathbf{x}) \otimes I_d)(I_n \otimes V))$  is singular, where  $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_n)$ .*

**Proof.** In vector form, the clustering equilibrium condition  $\mathbf{x}_i = \gamma_i V \sum_{j=1}^n A_{ij}(\mathbf{x})\mathbf{x}_j$  becomes  $\mathbf{x} = (\Gamma \otimes I_d)(A(\mathbf{x}) \otimes I_d)(I_n \otimes V)\mathbf{x}$ , where  $\mathbf{x}^T = [\mathbf{x}_1^T \dots \mathbf{x}_n^T]$ . Rewriting it as  $(I_n \otimes I_d - (\Gamma \otimes I_d)(A(\mathbf{x}) \otimes I_d)(I_n \otimes V))\mathbf{x} = 0$ , the algebraic equation has nontrivial solutions if and only if the matrix  $(I_n \otimes I_d - (\Gamma \otimes I_d)(A(\mathbf{x}) \otimes I_d)(I_n \otimes V))$  is singular. ■

**Remark 5** From the proof of Proposition 1, it follows that if  $\mathbf{x}_1, \dots, \mathbf{x}_n$  form an  $m$ -clustering equilibrium point, since  $A(\mathbf{x}) = A(-\mathbf{x})$ , then also the antipodal point  $-\mathbf{x}_1, \dots, -\mathbf{x}_n$  is an  $m$ -clustering equilibrium point. However, ‘bipartite’ versions of the clustering equilibrium (in which only some  $\mathbf{x}_i$  flip sign) are typically not equilibria.

While the attention matrix  $A(\mathbf{x})$  is a function of the state even at the equilibrium point (with the exception of a consensus state), its rank is however fixed for various classes of equilibria.

**Lemma 7** For a consensus state  $\mathbf{x}_i = \mathbf{v}_k \forall i$  we have  $A_{ij}(\mathbf{x}) = \frac{1}{n}$ , i.e., the attention matrix  $A(\mathbf{x})$  is the rank-1 uniform matrix  $A(\mathbf{x}) = \frac{1}{n}\mathbf{1}\mathbf{1}^T$ . For a bipartite consensus state  $\mathbf{x}_i = \pm \mathbf{v}_k \forall i$ , the attention matrix  $A(\mathbf{x})$  has rank 2. For an  $m$ -clustering equilibrium  $\mathbf{x}$  the rank of  $A(\mathbf{x})$  is  $m$ .

**Proof.** At a consensus equilibrium  $\mathbf{x}_i = \mathbf{v}_k$  for all  $i$ , and  $A_{ij}(\mathbf{x})$  is composed of all equal terms

$$A_{ij}(\mathbf{x}) = \frac{e^{\mathbf{v}_k^T Q^T K \mathbf{v}_k}}{\sum_{\ell=1}^n e^{\mathbf{v}_k^T Q^T K \mathbf{v}_\ell}} = \frac{1}{n}.$$

At a bipartite consensus point  $\mathbf{x}_i = \pm \mathbf{v}_k$ , split the  $n$  tokens into two sets  $\mathcal{V}_1$  and  $\mathcal{V}_2$ ,  $\mathcal{V}_1 \cup \mathcal{V}_2 = \{1, \dots, n\}$ , according to whether  $\mathbf{x}_i = \mathbf{v}_k$  or  $\mathbf{x}_i = -\mathbf{v}_k$  at equilibrium. Assume that  $n_1 = |\mathcal{V}_1|$  tokens are equal to  $\mathbf{v}_k$  and  $n_2 = |\mathcal{V}_2|$  equal to  $-\mathbf{v}_k$ , with  $n_1 + n_2 = n$ .

Four different entries appear in  $A_{ij}(\mathbf{x})$ , two due to the denominator (depending on whether  $\mathbf{x}_i = \mathbf{v}_k$  or  $\mathbf{x}_i = -\mathbf{v}_k$ ), and two due to the numerator (depending on whether  $\mathbf{x}_i$  and  $\mathbf{x}_j$  have the same sign). More specifically, denoting  $\alpha_1^k = e^{\mathbf{v}_k^T Q^T K \mathbf{v}_k}$  and  $\alpha_2^k = e^{-\mathbf{v}_k^T Q^T K \mathbf{v}_k}$ , then the entries of the attention matrix are

$$A_{ij}(\mathbf{x}) = \begin{cases} \frac{\alpha_1^k}{n_1\alpha_1^k + n_2\alpha_2^k} & \text{if } i \in \mathcal{V}_1, j \in \mathcal{V}_1 \\ \frac{\alpha_2^k}{n_1\alpha_1^k + n_2\alpha_2^k} & \text{if } i \in \mathcal{V}_1, j \in \mathcal{V}_2 \\ \frac{\alpha_2^k}{n_1\alpha_2^k + n_2\alpha_1^k} & \text{if } i \in \mathcal{V}_2, j \in \mathcal{V}_1 \\ \frac{\alpha_1^k}{n_1\alpha_2^k + n_2\alpha_1^k} & \text{if } i \in \mathcal{V}_2, j \in \mathcal{V}_2. \end{cases}$$

Letting  $\beta_1^k = n_1\alpha_1^k + n_2\alpha_2^k$ , and  $\beta_2^k = n_1\alpha_2^k + n_2\alpha_1^k$  and assuming w.l.o.g. that the first  $n_1$  agents are in  $\mathcal{V}_1$  and the last  $n_2$  in  $\mathcal{V}_2$ , the attention matrix is

$$A(\mathbf{x}) = \begin{bmatrix} \frac{1}{\beta_1^k}(\alpha_1^k & \dots & \dots & \alpha_1^k & \alpha_2^k & \dots & \alpha_2^k) \\ \vdots & & & \vdots & \vdots & & \vdots \\ \frac{1}{\beta_1^k}(\alpha_1^k & \dots & \dots & \alpha_1^k & \alpha_2^k & \dots & \alpha_2^k) \\ \frac{1}{\beta_2^k}(\alpha_2^k & \dots & \dots & \alpha_2^k & \alpha_1^k & \dots & \alpha_1^k) \\ \vdots & & & \vdots & \vdots & & \vdots \\ \frac{1}{\beta_2^k}(\alpha_2^k & \dots & \dots & \alpha_2^k & \alpha_1^k & \dots & \alpha_1^k) \end{bmatrix}, \quad (8)$$

from which it is obvious that the rank must be 2.

As for an  $m$ -clustering equilibrium point with vectors  $\mathbf{w}_1, \dots, \mathbf{w}_m$ , following a similar procedure it is easy to realize that since  $\mathbf{w}_i \neq \mathbf{w}_j$  and  $Q^T K$  is not symmetric, there are  $m^2$  different entries in the numerators of  $A_{ij}$ , and only  $m$  in the denominators. Rearranging

the entries as in (8), each row of  $A$  has exactly  $m$  different terms, and a block counting argument leads to  $\text{rank}(A) = m$ .  $\blacksquare$

## 5.2 Stability analysis

In this section we study the stability properties of three of the four classes of equilibria of the self-attention dynamics (1). Theorem 2 summarizes our main results. Before stating it, we need some extra notation. Consider a bipartite consensus equilibrium associated with the eigenvalue  $\mathbf{v}_k$  of  $V$ , and compute  $\alpha_i^k$  and  $\beta_i^k$ . Denote  $\delta_1^k = \frac{n_1\alpha_1^k - n_2\alpha_2^k}{\beta_1^k}$  and  $\delta_2^k = \frac{n_2\alpha_1^k - n_1\alpha_2^k}{\beta_2^k}$ .

**Theorem 2** *For the self-attention dynamics (1), under Assumption 1, the consensus equilibrium associated to the principal eigenvector  $\mathbf{v}_1$  of  $V$  is always asymptotically stable, while the other consensus equilibria  $\mathbf{v}_k$ ,  $k = 2, \dots, d$  are all unstable. A bipartite consensus equilibrium  $\mathbf{x}_i = \pm \mathbf{v}_k \forall i$  is asymptotically stable iff  $\delta_\ell^k \lambda_k > 0$ ,  $\ell = 1, 2$ , and the following inequalities are satisfied  $\forall j = 1, \dots, k-1, k+1, \dots, d$ :*

$$\begin{aligned} & \left(1 - \frac{\alpha_2^k(n_1^2 + n_2^2)}{2\alpha_1^k n_1 n_2}\right) \lambda_j - \left(1 - \frac{(\alpha_2^k)^2}{(\alpha_1^k)^2}\right) \lambda_k < 0 \\ & \delta_1^k \delta_2^k \lambda_k^2 + \left(\delta_2^k \frac{n_1 \alpha_1^k}{\beta_1^k} + \delta_1^k \frac{n_2 \alpha_1^k}{\beta_2^k}\right) \lambda_j \lambda_k \\ & + \frac{n_1 n_2}{\beta_1^k \beta_2^k} ((\alpha_2^k)^2 - (\alpha_1^k)^2) \lambda_j^2 < 0. \end{aligned} \quad (9)$$

All polygonal equilibria are unstable.

The proof is based only on Lyapunov indirect method. As for the multiagent Oja flow, it is broken down into various lemmas. Letting, as in Section 4,  $f_i(\mathbf{x})$  be the right-hand side of (1) and  $f(\mathbf{x})$  its vectorization, we can compute the Jacobian of (1) as follows.

**Lemma 8** *The Jacobian of (1),  $F(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[ \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} \right]$  has the following components:*

- *diagonal terms:*

$$\begin{aligned} \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_i} &= (I - \mathbf{x}_i \mathbf{x}_i^T) V \left( (I + \mathbf{x}_i \mathbf{x}_i^T Q^T K) A_{ii}(\mathbf{x}) \right. \\ &\quad + \sum_j \mathbf{x}_j (\mathbf{x}_j^T K^T Q - \mathbf{x}_i^T Q^T K A_{ii}(\mathbf{x})) \\ &\quad - \sum_\ell \mathbf{x}_\ell^T K^T Q A_{i\ell}(\mathbf{x}) \Big) A_{ij}(\mathbf{x}) \\ &\quad - \sum_j (\mathbf{x}_i^T V \mathbf{x}_j I + \mathbf{x}_i \mathbf{x}_j^T V) A_{ij}(\mathbf{x}); \end{aligned}$$

- *off-diagonal terms:*

$$\begin{aligned} \frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} &= (I - \mathbf{x}_i \mathbf{x}_i^T) V \left( I + \mathbf{x}_k \mathbf{x}_i^T Q^T K \right. \\ &\quad \left. - \sum_j \mathbf{x}_j \mathbf{x}_i^T Q^T K A_{ij}(\mathbf{x}) \right) A_{ih}(\mathbf{x}). \end{aligned}$$

**Proof.** For simplicity of notation we write  $A_{ij}$  instead of  $A_{ij}(\mathbf{x})$ . For the diagonal terms we have

$$\begin{aligned}\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_i} &= V \sum_j \left( \mathbf{x}_j \frac{\partial A_{ij}}{\partial \mathbf{x}_i} + A_{ij} \frac{\partial \mathbf{x}_j}{\partial \mathbf{x}_i} \right) \\ &\quad - \mathbf{x}_i^T V \sum_j A_{ij} \mathbf{x}_j - \mathbf{x}_i \frac{\partial}{\partial \mathbf{x}_i} \left( \mathbf{x}_i^T V \sum_j \mathbf{x}_j A_{ij} \right) \\ &= V \sum_j \mathbf{x}_j \frac{\partial A_{ij}}{\partial \mathbf{x}_i} + A_{ii} I - \mathbf{x}_i^T V \sum_j A_{ij} \mathbf{x}_j I \\ &\quad - \mathbf{x}_i \left( \sum_j \mathbf{x}_j^T V A_{ij} + \mathbf{x}_i^T V A_{ii} + \mathbf{x}_i^T V \sum_j \mathbf{x}_j \frac{\partial A_{ij}}{\partial \mathbf{x}_i} \right)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial A_{ij}}{\partial \mathbf{x}_i} &= \frac{\partial}{\partial \mathbf{x}_i} \frac{e^{\langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle}}{\sum_{\ell=1}^n e^{\langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle}} \\ &= \left( \mathbf{x}_j^T K^T Q e^{\langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle} \sum_{\ell} e^{\langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle} \right. \\ &\quad + \mathbf{x}_i^T Q^T K e^{\langle Q\mathbf{x}_i, K\mathbf{x}_i \rangle} \sum_{\ell} e^{\langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle} \delta_{ij} \\ &\quad \left. - e^{\langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle} \left( \sum_{\ell} \mathbf{x}_{\ell}^T K^T Q e^{\langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle} \right. \right. \\ &\quad \left. \left. + \mathbf{x}_i^T Q^T K e^{\langle Q\mathbf{x}_i, K\mathbf{x}_i \rangle} \right) \right) / \left( \sum_{\ell} e^{\langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle} \right)^2 \\ &= \left( \mathbf{x}_j^T K^T Q - \mathbf{x}_i^T Q^T K A_{ii} - \sum_{\ell} \mathbf{x}_{\ell}^T K^T Q A_{i\ell} \right) A_{ij} \\ &\quad + \mathbf{x}_i^T Q^T K A_{ii} \delta_{ij}\end{aligned}\tag{10}$$

with  $\delta_{ij} = 1$  if  $i = j$  and 0 otherwise. For the off-diagonal terms, instead we have

$$\begin{aligned}\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}_h} &= (I - \mathbf{x}_i \mathbf{x}_i^T) V \sum_j \frac{\partial}{\partial \mathbf{x}_h} (\mathbf{x}_j A_{ij}) \\ &= (I - \mathbf{x}_i \mathbf{x}_i^T) V \left( A_{ih} I + \sum_j \mathbf{x}_j \frac{\partial A_{ij}}{\partial \mathbf{x}_h} \right),\end{aligned}$$

where

$$\frac{\partial A_{ij}}{\partial \mathbf{x}_h} = \mathbf{x}_i^T Q^T K (A_{ih} \delta_{jh} - A_{ij} A_{ih}).\tag{11}$$

■

Since, from Lemma 7, in a consensus equilibrium  $A(\mathbf{v}_k) = \frac{1}{n} \mathbf{1} \mathbf{1}^T$ , it is a straightforward computation to show that  $F(\mathbf{v}_k)$  is still given by (7), and hence that  $\mathbf{v}_1$  is locally asymptotically stable while  $\mathbf{v}_k$ ,  $k = 2, \dots, d$ , are all unstable. This is stated in the following lemma.

**Lemma 9** *The consensus point  $\mathbf{x}_i = \mathbf{v}_1 \forall i$  is locally asymptotically stable for (1), while the remaining consensus equilibria  $\mathbf{x}_i = \mathbf{v}_k \forall i$ , with  $k = 2, \dots, d$ , are all unstable. At a consensus point  $\mathbf{v}_k$ , the eigenvalues of  $F(\mathbf{v}_k)$  are those indicated in Lemma 4.*

The analysis of a bipartite consensus is instead more complicated. As in the proof of Lemma 7, we assume that for the first  $n_1$  agents it is  $\mathbf{x}_i = \mathbf{v}_k$ , while for the last  $n_2$  it is  $\mathbf{x}_i = -\mathbf{v}_k$ . In this way the Jacobian has a bipartite structure that reflects that of (8), plus diagonal blocks.

**Lemma 10** *At a bipartite consensus equilibrium  $\mathbf{x}_i = \pm \mathbf{v}_k$ , the Jacobian can be expressed as*

$$F(\mathbf{v}_k) = \begin{bmatrix} F_d^1 & & & & \\ & \ddots & & & \\ & & F_d^1 & & \\ & & & F_d^2 & \\ & & & & \ddots \\ & & & & & F_d^2 \end{bmatrix} + \begin{bmatrix} F_o^{11} & \dots & F_o^{11} & F_o^{12} & \dots & F_o^{12} \\ \vdots & & \vdots & \vdots & & \vdots \\ F_o^{11} & \dots & F_o^{11} & F_o^{12} & \dots & F_o^{12} \\ F_o^{21} & \dots & F_o^{21} & F_o^{22} & \dots & F_o^{22} \\ \vdots & & \vdots & \vdots & & \vdots \\ F_o^{21} & \dots & F_o^{21} & F_o^{22} & \dots & F_o^{22} \end{bmatrix},$$

where

$$\begin{aligned} F_o^{11} &= \frac{\alpha_1^k}{\beta_1^k}(I - \mathbf{v}_k \mathbf{v}_k^T)V, \quad i, h \in \mathcal{V}_1 \\ F_o^{12} &= \frac{\alpha_2^k}{\beta_1^k}(I - \mathbf{v}_k \mathbf{v}_k^T)V, \quad i \in \mathcal{V}_1, h \in \mathcal{V}_2 \\ F_o^{21} &= \frac{\alpha_2^k}{\beta_2^k}(I - \mathbf{v}_k \mathbf{v}_k^T)V, \quad i \in \mathcal{V}_2, h \in \mathcal{V}_1 \\ F_o^{22} &= \frac{\alpha_1^k}{\beta_2^k}(I - \mathbf{v}_k \mathbf{v}_k^T)V, \quad i, h \in \mathcal{V}_2 \\ F_d^1 &= -\delta_1^k (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V), \quad i = h \in \mathcal{V}_1 \\ F_d^2 &= -\delta_2^k (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V), \quad i = h \in \mathcal{V}_2. \end{aligned}$$

The eigenvalues of  $F(\mathbf{v}_k)$  are

1.  $-2\delta_1^k \lambda_k$  of multiplicity  $n_1$  and  $-2\delta_2^k \lambda_k$  of multiplicity  $n_2$ ;
2.  $\gamma_{j,\pm}^k = \frac{1}{2} \left( a_j^k + d_j^k \pm \sqrt{(a_j^k - d_j^k)^2 + 4c_j^k b_j^k} \right)$  for  $j = 1, \dots, k-1, k+1, \dots, d$ , where

$a_j^k$ ,  $b_j^k$ ,  $c_j^k$  and  $d_j^k$  are given by

$$\begin{aligned} a_j^k &= -\delta_1^k \lambda_k + \lambda_j \frac{n_1 \alpha_1^k}{\beta_1^k}, & b_j^k &= \lambda_j \frac{n_2 \alpha_2^k}{\beta_1^k} \\ c_j^k &= \lambda_j \frac{n_1 \alpha_2^k}{\beta_2^k}, & d_j^k &= -\delta_2^k \lambda_k + \lambda_j \frac{n_2 \alpha_1^k}{\beta_2^k}. \end{aligned}$$

3.  $-\delta_1^k \lambda_k$  and  $-\delta_2^k \lambda_k$  of total multiplicity  $nd - n - 2d + 2$ .

The bipartite consensus equilibrium point is locally asymptotically stable iff  $\delta_\ell^k \lambda_k > 0$ ,  $\ell = 1, 2$ , and  $\gamma_{j,\pm}^k < 0$  for  $j = 1, \dots, k-1, k+1, \dots, d$ .

**Proof.** The formula for the Jacobian at a bipartite consensus point can be obtained from Lemma 8. After some tedious calculations one gets

- $i, h \in \mathcal{V}_1$

$$F_o^{11} = (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( I + \frac{2n_2 \alpha_2^k}{\beta_1^k} \mathbf{v}_k \mathbf{v}_k^T Q^T K \right) \frac{\alpha_1^k}{\beta_1^k},$$

- $i \in \mathcal{V}_1, h \in \mathcal{V}_2$

$$F_o^{12} = (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( I - \frac{2n_1 \alpha_1^k}{\beta_1^k} \mathbf{v}_k \mathbf{v}_k^T Q^T K \right) \frac{\alpha_2^k}{\beta_1^k},$$

- $i \in \mathcal{V}_2, h \in \mathcal{V}_1$

$$F_o^{21} = (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( I - \frac{2n_2 \alpha_1^k}{\beta_2^k} \mathbf{v}_k \mathbf{v}_k^T Q^T K \right) \frac{\alpha_2^k}{\beta_2^k},$$

- $i, h \in \mathcal{V}_2$

$$F_o^{22} = (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( I + \frac{2n_1 \alpha_2^k}{\beta_2^k} \mathbf{v}_k \mathbf{v}_k^T Q^T K \right) \frac{\alpha_1^k}{\beta_2^k},$$

- $i = h \in \mathcal{V}_1$

$$\begin{aligned} F_d^1 &= (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( \frac{4n_1 n_2 \alpha_2^k}{\beta_1^k} \mathbf{v}_k \mathbf{v}_k^T K^T Q \right) \frac{\alpha_1^k}{\beta_1^k} \\ &\quad - \frac{n_1 \alpha_1^k - n_2 \alpha_2^k}{\beta_1^k} (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V), \end{aligned}$$

- $i = h \in \mathcal{V}_2$

$$\begin{aligned} F_d^2 &= (I - \mathbf{v}_k \mathbf{v}_k^T) V \left( \frac{4n_1 n_2 \alpha_2^k}{\beta_2^k} \mathbf{v}_k \mathbf{v}_k^T K^T Q \right) \frac{\alpha_1^k}{\beta_2^k} \\ &\quad + \frac{n_1 \alpha_2^k - n_2 \alpha_1^k}{\beta_2^k} (\mathbf{v}_k^T V \mathbf{v}_k I + \mathbf{v}_k \mathbf{v}_k^T V). \end{aligned}$$

The expressions in the statement of the lemma follow if one observes that  $(I - \mathbf{v}_k \mathbf{v}_k^T) V \mathbf{v}_k = 0$ . For  $F(\mathbf{v}_k)$  there are 3 classes of eigenvectors

1. First class:  $\mathbf{p}^\ell = [\underbrace{0 \dots 0}_{\ell-1} \mathbf{v}_k^T \underbrace{0 \dots 0}_{n-\ell}]^T$ . There are  $n$  such eigenvectors, and they are obviously all orthogonal to each other. Since  $(I - \mathbf{v}_k \mathbf{v}_k^T) V \mathbf{v}_k = 0$ , it is  $F_o^{ij} \mathbf{v}_k = 0$  for all  $i, j = 1, 2$ , while  $F_d^1 \mathbf{v}_k = -2\lambda_k \delta_1^k \mathbf{v}_k$  and  $F_d^2 \mathbf{v}_k = -2\lambda_k \delta_2^k \mathbf{v}_k$ . This means that

$$F(\mathbf{v}_k) \mathbf{p}^\ell = \begin{cases} -2\lambda_k \delta_1^k \mathbf{p}^\ell, & \ell = 1, \dots, n_1 \\ -2\lambda_k \delta_2^k \mathbf{p}^\ell, & \ell = n_1 + 1, \dots, n \end{cases}$$

i.e.,  $\mathbf{p}^\ell$  is an eigenvector of  $F(\mathbf{v}_k)$ ,  $\ell = 1, \dots, n$ .

2. Second class: Consider the  $d-1$  vectors  $\mathbf{q}^h$  s.t.  $\mathbf{q}^h = [(\underbrace{\mathbf{w}_1^h)^T \dots (\mathbf{w}_1^h)^T}_{n_1 \text{ times}} \underbrace{(\mathbf{w}_2^h)^T \dots (\mathbf{w}_2^h)^T}_{n_2 \text{ times}}]^T$ , with  $\mathbf{w}_1^h = \sum_{j=1}^d \eta_j^{h,1} \mathbf{v}_j$  and  $\mathbf{w}_2^h = \sum_{j=1}^d \eta_j^{h,2} \mathbf{v}_j$ . We have  $F_o^{ij} \mathbf{w}_\ell^h = \frac{\alpha_j}{\beta_i} \sum_{m \neq k} \lambda_m \eta_m^{h,\ell} \mathbf{v}_m$  for  $i, j, \ell = 1, 2$  and  $F_d^i \mathbf{w}_i^h = -\delta_i^k \lambda_k (\mathbf{w}_i^h + \eta_k^{h,i} \mathbf{v}_k)$ ,  $i = 1, 2$ . The  $\mathbf{q}^h$  are obtained solving the algebraic equation  $F \mathbf{q}^h = \gamma \mathbf{q}^h$  where also the eigenvalue  $\gamma$  is an unknown. Expanding we obtain a block of  $n_1$  identical equations

$$\begin{aligned} & -\delta_1^k \lambda_k (\mathbf{w}_1^h + \eta_k^{h,1} \mathbf{v}_k) + \sum_{j \neq k} \lambda_j \left( \frac{n_1 \alpha_1^k}{\beta_1^k} \eta_j^{h,1} \right. \\ & \left. + \frac{n_2 \alpha_2^k}{\beta_1^k} \eta_j^{h,2} \right) \mathbf{v}_j = \gamma \sum_{j=1}^d \eta_j^{h,1} \mathbf{v}_j \end{aligned}$$

and another of  $n_2$  identical equations

$$\begin{aligned} & -\delta_2^k \lambda_k (\mathbf{w}_2^h + \eta_k^{h,2} \mathbf{v}_k) + \sum_{j \neq k} \lambda_j \left( \frac{n_1 \alpha_2^k}{\beta_2^k} \eta_j^{h,1} \right. \\ & \left. + \frac{n_2 \alpha_1^k}{\beta_2^k} \eta_j^{h,2} \right) \mathbf{v}_j = \gamma \sum_{j=1}^d \eta_j^{h,2} \mathbf{v}_j \end{aligned}$$

whose solution provides both the desired eigenvector  $\mathbf{w}_i^h$  and the associated eigenvalues  $\gamma$ . As can be seen projecting along  $\mathbf{v}_k$ , these equations have solution only if  $\eta_k^{h,\ell} = 0$  i.e., if both  $\mathbf{w}_1^h$  and  $\mathbf{w}_2^h$  are orthogonal to  $\mathbf{v}_k$ :  $\mathbf{v}_k^T \mathbf{w}_\ell^h = 0$ . (similarly to the second class of eigenvectors in the multiagent Oja flow case, see Lemma 4). Projecting these equations along the eigenvector  $\mathbf{v}_j$  and rearranging

$$\begin{aligned} & \overbrace{\left( -\delta_1^k \lambda_k + \lambda_j \frac{n_1 \alpha_1^k}{\beta_1^k} - \gamma \right) \eta_j^{h,1}}^{=a_j^k} + \overbrace{\left( \lambda_j \frac{n_2 \alpha_2^k}{\beta_1^k} \right) \eta_j^{h,2}}^{=b_j^k} = 0 \\ & \underbrace{\left( \lambda_j \frac{n_1 \alpha_2^k}{\beta_2^k} \right) \eta_j^{h,1}}_{=c_j^k} + \underbrace{\left( -\delta_2^k \lambda_k + \lambda_j \frac{n_2 \alpha_1^k}{\beta_2^k} - \gamma \right) \eta_j^{h,2}}_{=d_j^k} = 0. \end{aligned}$$

Notice that  $a_j^k$ ,  $b_j^k$ ,  $c_j^k$  and  $d_j^k$  are independent of the index  $h$ , hence, to avoid repeated identical algebraic equations, we can take  $h = j$  and drop one index in the

$\eta_j^{h,i}$  variables:  $\eta_j^{h,i} = \eta_j^i$ , obtaining

$$\begin{aligned}(a_j^k - \gamma)\eta_j^1 + b_j^k\eta_j^2 &= 0 \\ c_j^k\eta_j^1 + (d_j^k - \gamma)\eta_j^2 &= 0,\end{aligned}$$

which leads to the formula for the eigenvalues  $\gamma_{j,\pm}^k = \frac{1}{2}\left(a_j^k + d_j^k \pm \sqrt{(a_j^k - d_j^k)^2 + 4c_j^k b_j^k}\right)$ . Since  $b_j^k c_j^k > 0$ , the two solutions are always real. The relationship between the components of the two vectors  $\mathbf{w}_\ell^j$  is then  $\eta_j^2 = -\frac{a_j^k - \gamma_{j,\pm}^k}{b_j^k}\eta_j^1$ . Notice that for each  $\mathbf{q}^j$  there are two eigenvalues  $\gamma_{j,\pm}^k$ , for a total of  $2(d-1)$  eigenvalues in this class.

3. The third class contains the  $nd - n - 2d + 2$  remaining eigenvectors subdivided into two groups:  $\mathbf{r}_1 = [(\mathbf{z}^1)^T \dots (\mathbf{z}^{n_1})^T \ 0 \ \dots \ 0]^T$  and  $\mathbf{r}_2 = [0 \ \dots \ 0 \ (\mathbf{z}^{n_1+1})^T \ \dots \ (\mathbf{z}^n)^T]^T$  where  $\mathbf{z}^i$  s.t.  $\mathbf{v}_k^T \mathbf{z}^i = 0$  and  $(\mathbf{w}_\ell^j)^T \mathbf{z}^i = 0$  for all  $i = 1, \dots, n$ ,  $j = 1, \dots, k-1, k+1, \dots, d$  and  $\ell = 1, 2$ . The  $\mathbf{z}^i$  are chosen so that  $\sum_{i=1}^{n_1} \mathbf{z}^i = 0$  and  $\sum_{i=n_1+1}^n \mathbf{z}^i = 0$ . Computing:  $F_o^{ij} \mathbf{z}^\ell = \frac{\alpha_j}{\beta_i} V \mathbf{z}^\ell$  for all  $i, j = 1, 2$ , and  $F_d^i \mathbf{z}^\ell = -\delta_i \lambda_k \mathbf{z}^\ell$ ,  $i = 1, 2$ . Hence

$$F(\mathbf{v}_k) \mathbf{r}_1 = \begin{bmatrix} \frac{\alpha_1^k}{\beta_1^k} \sum_{j=1}^{n_1} V \mathbf{z}^j - \delta_1^k \lambda_k \mathbf{z}^1 \\ \vdots \\ \frac{\alpha_1^k}{\beta_1^k} \sum_{j=1}^{n_1} V \mathbf{z}^j - \delta_1^k \lambda_k \mathbf{z}^{n_1} \\ \frac{\alpha_2^k}{\beta_2^k} \sum_{j=1}^{n_1} V \mathbf{z}^j \\ \vdots \\ \frac{\alpha_2^k}{\beta_2^k} \sum_{j=1}^{n_1} V \mathbf{z}^j \end{bmatrix} = -\delta_1^k \lambda_k \mathbf{r}_1$$

and, similarly,  $F(\mathbf{v}_k) \mathbf{r}_2 = -\delta_2^k \lambda_k \mathbf{r}_2$ .

Concerning stability, notice that the eigenvalues of  $F(\mathbf{v}_k)$  in the first and third class depend exclusively from the sign of  $\lambda_k$  and  $\delta_\ell^k$ , while eigenvalues depending on the difference  $\lambda_j - \lambda_k$  no longer appear directly, even though  $\lambda_j$  and  $\lambda_k$  enter into the complicated expressions of the second class, which varies with the cardinality of the splitting  $\mathcal{V}_1/\mathcal{V}_2$  in the bipartite consensus equilibria associated to  $\mathbf{v}_k$ . What can be concluded straightforwardly is that a bipartite consensus is stable iff  $\delta_\ell^k \lambda_k > 0$ ,  $\ell = 1, 2$  and  $\gamma_{j,\pm}^k$  are all in the left half plane. ■

**Lemma 11** *The polygonal equilibria are all unstable for (1).*

**Proof.** The idea of the proof is similar to that used in the multiagent Oja flow. A polygonal equilibrium point  $\mathbf{x} = \mathbf{s} = [\mathbf{s}_1 \ \dots \ \mathbf{s}_n] \in (\mathbb{S}^{d-1})^n$  is s.t.  $V \sum_{j=1}^n A_{ij}(\mathbf{s}) \mathbf{s}_j = 0$ . Let us compute the linearization of (1) at  $\mathbf{s}$ , obtained perturbing  $\mathbf{s}$  with a perturbation

$\mathbf{u} = [\mathbf{u}_1 \dots \mathbf{u}_n]$  with  $\mathbf{u}_i \in T_{\mathbf{s}_i} \mathbb{S}^{d-1}$ . Retaining only the first order terms:

$$\begin{aligned}\dot{\mathbf{u}}_i &= \frac{1}{n} (\mathbf{u}_i \mathbf{s}_i^T + \mathbf{s}_i \mathbf{u}_i^T) V \underbrace{\sum_{j=1}^n A_{ij}(\mathbf{s}) \mathbf{s}_j}_{=0} \\ &\quad + (I - \mathbf{s}_i \mathbf{s}_i^T) V \sum_j \sum_h (\mathbf{z}_h^{ij})^T \mathbf{u}_h \mathbf{s}_j \\ &\quad + (I - \mathbf{s}_i \mathbf{s}_i^T) V \sum_{j=1}^n A_{ij}(\mathbf{s}) \mathbf{u}_j + h.o.t.\end{aligned}$$

where the vectors  $(\mathbf{z}_h^{ij})^T = \frac{\partial A_{ij}(\mathbf{s})}{\partial \mathbf{u}_h}$  are computed in (10) and (11). Denote  $\xi^{ij} = \sum_h (\mathbf{z}_h^{ij})^T \mathbf{u}_h$  and observe that  $\xi^{ij}$  is a sum of bilinear forms  $\mathbf{s}_\ell B_k \mathbf{u}_h$  for some matrices  $B_k$  (see (10) and (11) for their specific expressions). From the matrix Cauchy-Schwartz inequality, for each of these bilinear forms it holds  $-\|B_k\|_2 \leq \mathbf{s}_\ell B_k \mathbf{u}_h \leq \|B_k\|_2$ , where  $\|B_k\|_2$  is independent of  $\mathbf{u}_h$ .

Expanding in a basis of eigenvectors of  $V$ :  $\mathbf{s}_i = \sum_{k=1}^d \zeta_k^i \mathbf{v}_k$  and  $\mathbf{u}_j = \sum_{k=1}^d \eta_k^j \mathbf{v}_k$ , we get

$$\begin{aligned}\dot{\mathbf{u}}_i &= \sum_k \dot{\eta}_k^i \mathbf{v}_k = \left( \left( I - \sum_k \zeta_k^i \mathbf{v}_k \sum_\ell \zeta_\ell^i \mathbf{v}_\ell^T \right) V \cdot \right. \\ &\quad \left. \cdot \sum_j \left( \xi^{ij} \sum_\ell \zeta_\ell^i \mathbf{v}_k + A_{ij}(\mathbf{s}) \sum_k \eta_k^j \mathbf{v}_k \right) \right).\end{aligned}$$

Assuming that the perturbation  $\mathbf{u}$  is aligned with  $\mathbf{v}_1$ , i.e., for all  $i$ ,  $\eta_1^i \neq 0$  and  $\eta_k^i = 0$  for  $k = 2, \dots, d$ , then, projecting along  $\mathbf{v}_1$ , yields

$$\dot{\eta}_1^i = (1 - (\zeta_1^i)^2) \lambda_1 \left( \sum_j \xi^{ij} \zeta_1^j + \sum_j A_{ij}(\mathbf{s}) \eta_1^j \right).$$

Denoting  $\boldsymbol{\eta}_1 = [\eta_1^1 \dots \eta_1^n]^T$  and  $\boldsymbol{\zeta}_1 = [\zeta_1^1 \dots \zeta_1^n]^T$  the collection of the  $\eta_1^i$  and  $\zeta_1^i$  components of all agents, the previous ODEs can be expressed in vector form as

$$\dot{\boldsymbol{\eta}}_1 = \lambda_1 (I - \Psi_1^2) (\Xi \boldsymbol{\zeta}_1 + A(\mathbf{s}) \boldsymbol{\eta}_1) \tag{12}$$

where  $\Psi_1 = \text{diag}(\boldsymbol{\zeta}_1)$ ,  $\Xi = [\xi^{ij}]$  is a matrix with lower and upper bounds independent of  $\mathbf{u}$ , and  $\boldsymbol{\zeta}_1$  is fixed.  $A(\mathbf{s})$  is a row stochastic matrix,  $\rho(A(\mathbf{s})) = 1$  and  $A(\mathbf{s})\mathbf{1} = \mathbf{1}$ . Since  $|\zeta_1^i| < 1$  because  $\mathbf{s}$  is not an eigenvector of  $V$  and  $\lambda_1 > 0$ , the system (12) diverges when  $\boldsymbol{\eta}_1 = \epsilon \mathbf{1}$  for some scalar  $\epsilon$ . Hence the polygonal equilibrium  $\mathbf{s}$  is unstable. ■

**Proof of Theorem 2.** The proof is the direct combination of Lemmas 9, 10 and 11, with the only observation that in the second class of eigenvalues of the bipartite consensus case of Lemma 10, the formula for the eigenvalues can be written equivalently as  $\gamma_{j,\pm}^k = \frac{1}{2} \left( a_j^k + d_j^k \pm \sqrt{(a_j^k + d_j^k)^2 - 4a_j^k d_j^k + 4c_j^k b_j^k} \right)$ , from which the condition  $\gamma_{j,\pm}^k < 0$  becomes  $a_j^k + d_j^k < 0$  and  $c_j^k b_j^k > a_j^k d_j^k \forall j = 1, \dots, k-1, k+1, \dots, d$ . From these, after some calculations, one gets (9). ■

**Remark 6** Since  $A(\mathbf{x}) = A(-\mathbf{x})$ , the global symmetry between any equilibrium point  $\mathbf{x}$  and its antipodal point  $-\mathbf{x}$  is preserved, hence  $\mathbf{x}$  and  $-\mathbf{x}$  have the same stability properties.

The characterization of Theorem 2 is weaker than that of Theorem 1 in several aspects, because the behavior of (1) is significantly more complex than that of (5). In particular, while in (5) there is an almost globally asymptotically stable attractor, the hallmark of (1) is its multistability. We list now some of the limitations of Theorem 2:

- The stability character of the bipartite consensus equilibria cannot be determined a priori, as cannot be the  $\mathbf{v}_k$  to which they align.
- The stability character of the clustering equilibria could not be verified analytically. In particular obtaining an explicit expression of the Jacobian linearization and of its eigenvalues seems out of reach for now. What we can see in simulations is that clustering equilibria (typically of low  $m$ ) can be locally asymptotically stable, but seem to be rarer than the bipartite consensus equilibria.
- The total number of coexisting attractors cannot be determined a priori.
- No Lyapunov-like function with globally nonincreasing derivative could be found for (1).

We also notice that for this multistable system computing the basin of attraction of the various equilibria seems a difficult problem. The simple case  $V = I$  treated in [1] in terms of hemispheres does not obey to Assumption 1.

## 6 Numerical examples

In this section we investigate some small-scale examples numerically, to get some insight into the behavior of the system (1).

**Example 1** For  $d = 3$  and  $n = 10$ , and for randomly chosen  $Q, K$  and  $V$ , examples of trajectories are given in Fig. 1. In panels (a) and (b) the agents converge to consensus equilibria, in panel (c) to a bipartite consensus equilibrium and in panel (d) to a 3-clustering equilibrium. In the bipartite consensus case, the stable equilibrium point is aligned with  $\mathbf{v}_1$ . Notice how sometimes the transient of a trajectory can be long and irregular (see example of panel (b)), which reminds somehow of the idea of a transient “metastability” mentioned in [8].

**Example 2** We consider now an example with  $d = 20$  and  $n = 100$ . We generate randomly 50 instances of the matrices  $Q, K$  and  $V$ , and for each triplet we perform 100 simulations, all leading to an equilibrium point. These equilibria are classified into consensus, bipartite consensus (specifying also to which eigenvector  $\mathbf{v}_k$  they align with) and clustering, specifying also  $m$ , the number of clusters. The resulting values are shown in Fig. 2. Recall that a 1-clustering is a consensus equilibrium not aligned with any eigenvector  $\mathbf{v}_k$ . A 2-clustering equilibrium is instead in general not a bipartite consensus. In more than 50% of the instances multistability appears.

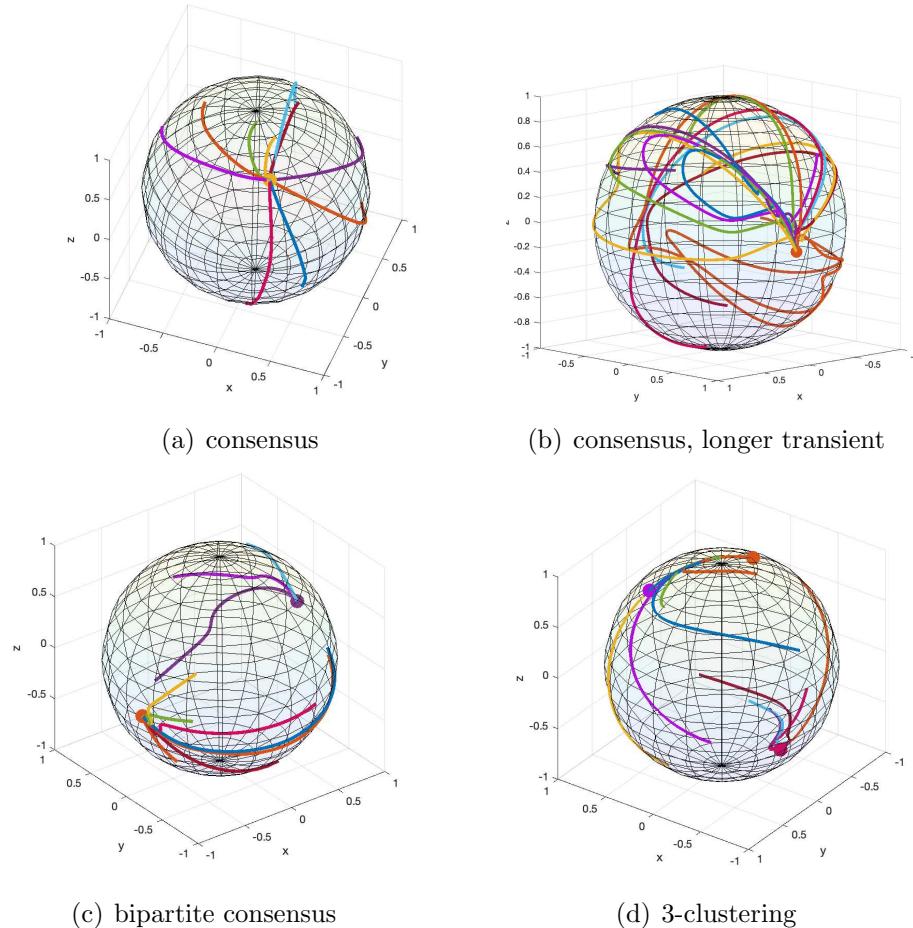


Figure 1: Example 1, with  $d = 3$  and  $n = 10$ . The solid dot is the endpoint of a trajectory.

**Example 3** In this example we aim to check the local stability of all consensus and bipartite consensus equilibria associated to all eigenvectors  $\mathbf{v}_k$ ,  $k = 1, \dots, d$ . Due to their explosion in number ( $2^n$ ) this can be done exhaustively only for small scale systems. Here we choose  $d = 4$  and  $n = 10$ . Fig. 3(a) shows that in 100 instances we tested, out of  $d2^n = 4096$  such equilibria, in some cases nearly half can be stable. In some other cases, instead, only the consensus aligned with  $\mathbf{v}_1$  (and its antipodal point) are instead stable, depending on the choice of  $V$ ,  $Q$  and  $K$ . Interestingly, the stable bipartite consensus equilibria are always aligned with the principal eigenvector  $\mathbf{v}_1$  or with the least (i.e., most negative) eigenvector  $\mathbf{v}_4$ . Whenever the latter case occurs, it is always  $|\lambda_4| > \lambda_1$ . Convergence to  $\mathbf{v}_1$  and  $\mathbf{v}_4$  can coexist in a system.

In Fig. 3(b) we consider instead a larger system,  $d = 10$  and  $n = 100$ , and for each  $\mathbf{v}_k$  we sample 100 bipartite consensus equilibria for each eigenvector  $\mathbf{v}_k$ . The situation is very similar: out of a total of 1000 equilibria, a fraction varying between 1 and 200 is stable, and convergence to bipartite consensus aligned with  $\mathbf{v}_1$  dominates, followed by  $\mathbf{v}_{10}$ . None of the other eigenvectors has any stable equilibrium.

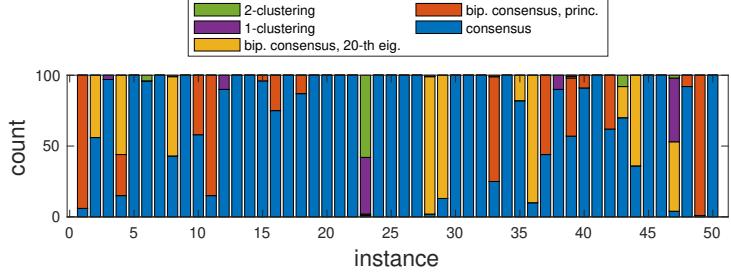


Figure 2: Example 2, numerical classification of stable equilibria.

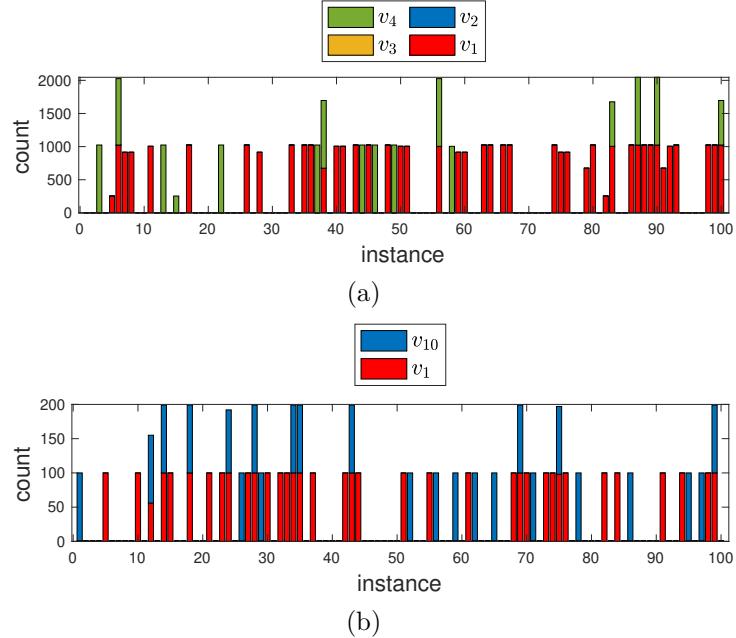


Figure 3: Example 3. (a): exhaustive count of all stable consensus + bipartite consensus equilibria; (b): random sampling stable bipartite consensus equilibria.

## 7 Extensions of the model

In formulating the model (1) we made a series of simplifying assumptions, which are now commented upon.

- *$V$  symmetric and with a simple, positive principal eigenvalue.* Numerically we see that this assumption can be relaxed as long as the principal eigenvalue of  $V$  remains real and simple. When a complex conjugate pair becomes the principal eigenvalue of  $V$ , then the self-attention dynamics may converge to a stable limit cycle. It remains to understand whether bipartite consensus or clustering equilibria are still present in this case, and what is their stability character.
- *A scaling factor  $\beta$  is disregarded* in the inner product leading to the attention matrix. This scaling factor is sometimes defined as  $\beta = \frac{1}{\sqrt{d}}$ , but in principle it can be interpreted as an inverse temperature. Including it, the attention matrix

becomes

$$A_{ij}(\mathbf{x}) = \frac{e^{\beta \langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle}}{\sum_{\ell=1}^n e^{\beta \langle Q\mathbf{x}_i, K\mathbf{x}_{\ell} \rangle}}. \quad (13)$$

One can study the behavior of (1) in the various possible regimes of  $\beta$ , see [8]. As stated in next proposition, when  $\beta \rightarrow 0$  we recover the multiagent Oja flow (5).

**Proposition 2** *The self-attention model (1) with the attention coefficients (13) collapses into the multiagent Oja flow (5) when  $\beta \rightarrow 0$ .*

**Proof.** Just observe that when  $\beta \rightarrow 0$ ,  $e^{\beta \langle Q\mathbf{x}_i, K\mathbf{x}_j \rangle} \rightarrow 1$ , hence  $A_{ij}(\mathbf{x}) \rightarrow \frac{1}{n}$ , regardless of  $\mathbf{x}$ . ■

- The model (1) uses a “single-head” attention mechanism, instead of a “multi-head” attention. A multihead self-attention dynamics looks like

$$\dot{x}_i = (I - x_i x_i^T) \sum_{h=1}^H V_h \sum_{j=1}^n A_{h,ij}(x) x_j.$$

It is trivial to show that consensus is still an asymptotically stable equilibrium point, with the single principal eigenvector of  $V$  replaced by a combination of principal eigenvectors of all  $V_h$ . The analysis of the other equilibria and of their stability is instead more complex and will be discussed in another venue.

- *Continuous-time instead of discrete-time.* A similar analysis can be carried out in discrete-time. In fact, the discrete-time model can be considered an Euler discretization of the continuous-time model [8, 1].
- *Time-invariant  $Q$ ,  $K$  and  $V$ ,* instead of time-varying. In the time-varying case, the analysis becomes more challenging, because asymptotic stability must be shown in a uniform sense. See [1] for some progress in this direction.
- *No feedforward neural network.* This is impossible to include in the continuous-time model. See again [1] for comments on what happens when it is added in discrete-time.

## 8 Conclusion

For the self-attention dynamical model of a transformer, in this paper we carried out a thorough analysis of the landscape of equilibria and investigated their stability properties. A feature that emerges is that multistability often occurs, associated typically, but not exclusively, to consensus (or consensus-like equilibria, like bipartite consensus). Another feature is that these stable consensus-like equilibria are aligned with the eigenvectors of the value matrix  $V$ , typically with the principal eigenvector, but sometimes also with other eigenvectors. If this property is confirmed also in more realistic models, it suggests that each layer of a transformer may act by tilting a token vector towards one of the eigenvectors of the value matrix, a property that we plan to verify experimentally in the near future.

## 9 Acknowledgments

This work was done while the author was on sabbatical at MIT. The hospitality of Ali Jadbabaie and IDSS is gratefully acknowledged.

## 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] M. Caponigro, A. C. Lai, and B. Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete and Continuous Dynamical Systems-Series A*, 35(9):4241–4268, 2015.
- [4] 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.
- [5] G. J. Dovonon, M. M. Bronstein, and M. J. Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.
- [6] S. Dutta, T. Gautam, S. Chakrabarti, and T. Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.
- [7] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.
- [8] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.
- [9] U. Helmke and J. B. Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.
- [10] K. Hornik and C.-M. Kuan. Convergence analysis of local feature extraction algorithms. *Neural Networks*, 5(2):229–240, 1992.
- [11] N. Karagodin, Y. Polyanskiy, and P. Rigollet. Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, 37:115652–115681, 2024.
- [12] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

- [13] T. Lin, Y. Wang, X. Liu, and X. Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.
- [14] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.
- [15] R. Mahony and P.-A. Absil. The continuous-time rayleigh quotient flow on the sphere. *Linear algebra and its applications*, 368:343–357, 2003.
- [16] J. Markdahl, J. Thunberg, and J. Gonçalves. Almost global consensus on the  $n$ -sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2017.
- [17] T. Nguyen, T. Nguyen, and R. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36:80233–80256, 2023.
- [18] 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.
- [19] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.
- [20] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
- [21] 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.
- [22] T. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural networks. *Neural Networks*, 2(459-473):8, 1989.
- [23] M. Scholkemper, X. Wu, A. Jadbabaie, and M. T. Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.
- [24] H. Shi, J. Gao, H. Xu, X. Liang, Z. Li, L. Kong, S. Lee, and J. T. Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022.
- [25] J. Thunberg, J. Markdahl, F. Bernard, and J. Goncalves. A lifting method for analyzing distributed synchronization on the unit sphere. *Automatica*, 96:253–258, 2018.
- [26] B. Van Dijk, T. Kouwenhoven, M. R. Spruit, and M. J. van Duijn. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. *arXiv preprint arXiv:2310.19671*, 2023.

- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [28] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.
- [29] X. Wu, A. Ajorlou, Y. Wang, S. Jegelka, and A. Jadbabaie. On the role of attention masks and layernorm in transformers. *Advances in Neural Information Processing Systems*, 37:14774–14809, 2024.
- [30] W.-Y. Yan, U. Helmke, and J. B. Moore. Global analysis of oja’s flow for neural networks. *IEEE Transactions on Neural Networks*, 5(5):674–683, 1994.
- [31] S. Yoshizawa, U. Helmke, and K. Starkov. Convergence analysis for principal component flows. *International Journal of Applied Mathematics and Computer Science*, 11(1):223–236, 2001.
- [32] S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.
- [33] Z. Zhang, S. Al-Abri, and F. Zhang. Opinion dynamics on the sphere for stable consensus and stable bipartite dissensus. *IFAC-PapersOnLine*, 55(13):288–293, 2022.
- [34] Z. Zhang, Y. Li, S. Al-Abri, and F. Zhang. Mixed opinion dynamics on the unit sphere for multi-agent systems in social networks. In *2025 American Control Conference (ACC)*, pages 4824–4829. IEEE, 2025.